Want Better Synthetic Data? Steer It: Activation Steering for Low-Resource Language Generation

Jan Cegin♠, Daniil Gurgurov†, Yusser Al Ghussin†, Simon Ostermann†
♠ Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia
jan.cegin@kinit.sk
† German Research Institute for Artificial Intelligence (DFKI), Saarbrücken, Germany
{daniil.gurgurov, yusser.al_ghussin, simon.ostermann}@dfki.de

Abstract

Large language models (LLMs) have become an effective tool for synthetic data generation, including for low-resource languages, where generated data can improve downstream task performance. Current best-performing approaches typically rely on few-shot prompting with target-language examples, which increases inference costs and may reduce diversity through lexical anchoring. In this work, we investigate activation steering as an alternative for low-resource synthetic data generation. We study two steering strategies: *Language* Steering, which targets the linguistic identity of a language, and *Quality* Steering, which captures well-formedness by contrasting human-written and backtranslated text representations. We evaluate these methods across four open-source LLMs, multiple layers, and 11 typologically diverse languages by generating sentiment and topic classification data and finetuning smaller classifiers. Steering is applied in both zero-shot and few-shot prompting settings and compared against non-steered counterparts. Our results show that steering on early layers consistently improves the diversity of generated data while often yielding stronger downstream model performance, particularly for low-resource languages.

1 Introduction

With the emergence of LLMs, various studies have demonstrated their ability to generate label-adherent and well-formed text (Cegin et al., 2023; Ubani et al., 2023). These two capabilities have made them an ideal tool for data *augmentation* and synthetic data *generation*, as demonstrated across domains such as sentiment analysis (Onan, 2023), intent recognition (Cegin et al., 2024), and topic classification (Piedboeuf and Langlais, 2023). For synthetic data generation, the workflow usually consists of prompting an LLM to produce a set number of samples with given labels in a particular language. These are then often used for fine-tuning of downstream encoder models.

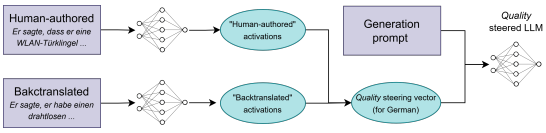


Figure 1: Example showing how *Quality* steering vectors are created. This example shows a contrastive collection of activations for German, from which a steering vector is created. This steering vector is used before generating data from the target LLM and language.

Recent works have extended these approaches to low-resource languages, often improving downstream model performance (Anikina et al., 2025; Pranida et al., 2025; Cegin et al., 2026), where data scarcity is prominent (Joshi et al., 2020). The strongest results generally come from few-shot prompting with examples from the target language (Anikina et al., 2025). However, this increases inference costs and may reduce diversity through lexical anchoring.

In parallel, research on representation engineering or activation steering has shown that modifying internal model activations can steer LLM behavior toward desired semantic or stylistic properties (Zou et al., 2023; Turner et al., 2024). By injecting steering vectors into hidden states, prior work has controlled attributes such as truthfulness (Li et al., 2023), sentiment (Rimsky et al., 2024), and toxicity (Turner et al., 2024) without additional fine-tuning. Compared to few-shot prompting, activation steering provides an efficient alternative that may better capture various abstract properties (Ostermann et al., 2026). Additionally, it can be combined with any type of prompt, zero- or few-shot (Radevski et al., 2026).

In this paper, we investigate activation steering for low-resource synthetic data generation. We generate synthetic datasets for sentiment and topic classification tasks across 11 typologically diverse

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 29385–29412
October 24-29, 2026 ©2026 Association for Computational Linguistics

languages and evaluate them through downstream fine-tuning performance. We study two steering strategies: *Language* Steering, which targets linguistic identity per previous studies (Gurgurov et al., 2026, 2025b; Ghussin et al., 2026a,b), and *Quality* Steering, which isolates well-formedness by contrasting human-written and backtranslated text. To our knowledge, this is the first work to derive and use *Quality* steering vectors. Across four open-source LLMs and multiple layers, we apply these steering vectors to both zero- and few-shot prompts (Anikina et al., 2025) and analyze the resulting data in terms of diversity and representational properties, compared to no-steering baselines. Code can be found at `https://github.com/kinit-sk/steering-synth-gen`.

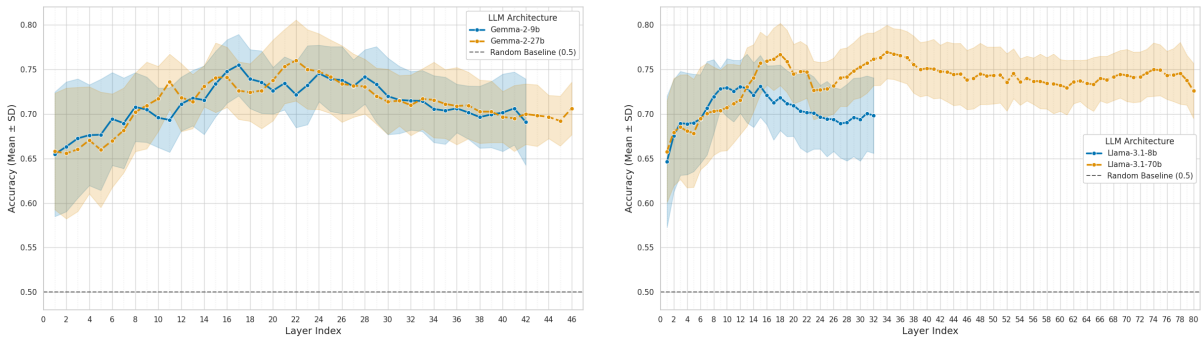
Our main contributions are:

- For the first time, we apply *Language* steering vectors for synthetic data generation and introduce and utilize *Quality* steering vectors by contrasting human-authored texts with back-translated texts.
- We analyze the cosine similarity between *Quality* and *Language* steering vectors, demonstrating that their alignment is highly LLM-dependent and most polarized in earlier layers. Crucially, the majority of evaluated languages display strong negative similarity, implying that, in general, *Quality* steering vectors are not tied to a specific language.
- We show that applying *Quality* and *Language* steering to early layers improves LLM performance in generating synthetic data, as measured by downstream performance, while increasing the diversity of the generated data for both zero- and few-shot prompting.

2 Related Work

Synthetic data. LLMs are increasingly being used to create semantically new samples adhering to a given label (Ubani et al., 2023; Cegin et al., 2025). LLM-based augmentation and data generation have been used for a variety of tasks such as automated scoring (Fang et al., 2023), low-resource language generation (Ghosh et al., 2023), intent classification (Sahu et al., 2022), sentiment analysis (Piedboeuf and Langlais, 2023; Onan, 2023), content recommendation (Liu et al., 2024), and health symptoms classifications (Dai et al., 2023). As LLMs are better generators than classifiers of data in low-resource languages (Pecher et al., 2026), recent studies have focused on generating such label-adhering data in a variety of low-resource languages like Vietnamese (Feng et al., 2021), Marathi (Tran et al., 2026), or various African languages (Belay et al., 2026), and for a variety of tasks and domains such as QA (Namboori et al., 2023), fact-checking (Chung et al., 2025), NER (Liu et al., 2021), or text classification (Glenn et al., 2023). A recent study has demonstrated that generating data rather than translating it leads to better model performance (Pranida et al., 2025). Other studies focused on finding the best approaches for generating data in low-resource languages, such as enhancing generator selection (Cegin et al., 2026) or finding the best prompting techniques (Anikina et al., 2025). As such, the current state-of-the-art technique (Anikina et al., 2025) uses few-shot examples in the prompt itself, leading to increased performance at the cost of additional inference.

Activation steering. A separate line of work has explored activation steering, where model behavior is modified by intervening directly on intermediate representations during the forward pass rather than through prompting or fine-tuning (Zou et al., 2023; Turner et al., 2024; Ostermann et al., 2026). The most widely used family of methods is difference-based steering, which derives steering vectors from contrastive examples (Turner et al., 2024; Rimsky et al., 2024; Marks and Tegmark, 2023). Beyond this, optimization-based approaches such as ReFT (Wu et al., 2024) and its rank-1 variant (Wu et al., 2025) learn low-rank interventions on hidden states, while dictionary-learning methods use sparse autoencoders to decompose activations into interpretable, individually steerable features (Gao et al., 2024; Cunningham et al., 2024). These techniques have been shown to reliably control a wide range of behaviors, such as sentiment, topic, and style (Turner et al., 2024; Konen et al., 2024), truthfulness and sycophancy (Li et al., 2023; Panickssery et al., 2023), and refusal (Arditi et al., 2024). Particularly relevant to our setting, steering has also been applied to multilingual control: Previous work has identified language-specific neurons (Zhao et al., 2024; Tang et al., 2024; Gurgurov et al., 2025b), SAE features (Chou et al., 2025; Deng et al., 2025), and language vectors (Wong et al., 2026; Gurgurov et al., 2026; Ghussin et al., 2026a) that can be ac-



(a) Gemma 2 model results.

(b) Llama 3.1 model results.

Figure 2: Linear probing results for human-authored vs. backtranslated textual pairs aggregated over languages.

tivated or ablated to change the output language while preserving semantics quality. Recent work has further shown that language-vector steering can be used for culturally aware multilingual inference: Ghussin et al. (2026b) reports improvements for cultural knowledge retrieval. While these studies establish that linguistic identity is encoded as a steerable direction in activation space, they focus on language control or downstream improvement on cultural benchmarks rather than the *generation quality* of the produced text. Our work addresses this gap by using both language and quality steering vectors to produce better synthetic data for low-resource languages.

3 Methodology and Experiments

Our methodology consists of a simple procedure where we (1) collect activations from the residual stream from given layers and create the steering vectors from the collected activations, (2) apply the steering vector to the LLM, (3) generate data with labels in a given language, and (4) finetune a downstream model and evaluate it on human test data. This is done for 11 different languages.

These 11 languages in our study are selected to cover diverse typological properties and ensure overlap with the evaluation datasets. They span Indo-European (Germanic: Danish, German; Slavic: Czech, Slovak, Slovenian), Afro-Asiatic (Semitic: Amharic, Hebrew, Maltese), and Austronesian (Indonesian, Javanese, Sundanese) families (Nordhoff and Hammarström, 2011). This design allows us to test steering robustness across varied morphological systems, from agglutinative to fusional structures, as well as across different writing systems (Ge'ez, Hebrew, Latin). This diversity supports evaluating whether the learned steering manifold generalizes beyond orthographic form.

3.1 Computing Steering Vectors

Language Vectors. We use the FLORES (NLLB Team et al., 2024) and BOUQUET (Andrews et al., 2025) datasets, both of which contain human-authored multilingual texts. FLORES provides professionally translated parallel sentences across all 11 target languages listed in Appendix E, enabling control over semantic content while isolating linguistic identity (Ghussin et al., 2026b). BOUQUET is added to increase lexical diversity and include more naturalistic language. Concatenating both datasets also ensures sufficient token coverage for stable mean activations.

To construct *Language* steering vectors, we employ a one-vs-rest contrastive approach over the residual stream activations of all transformer blocks (Ghussin et al., 2026b). For each of the 11 target languages, we compute a mean activation vector from the dataset by averaging across all non-padding tokens, yielding a language-specific mean representation in the model's latent space for every layer. The steering vector is then defined as the difference between the target language mean representation and the mean representation of all remaining languages, followed by normalization.

Quality Vectors. We follow the contrastive activation extraction approach of Rimsky et al. (2024), which requires paired examples representing opposing properties (in our case: human vs. generated samples). To construct these pairs, we generate backtranslations for FLORES and BOUQUET using the distilled NLLB model (Koishekenov et al., 2023). We treat the original human-authored text as the desired representation and the backtranslated text as its contrastive counterpart, following prior findings that synthetic text is generally lower quality than human-authored data (Schaffelder and

Table 1: Mean F1 difference across languages, models, steering methods, and layers aggregated over tasks, with positive and negative differences compared to a **zero-shot prompt** with no steering. The best alpha values and results were selected based on validation set results per layer, task, and language combination.

Method	LLM	Layer	am	cs	da	de	he	id	jv	mt	sk	sl	su	Average
Language	Gemma 2 27b	L10	2.74	0.97	-0.65	3.71	-0.30	2.71	1.85	0.45	3.70	3.06	6.86	2.28
		L22	3.22	0.75	-0.08	1.01	0.36	1.76	-0.33	1.16	3.90	2.01	5.11	1.72
		L34	4.34	0.86	0.23	0.71	0.83	2.67	2.78	-1.10	3.24	3.06	6.20	2.17
	Gemma 2 9b	L9	-1.91	4.28	2.07	1.99	5.80	0.28	-1.26	3.64	6.71	9.51	2.61	3.07
		L20	-1.58	3.59	2.41	0.53	2.62	0.88	0.21	3.38	8.02	5.98	3.96	2.73
		L31	3.22	1.71	2.17	1.49	2.99	0.70	0.40	2.96	6.56	6.27	2.52	2.82
	Llama3.1 70b	L17	3.47	-0.62	1.27	2.62	1.32	2.74	3.53	1.91	9.83	3.33	-0.46	2.63
		L37	4.65	1.29	1.88	3.61	2.40	1.35	5.43	0.53	7.35	2.48	2.96	3.08
		L57	3.43	0.34	1.77	2.61	1.82	2.52	4.19	1.49	3.92	3.78	4.60	2.77
	Llama3.1 8b	L7	8.50	0.74	1.18	1.48	0.81	2.47	2.47	4.49	4.01	4.01	-0.14	2.73
		L15	9.28	2.94	3.40	2.27	-0.05	4.78	0.81	4.25	3.84	4.59	-1.17	3.18
		L23	10.56	2.20	3.16	-0.63	-1.13	1.77	-0.08	3.08	4.86	2.66	0.14	2.42
Quality	Gemma 2 27b	L10	3.01	0.86	-0.02	2.71	0.93	2.30	1.38	1.40	3.02	3.92	5.65	2.29
		L22	2.06	0.56	-0.55	0.98	1.21	2.36	1.79	1.65	3.57	3.49	5.35	2.04
		L34	2.82	0.26	-0.54	1.45	0.20	2.31	3.44	0.28	4.68	2.77	4.28	2.00
	Gemma 2 9b	L9	0.34	8.04	2.19	3.90	4.63	1.83	0.26	1.76	11.77	10.25	5.12	4.55
		L20	3.34	3.66	3.31	2.69	1.34	2.66	0.52	2.82	6.81	5.42	4.39	3.36
		L31	2.25	4.22	2.22	2.10	1.81	1.77	-0.26	3.27	13.58	4.96	3.01	3.54
	Llama3.1 70b	L17	5.41	1.26	1.02	1.87	1.56	1.00	1.08	2.46	7.30	1.89	2.95	2.53
		L37	4.83	0.44	0.60	0.69	1.53	1.71	-1.84	2.12	7.51	1.71	2.73	2.00
		L57	4.96	1.17	1.31	2.93	0.92	0.92	0.26	2.03	6.08	1.68	0.90	2.11
	Llama3.1 8b	L7	13.13	2.97	3.89	1.04	1.26	1.63	1.79	5.38	4.31	4.46	0.83	3.70
		L15	11.36	1.98	2.17	-0.15	0.24	1.87	1.60	4.96	3.34	2.85	1.49	2.88
		L23	10.85	2.73	1.85	0.32	-2.65	2.68	0.74	4.09	3.41	3.13	1.43	2.60

Gatt, 2026). Each *Quality* steering vector is created language-specifically from the contrastive activations of human-authored and backtranslated texts from that given language. The aim is to capture subtle representational differences between human-written and synthetic text. An example is shown in Figure 1. All steering vectors are derived from task-agnostic data. Additional details are provided in Appendix F.

