STEEREVAL: Inference-time Interventions Strengthen Multilingual Generalization in Neural Summarization Metrics

Anonymous ACL submission

Abstract

An empirical bottleneck for Natural Language Generation tasks such as summarization is the lack of accurate and robust *evaluation metrics for many languages*, which hinders progress. While Large Language Models (LLMs) are increasingly adopted as evaluators with the advent of LLM-as-a-Judge approaches, their agreement with human judgments remains low across many languages. We take inspiration from recent studies suggesting that multilingual language models often use English as an internal pivot language, and that misalignment with this pivot can lead to degraded downstream performance during text generation. Motivated by this hypothesis, we ask whether steering the activations of neural metrics with respect to an English pivot can improve their correlation with human judgments in multilingual summarization evaluation. We experiment with encoder- and decoder-based metrics across five languages. We find that activation steering is effective for neural metrics, and that a single, language-independent steering vector improves metric–human agreement. Notably, steering also results in improvements when applied to COMET, a small encoder-based metric, showing notable human agreement despite its small scale.

1 Introduction

Metrics for Natural Language Generation (NLG) tasks such as summarization have become increasingly model-based (Zhang et al., 2020; Rei et al., 2020; Sellam et al., 2020), with the latest trend leveraging Large Language Models (LLMs) as evaluators (Liu et al., 2023; Fu et al., 2024), a paradigm know as LLM-as-a-Judge (Zheng et al., 2023).

When built on multilingual backbones, LLM evaluators can generalize across languages to some extent (Zhang et al., 2020; Rei et al., 2020; Sellam et al., 2020). However, preliminary studies on model-based multilingual evaluation suggest they often do not correlate well with human judgments (Mondshine et al., 2025), limiting their reliability as proxies for human evaluation.

We hypothesize this gap to be partly due to the way multilingual models process different languages: recent work has shown that multilingual LLMs internally rely on English as a pivot language (Wendler et al., 2024), and that the degree of alignment between a language's internal representations and this pivot varies across languages (Wang et al., 2025a). Such misalignment has been linked to degraded performance in multilingual factual recall (Lu et al., 2025) and other discriminative and generative tasks (Wang et al., 2025a; Bafna et al., 2025), suggesting that explicitly controlling this alignment could improve performance. Wang et al. (2025a), for example, directly intervene on internal representations by steering them toward (or away from) English, showing this intervention to be beneficial across a wide variety of NLG tasks. Lu et al. (2025) further show how alignment to the English pivot improves factual recall across languages. Such work, however, is typically limited to discriminative and generative tasks.

However, whether such activation steering can improve performance on *evaluation* tasks, and specifically on model-based metrics such as LLM-as-a-Judge, remains unexplored. This is a particularly important gap: reliable automatic metrics for multilingual evaluation are scarce, making the quality of model-based evaluation a bottleneck for progress across languages. As such, in this work we explicitly ask: To what degree does steering the activations of low-resource languages with respect to English strengthen *LLM-as-a-Judge* performance?

To answer this question, we focus on the task of summarization evaluation and first conduct a meta-analysis of LLM-as-a-Judge summarization metrics, assessing their reliability and alignment with human judgments. Corroborating prior work, we