To construct *Quality* steering vectors, we use TransformerLens (Nanda and Bloom, 2022) to extract residual stream activations from transformer blocks without further training. We compute token-wise global averages for each example and define the steering vector as the difference between the mean activations of the contrastive pairs. Each of the *Quality* steering vectors is computed for a specific language (from data from that specific language). Details can be found in Appendix D.

3.2 Using Steering Vectors for Generation

To evaluate the effectiveness of steering vectors for synthetic data generation, we apply them at equivalent relative depths across four instruction-tuned LLMs: Gemma-2-9B, Gemma-2-27B, Llama-3.1-8B, and Llama-3.1-70B (Team et al., 2024; AI@Meta, 2024). Steering is applied at approximately 21%, 48%, and 74% of model depth, corresponding to early, middle, and late processing stages in the Transformer hierarchy (Bartoszcze et al., 2025). Early layers should be primarily associated with the consolidation of syntactic and linguistic identity (Tenney et al., 2019); the middle layers should represent the peak of conceptual abstraction (Geva et al., 2021); and the later layers, where the model transitions from abstract conceptual processing toward task-specific refinement and next-token probability mapping (Belrose et al., 2025). These layers were additionally selected based on stable linear probe performance in distinguishing human-authored from synthetic text (Section 4.1). Further details are in Appendix I.

We further investigate the effect of steering strength α. For Gemma models, we evaluate $\alpha \in \{25, 50, 75, 100\}$, while for Llama models we use $\alpha \in \{1, 2, 3, 4\}$. We observe substantial differences in sensitivity between the model families: Llama models frequently collapse at higher α values (e.g., repetition or empty outputs), whereas Gemma models require larger intervention strengths, remaining stable even at $\alpha = 100$. We attribute this to architectural differences, particularly Gemma-2's use of logit soft-capping and Query-Key normalization (Team et al., 2024), as well as its larger residual stream norms.

Table 2: Mean F1 difference across languages, models, steering methods, and layers aggregated over tasks, with positive and negative differences compared to a **few-shot prompt** with no steering. The best alpha values and results were selected based on validation set results per layer, task, and language combination.

Method	LLM	Layer	am	cs	da	de	he	id	jv	mt	sk	sl	su	Average
Language	Gemma 2 27b	L10	1.58	2.48	2.04	1.91	3.71	3.48	0.41	3.37	0.49	5.09	-0.30	2.20
		L22	-0.64	4.51	0.37	1.86	1.87	1.66	-0.02	2.68	-0.20	5.75	0.55	1.67
		L34	0.97	4.47	1.72	1.73	0.80	2.29	0.46	2.05	1.85	3.94	0.16	1.86
	Gemma 2 9b	L9	-3.42	4.31	1.70	2.63	1.22	-1.49	2.84	1.93	5.40	-1.16	-0.49	1.22
		L20	-0.07	4.79	5.63	1.70	6.82	1.56	4.07	-0.52	3.36	1.39	0.46	2.65
		L31	-0.70	2.24	4.56	2.57	1.56	0.12	4.41	1.03	3.50	0.57	0.53	1.85
	Llama3.1 70b	L17	2.10	-0.12	-0.13	3.75	3.55	5.29	0.38	1.12	6.67	0.98	2.04	2.33
		L37	0.32	1.37	1.35	1.25	4.67	3.09	0.98	3.23	3.65	0.82	4.48	2.29
		L57	1.22	1.44	1.67	1.44	4.41	1.72	1.68	1.42	1.73	1.73	4.26	2.07
	Llama3.1 8b	L7	1.62	2.01	3.43	2.26	1.06	1.80	-0.47	2.95	-0.52	1.80	-0.59	1.40
		L15	1.39	2.23	2.15	1.75	-0.10	2.46	1.34	2.84	3.45	2.30	-0.23	1.78
		L23	1.89	2.84	2.41	2.19	0.43	2.61	2.28	2.23	2.91	2.30	0.81	2.08
Quality	Gemma 2 27b	L10	-0.13	2.23	2.52	1.25	2.22	2.26	0.10	3.68	2.31	5.33	0.70	2.04
		L22	4.25	4.07	1.81	2.06	2.20	3.01	-0.46	2.97	1.26	5.80	0.45	2.49
		L34	1.17	3.71	1.22	1.10	3.66	1.84	0.26	3.55	1.80	3.47	0.67	2.04
	Gemma 9b	L9	0.74	3.24	2.75	3.95	8.91	1.10	4.03	3.19	2.08	1.45	1.50	2.99
		L20	-0.34	1.57	2.26	1.66	7.62	1.31	1.52	0.68	1.62	-0.02	2.31	1.84
		L31	-0.19	2.21	3.77	2.30	4.08	0.87	0.52	-1.14	4.38	0.36	1.03	1.65
	Llama3 70b	L17	1.55	1.65	1.14	1.97	5.42	3.02	1.50	3.46	6.52	2.16	3.63	2.91
		L37	0.96	1.11	1.16	1.90	3.12	0.66	1.56	-0.07	2.89	2.34	4.77	1.86
		L57	-0.09	1.46	0.97	0.88	2.39	1.56	1.26	2.44	2.75	1.65	5.48	1.89
	Llama3 8b	L7	2.39	2.18	3.95	2.37	0.91	1.00	0.62	3.13	0.79	0.54	0.37	1.66
		L15	1.49	0.93	0.31	1.75	0.48	1.47	-0.23	3.21	-1.08	-0.04	-0.28	0.73
		L23	4.89	1.00	-0.89	2.06	-0.49	1.18	0.23	2.97	-0.25	1.99	0.31	1.18

3.3 Finetuning Downstream Models

In our experiments, downstream model performance serves as the primary indicator of synthetic data quality, following prior work (Anikina et al., 2025; Cegin et al., 2026). We evaluate the effect of *Language* and *Quality* steering on synthetic data for topic classification (multi-class) and sentiment analysis (binary). Due to the limited availability of multilingual benchmarks, we use SIB-200 (Adelani et al., 2024) and the sentiment dataset collection of (Brychcín and Habernal, 2013; Gurgurov et al., 2024, 2025a). For each combination of LLM, language, layer, α, and task, we generate synthetic datasets, producing 50 samples per label for sentiment and 20 samples per label for topic classification, similar to prior setups (Anikina et al., 2025).

For downstream evaluation, we fine-tune XLM-R (Conneau et al., 2019) (*facebookAI/xlm-roberta-base*) with early stopping. We compare steering-based generation against two baselines: (1) zero-shot prompting without demonstrations, and (2) state-of-the-art few-shot prompting with human demonstrations (Anikina et al., 2025). Prompt templates are found in Appendix H, with further implementation details provided in Appendix G. For each combination of language, prompt, steering method, layer, and model, we report the downstream model performance achieved by the model with the best validation set loss. This led to various alphas for each cell in Tables 1 and 2, due to which we report their corresponding alpha values in Tables 12 and 13. We also investigate a case where we choose the same alpha per layer, steering method, and LLM generator to see if steering is applicable in cases where there is no validation set in Section 4.3.

4 Results and Discussion

4.1 Pre-Experiment: Linear Probing Results

We first conduct a linear probe analysis of human-authored (Flores + BOUQET) vs. back-translated pairs. This diagnostic step is done to justify our use of linear steering vectors for the *quality* approach; if a simple logistic regression can distinguish between these distributions with high accuracy, it confirms that the "quality" concept may be encoded as a specific direction in the latent space that we can manipulate (Park et al., 2024). The details on how the linear probe is computed can be found in Appendix C.

As seen in the aggregated results over all languages in Figure 2, all tested models across both the Gemma-2 and Llama-3.1 families exhibit accuracies significantly above the 0.5 random baseline from the very first layer. While the probe accuracy peaks in the middle-to-late layers, the "qual-

Table 3: Relative diversity metric differences across models, steering methods, and layers, aggregated over tasks and different alphas. Relative positive and negative changes.

(a) **Zero-shot** setup

Model + Steering	Layer	LexDiv	EmbDiv	IsoRad.	Homogen.
Gemma 2 27b (lang.)	L10	0.06	-1.35	-1.02	-1.24
	L22	-0.00	-2.16	-0.98	0.55
	L34	0.42	-1.17	-0.26	3.06
Gemma 2 27b (qual.)	L10	0.83	-0.01	-0.62	1.45
	L22	0.30	-1.55	-0.92	-0.51
	L34	-0.16	-2.38	-0.60	1.96
Gemma 2 9b (lang.)	L9	2.12	23.51	1.81	12.50
	L20	10.72	13.57	1.24	6.07
	L31	7.68	16.54	0.99	4.80
Gemma 2 9b (qual.)	L9	21.78	52.83	4.09	11.33
	L20	17.07	18.22	0.10	5.52
	L31	22.15	32.59	0.66	6.80
Llama3.1 70b (lang.)	L17	-29.03	55.32	1.93	13.13
	L37	-2.43	-2.94	-0.85	1.60
	L57	-0.97	-4.14	-1.61	-0.47
Llama3.1 70b (qual.)	L17	14.98	28.64	1.90	6.91
	L37	10.33	8.37	-0.06	7.26
	L57	2.40	-6.37	-1.57	1.18
Llama3.1 8b (lang.)	L7	69.28	68.10	3.24	5.77
	L15	44.03	58.97	2.45	3.34
	L23	32.80	20.41	2.22	5.27
Llama3.1 8b (qual.)	L7	43.08	35.96	3.35	8.95
	L15	33.73	-14.71	1.21	5.63
	L23	36.48	-2.27	2.28	7.73

(b) **Few-shot** setup

Model + Steering	Layer	LexDiv	EmbDiv	IsoRad.	Homogen.
Gemma 2 27b (lang.)	L10	0.33	-3.66	0.38	1.22
	L22	0.92	-4.92	0.15	0.26
	L34	0.99	-2.61	0.25	0.11
Gemma 2 27b (qual.)	L10	1.37	-3.48	0.33	1.00
	L22	1.51	-4.45	0.32	0.43
	L34	0.75	-5.62	0.49	0.65
Gemma 2 9b (lang.)	L9	-10.57	11.69	0.30	-4.56
	L20	6.77	11.91	0.09	-7.19
	L31	4.13	2.74	0.24	-2.80
Gemma 2 9b (qual.)	L9	10.82	23.05	1.29	0.19
	L20	5.98	22.17	0.95	0.59
	L31	4.15	14.53	0.20	-1.90
Llama3.1 70b (lang.)	L17	-29.66	43.96	1.37	5.36
	L37	-1.56	4.48	0.52	2.40
	L57	-1.35	4.83	0.29	1.21
Llama3.1 70b (qual.)	L17	8.15	15.55	0.94	2.62
	L37	10.10	12.51	1.56	5.66
	L57	1.39	0.10	0.16	0.37
Llama3.1 8b (lang.)	L7	43.19	57.04	1.45	-6.83
	L15	5.40	39.75	0.53	-0.90
	L23	-6.11	11.37	0.01	-1.06
Llama3.1 8b (qual.)	L7	14.00	1.84	-0.08	-2.83
	L15	-17.68	-44.81	-1.61	3.75
	L23	-11.14	-30.27	-1.09	-0.16

ity" signal is already well-defined by the earlier layers. The Gemma-2-9b model peaks at Layer 17 with an accuracy of approximately 0.75, while the larger Gemma-2-27b model peaks at Layer 22 with a slightly higher accuracy of approximately 0.76. The Llama-3.1-70b model demonstrates superior separability compared to its 8b counterpart, reaching a sustained plateau of approximately 0.77 accuracy between layers 35 and 40.

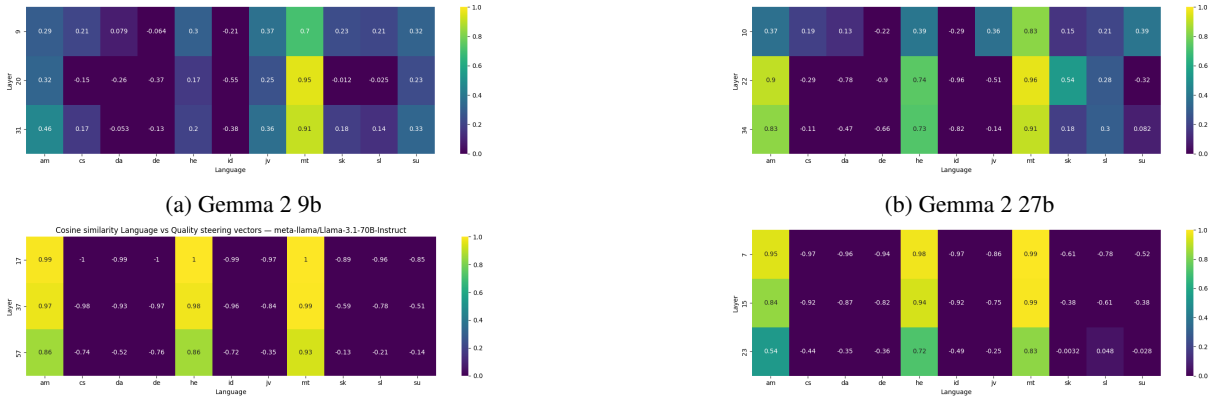
4.2 Downstream Model Performance Evaluation

We report the results for only the best alpha per cell in Tables 1 and 2 in terms of downstream model performance over both tasks combined. Additional results about how different alpha values affect downstream model performance can be found in the Appendix K. For statistical tests, we used Mann-Whitney-U (Mann and Whitney, 1947) with *p=0.05*. Full F1 baseline results for both zero- and few-shot settings can be found in Appendix J.

Steering vectors applied with zero-shot prompts Aggregated results across tasks are shown in Table 1, with per-task results provided in Appendix L. Overall, *Quality* steering consistently outperformed *Language* steering in terms of downstream F1 gains, suggesting that steering toward a "human-authored" manifold is more beneficial than steering toward generic linguistic identity alone.

Earlier layers proved the most effective intervention points, particularly for *Quality* steering. Across all task-aggregated experiments, *Quality* steering improved performance in 97.73% of cases for early layers, compared to 95.45% and 93.18% for middle and late layers. *Language* steering showed an inverse trend, with gains in 86.36%, 88.64%, and 90.91% of cases, respectively. Statistically significant improvements for *Quality* steering occurred in 45.46% of cases for early layers, versus 45.46% and 36.36% for middle and late layers. For *Language* steering, statistically significant gains were observed in 34.09%, 36.36%, and 35.23% of cases, respectively. Per-task results are provided in Appendix L.

Across languages, Slovak, and Slovenian showed the largest gains, while Javanese and Danish were comparatively resistant to steering, particularly for Gemma models. Both were languages with a considerable number of negative outcomes. Among models, Gemma-2-9B and Llama-3.1-8B were the most responsive to steering, producing the largest absolute F1 improvements. Overall, Llama models benefited more consistently from steering, whereas Gemma-2-27B appeared more conservative, suggesting that larger models may require stronger or more precise interventions.



(a) Gemma 2 9b

(b) Gemma 2 27b

(c) Llama 3.1 70b

(d) Llama 3.1 8b

Figure 3: Cosine similarity between quality and language vectors for different LLMs used in this study.

Steering vectors applied with few-shot prompts The aggregated few-shot results are shown in Table 2, with task-specific results provided in Appendix L. Compared to the zero-shot setting, the advantage of *Quality* steering over *Language* steering becomes less pronounced, and the strong preference for early-layer interventions weakens when demonstrations are included in the prompt. Nevertheless, early-layer *Quality* steering remains the most consistent approach overall. Notably, *Language* steering has no clear winner for layer selection, while *Quality* steering yields best performance gains at earlier layers.

Quality steering was again most effective at early layers, improving downstream performance in 97.73% of cases, compared to 81.82% and 86.36% for middle and late layers. *Language* steering showed an inverse trend, with gains in 77.27%, 84.09%, and 97.73% of cases, respectively. Average gains decreased for most models, except for Llama-3.1-70B and Gemma 2 27B, which appeared to benefit from combining steering with few-shot demonstrations. Statistically significant improvements for *Quality* steering occurred in 36.36% of cases for early layers, versus 21.59% and 25% for middle and late layers. For *Language* steering, statistically significant gains were observed in 29.55%, 31.82%, and 26.14% of cases, respectively. Together with the zero-shot results, these findings indicate that early-layer steering consistently outperforms deeper interventions. Per-task results are provided in Appendix L.

Among models, Gemma-2-27B and Llama-3.1-70B were the most robust, showing predominantly positive performance shifts. In contrast, Gemma-2-9B behaved more inconsistently in the few-shot setting, exhibiting substantially more negative outcomes than in the zero-shot experiments.

Comparison of zero-shot vs. few-shot Comparing zero-shot and few-shot prompting reveals a clear saturation effect. In the zero-shot setting, steering often produced substantial gains, including double-digit F1 improvements, whereas the same configurations yielded smaller improvements under few-shot prompting. This suggests that few-shot demonstrations already shift the model toward the "human-authored" manifold, reducing the additional effect of steering. Nevertheless, *Quality* steering applied to early layers continued to provide consistent improvements.

The strong "early layers are best" trend observed in zero-shot generation also becomes less rigid in the few-shot setting. Larger models, in particular, showed greater robustness and occasionally benefited from steering at deeper layers.

Overall, our results indicate that activation steering is most impactful in zero-shot settings, where it can compensate for the absence of stylistic guidance in the prompt. In few-shot scenarios, steering remains beneficial, though with smaller gains. This shows that, regardless of the original prompt used, steering is very often beneficial for downstream model performance when applied to the used prompting technique, with *Quality* steering offering higher and more consistent increases.

4.3 Using Only Best Alpha Per Layer

Our evaluation setup uses a validation split of the given dataset and language to select the best alpha on the given layer via validation loss. To investigate whether we can select a single alpha value per layer, steering method, and generator LLM, we use the

same alpha value for the entire layer (e.g., all data generated using Llama3 8b had the same alpha applied for the *Quality* steering vector at layer 7). The results for the zero-shot and few-shot settings can be found in Tables 14 and 15, respectively.

For the zero-shot setting, early layers proved the most effective intervention points again. Across all task-aggregated experiments, *Quality* steering improved performance in 79.54% of cases for early layers, compared to 68.18% and 70.45% for middle and late layers. *Language* steering showed a similar trend, with gains in 72.73%, 70.45%, and 63.64% of cases, respectively. This is an absolute decrease in the number of positive cases by approximately 20% to 28% compared to using the best alpha per case. Statistically significant improvements have dropped by approximately 10% on average.

We can see a similar trend in the few-shot setting, where the absolute number of positive increases drops by 25% to 30%. However, the number of statistically significant cases remains similar. For languages, we can see that using the correct alpha for generation becomes even more important — languages like Danish, Hebrew, or Javanese yield overall slight decreases in model performance. This underscores the importance of validation splits during evaluation for selecting the best alpha.

5 Evaluating Diversity of Generated Data

As shown in Section 4.2, both *Quality* and *Language* steering generally improve downstream model performance. To better understand these gains, we additionally analyze the diversity of the generated data, which has previously been linked to improved downstream robustness and performance (Cegin et al., 2026).

We evaluate four diversity metrics: (1) *Lexical diversity*, measuring the ratio of unique character 3-grams; (2) *Embedding diversity*, capturing the average pairwise cosine distance between text embeddings; (3) *Homogeneity*, measuring the structural consistency of the embedding distribution; and (4) *Isocontour radius*, estimating the overall spread of embedding representations. Details of the computation are provided in Appendix H.

We compute these metrics across all datasets. The aggregated results are in Table 3a for zero-shot and in Table 3b for few-shot prompting, with additional visualizations in Appendix O.

In the zero-shot setting, steering generally increases output diversity, particularly for smaller models and when applied to early layers. Most configurations show gains across all diversity metrics, with the notable exception of Gemma-2-27B, which also showed limited responsiveness in downstream evaluation. Increases in lexical and embedding diversity indicate that steering reduces repetitive generations and broadens semantic coverage, while higher homogeneity suggests that this increased diversity remains structurally coherent rather than noisy. Early-layer steering appears especially effective because it influences representations before higher-level semantic processing has stabilized, allowing the model to propagate the intervention throughout the generation. This aligns with the stronger downstream performance improvements observed for earlier-layer steering.

In the few-shot setting, diversity gains become smaller and less consistent, suggesting that demonstrations already constrain the model's latent space. In some cases, deeper-layer *Quality* steering even reduces diversity, particularly for Gemma-2-27B. Smaller models also tend to show decreases in homogeneity under few-shot steering, whereas larger models such as Llama-3.1-70B often maintain or improve structural consistency. Interestingly, *Language* steering frequently produces larger increases in embedding diversity than *Quality* steering in Llama models, possibly because language vectors are less aligned with the representations already induced by few-shot examples.

Although increased diversity does not always guarantee better downstream performance, our results show that steering vectors consistently encourage the generation of more varied data without explicit diversity optimization. This likely contributes to the improved robustness and downstream effectiveness that we observed.

6 Similarity of *Language* and *Quality* Vectors

We provide cosine similarity between *Language* and *Quality* vectors per LLM in Figure 3.

We see a distinct difference between LLMs in Gemma and Llama families. For Gemma 2 9b, most languages show low to moderate positive similarity, except Maltese. The polarity is much stronger in Gemma 2 27b. In contrast, both Llama models show a remarkably similar, highly polarized pattern for the same languages. Amharic, Hebrew, and Maltese are all nearly perfectly positively correlated, while other languages are nearly perfectly

negatively correlated.

The languages appear to cluster into two distinct groups based on vector alignment. For the positive cluster, the *Quality* and *Language* are essentially in the same direction in the model's latent space. This explains why Amharic showed strong F1 gains under both steering types in most of our experiments. Notably, all of these languages are Semitic, which could indicate that for specific language groups the model sees *Quality* and *Language* as nearly identical. For the majority of languages, especially European and Southeast Asian ones, the *Quality* direction is often the mathematical opposite of *Language* direction in Llama models.

The polarities (-1.0 or 1.0) are most consistent in the early layers across all models. This reinforces our earlier conclusion that early-layer interventions are more effective because they target these highly defined, clear directions before the representations become more "mixed" or diffuse in later layers. The polarities indicate that the *Quality* direction is often the inverse of the *Language* direction. While both of these steering vectors can help in downstream model performance, this indicates that *Quality* steering vectors are (in general) not tied to a specific language and seem to be language-agnostic. This also explains the volatility of improvements from *Language* steering, as for most of the languages, it is steering the model in the exact opposite direction of the *Quality* steering.

7 LLM-as-Judge Evaluation

We also evaluated the generated data in terms of label adherence, language adherence, and fluency via LLM-as-Judge using Gemma 3 27b. The aggregated results and details can be found in Appendix P.

In general, we see only slight changes for *Quality* steering for both zero- and few-shot setups. We see that the early layers lead to often smaller relative decreases in all categories (in 61.1 % of cases), as such steering seems to produce only noisier, but still valid, samples that are beneficial for downstream model performance. However, the decreases were not significant.

Language steering had even more negative cases (66.66%) and produced stronger outliers in the Llama 3.1 70b model in particular, where early layers lead to an approximate 10% relative decrease in all 3 metrics. Similar changes can also be seen in other models, indicating that *Language* steering can hurt the perceived validity (by the LLM) of the data.

8 Conclusion

In this work, we investigated activation steering for synthetic data generation in low-resource languages. Across four open-source LLMs, 11 languages, and two classification tasks, we showed that both *Language* and *Quality* steering can improve downstream performance while consistently increasing the diversity of generated data. Our results further demonstrate that steering is most effective when applied to earlier transformer layers, particularly in zero-shot settings. This finding is consistent with the cultural steering results of Ghussin et al. (2026b), who likewise observe that the effectiveness of vectors is strongly layer-dependent, with earlier-layer steering yielding the most reliable gains.

Overall, *Quality* steering proved to be the more reliable approach, suggesting that steering models toward a "human-authored" notion is especially beneficial for synthetic data generation. We additionally found that the relationship between Language and *Quality* vectors is highly model- and language-dependent, highlighting structured multilingual representations in LLM latent spaces.

These findings position activation steering as an efficient alternative or complement to few-shot prompting for multilingual synthetic data generation, especially in low-resource settings.

Acknowledgments

This work was partially supported by AI-CODE (GA No. 101135437), by the European Union under the project lorAI - Low Resource Artificial Intelligence, GA No. 101136646, by NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I01-03-V04-00068 (GEPERO), and by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project TRAILS (01IW24005).

This work was computationally supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254) and by the use of the supercomputer PERUN, with the support of the European Union from the funds of the Recovery and Resilience Plan of the Slovak Republic within the framework of project No. 17I03-04-P03-00001.

Limitations

While our work demonstrates the efficacy of activation steering for low-resource synthetic data generation, several limitations remain to be addressed in future work.

Task and Generative Scope: Our evaluation is bounded by two text classification tasks: sentiment analysis and topic classification. While classification is a standard benchmark for evaluating synthetic data utility (Anikina et al., 2025; Cegin et al., 2026), it does not fully test the generative boundaries of activation steering.

Dependence on Machine Translation Quality: The extraction of our Quality steering vector relies on constructing contrastive text pairs via two rounds of backtranslation using the NLLB-200-600M distilled model. We did not investigate how many specific rounds or which translation models are the best.

Hyperparameter Sensitivity and Model Dependency: Our results highlight that the optimal steering intensity (α) and layer selection are heavily model- and language-dependent. For instance, the optimal scaling factors vary by orders of magnitude between the Gemma and Llama families (e.g., $\alpha = 100.0$ for Gemma 2 27B vs. $\alpha = 1.0$ for Llama 3.1 70B). At present, we lack a predictive, analytical framework to establish the perfect layer and multiplier combination a priori, limiting the immediate plug-and-play deployability of this method.

LLM-as-Judge: Our evaluation used LLM-as-Judge, which can be misleading and produce invalid annotations.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Mark Duppenthaler, Nathanial Paul Ekberg, Cynthia Gao, Daniel Edward Licht, Jean Maillard, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenko, Albert Ventayol-Boada, and Shireen Yates. 2025. BOUQuET : dataset, benchmark and open initiative for universal quality evaluation in translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27515–27535, Suzhou, China. Association for Computational Linguistics.

Tatiana Anikina, Jan Cegin, Jakub Simko, and Simon Ostermann. 2025. A rigorous evaluation of LLM data generation strategies for low-resource languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8282–8303, Suzhou, China. Association for Computational Linguistics.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083.

Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. 2025. Representation engineering for large-language models: Survey and research challenges. *Preprint*, arXiv:2502.17601.

Tadesse Destaw Belay, Shahriar Kabir Nahin, Israel Abebe Azime, Ocean Monjur, Shamsuddeen Hassan Muhammad, Seid Muhie Yimam, and Anshuman Chhabra. 2026. Afrilangtutor: Advancing language tutoring and culture education in low-resource languages with large language models. *Preprint*, arXiv:2604.20996.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2025. Eliciting latent predictions from transformers with the tuned lens. *Preprint*, arXiv:2303.08112.

Tomáš Brychcín and Ivan Habernal. 2013. Unsupervised improving of sentiment analysis using global target context. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 122–128, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Jan Cegin, Branislav Pecher, Jakub Simko, Ivan Srba, Maria Bielikova, and Peter Brusilovsky. 2024. Effects of diversity incentives on sample diversity and downstream model performance in LLM-based text augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13148–13171, Bangkok, Thailand. Association for Computational Linguistics.

Jan Cegin, Branislav Pecher, Jakub Simko, Ivan Srba, Maria Bielikova, and Peter Brusilovsky. 2025. Use random selection for now: Investigation of few-shot selection strategies in LLM-based text augmentation.

In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5533–5550, Suzhou, China. Association for Computational Linguistics.

Jan Cegin, Branislav Pecher, Ivan Srba, and Jakub Simko. 2026. RoSE: Round-robin synthetic data evaluation for selecting LLM generators without human test sets. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5530–5545, Rabat, Morocco. Association for Computational Linguistics.

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2023. ChatGPT to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1889–1905, Singapore. Association for Computational Linguistics.

Cheng-Ting Chou, George Liu, Jessica Sun, Cole Blondin, Kevin Zhu, Vasu Sharma, and Sean O'Brien. 2025. Causal language control in multilingual transformers via sparse feature steering. *Preprint*, arXiv:2507.13410.

Yi-Ling Chung, Aurora Cobo, and Pablo Serna. 2025. Beyond translation: Llm-based data generation for multilingual fact-checking. *arXiv preprint arXiv:2502.15419*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Hoagy Cunningham, Aidan Ewart, Logan Riggs Smith, Robert Huben, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Aug-gpt: Leveraging chatgpt for text data augmentation. *Preprint*, arXiv:2302.13007.

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. 2025. Unveiling language-specific features in large language models via sparse autoencoders. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4563–4608, Vienna, Austria. Association for Computational Linguistics.

Luyang Fang, Gyeong-Geon Lee, and Xiaoming Zhai. 2023. Using gpt-4 to augment unbalanced data for automatic scoring. *Preprint*, arXiv:2310.18365.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. ArXiv:2406.04093.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sreyan Ghosh, Chandra Kiran Evuru, Sonal Kumar, S Ramaneswaran, S Sakshi, Utkarsh Tyagi, and Dinesh Manocha. 2023. Dale: Generative data augmentation for low-resource legal nlp. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Sentosa, Singapore.

Yusser Al Ghussin, Daniil Gurgurov, Tanja Baeumel, Josef van Genabith, Patrick Schramowski, and Simon Ostermann. 2026a. Multilingual steering by design: Multilingual sparse autoencoders and principled layer selection. *Preprint*, arXiv:2605.23036.

Yusser Al Ghussin, Daniil Gurgurov, Yasser Hamidullah, Josef van Genabith, Cristina España-Bonet, and Simon Ostermann. 2026b. Dfki-mlt at semeval-2026 task 7: Steering multilingual models towards cultural knowledge. *Preprint*, arXiv:2605.23069.

Parker Glenn, Alolika Gon, Nikhil Kohli, Sihan Zha, Parag Pravin Dakle, and Preethi Raghavan. 2023. Jetsons at the FinNLP-2023: Using synthetic data and transfer learning for multilingual ESG issue classification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 133–139, Macao. -.

Daniil Gurgurov, Yusser Al Ghussin, Tanja Baeumel, Cheng-Ting Chou, Patrick Schramowski, Marius Mosbach, Josef van Genabith, and Simon Ostermann. 2026. Clas-bench: A cross-lingual alignment and steering benchmark. *arXiv preprint arXiv:2601.08331*.

Daniil Gurgurov, Mareike Hartmann, and Simon Ostermann. 2024. Adapting multilingual LLMs to low-resource languages with knowledge graphs via adapters. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 63–74, Bangkok, Thailand. Association for Computational Linguistics.

Daniil Gurgurov, Rishu Kumar, and Simon Ostermann. 2025a. GrEmLIn: A repository of green baseline embeddings for 87 low-resource languages injected with multilingual graph knowledge. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1204–1221, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6282–6293.

Yeskendir Koishekenov, Alexandre Berard, and Vassilina Nikoulina. 2023. Memory-efficient NLLB-200: Language-specific expert pruning of a massively multilingual machine translation model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3567–3585, Toronto, Canada. Association for Computational Linguistics.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style vectors for steering generative large language model. *Preprint*, arXiv:2402.01618.