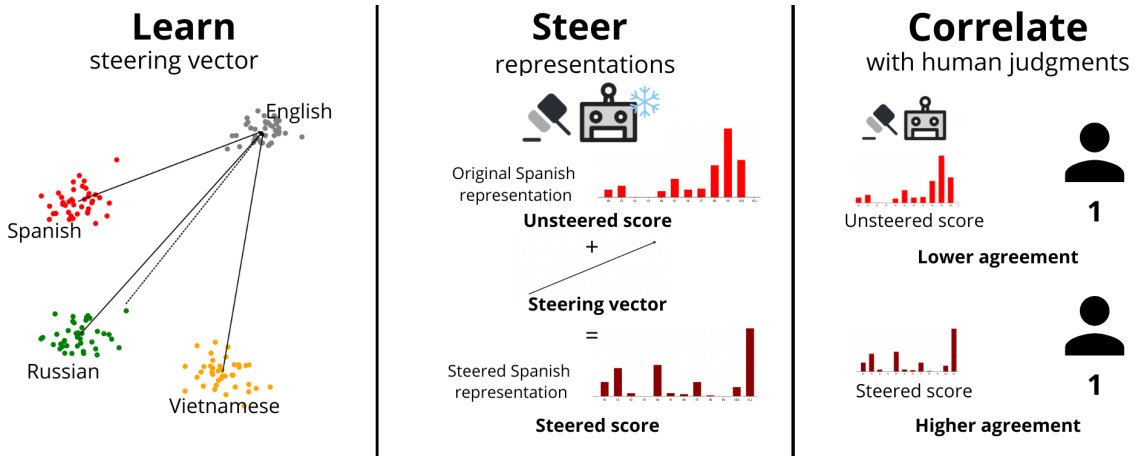

Figure 1: Illustration of our STEEREVAL method. We first use parallel data to learn cross-lingual vectors of mappings from *Source language X* → *English* or a single multilingual average vector. We then apply the steering to neural metrics. Finally, we show that such steering results in better correlation with human judgments.

find modest and inconsistent correlation with human judgments (Mondshine et al., 2025), showing the need for strengthened multilingual evaluators. To this end, we investigate the effect of steering multilingual LLM-as-a-Judge models with respect to a high-resource pivot language at test time, finding it to improve metric reliability. We further demonstrate that language-specific steering vectors may not be required, and a single, multilingual vector can lead to the largest improvements. Finally, we show that small encoder-based metrics such as COMET (Rei et al., 2020) show surprisingly competitive performance, and that this performance can further be improved through steering. Our contributions are as follows:

- **Multilingual metric meta-analysis.** We perform a meta-analysis of model-based summarization metrics across five typologically diverse languages. We examine two LLM-as-a-judge methods, using four backbone models and two prompt variations. We also examine COMET, an encoder-based metric which shows competitive performance for summarization.

- **Activation steering for multilingual metrics.** We apply activation steering to two popular LLM-as-a-judge methods – based respectively on next token probabilities (Liu et al., 2023) and average log likelihood (Fu et al., 2024) — where we show activation steering to strengthen correlation with human judgments for both approaches. We also extend steering to encoder-based models, largely unexplored in the broader activation steering literature, where we also observe consistent correlation gains. Our intervention consists of a simple and efficient vector-based addition, occurs at inference time, and requires no model retraining, making it an efficient way to strengthen metrics.

- **Multilingual intervention**. We measure the cross-lingual similarity between *Source language X* → *English* intervention vectors in LLMs, and find these vectors to be largely similar across languages, corroborating prior work in the generalizability of refusal vectors (Wang et al., 2025b) and, more broadly, shared language geometries (Chi et al., 2020; Dumas et al., 2025). Building on this finding, we explore multilingual steering, where we average over language-specific vectors to obtain a single multilingual steering vector. We show this solution outperforms language-specific steering across many settings.

2 Related Work

2.1 Multilingual summarization evaluation

Automatic evaluation of natural language generation (NLG) and summarization is a longstanding challenge (Sai et al., 2022; Celikyilmaz et al., 2021). Most existing work has focused on English. For instance, Fabbri et al. (2021) evaluate 14 summarization metrics across four quality dimensions, while subsequent studies show that correlation with human judgments of reference-based metrics can drop when scoring LLM-generated outputs (Casola

et al., 2025). As evaluation has increasingly shifted towards the LLM-as-a-Judge paradigm, Bavaresco et al. (2025) conducted a large-scale evaluation across multiple tasks, including summarization, and demonstrated high variability in metric behavior across different LLM judges.

However, the extent to which these findings generalize beyond English remains understudied. Clark et al. (2023) introduce a multilingual dataset in which model outputs are annotated in a binary fashion for specific desirable properties (e.g., conciseness), and shows it can be used to train strong ad-hoc evaluation metrics. Similarly, Hada et al. (2024) propose a multilingual dataset with human judgments using either binary labels or a 0–2 scale, and use it to assess LLMs as evaluators across five dimensions. More focused studies include Barnes et al. (2025), who analyze LLM-as-a-judge performance for Spanish and Basque, highlighting variation in evaluators' behaviors. Finally, Mondshine et al. (2025) recently introduced an abstractive summarization dataset for meta-evaluating automatic metrics across eight diverse languages, focusing on n-gram- and small model-based metrics, concluding that correlations with human judgments remain relatively low in multilingual settings.