Yi-An Lai, Xuan Zhu, Yi Zhang, and Mona Diab. 2020. Diversity, density, and homogeneity: Quantitative characteristic metrics for text collections. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1739–1746, Marseille, France. European Language Resources Association.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: eliciting truthful answers from a language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846, Online. Association for Computational Linguistics.

Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024. Once: Boosting content-based recommendation with both open- and closed-source large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 452–461, New York, NY, USA. Association for Computing Machinery.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

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

Amani Namboori, Shivam Sadashiv Mangale, Andy Rosenbaum, and Saleh Soltan. 2023. Gemquad: Generating multilingual question answering datasets from large language models using few shot learning. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*.

Neel Nanda and Joseph Bloom. 2022. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Sebastian Nordhoff and Harald Hammarström. 2011. Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources. In *First International Workshop on Linked Science 2011-In conjunction with the International Semantic Web Conference (ISWC 2011)*.

Aytuğ Onan. 2023. Srl-aco: A text augmentation framework based on semantic role labeling and ant colony optimization. *Journal of King Saud University - Computer and Information Sciences*, 35(7):101611.

Simon Ostermann, Daniil Gurgurov, Tanja Baeumel, Michael A. Hedderich, Sebastian Lapuschkin, Wojciech Samek, and Vera Schmitt. 2026. From weights to activations: Is steering the next frontier of adaptation? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29854–29879, San Diego, California, United States. Association for Computational Linguistics.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. *Preprint*, arXiv:2311.03658.

Branislav Pecher, Jan Cegin, Robert Belanec, Ivan Srba, Jakub Simko, and Maria Bielikova. 2026. Better as generators than classifiers: Leveraging LLMs and synthetic data for low-resource multilingual classification. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2840–2857, Rabat, Morocco. Association for Computational Linguistics.

Frédéric Piedboeuf and Philippe Langlais. 2023. Is ChatGPT the ultimate data augmentation algorithm? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15606–15615, Singapore. Association for Computational Linguistics.

Salsabila Zahirah Pranida, Rifo Ahmad Genadi, and Fajri Koto. 2025. Synthetic data generation for culturally nuanced commonsense reasoning in low-resource languages. *Preprint*, arXiv:2502.12932.

Gorjan Radevski, Kiril Gashteovski, Giwon Hong, Carolin Lawrence, and Goran Glavaš. 2026. Compositional steering of large language models with steering tokens. *arXiv preprint arXiv:2601.05062*.

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

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Max Schaffelder and Albert Gatt. 2026. Synthetic eggs in many baskets: The impact of synthetic data diversity on LLM fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 7265–7293, San Diego, California, United States. Association for Computational Linguistics.

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

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Van-Hien Tran, Huy Hien Vu, Hideki Tanaka, and Masao Utiyama. 2026. Representation-aware prompting for zero-shot Marathi text classification: IPA, Romanization, repetition. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 436–443, Rabat, Morocco. Association for Computational Linguistics.

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

Solomon Ubani, Suleyman Olcay Polat, and Rodney Nielsen. 2023. Zeroshotdataaug: Generating and augmenting training data with chatgpt. *Preprint*, arXiv:2304.14334.

Sing Hieng Wong, Hassan Sajjad, and AB Siddique. 2026. Langfir: Discovering sparse language-specific features from monolingual data for language steering. *arXiv preprint arXiv:2604.03532*.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2025. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2024. ReFT: Representation finetuning for language models. *arXiv:2404.03592*.

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

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Ethical Considerations

Based on a thorough ethical assessment performed on the basis of intra-institutional ethical guidelines and checklists tailored to the use of data and algorithms, we see no ethical concerns pertaining directly to the conduct of this research. Although the production of new data through LLMs bears several risks, such as the introduction of biases, the small size of the produced dataset, sufficient for experimentation, is, at the same time, insufficient for any major machine learning endeavours where such biases could be transferred.

We follow the license terms for all the models and datasets we used (such as the one required for the use of the Llama-3.1 model) – all models and datasets allow their use as part of the research.

We used generative AI for grammatical checks and fixes of the paper, and have verified all the texts changed by the generative AI.

B Computational Resources

Our experiments were run on a computational cluster with H200 GPUs, 2x AMD EPYC 9745 CPUs, and 128 GB of RAM. Our generation experiments took approximately 6-8 minutes per one combination of LLM, steering vector, layer, task, language, and alpha value. Given that we had a total of 2112 combinations, our generation experiments took approximately 260 GPU hours. For the fine-tuning experiments, we used approximately 240 GPU hours. In total, for our experiments, we used approximately 500 GPU hours.

C Linear Probe Details

To quantify the linear separability of the quality signal within the model's latent representations, we implemented a layer-wise probing diagnostic using logistic regression. For each input sequence in our human-authored and backtranslated distributions, we extracted the hidden states from the model's residual stream. We then applied mean pooling across non-padding tokens to generate representative sequence-level embeddings for every layer. These activations were sanitized by clipping outliers and then L2-normalized to facilitate stable classifier training. Using a stratified 80/20 train-test split, we trained a logistic regression classifier on the embeddings of each layer independently and measured the resulting classification accuracy. As illustrated in Figure 2, the layers exhibiting peak accuracy represent the regions where the quality concept is most robustly and linearly encoded, providing the optimal candidates for activation steering interventions.

D Steering Vectors Computation Details

D.1 *Language* Steering Vector

To construct the language steering vectors, we employ a one-vs-rest contrastive methodology across the residual stream of all transformer blocks. This approach ensures that the resulting vector captures features specific to the target language rather than shared cross-lingual properties or general model priors.

For each target language $L \in \mathcal{L}$, where $\mathcal{L}$ is our set of 11 typologically diverse languages, we pass the dataset through the model. Using the TransformerLens framework, we extract the residual stream activations for every layer l. We compute the mean activation vector $\mu_{l,L}$ by averaging across all non-padding tokens N_L in the language-specific dataset:

$$\mu_{l,L} = \frac{1}{N_L} \sum_{i,j} \mathbf{a}_{l,i,j} \cdot m_{i,j}$$

. This process results in a language-specific centroid in the latent space for every layer of the model.

To isolate the linguistic identity of language L, we define the steering vector $\mathbf{v}_{l,L}$ as the difference between the language's mean activation and the mean activation of all other languages in our study:

$$\mathbf{v}_{l,L} = \mu_{l,L} - \frac{1}{|\mathcal{L}| - 1} \sum_{L' \in \mathcal{L}, L' \neq L} \mu_{l,L'}$$

By contrasting the target language against a global multilingual mean, we filter out common semantic and structural activations, focusing the vector on the unique "fingerprint" of the target language.

Finally, we apply layer-wise L2 normalization to the difference vectors:

$$\hat{\mathbf{v}}_{l,L} = \frac{\mathbf{v}_{l,L}}{\|\mathbf{v}_{l,L}\|_2}$$

This normalization step is critical for maintaining stability during the generation phase, as it ensures that the steering magnitude α remains consistent across different layers and languages regardless of the original activation scales.

D.2 *Quality* Steering Vector

For each target language $L \in \mathcal{L}$, we utilize two matching corpora representing a binary quality dimension $\mathcal{D} = \{d_{\text{good}}, d_{\text{bad}}\}$, where d_{good} consists of authentic, human-authored text and d_{bad} comprises corresponding backtranslations. We process each corpus through the model independently to capture the token-level post-residual block hidden states $\mathbf{a}_{l,i,j}$ across all layers l. The direct mean activation vector $\mu_{l,L,d}$ for an attribute dimension $d \in \mathcal{D}$ within language L is accumulated over its respective unpadded token count $N_{L,d}$:

$$\mu_{l,L,d} = \frac{1}{N_{L,d}} \sum_{i,j} \mathbf{a}_{l,i,j} \cdot m_{i,j}$$

To extract a clean signal, the raw quality direction $\mathbf{v}_{l,L,\text{qual}}$ is isolated by subtracting the centroid of the corrupted dataset from the centroid of the human-authored dataset:

$$\mathbf{v}_{l,L,\text{qual}} = \mu_{l,L,d_{\text{good}}} - \mu_{l,L,d_{\text{bad}}}$$

By establishing this explicit pairwise vector difference within the same language, task-neutral semantic properties cancel out, leaving a vector that should isolate the geometric directional shift toward more human-like data.

Following the same stability protocol as our language vectors, we enforce a layer-wise L2 normalization step to yield the final quality steering direction:

$$\hat{\mathbf{v}}_{l,L,\text{qual}} = \frac{\mathbf{v}_{l,L,\text{qual}}}{\|\mathbf{v}_{l,L,\text{qual}}\|_2}$$

E Language Abbreviations

Language abbreviations for each language used in the study can be found in Table 4.

Code	Language
am	Amharic
cs	Czech
de	German
da	Danish
he	Hebrew
id	Indonesian
jv	Javanese
mt	Maltese
sk	Slovak
sl	Slovenian
su	Sundanese

Table 4: Language abbreviations.

F Backtranslation Details

For the backtranslation and creation of parallel synthetic data to human-authored data, we employ the NLLB 600M distilled model [1]. To ensure enough distinction between the human-authored and synthetic data, we use two rounds of backtranslation from the source language to target language 1, then to target language 2, back to target language 1, and then back to the source language. We use English as target language 1 and Chinese as target language 2 as the two target languages, due to their resourcefulness. We also investigated additional corruptions of the synthetic data (e.g., random token swaps), but decided against it given the sufficient linear probe results from Section 4.1.

G Downstream Fine-tuning of XLM-R

For the downstream evaluation, we fine-tune the XLM-R (Conneau et al., 2019) *FacebookAI/xlm-roberta-base* model with a batch size of 16 and employ early stopping with a patience of 10 epochs to prevent overfitting. We perform hyperparameter optimisation to determine the optimal learning rate and set it to 1e-5. *AdamW* is used as an optimiser. We balance the generated datasets to have the same number of samples per label. We perform finetuning 20 times with different seeds for each generated data. We normalise all inputs by converting them to lowercase and removing punctuation.

H Diversity Metrics

To quantitatively evaluate the geometric and linguistic characteristics of the generated text, we employ four distinct diversity metrics. Before metric computation, all generated texts undergo a standardized normalization pipeline, which normalizes text into the NFKD variant, strips diacritics and non-textual symbols (such as emojis), and collapses consecutive whitespaces.

For dense vector calculations, text sequences are mapped into a latent embedding space using the Qwen3-Embedding-4B [2] encoder model.

Lexical diversity: Calculated as the ratio of unique character-level trigrams to total trigrams across the corpus. It captures surface-level variety; higher scores indicate diverse, natural phrasing,

[1] https://huggingface.co/facebook/nllb-200-distilled-600M

[2] https://huggingface.co/Qwen/Qwen3-Embedding-4B

while lower scores signify repetitive or formulaic text.

Embedding diversity: Computed by grouping the dataset by task labels, calculating the mean pairwise cosine distance within each group g, and taking the macro-average across all unique labels G. It measures categorical semantic dispersion, showing whether a model generates a broad range of distinct concepts or collapses into a narrow thematic subspace.

Isocontour radius (Lai et al., 2020): Determines the geometric boundary of the generated dataset by measuring the Euclidean distance of all individual embeddings to the global dataset centroid μ. The metric outputs the 90th percentile of this distance distribution. It represents the total volume of the latent operational envelope; a larger radius indicates that the steering intervention has pushed the generations into wider, peripheral regions of the latent space.

Homogeneity (Lai et al., 2020): Evaluates structural density uniformity using a similarity-based Kernel Density Estimation (KDE). It applies an RBF kernel ($\sigma = 0.5$) to the pairwise cosine similarities to compute a localized density score for each point. It tracks structural consistency across the manifold; a lower score indicates a highly uniform, evenly distributed layout, whereas a higher score signals structural imbalance and tight data clustering.

I Prompt Templates and Generation Details

For generating synthetic data, we used this general prompt template for the **zero-shot** setting:

Topic: *"Generate a long wiki-style sentence on the topic of label in language language. Output only the text in language."*

Sentiment: *"Generate a review with label sentiment in language language. Output only the text and nothing else."*

For the **few-shot** setting, we used 5 shots in the target language and label, with these prompts:

Topic: *"Generate a long wiki style sentence on the topic of label in language language based on examples. Output only the text in language. Examples: examples"*

Sentiment: *"Generate a review with label sentiment in language language based on examples. Output only the text and nothing else. Examples: examples"*

Sampling parameters used for generation were: *temperature=0.8, top_p=0.9, max_tokens=512, freq_penalty=0.0*. We excluded duplicates to ensure unique samples were collected.

Specific versions of LLMs used for generations were: Llama-3.1-70b-instruct [3], Llama-3.1-8b-instruct [4], gemma-2-27b-it [5], gemma-2-9b-it [6]. The LLMs were used with *bfloat16* precision.

J Baseline Results for Zero- and Few-shot Setting

We provide the baseline result for topic and sentiment detection for both zero-shot in Table 6 and for the few-shot setting in Table 7.

K Alpha Values Analysis

We provide an analysis of alpha values in all settings in Table 16 for zero-shot and in Table 17 for the few-shot setting.

For generic language vectors, increasing alpha often behaves like a poison pill. As alpha increases, variance widens significantly, and medians frequently trend downward. *Quality* vectors handle high alpha values much better. Instead of collapsing, higher alphas either monotonically improve performance or reach a safe plateau with tight variances. We see the best performance for alpha values at *50* or *75* for Gemma 2 models and *2* or *3* for Llama3.1 models, indicating that the best downstream model performance comes at moderate-to-strong steering.

Larger models are more brittle to bad steering, as for both Llama 3.1 70b and Gemma 2 27b, choosing the wrong vector (*Language*) or the wrong layer/alpha combination leads to dramatic, sweeping performance drops, most notably for high alpha values in the Llama3.1 70b model. They require precise steering. On the other hand, Llama 3.1 8b and Gemma 2 9b show remarkably clean, uniform positive shifts across almost all alpha values in their early and middle layers, indicating that smaller models might have more centralized "quality pathways" that are easier to uniformly amplify.

[3] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

[4] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[5] https://huggingface.co/google/gemma-2-27b-it

[6] https://huggingface.co/google/gemma-2-9b-it

L Downstream Model Performance Per Task

We provide a per-task breakdown for the zero-shot setting for the topic detection task in Table 8 and for the sentiment detection task in Table 9. For the few-shot setting, the topic detection task can be found in Table 10 and the sentiment detection task in Table 11. All of these tables also contain visualizations showing which specific cases had statistically significant increases in downstream model performance.

In terms of the zero-shot setting, we see similar results to those observed in Section 4.2, as the early layers *Quality* steering vector outperforms other approaches, except for the Llama 3.1 70b model, where *Language* vectors applied to earlier layers slightly outperform it. The best performing method is steering at earlier layers, as it offers an increase in downstream model performance in 79.55% of cases for both topic and sentiment detection. Gemma models see a larger increase in sentiment detection task, while Llama models see a larger increase in downstream model performance in the topic detection task.

For the few-shot setting, again, the *Quality* steering vectors outperform the *Language* steering vectors, in particular at the earlier layers. For topic detection, steering via early layer *Quality* vectors results in increased downstream performance in 70.45% of cases for topic detection and 84.09% of cases for sentiment detection. The noisier sentiment detection seems to benefit more from the "human-likeness" introduced by the *Quality* steering vectors.

M Downstream Model Performance With Alpha Selected Per Layer

We provide results for cases where the best alpha is selected not per case, but per layer of the given LLM generator and steering method. Results can be found in Table 14 for zero-shot and Table 15.

N Detailed Downstream Model Visualizations

We provide detailed visualizations for different languages, LLMs, steering vectors, and layers in Table 18 for zero-shot and in Table 19 for few-shot.

O Detailed Diversity Visualizations

We provide a detailed diversity visualization for each of our 4 metrics and approaches for both the zero-shot baseline and the few-shot baseline in Tables 20 and 21, respectively.

P LLM-as-Judge Setup

For the LLM-as-a-judge setup, we leveraged the Gemma-3-27b-it model [7], with temperature set to 0 and maximum new tokens set to 64. Aggregated results per language and task can be found in Table 5. We used the following prompt:

You are an expert linguistic evaluator. Analyze the following synthetic text sample and grade it strictly according to the defined 1-5 scale for Fluency, Label Adherence, and Language Adherence.

EVALUATION CRITERIA: 1. Fluency: Grammatical correctness and natural flow. 2. Label Adherence: How accurately the text matches the intended target label, prompt, or concept. 3. Language Adherence: How purely the text uses the designated target language and its correct script.

INPUT DATA: - Target Language: language - Target Label/Concept: label description in natural language - Generated Text: sample

Output your scores strictly as a JSON object matching the requested schema. Return ONLY the raw JSON object. Do not include markdown code block formatting or any introductory/concluding text.

[7] https://huggingface.co/google/gemma-3-27b-it

Table 5: Relative performance metric differences in % (Fluency, Language Adherence, Label Adherence) across models, steering methods, and layers, aggregated over tasks and languages compared to the baseline of no steering. Using the best alpha from downstream model performance. Relative positive and negative changes.

(a) **Zero-shot** setup

Model + Steering	Layer	Fluency	LangAdh.	LabelAdh.
Gemma 2 27b (lang.)	L10	0.34	-0.25	-0.04
	L22	0.28	-0.26	0.05
	L34	0.47	0.19	0.10
Gemma 2 27b (qual.)	L10	0.69	0.01	0.31
	L22	0.18	0.09	-0.13
	L34	0.38	0.10	-0.28
Gemma 2 9b (lang.)	L9	-7.47	-0.91	-4.29
	L20	-5.52	-0.23	-3.36
	L31	-2.25	-0.40	-2.10
Gemma 2 9b (qual.)	L9	-1.89	0.41	-2.32
	L20	-0.16	1.55	-1.30
	L31	-3.93	0.86	-3.13
Llama3.1 70b (lang.)	L17	-10.61	-5.54	-8.34
	L37	-0.90	0.31	0.12
	L57	0.17	0.41	0.14
Llama3.1 70b (qual.)	L17	-2.47	-0.44	-1.88
	L37	-0.62	-0.54	-0.48
	L57	0.91	0.20	0.76
Llama3.1 8b (lang.)	L7	4.84	0.38	-0.05
	L15	-2.13	-1.02	-2.67
	L23	2.35	0.52	-0.66
Llama3.1 8b (qual.)	L7	1.96	-0.52	0.37
	L15	4.38	-0.06	3.53
	L23	5.96	0.16	3.72

(b) **Few-shot** setup

Model + Steering	Layer	Fluency	LangAdh.	LabelAdh.
Gemma 2 27b (lang.)	L10	0.84	-0.08	0.85
	L22	0.70	-0.09	-0.00
	L34	0.36	-0.35	-0.17
Gemma 2 27b (qual.)	L10	-0.81	-0.29	-0.56
	L22	-0.43	-0.14	-0.40
	L34	-0.51	-0.32	0.07
Gemma 2 9b (lang.)	L9	-4.76	-1.11	-3.07
	L20	-6.01	-0.36	-5.28
	L31	-2.46	0.00	-3.03
Gemma 2 9b (qual.)	L9	-3.80	-0.72	-3.20
	L20	-1.83	0.88	-2.73
	L31	-4.61	-0.11	-3.45
Llama3.1 70b (lang.)	L17	-10.42	-3.41	-7.64
	L37	-0.23	0.12	-0.70
	L57	0.37	0.25	-0.36
Llama3.1 70b (qual.)	L17	-1.73	-0.39	-2.03
	L37	-1.16	-0.44	-1.50
	L57	0.69	0.01	0.48
Llama3.1 8b (lang.)	L7	-6.18	0.38	-9.39
	L15	-10.27	-2.14	-5.66
	L23	-5.78	-0.51	-3.71
Llama3.1 8b (qual.)	L7	-2.83	-0.75	-2.39
	L15	-0.98	-0.19	1.18
	L23	0.14	-0.48	0.66

Table 6: Mean F1 scores across languages and tasks for baseline **zero-shot** setting with no steering.

LLM	Language Task	am	cs	da	de	he	id	jv	mt	sk	sl	su	Average
Gemma 2 27b	Sentiment	74.53	83.55	95.27	64.10	75.69	84.07	45.47	54.82	69.53	73.23	67.35	71.60
	Topic	48.28	59.26	68.93	64.82	62.99	65.99	60.43	41.33	64.46	63.84	61.01	60.12
Gemma 2 9b	Sentiment	79.52	80.78	94.97	71.50	71.98	87.07	62.53	53.12	66.06	68.64	73.58	73.62
	Topic	49.05	61.48	64.54	62.06	60.26	66.35	60.19	38.14	60.98	63.56	62.40	59.00
Llama3.1 70b	Sentiment	77.21	82.94	91.72	68.91	74.46	86.87	61.79	57.17	71.04	79.69	74.29	75.10
	Topic	50.21	75.82	76.57	73.54	67.18	70.16	49.64	39.01	71.93	76.71	61.85	64.78
Llama3.1 8b	Sentiment	53.66	85.10	94.05	71.18	76.77	87.93	62.10	51.78	82.53	81.89	78.88	75.08
	Topic	33.82	71.06	71.55	75.84	71.24	68.04	62.72	46.41	71.65	69.79	65.18	64.30

Table 7: Mean F1 scores across languages and tasks for baseline **few-shot** setting with no steering.

LLM	Language Task	am	cs	da	de	he	id	jv	mt	sk	sl	su	Average
Gemma 2 27b	Sentiment	74.93	68.14	91.64	61.12	58.20	80.64	60.01	59.25	75.24	59.76	80.36	69.94
	Topic	56.70	76.64	75.06	79.39	69.56	74.96	67.42	48.30	76.30	73.02	66.14	69.41
Gemma 2 9b	Sentiment	81.69	75.47	87.32	59.17	55.99	87.20	54.91	57.81	72.60	82.43	78.87	72.13
	Topic	56.14	76.33	74.82	76.07	73.53	74.62	62.76	51.02	77.48	74.51	68.82	69.65
Llama3.1 70b	Sentiment	82.72	83.30	94.68	56.36	70.67	77.89	56.34	57.13	65.10	73.90	74.65	72.07
	Topic	62.65	77.92	75.50	80.55	74.85	81.03	69.51	48.74	78.90	76.78	68.74	72.29
Llama3.1 8b	Sentiment	76.53	81.48	85.74	69.01	78.52	87.20	56.67	55.94	80.13	72.16	80.24	74.87
	Topic	52.06	79.78	76.74	76.04	74.93	76.01	69.14	48.15	77.94	76.30	68.11	70.47

Table 8: Mean F1 difference across languages, models, steering methods, and layers for **topic** detection, with positive and negative differences compared to **zero-shot** prompt with no steering. Cells with * denote statistically significant increases over baseline.

Method	LLM	Layer	am	cs	da	de	he	id	jv	mt	sk	sl	su	Average
Language	Gemma 27b	L10	3.35*	1.30	-1.23	1.31	-0.13	3.69*	-0.81	1.12	0.81	0.01	0.74	0.92
		L22	4.32*	1.22	0.21	1.24	0.47	2.14	-0.76	2.65	1.40	1.28	0.29	1.31
		L34	6.49*	0.84	0.65	0.07	1.40	3.36*	-1.54	0.23	1.51	0.45	1.69	1.38
	Gemma 9b	L9	-4.49	5.34*	4.24*	2.80	8.46*	-0.27	-2.92	5.03*	6.79*	4.82*	0.19	2.73
		L20	-4.59	1.77	5.19*	2.26*	3.90*	1.03	0.91	4.08*	4.32*	1.04	0.62	1.87
		L31	4.84*	2.69*	4.14*	2.74*	1.68	0.88	0.53	4.41*	3.20*	2.21	0.76	2.55
	Llama3 70b	L17	2.24	-2.07	-0.28	3.54	0.36	3.82*	6.97*	1.70	7.77*	-0.74	-3.17	1.83
		L37	4.78*	-0.13	1.23	4.38*	1.65	2.17	10.96*	0.78	3.42*	1.31	-0.38	2.74
		L57	3.09	-1.83	1.04	2.51	1.02	3.42	8.96*	0.51	0.20	2.61	1.67	2.11
	Llama3 8b	L7	10.79*	0.87	1.39	0.73	0.73	2.95	4.97*	4.06*	4.79*	4.67*	0.67	3.33
		L15	13.84*	5.77*	5.34*	2.69	0.08	9.30*	1.28	1.25	3.78*	6.06*	-1.81	4.33
		L23	7.99*	4.39*	5.31*	0.34	-2.68	3.32*	0.43	2.49	6.00*	4.88*	-0.47	2.91
Quality	Gemma 27b	L10	4.66*	1.18	0.14	1.77	1.00	1.70	0.32	2.40*	1.03	1.30	0.68	1.47
		L22	4.00*	0.44	-0.86	1.62	1.23	2.82*	-0.91	3.90*	1.27	2.47	0.24	1.47
		L34	5.17*	-0.24	-0.84	1.68	0.91	2.35*	-0.24	0.23	2.04	0.87	-0.28	1.06
	Gemma 9b	L9	0.44	10.70*	4.39*	4.82*	4.19*	1.32	0.80	2.83*	6.95*	8.54*	4.15*	4.47
		L20	2.68*	2.64*	6.40*	5.08*	1.94	4.11*	0.87	3.91*	3.21*	6.96*	2.35	3.65
		L31	2.27	3.89*	4.37*	4.01*	0.85	2.15	-0.49	4.23*	13.41*	9.07*	0.64	4.04
	Llama3 70b	L17	8.35*	-0.62	0.41	2.37	0.47	1.07	3.42*	1.31	3.57*	1.58	-0.94	1.91
		L37	4.76*	-1.67	0.28	-0.80	1.12	2.74*	-2.24	-0.85	4.28*	0.77	-2.42	0.54
		L57	6.17*	-0.44	1.57*	1.91	0.15	1.12	1.66	1.74	-0.63	0.95	-3.75	0.95
	Llama3 8b	L7	11.66*	5.31*	6.77*	1.26	1.52	3.71*	3.47*	5.89*	5.13*	5.45*	0.76	4.63
		L15	7.69*	2.80*	3.65*	-0.71	0.62	3.06*	2.56*	2.60	4.01*	3.63*	2.19*	2.92
		L23	9.53*	4.39*	3.24*	0.36	-5.22	4.72*	1.65	1.46	4.36*	4.44*	1.74*	2.79

Table 9: Mean F1 difference across languages, models, steering methods, and layers for **sentiment** detection, with positive and negative differences compared to a **zero-shot** prompt with no steering. Cells with * denote statistically significant increases over baseline.

Method	LLM	Layer	am	cs	da	de	he	id	jv	mt	sk	sl	su	Average
Language	Gemma 27b	L10	2.13	0.65	-0.08	6.11*	-0.47	1.72	4.51	-0.22	6.59*	6.12	12.98*	3.64
		L22	2.12	0.28	-0.37	0.79	0.26	1.38	0.10	-0.32	6.39*	2.74	9.94*	2.12
		L34	2.20	0.88	-0.20	1.36	0.26	1.98	7.11*	-2.43	4.96	5.68*	10.71*	2.96
	Gemma 9b	L9	0.67	3.21	-0.09	1.17	3.13	0.82	0.41	2.26	6.62*	14.21*	5.03*	3.40
		L20	1.43	5.41*	-0.37	-1.19	1.35	0.73	-0.50	2.69	11.72*	10.92*	7.30*	3.59
		L31	1.59	0.73	0.20	0.23	4.30*	0.51	0.27	1.51	9.92*	10.34*	4.27*	3.08
	Llama3 70b	L17	4.71*	0.83	2.82*	1.71	2.29	1.66	0.09	2.11	11.89*	7.41*	2.26	3.43
		L37	4.51*	2.72*	2.52*	2.84*	3.15*	0.52	-0.11	0.28	11.27*	3.65	6.30*	3.42
		L57	3.78*	2.52*	2.50*	2.72	2.62*	1.62	-0.58	2.47	7.63*	4.94*	7.53*	3.43
	Llama3 8b	L7	6.21*	0.60	0.97	2.23	0.89	1.99	-0.03	4.91*	3.24*	3.36*	-0.95	2.13
		L15	4.73*	0.10	1.45	1.85	-0.17	0.25	0.34	7.24*	3.90*	3.13	-0.53	2.03
		L23	13.12*	0.00	1.01	-1.59	0.42	0.22	-0.58	3.68*	3.71*	0.44	0.75	1.93
Quality	Gemma 27b	L10	1.37	0.54	-0.18	3.66	0.86	2.90	2.44	0.40	5.01	6.54*	10.63*	3.11
		L22	0.13	0.69	-0.24	0.33	1.20	1.89	4.49	-0.61	5.88*	4.51*	10.46*	2.61
		L34	0.46	0.77	-0.25	1.23	-0.52	2.27	7.11*	0.34	7.31*	4.67	8.84*	2.93
	Gemma 9b	L9	0.23	5.37 *	-0.02	2.98*	5.08*	2.33*	-0.28	0.70	16.60*	11.95*	6.09*	4.64
		L20	4.00*	4.68*	0.21	0.31	0.73	1.21	0.17	1.72	10.42*	3.88	6.44*	3.07
		L31	2.24	4.56*	0.06	0.18	2.77	1.38	-0.03	2.32	13.74*	0.85	5.38*	3.04
	Llama3 70b	L17	2.48*	3.15*	1.64	1.36	2.64*	0.94	-1.26	3.62*	11.04*	2.20	6.84*	3.15
		L37	4.90*	2.56*	0.91	2.17	1.94	0.68	-1.43	5.10*	10.74*	2.65	7.88*	3.47
		L57	3.75*	2.78*	1.06	3.95*	1.70	0.71	-1.14	2.31	12.79*	2.42	5.55*	3.26
	Llama3 8b	L7	14.60*	0.63	1.01*	0.81	1.00*	-0.45	0.11	4.88*	3.48*	3.46*	0.89	2.77
		L15	15.02*	1.17*	0.68	0.42	-0.13	0.67	0.64	7.31*	2.66*	2.07*	0.79	2.85
		L23	12.16*	1.07*	0.47	0.28	-0.07	0.64	-0.18	6.73*	2.46*	1.82	1.11	2.41

Table 10: Mean F1 difference across languages, models, steering methods, and layers for **topic** detection, with positive and negative differences compared to a **few-shot** prompt with no steering. Cells with * denote statistically significant increases over baseline.