2.2 Activation steering

The linear representation hypothesis (Park et al., 2024) proposes that high-level concepts are represented as directions in the activation space of language models. Motivated by this hypothesis, an increasing body of work has made use of a test-time method known as activation steering to control LLM behaviors. Such works typically isolate the direction that represents a given concept by way of vector arithmetic, where the isolated vector can then be added to activations at test time (Turner et al., 2024). Such a steering method has been applied to induce or reduce characteristics such as truthfulness (Marks and Tegmark, 2024), toxicity, sentiment (Turner et al., 2024), refusal (Arditi et al., 2024), and language steering (Sterz et al., 2025). Beyond steering vectors, Wang et al. (2025a) show that a learned linear map between low- and high-resource language representations can serve as an effective steering mechanism. In this paper, we focus on the vector-based intervention approach, which we apply to strengthening multilingual summarization evaluation on non-English languages.

3 Experimental Setup

3.1 Vector-based metric steering

Figure 1 provides an overview of our steering methodology. Specifically, we perform a vector-based intervention via the following steps.

Steering direction learning In line with the approach discussed in Section 2.2, we derive a steering direction between a source language and English from parallel data. Specifically, given representations in the source (lower-resourced) language h_s and English (the target language) h_t, we compute a per-layer steering vector by subtracting the mean of h_s from the mean of h_t over all time steps, capturing the direction from the source to the target language in activation space:

$$\bar{h}_v = \bar{h}_t - \bar{h}_s$$

Note that the parallel data used to learn the steering vector is disjoint from the summarization evaluation data, and any parallel corpus can in principle be used for this purpose.

Test time intervention Keeping the language model frozen, we then steer its internal representations. Specifically, at test time, we add the source-to-target vectors to activations in a given layer,

$$h_{q,l}^{mix} = h_s + \sigma \bar{h}_v$$

where σ is a scalar parameter controlling the strength and sign of the steering, q the final representation of the input at test time, and l the layer the steering is applied to; we steer a single layer. We derive representations from and apply steering to the residual stream.

For each language, we jointly tune the steering strength σ and the application layer l on a language-specific development set with human judgments, and then evaluate the selected configuration on the corresponding disjoint test set. Following prior work (Lu et al., 2025; Bafna et al., 2025), we restrict the layer search space to the middle and final layers, as these exhibit the greatest representational mismatch with English for other tasks. For COMET, we steer the pooling-space embedding directly (no per-layer selection). We search over $\sigma \in [-5, 5]$ with a step size of 1. Hyperparameters are selected via joint grid search to maximize the average Pearson correlation across six evaluation dimensions (see Section 3.2).

Performance testing We evaluate the metric using the steered representations and analyze its agreement with human judgments. Since the Seahorse dataset (Clark et al., 2023) (which we discuss in Section 3.2) is annotated with binary labels, we use Pearson correlation and ROC-AUC, following the original paper.

3.2 Datasets

Steering direction learning To learn the language alignments, we leverage the FLORES multi-parallel dataset (Goyal et al., 2021), which consists of high-quality human-translated sentences primarily from Wikipedia. We follow Wang et al. (2025a) and use 500 parallel sentence pairs.

Parameter tuning and performance testing We use the Seahorse multilingual summarization dataset (Clark et al., 2023) as key benchmark, a high-quality human evaluation dataset containing binary scores for six languages i.e., English, German, Spanish, Russian, Turkish, and Vietnamese. Source texts are extracted from four datasets in the GEM benchmark (Gehrmann et al., 2021): XSum (Narayan et al., 2018) (for English only), XL-Sum (Hasan et al., 2021), MLSum (Scialom et al., 2020), and WikiLingua (Ladhak et al., 2020). Summaries are the outputs of eight abstractive summarization systems based on T5 (Raffel et al., 2023) (only for English), mT5 (Xue et al., 2021) and PaLM (Chowdhery et al., 2022); human references from the respective datasets, when available, are also treated as hypotheses. The dataset is large-scale, with language-specific test sets of 2,600+ annotated outputs. Each summary is annotated by trained raters on six binary dimensions: whether it is a) comprehensible, b) non-repetitive, c) grammatical, d) concise, e) whether its content is fully attributable to the source article, and f) whether it captures the main ideas of the source article. We preserve the original split and use the dataset validation set to tune the method hyperparameters and its test set for the final evaluation.[1]

3.3 Metrics

We primarily examine the following metrics:

Direct prompting We directly prompt a multilingual model to produce a score for each evaluation dimension. However, directly extracting the score from the model output can be unreliable, as responses may include additional text or justifications, an issue even more complex in multilingual settings. Thus, we rely on token probabilities. We follow G-Eval (Liu et al., 2023) and compute the score as a weighted sum over the token probabilities corresponding to the evaluation scale at the first generated token. The score is computed as:

$$\text{score} = \sum_i p(s_i) \times s_i$$

where $s_i \in \{0, 1\}$, mimicking the dataset annotation, and $p(s_i)$ is the token probability.

GPTScore (prompt-conditioned log-likelihood) Following Fu et al. (2024), we evaluate a hypothesis using the model's conditional generation probabilities. Given a prompt that specifies a desired property (e.g., conciseness), the model is instructed to generate a summary exhibiting that property. The summary to be evaluated is thus scored by computing the average log-probability of its tokens under the model, conditioned on the prompt. Specifically, we compute the final score as:

$$\text{score} = \frac{1}{N} \sum_{t=1}^{N} \log p(h_t \mid h_{<t}, P, \theta)$$

where P is the prompt, N is the number of tokens in the summary, and θ the model parameters.

For both LLM-based metrics, we use Llama-3.2-3B[2] (Grattafiori et al., 2024) and aya-expanse-8b[3] (Dang et al., 2024), as backbones and further experiment with their larger counterparts i.e., Llama-3.1-8B Instruct[4] and aya-expanse-32b[5] in Section 5. These models are chosen since they are open, instruction-tuned, diverse in size, have broad language coverage, and have shown good performance in various multilingual tasks. We prompt each model both in English and with a machine translated version in the target language. Appendix 3 details the metrics' implementation, including the prompts.

COMET Due to prior work reporting state-of-the-art performance on many languages by directly applying COMET on multilingual summarization evaluation (Mondshine et al., 2025; Krubiński and Pecina, 2022), we also analyze the effect of steering on COMET, which is a multilingual neural metric

[1] We provide dataset statistics in Appendix D.

[2] *meta-llama/Llama-3.2-3B*

[3] *CohereLabs/aya-expanse-8b*

[4] *meta-llama/Llama-3.1-8B-Instruct*

[5] *CohereLabs/aya-expanse-32b*

typically used in machine translation. We rely on the *wmt22-comet-da* model, which is built on top of XLM-R (Conneau et al., 2020) and follow the experimental design of Mondshine et al. (2025) to adapt the machine-translation metric to summarization: we replace the source input with an empty string, use the system-generated summary as the hypothesis, and the human-written summary as the reference. Note that COMET, in contrast to LLM-based judges, requires a reference summary. In the Seahorse dataset, reference summaries are themselves treated as hypotheses and scored by human judges. We therefore compute COMET scores only for the subset of model-generated hypotheses that have an associated reference, and results are not directly comparable to those from the LLMs.

4 Results

Tables 1 and 2 summarizes the main results of our LLM-as-a-Judge and COMET steering experiments respectively. We primarily report results using Pearson correlation and focus on English prompting in the main text; detailed results when prompting in the source language are in Appendix E, while results using AUC are presented in Appendix F. We report the average correlation over the six evaluation dimensions per language; per-dimension results are in Appendix A. We focus on language-specific steering and later discuss the zero-shot multilingual average vector setup in Section 5.

Baselines show relatively modest and variable correlation with human judgments. Across all settings, baseline systems achieve relatively low positive Pearson correlation, showing that summary evaluation remains challenging in a multilingual setting and that metrics designed for English do not necessarily transfer across languages, corroborating previous results (Clark et al., 2023; Mondshine et al., 2025). Larger models do not consistently outperform smaller ones: for GPTScore, aya-expanse-32b shows comparable baseline performance to its smaller (8b) counterpart and, interestingly, lower performance than Llama-3.1-8b across all languages. No single configuration is superior across all settings: the choice of metric, backbone model, and prompt language interact in non-trivial ways. While comparisons with LLM-based scores should be taken with care given the modified test set (see Section 3.3), COMET typically achieves a strong baseline correlations across all languages, including for languages like German that show consistently low correlations for LLM-based metrics. This is surprising given that COMET is a much smaller model trained specifically for translation quality estimation rather than summarization evaluation. We hypothesize this improvement might be due to the specific supervised training and to the use of a reference, not available to the LLMs in our prompts.

Steering consistently improves performance. Language-specific steering typically improves over the baseline, with gains in the vast majority of languages, metrics, and backbone models. While the advantage is generally incremental, the method shows consistent improvement across the majority of the configurations, which is notable given the simplicity and generality of our inference-time steering method. Steering typically improves over the baseline across LLM-based metrics, though the magnitude and consistency of gains vary considerably depending on the configuration. Improvements are also observed for COMET, though these are less consistent over languages when using language-specific steering.