Method	LLM	Layer	am	cs	da	de	he	id	jv	mt	sk	sl	su	Average
Language	Gemma 27b	L10	0.79	1.67	2.27	-0.60	0.11	2.29	-0.39	2.85*	1.73	3.75*	-0.12	1.30
		L22	-1.95	1.27	1.23	0.15	3.00*	1.09	-0.53	3.11*	0.20	2.49*	0.26	0.94
		L34	0.94	1.54	2.49	0.03	1.15	1.06	-0.20	2.89	0.40	3.40*	-0.25	1.22
	Gemma 9b	L9	-3.51	-0.67	2.92*	3.36*	0.05	0.09	2.82	0.82	-0.88	1.16	-1.92	0.39
		L20	0.34	1.90	4.08*	1.36	0.87	1.23	2.30*	-0.60	0.54	1.11	-1.21	1.08
		L31	-1.10	1.98	2.49	3.86*	-0.14	1.14	2.16	2.07	0.05	2.01	-0.22	1.30
	Llama3 70b	L17	3.34*	-1.02	-0.60	-1.43	2.19	0.92	3.42*	-0.72	0.63	-0.17	0.51	0.64
		L37	2.14	2.12	2.28	-0.17	2.65*	0.00	2.89*	2.95*	1.62	-0.05	4.27*	1.88
		L57	2.70	1.58	3.27*	0.31	3.56*	0.04	3.27*	0.87	2.76	0.04	4.32 *	2.06
	Llama3 8b	L7	2.20	0.64	-1.20	3.71*	1.42	1.97	0.65	3.69*	-2.29	3.01	0.70	1.32
		L15	2.98*	1.10	-1.90	2.97	-0.42	3.92*	2.03	2.70	2.08	2.96	0.41	1.71
		L23	2.39	2.61	0.51	4.02*	0.57	4.20*	4.42*	2.29	1.07	2.81	1.44	2.39
Quality	Gemma 27b	L10	-1.00	0.95	2.18	-0.36	3.05*	1.28	-0.06	2.93*	0.93	3.87*	-0.32	1.22
		L22	4.22*	2.62	2.23	0.74	2.51	1.65	-0.34	3.73*	0.70	3.23	-0.19	1.92
		L34	0.60	1.28	1.60	-0.98	1.83	2.06	-1.35	4.76*	0.91	2.67	-0.09	1.21
	Gemma 9b	L9	1.41	1.64	2.47	3.70*	0.76	3.05*	2.51	2.91*	-0.11	2.18	-0.30	1.84
		L20	-1.11	1.52	2.53	3.10*	0.07	2.11	2.95	0.07	0.08	1.91	3.71*	1.54
		L31	-0.72	1.51	2.44	3.37*	-0.08	1.33	0.99	-2.91	0.86	0.98	-0.29	0.68
	Llama3 70b	L17	5.20*	3.40*	1.79	0.45	2.47	0.87	1.78	3.84*	2.54	2.97	3.81*	2.65
		L37	3.72*	0.65	1.93	0.79	1.28	-0.84	3.06*	-0.32	1.06	1.92	4.22*	1.59
		L57	0.95	0.75	1.55	0.19	3.27*	-0.21	2.73	2.04	2.84*	1.82	4.24*	1.83
	Llama3 8b	L7	3.54*	2.97	2.82	4.06*	1.34	1.22	1.22	4.02*	0.19	0.65	0.14	2.02
		L15	2.21	1.07	-1.46	2.26	0.80	2.42	1.01	5.30*	-0.67	3.26*	-0.52	1.43
		L23	7.47*	0.50	-2.28	2.80	-0.84	2.04	0.91	4.30*	-0.87	3.82*	-0.05	1.62

Table 11: Mean F1 difference across languages, models, steering methods, and layers for **sentiment** detection, with positive and negative differences compared to a **few-shot** prompt with no steering. Cells with * denote statistically significant increases over baseline.

Method	LLM	Layer	am	cs	da	de	he	id	jv	mt	sk	sl	su	Average
Language	Gemma 27b	L10	2.36	3.30*	1.81	4.42*	7.30*	4.66*	1.22	3.89*	-0.75	6.43*	-0.47	3.11
		L22	0.68	7.76*	-0.49	3.57*	0.75	2.24	0.49	2.26	-0.60	9.00*	0.85	2.41
		L34	1.00	7.40*	0.94	3.42	0.45	3.52*	1.13	1.21	3.30	4.48*	0.57	2.49
	Gemma 9b	L9	-3.33	9.28*	0.49	1.89	2.39	-3.06	2.85	3.03*	11.69*	-3.48	0.95	2.06
		L20	-0.49	7.68*	7.17*	2.04	12.78*	1.88	5.85*	-0.44	6.18*	1.66	2.13	4.22
		L31	-0.31	2.51	6.64*	1.27	3.25*	-0.89	6.66*	-0.02	6.96*	-0.88	1.27	2.41
	Llama3 70b	L17	0.85	0.79	0.34	8.93*	4.91*	9.66*	-2.66	2.95	12.71*	2.13	3.57*	4.02
		L37	-1.50	0.62	0.41	2.68	6.69*	6.19*	-0.94	3.52*	5.68*	1.68	4.70*	2.70
		L57	-0.25	1.31	0.08	2.58	5.27*	3.40*	0.09	1.98	0.71	3.43*	4.20*	2.07
	Llama3 8b	L7	1.04	3.37*	8.05*	0.82	0.70	1.62*	-1.59	2.21*	1.25	0.60	-1.88	1.47
		L15	-0.20	3.36*	6.21*	0.53	0.22	1.00	0.66	2.99*	4.83*	1.65	-0.87	1.85
		L23	1.40	3.07*	4.32*	0.37	0.29	1.01	0.15	2.18	4.76*	1.79	0.18	1.77
Quality	Gemma 27b	L10	0.74	3.52*	2.86	2.86	1.38	3.24*	0.25	4.43*	3.69*	6.78*	1.72	2.86
		L22	4.29*	5.52*	1.38	3.38	1.90	4.38*	-0.57	2.21	1.82	8.37*	1.09	3.07
		L34	1.74	6.14*	0.85	3.18*	5.48*	1.63	1.87	2.34	2.69*	4.28*	1.43	2.87
	Gemma 9b	L9	0.06	4.85*	3.04	4.19*	17.06*	-0.86	5.54*	3.47*	4.28*	0.72	3.30	4.15
		L20	0.43	1.62	1.99	0.21	15.16*	0.52	0.09	1.29	3.17*	-1.95	0.90	2.13
		L31	0.35	2.91*	5.10*	1.23	8.23*	0.42	0.05	0.62	7.89*	-0.26	2.35	2.63
	Llama3 70b	L17	-2.10	-0.10	0.48	3.49*	8.36*	5.18*	1.22	3.09*	10.49*	1.36	3.46*	3.17
		L37	-1.81	1.57	0.38	3.01	4.97*	2.16	0.07	0.18	4.72*	2.76	5.32*	2.12
		L57	-1.13	2.18	0.40	1.56	1.51	3.32*	-0.20	2.83*	2.65	1.49	6.72*	1.94
	Llama3 8b	L7	1.23	1.38	5.08*	0.67	0.49	0.79	0.02	2.24*	1.40	0.44	0.60	1.30
		L15	0.77	0.78	2.09*	1.23	0.16	0.52	-1.46	1.12	-1.50	-3.33	-0.05	0.03
		L23	2.32*	1.50*	0.49	1.31	-0.14	0.33	-0.46	1.65	0.36	0.17	0.67	0.75

Table 12: Selected best α values based on validation sets for each steering method, model, layer, language, and task for the **zero-shot** setting.

		Task						sent												topic				
Method	LLM	Language	am	cs	da	de	he	id	jv	mt	sk	sl	su	am	cs	da	de	he	id	jv	mt	sk	sl	su
Language	Gemma 27b	L10	75	75	25	75	25	25	100	25	50	100	75	50	75	25	100	50	100	50	25	25	75	75
		L22	75	100	25	75	50	50	100	50	25	50	75	100	100	75	25	75	75	100	75	25	25	50
		L34	75	25	100	25	75	25	75	25	50	50	75	75	75	100	50	100	75	100	25	25	100	75
	Gemma 9b	L9	25	75	50	100	50	75	25	75	75	100	100	50	100	25	75	75	25	25	100	100	100	100
		L20	50	100	100	100	75	100	25	50	25	100	75	25	75	50	25	100	25	75	25	75	25	50
		L31	75	100	100	75	100	75	25	25	75	25	25	25	100	75	50	100	75	100	25	100	100	25
	Llama3 70b	L17	2	1	1	2	2	1	2	3	3	3	2	1	1	1	2	1	1	3	3	2	1	1
		L37	1	3	2	1	1	3	3	3	2	3	2	4	2	3	3	2	2	4	3	3	1	4
		L57	3	1	1	4	2	2	3	2	1	2	2	2	1	2	1	2	4	4	2	2	3	4
	Llama3 8b	L7	2	3	2	4	3	4	1	4	2	1	2	1	3	2	2	1	3	3	3	2	1	1
		L15	4	4	2	1	4	4	2	3	3	2	1	1	2	4	4	1	3	2	4	4	3	2
		L23	2	2	2	1	1	1	1	1	3	4	1	1	4	4	1	2	1	1	4	1	4	3
Quality	Gemma 27b	L10	100	75	100	100	75	25	75	100	75	100	100	100	50	100	75	50	25	25	50	25	75	75
		L22	50	75	100	100	75	50	100	25	25	100	25	50	25	75	75	100	50	25	100	75	75	50
		L34	75	75	100	50	25	75	100	75	50	50	75	75	100	25	50	100	100	50	100	75	50	75
	Gemma 9b	L9	50	50	25	100	100	75	50	50	100	75	50	75	100	50	50	75	25	75	25	75	25	75
		L20	50	100	50	100	75	25	25	75	25	50	75	50	25	50	50	25	100	75	50	100	100	25
		L31	75	100	100	100	50	100	25	50	100	75	50	75	75	50	75	50	75	75	75	100	25	100
	Llama3 70b	L17	3	4	3	1	4	3	1	3	4	1	3	4	2	1	4	4	2	4	4	3	4	4
		L37	1	3	2	3	1	3	4	1	4	1	4	3	4	2	2	2	4	3	2	3	1	3
		L57	3	2	4	1	3	4	4	1	3	3	2	2	4	2	2	1	2	1	2	2	3	2
	Llama3 8b	L7	1	2	4	2	2	1	2	4	3	1	2	3	3	3	4	3	4	2	4	1	2	3
		L15	3	1	2	1	1	4	1	3	1	3	1	2	4	2	2	2	2	2	4	1	4	3
		L23	2	2	2	1	2	3	3	4	2	1	3	2	4	3	3	1	4	4	2	2	2	2

Table 13: Selected best α values based on validation sets for each steering method, model, layer, language, and task for the **few-shot** setting.

		Task						sent												topic				
Method	LLM	Language	am	cs	da	de	he	id	jv	mt	sk	sl	su	am	cs	da	de	he	id	jv	mt	sk	sl	su
Language	Gemma 27b	L10	25	25	75	25	25	75	100	50	25	75	50	25	25	75	50	50	100	50	50	100	75	100
		L22	50	25	50	25	100	75	25	100	50	50	25	50	25	75	100	50	25	75	75	50	100	100
		L34	25	75	25	50	100	25	100	100	75	100	25	25	100	75	100	25	25	50	75	25	100	75
	Gemma 9b	L9	50	100	25	100	50	25	50	75	100	25	25	50	75	50	25	75	50	50	50	50	25	25
		L20	25	100	75	50	100	100	25	100	50	75	25	50	100	75	25	100	50	100	25	50	50	75
		L31	25	25	50	25	25	50	25	25	50	100	75	50	75	75	100	75	75	100	75	25	100	50
	Llama3 70b	L17	1	1	1	3	3	3	2	4	3	1	1	2	2	1	1	2	2	1	1	1	2	1
		L37	1	1	4	3	1	3	4	1	4	1	4	2	2	2	3	3	2	1	2	1	2	2
		L57	2	2	1	3	3	1	3	1	3	4	3	1	1	2	1	1	2	4	3	2	3	1
	Llama3 8b	L7	3	4	3	4	2	4	3	2	4	2	1	1	3	2	2	4	2	1	2	2	2	1
		L15	2	4	3	2	1	4	3	2	4	2	1	2	3	4	3	4	3	2	1	4	3	2
		L23	1	4	1	1	1	3	3	1	4	3	1	2	4	3	3	3	3	4	2	1	2	4
Quality	Gemma 27b	L10	75	50	75	75	25	75	50	50	75	75	25	100	75	75	75	50	75	75	75	75	50	25
		L22	100	25	50	75	25	75	25	25	75	100	50	25	75	25	75	25	50	100	25	50	100	100
		L34	100	25	25	100	100	100	75	100	100	100	75	75	25	100	75	50	50	50	50	100	100	100
	Gemma 9b	L9	25	100	25	100	100	25	100	25	100	50	50	25	25	25	50	25	50	75	25	75	50	100
		L20	50	75	75	75	100	25	25	25	75	25	50	75	100	25	100	75	25	75	50	25	25	100
		L31	75	100	75	25	100	75	25	50	100	75	50	50	75	50	50	100	50	50	75	50	100	100
	Llama3 70b	L17	1	1	1	1	4	4	1	4	4	2	4	1	3	2	2	1	4	3	2	1	3	3
		L37	1	1	2	3	1	2	4	3	3	4	1	2	3	1	3	3	2	4	1	4	4	1
		L57	1	2	4	2	3	3	2	2	3	4	2	4	4	4	3	2	4	1	2	3	4	2
	Llama3 8b	L7	2	1	2	1	1	2	1	2	4	3	4	3	2	4	1	4	2	3	2	4	3	2
		L15	3	3	4	2	3	1	1	1	1	2	2	3	2	3	1	1	1	1	3	2	3	3
		L23	2	3	3	2	1	1	1	4	1	2	3	4	4	2	4	2	2	1	4	2	1	4

Table 14: Mean F1 difference across languages, models, steering methods, and layers aggregated over tasks, with positive and negative differences compared to a **zero-shot prompt** with no steering. Best alpha selected per layer as the sum of differences.

Method	LLM	Layer	am	cs	da	de	he	id	jv	mt	sk	sl	su	Average
Language	Gemma 2 27b	L10 ($\alpha = 100.0$)	0.17	0.28	-2.10	2.24	-1.16	2.49	1.84	-0.56	1.52	2.66	2.33	0.88
		L22 ($\alpha = 75.0$)	2.36	-0.22	-0.13	0.73	-0.52	0.77	-2.67	0.38	1.83	0.07	4.92	0.68
		L34 ($\alpha = 75.0$)	4.34	-0.51	-0.43	-1.37	-0.36	2.17	1.96	-1.71	-1.72	-0.07	6.20	0.77
	Gemma 2 9b	L9 ($\alpha = 50.0$)	-3.25	1.48	1.09	0.43	2.29	0.17	-3.63	0.45	4.08	6.88	0.16	0.92
		L20 ($\alpha = 50.0$)	-1.72	-0.25	1.88	-1.04	-3.54	-0.37	-0.86	2.75	6.32	3.76	3.28	0.93
		L31 ($\alpha = 25.0$)	2.30	-0.70	1.69	0.50	1.73	-0.46	-2.22	2.96	-0.59	5.62	2.52	1.21
	Llama3.1 70b	L17 ($\alpha = 2.0$)	3.25	-3.03	-0.91	2.62	0.98	1.19	2.42	0.79	9.67	1.27	-0.57	1.61
		L37 ($\alpha = 3.0$)	3.34	-0.75	1.39	3.43	0.66	0.97	4.10	0.53	4.82	2.22	0.85	1.96
		L57 ($\alpha = 2.0$)	2.97	-0.94	1.69	0.44	1.82	1.89	-0.97	1.49	1.87	3.35	1.56	1.38
	Llama3.1 8b	L7 ($\alpha = 2.0$)	8.39	-1.20	1.18	0.77	-0.77	1.52	-0.43	1.68	4.01	3.01	-0.24	1.63
		L15 ($\alpha = 1.0$)	4.52	1.25	2.13	1.52	-0.38	-0.62	0.57	0.38	2.32	2.11	-2.48	1.03
		L23 ($\alpha = 1.0$)	6.43	1.16	1.91	-0.63	-2.51	1.77	-0.08	2.03	4.36	0.69	-0.59	1.32
Quality	Gemma 2 27b	L10 ($\alpha = 100.0$)	3.01	0.16	-0.02	2.32	-0.91	0.49	-0.80	0.67	1.92	3.66	4.91	1.40
		L22 ($\alpha = 100.0$)	-0.46	-0.31	-1.54	-0.39	0.05	0.98	1.79	1.34	1.04	1.18	3.44	0.65
		L34 ($\alpha = 50.0$)	-0.32	-0.35	-1.03	1.45	-0.60	0.43	-0.94	-1.82	3.83	2.77	3.85	0.66
	Gemma 2 9b	L9 ($\alpha = 75.0$)	-0.01	5.20	1.29	1.75	4.62	1.71	-0.10	1.28	11.70	9.12	4.51	3.73
		L20 ($\alpha = 50.0$)	3.34	0.36	3.31	2.12	-0.14	1.54	-0.42	1.47	1.57	5.04	1.60	1.80
		L31 ($\alpha = 75.0$)	2.25	3.86	1.08	0.88	-0.06	0.89	-0.96	2.05	9.07	1.86	0.72	1.97
	Llama3.1 70b	L17 ($\alpha = 3.0$)	2.58	0.59	0.37	0.25	0.65	0.54	-1.24	2.31	7.04	-0.44	1.65	1.30
		L37 ($\alpha = 1.0$)	4.06	-1.74	-1.65	-1.01	1.44	0.83	-4.45	1.75	4.64	1.71	-0.26	0.48
		L57 ($\alpha = 2.0$)	4.47	0.32	0.72	0.89	0.56	0.14	-1.09	1.30	2.43	-0.29	0.90	0.94
	Llama3.1 8b	L7 ($\alpha = 2.0$)	7.49	2.91	2.08	0.77	0.82	1.45	1.79	3.20	3.14	3.03	-0.11	2.42
		L15 ($\alpha = 3.0$)	9.19	0.64	1.05	-1.57	-1.00	-0.63	0.47	4.06	1.01	1.62	1.20	1.46
		L23 ($\alpha = 2.0$)	10.85	2.01	1.17	-0.35	-3.06	1.37	-0.49	3.18	3.41	2.25	0.20	1.87

Table 15: Mean F1 difference across languages, models, steering methods, and layers aggregated over tasks, with positive and negative differences compared to a **few-shot prompt** with no steering. Best alpha selected per layer as the sum of differences.

Method	LLM	Layer	am	cs	da	de	he	id	jv	mt	sk	sl	su	Average
Language	Gemma 2 27b	L10 ($\alpha = 25.0$)	1.58	2.48	0.63	1.73	3.25	1.36	-0.96	2.27	-0.39	3.55	-1.76	1.25
		L22 ($\alpha = 50.0$)	-0.64	2.48	0.00	-0.59	1.85	1.04	-1.19	0.00	-0.20	5.51	-1.35	0.63
		L34 ($\alpha = 25.0$)	0.97	2.58	0.34	0.35	-0.26	2.29	-0.69	1.66	-0.75	2.32	-0.87	0.72
	Gemma 2 9b	L9 ($\alpha = 25.0$)	-4.14	0.34	0.24	1.48	-0.54	-1.54	0.51	-1.97	2.06	-1.16	-0.49	-0.47
		L20 ($\alpha = 50.0$)	-0.17	0.96	1.19	0.87	-0.02	-0.21	0.97	-2.13	3.36	-0.47	-1.45	0.26
		L31 ($\alpha = 50.0$)	-2.35	0.25	3.99	-0.77	0.02	0.03	0.20	-1.00	2.44	-1.08	0.49	0.20
	Llama3.1 70b	L17 ($\alpha = 1.0$)	1.85	-0.23	-0.13	0.91	1.32	-0.08	0.32	0.30	-1.20	0.43	2.04	0.50
		L37 ($\alpha = 2.0$)	-0.70	1.28	0.89	0.37	3.46	1.26	0.08	2.21	1.61	0.51	2.85	1.26
		L57 ($\alpha = 1.0$)	0.63	0.68	0.82	0.93	2.96	1.41	-0.24	0.21	0.06	1.42	3.91	1.16
	Llama3.1 8b	L7 ($\alpha = 2.0$)	-1.23	-0.24	2.50	1.23	0.33	1.49	-1.66	2.95	-1.21	1.80	-2.25	0.34
		L15 ($\alpha = 2.0$)	1.39	0.12	-1.09	1.32	-1.07	-0.63	0.70	2.36	-0.38	1.83	-0.90	0.33
		L23 ($\alpha = 3.0$)	-0.14	-0.07	1.77	1.73	0.35	2.61	0.70	0.13	-0.17	1.10	-1.28	0.61
Quality	Gemma 2 27b	L10 ($\alpha = 75.0$)	-0.45	2.18	2.52	1.25	0.72	2.26	-0.28	2.11	2.31	4.73	-1.30	1.46
		L22 ($\alpha = 75.0$)	0.04	3.68	-0.63	2.06	-0.45	2.66	-0.95	0.81	0.99	3.32	-1.72	0.89
		L34 ($\alpha = 25.0$)	0.97	2.58	0.34	0.35	-0.26	2.29	-0.69	1.66	-0.75	2.32	-0.87	0.72
	Gemma 2 9b	L9 ($\alpha = 25.0$)	0.74	0.87	2.75	-0.20	1.75	0.37	1.57	3.19	0.61	0.78	-0.09	1.12
		L20 ($\alpha = 25.0$)	-3.42	-0.72	1.07	-0.05	0.01	1.31	1.13	0.48	-0.24	-0.02	-1.92	-0.22
		L31 ($\alpha = 50.0$)	-1.51	-0.45	2.68	1.31	0.27	0.62	-2.70	-1.20	3.75	-0.67	-0.11	0.18
	Llama3.1 70b	L17 ($\alpha = 1.0$)	1.55	0.68	0.31	1.45	2.83	0.74	1.37	-0.46	1.79	1.12	2.51	1.26
		L37 ($\alpha = 1.0$)	-0.70	0.70	0.80	-0.03	2.84	-0.58	-0.28	-0.44	0.42	-0.09	4.77	0.67
		L57 ($\alpha = 2.0$)	-1.40	1.02	0.46	-0.90	1.25	-0.52	0.45	2.44	1.89	0.43	5.48	0.96
	Llama3.1 8b	L7 ($\alpha = 2.0$)	1.03	1.82	1.70	1.03	0.72	1.00	-1.36	3.13	0.04	-0.67	0.05	0.77
		L15 ($\alpha = 3.0$)	1.49	0.32	-0.08	1.56	-1.37	-0.03	-2.39	3.06	-6.14	-0.39	-0.90	-0.44
		L23 ($\alpha = 4.0$)	4.05	-0.43	-1.72	1.49	-1.53	0.71	-1.08	2.97	-2.02	-0.71	0.20	0.17

Table 16: Boxplot visualizations of relative increases/decreases in % for F1 metrics for LLM + steering methods for alpha values in **zero-shot** setting. The X axis in the figures represents the different alpha values for that LLM + steering combination aggregated over languages and tasks, and the Y axis represents the relative difference increase for F1 downstream model performance.

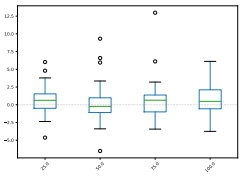
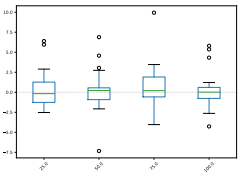
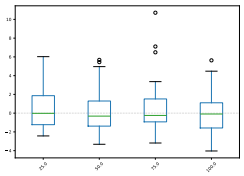
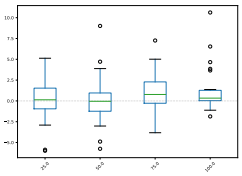
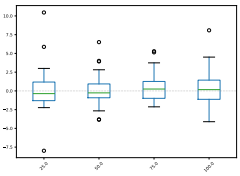
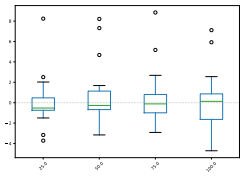
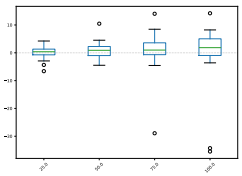
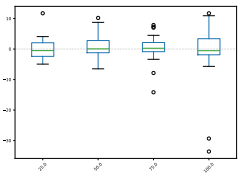
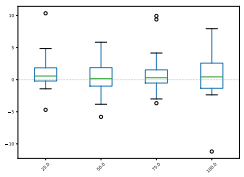
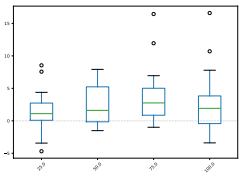
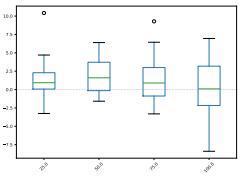
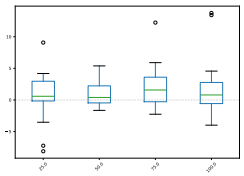
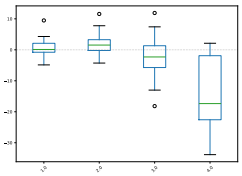
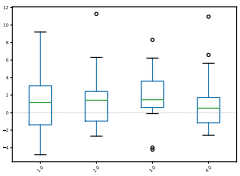
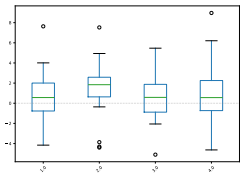
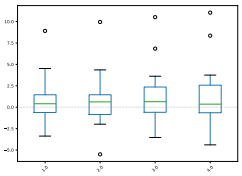
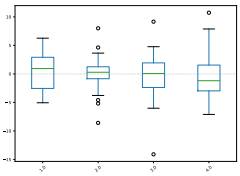
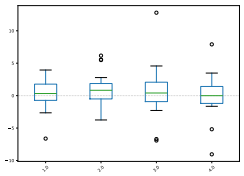
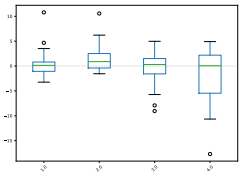
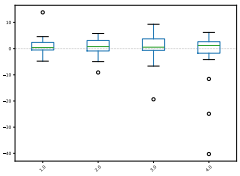
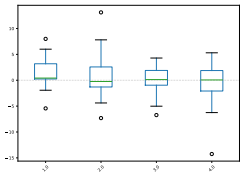
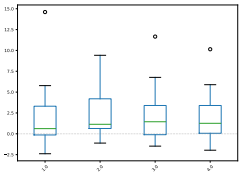
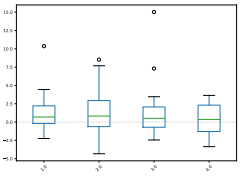
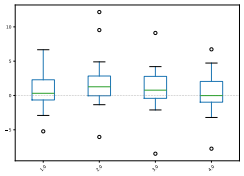
LLM + steering	Early layer	Middle layer	Late layer
Gemma 2 27b (lang.)			
Gemma 2 27b (qual.)			
Gemma 2 9b (lang.)			
Gemma 2 9b (qual.)			
Llama 3.1 70b (lang.)			
Llama 3.1 70b (qual.)			
Llama 3.1 8b (lang.)			
Llama 3.1 8b (qual.)			

Table 17: Boxplot visualizations of relative increases/decreases in % for F1 metrics for LLM + steering methods for alpha values in **few-shot** setting. The X axis in the figures represents the different alpha values for that LLM + steering combination aggregated over languages and tasks, and the Y axis represents the relative difference increase for F1 downstream model performance.

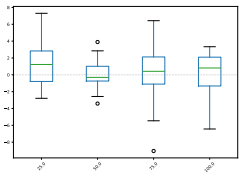
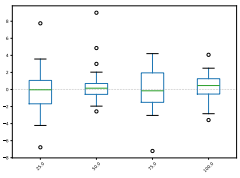
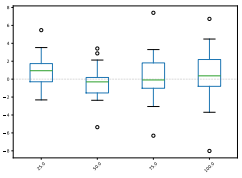
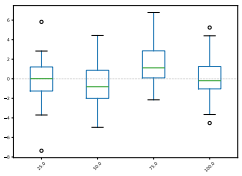
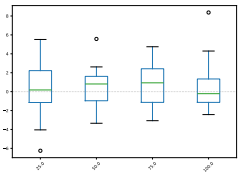
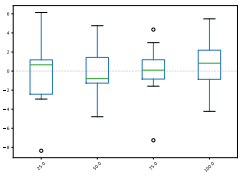
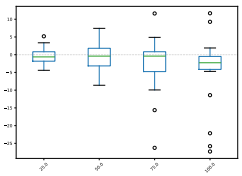
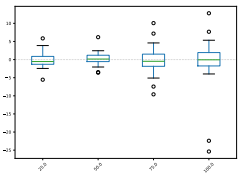
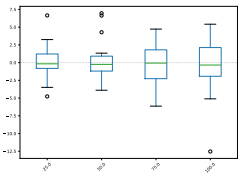
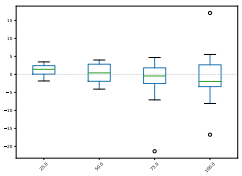
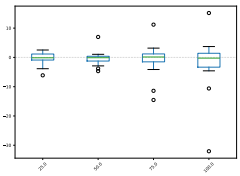
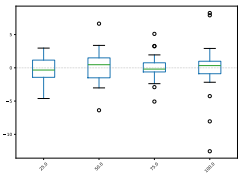
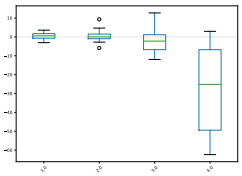
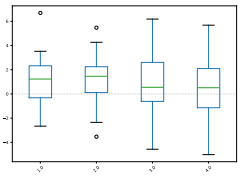
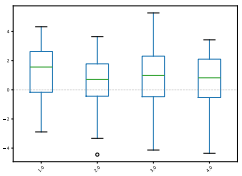
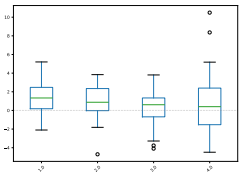
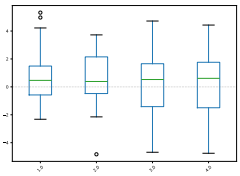
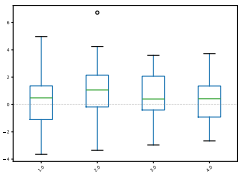
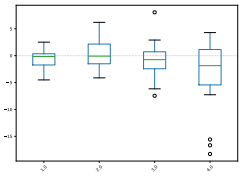
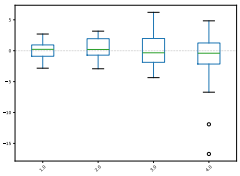
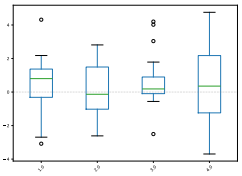
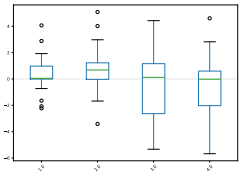
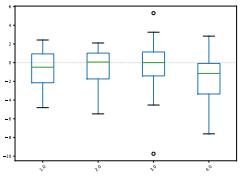
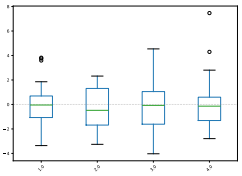
LLM + steering	Early layer	Middle layer	Late layer
Gemma 2 27b (lang.)			
Gemma 2 27b (qual.)			
Gemma 2 9b (lang.)			
Gemma 2 9b (qual.)			
Llama 3.1 70b (lang.)			
Llama 3.1 70b (qual.)			
Llama 3.1 8b (lang.)			
Llama 3.1 8b (qual.)			

Table 18: Boxplot visualizations of relative increases/decreases in % for F1 metrics for LLM + steering methods for various languages in **zero-shot** setting. The boxplots are for α values from Table 1. The X axis in the figures represents the languages, and the Y axis represents the relative difference increase for F1 downstream model performance.

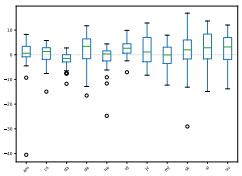
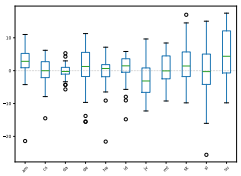
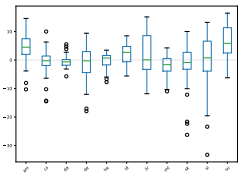
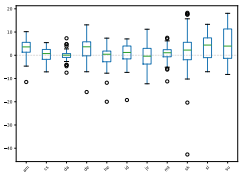
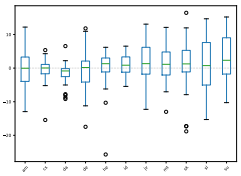
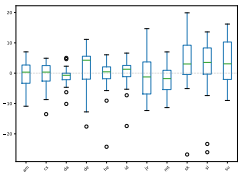
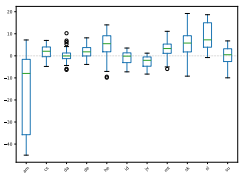
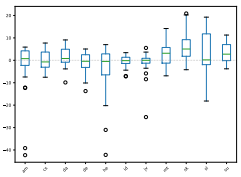
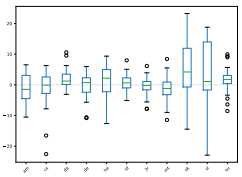
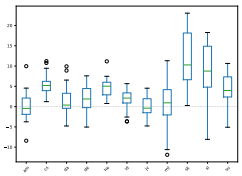
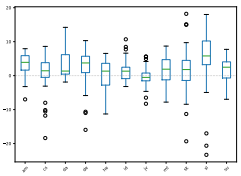
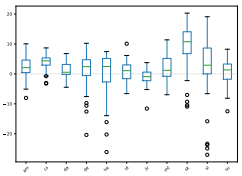
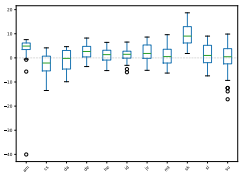
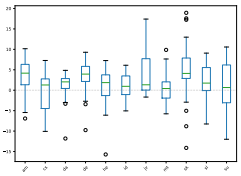
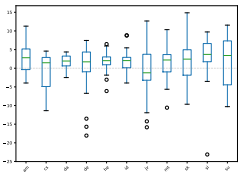
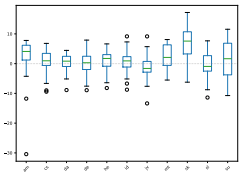
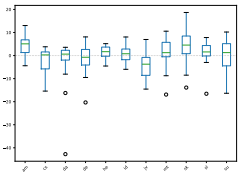
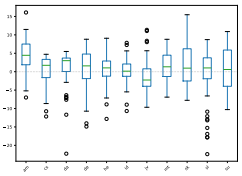
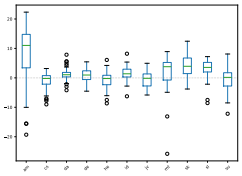
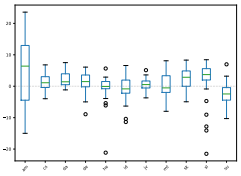
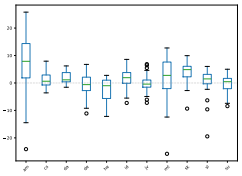
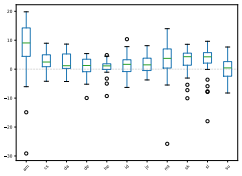
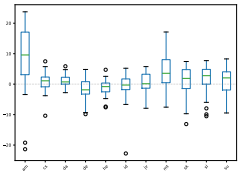
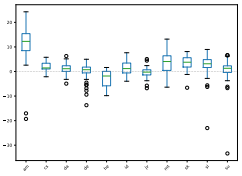
LLM + steering	Early layer	Middle layer	Late layer
Gemma 2 27b (lang.)			
Gemma 2 27b (qual.)			
Gemma 2 9b (lang.)			
Gemma 2 9b (qual.)			
Llama 3.1 70b (lang.)			
Llama 3.1 70b (qual.)			
Llama 3.1 8b (lang.)			
Llama 3.1 8b (qual.)			

Table 19: Boxplot visualizations of relative increases/decreases in % for F1 metrics for LLM + steering methods for various languages in **few-shot** setting. The boxplots are for α values from Table 1. The X axis in the figures represents the languages, and the Y axis represents the relative difference increase for F1 downstream model performance.

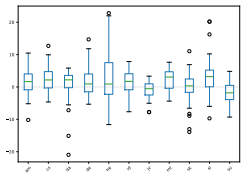
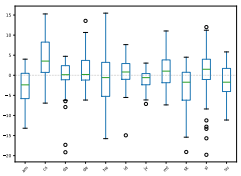
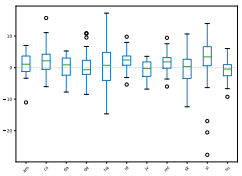
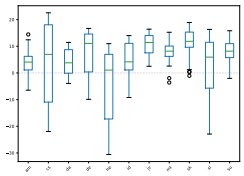
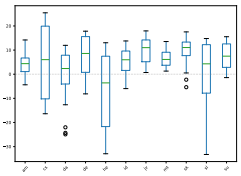
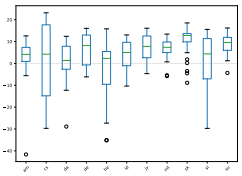
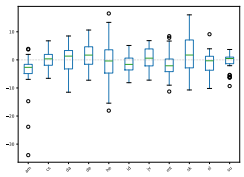
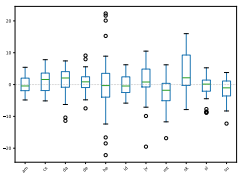
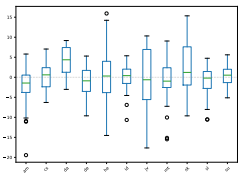
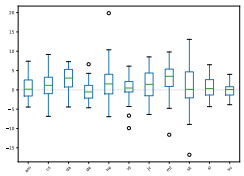
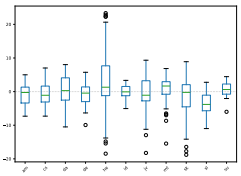
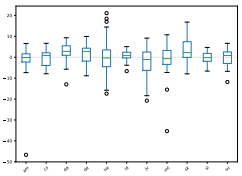
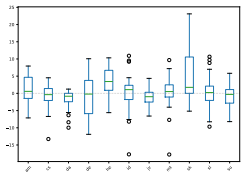
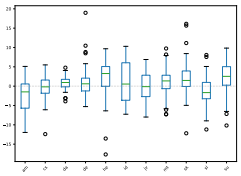
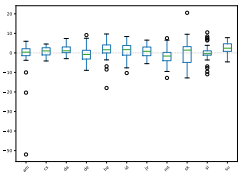
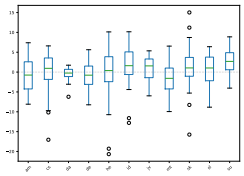
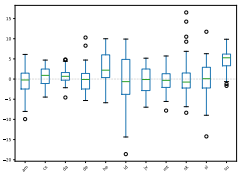
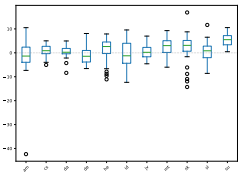
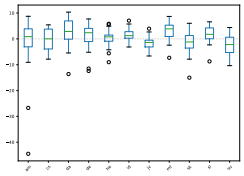
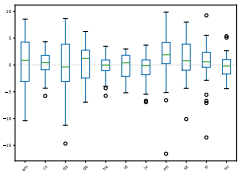
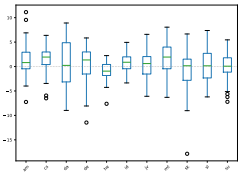
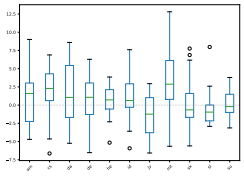
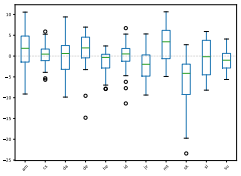
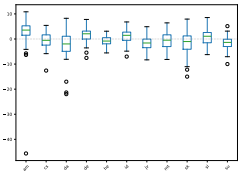
LLM + steering	Early layer	Middle layer	Late layer
Gemma 2 27b (lang.)			
Gemma 2 27b (qual.)			
Gemma 2 9b (lang.)			
Gemma 2 9b (qual.)			
Llama 3.1 70b (lang.)			
Llama 3.1 70b (qual.)			
Llama 3.1 8b (lang.)			
Llama 3.1 8b (qual.)			

Table 20: Relative increases/decreases in % for diversity metrics aggregated over LLMs and alphas for various languages and layers for **zero-shot** baseline. The X axis in the figures represents the languages, and the Y axis represents the relative difference increase for that given metric.

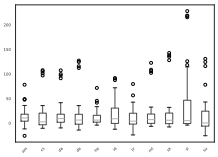
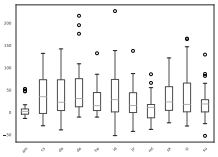
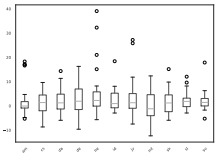
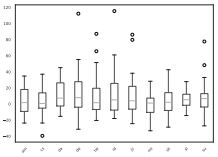
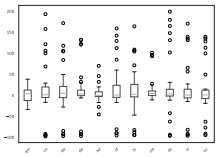
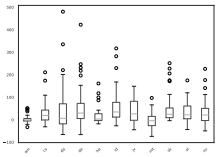
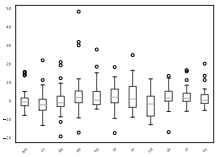
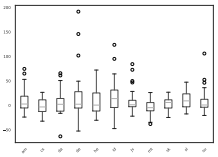
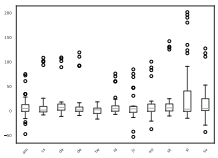
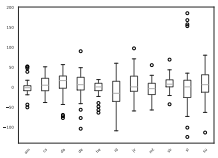
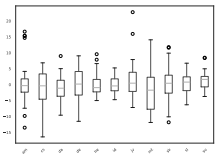
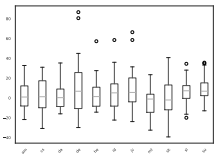
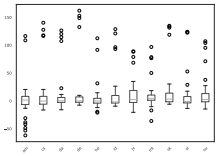
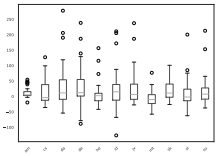
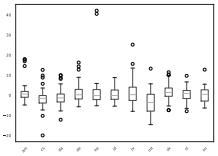
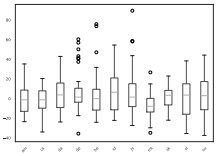
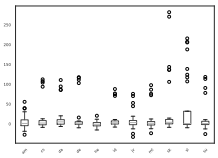
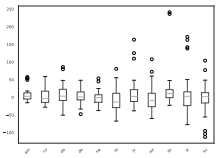
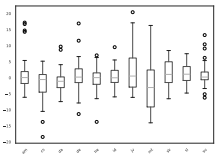
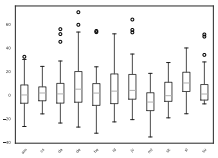
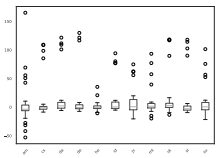
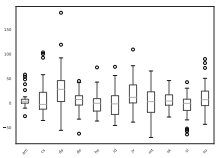
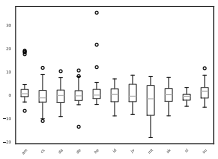
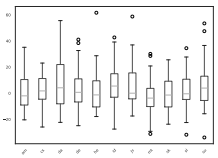
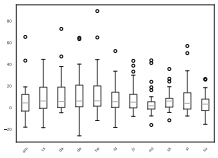
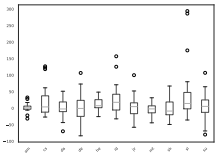
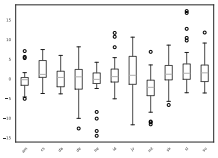
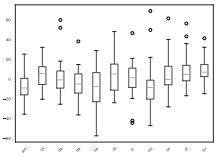
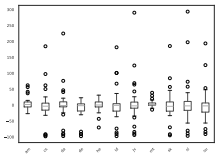
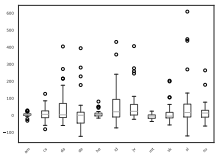
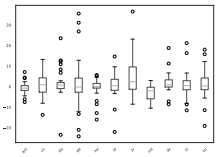
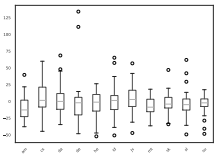
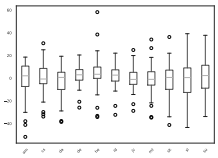
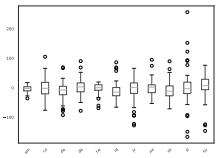
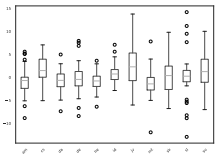
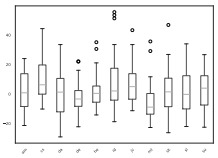
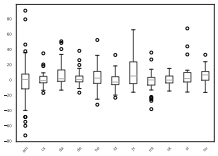
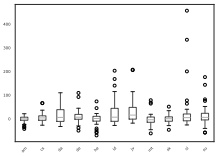
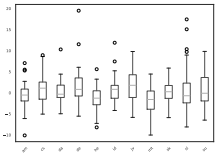
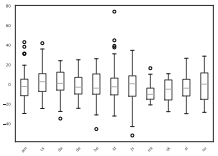
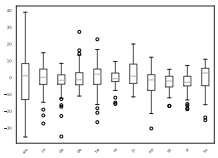
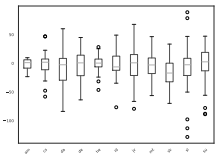
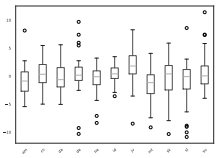
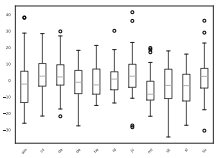
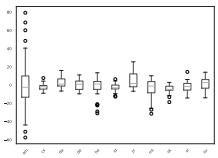
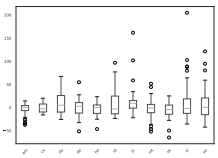
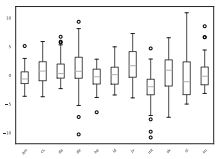
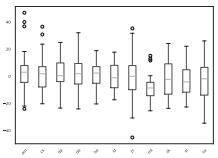
Layers + steering	Lexical div.	Embbeding div.	Isocontour rad.	Homogeneity
Early layers (qual.)				
Early layers (lang.)				
Middle layers (qual.)				
Middle layers (lang.)				
Later layers (qual.)				
Later layers (lang.)				

Table 21: Relative increases/decreases in % for diversity metrics aggregated over LLMs and alphas for various languages and layers for **few-shot** baseline. The X axis in the figures represents the languages, and the Y axis represents the relative difference increase for that given metric.

Layers + steering	Lexical div.	Embbeding div.	Isocontour rad.	Homogeneity
Early layers (qual.)				
Early layers (lang.)				
Middle layers (qual.)				
Middle layers (lang.)				
Later layers (qual.)				
Later layers (lang.)