GPTScore tends to be more stable across backbone models, and benefits most consistently from steering. GPTScore tends to show more consistent results than Direct Prompting across backbone models. We hypothesize this stability might be due to the log-likelihood signal being more robust across model sizes when compared to prompting for direct scoring. Interestingly, the Llama-3B model shows typically even better GPTScore baseline performance than aya-8B across languages despite the difference in size, suggesting that GPTScore may be less sensitive to model capacity than Direct Prompting. Steering improves GPTScore across nearly all language and model combinations: aya-expanse-8b and aya-expanse-32b show gains across all five languages with no degradations, while Llama-3.2-3B and Llama-3.1-8B-Instruct show only a single marginal degradation each, in Vietnamese. This suggests that the conditional log-likelihood scoring of multilingual models is particularly steerable, regardless of model size.

Direct prompting requires higher model capacity, and steering results are more mixed. In contrast to GPTScore, direct prompting appears to require a model with higher capacity to pro-

		de	es	ru	tr	vi
GPTScore						
`Llama-3.2-3B`	Baseline (EN-prompt)	$.022_{\pm.132}$	$.213_{\pm.141}$	$.270_{\pm.062}$	$.202_{\pm.108}$	$\underline{.139}_{\pm.092}$
	+ vector steer (EN-prompt)	$.022_{\pm.127}$	$.215_{\pm.142}\uparrow$	$.276_{\pm.062}\uparrow$	$\underline{.211}_{\pm.100}\uparrow$	$.138_{\pm.095}\downarrow$
	+ avgvec zero-shot (EN-prompt)	$\underline{.024}_{\pm.127}\uparrow$	$\underline{.221}_{\pm.149}\uparrow$	$\underline{.281}_{\pm.062}\uparrow$	$.210_{\pm.104}\uparrow$	$\underline{.139}_{\pm.097}$
`aya-expanse-8b`	Baseline (EN-prompt)	$.017_{\pm.143}$	$.150_{\pm.088}$	$.240_{\pm.067}$	$.190_{\pm.069}$	$.181_{\pm.063}$
	+ vector steer (EN-prompt)	$.018_{\pm.145}\uparrow$	$.155_{\pm.088}\uparrow$	$.247_{\pm.067}\uparrow$	$.198_{\pm.070}\uparrow$	$.185_{\pm.063}\uparrow$
	+ avgvec zero-shot (EN-prompt)	$\underline{.021}_{\pm.143}\uparrow$	$\underline{.156}_{\pm.089}\uparrow$	$\underline{.248}_{\pm.067}\uparrow$	$\underline{.199}_{\pm.070}\uparrow$	$\underline{.188}_{\pm.063}\uparrow$
`Llama-3.1-8B-Instruct`	Baseline (EN-prompt)	$.036_{\pm.138}$	$.228_{\pm.133}$	$.282_{\pm.059}$	$.242_{\pm.100}$	$\underline{.176}_{\pm.081}$
	+ avgvec zero-shot (EN-prompt)	$\underline{.040}_{\pm.133}\uparrow$	$\underline{\mathbf{.238}}_{\pm.143}\uparrow$	$\underline{\mathbf{.294}}_{\pm.058}\uparrow$	$\underline{\mathbf{.250}}_{\pm.096}\uparrow$	$.175_{\pm.085}\downarrow$
`aya-expanse-32b`	Baseline (EN-prompt)	$.007_{\pm.139}$	$.164_{\pm.084}$	$.244_{\pm.067}$	$.190_{\pm.070}$	$.195_{\pm.053}$
	+ avgvec zero-shot (EN-prompt)	$\underline{.011}_{\pm.140}\uparrow$	$\underline{.169}_{\pm.085}\uparrow$	$\underline{.250}_{\pm.065}\uparrow$	$\underline{.196}_{\pm.070}\uparrow$	$\underline{.199}_{\pm.053}\uparrow$
Direct Prompting						
`Llama-3.2-3B`	Baseline (EN-prompt)	$\underline{.050}_{\pm.050}$	$-.004_{\pm.045}$	$.014_{\pm.039}$	$.007_{\pm.043}$	$-.001_{\pm.040}$
	+ vector steer (EN-prompt)	$.049_{\pm.048}\downarrow$	$\underline{-.001}_{\pm.044}\uparrow$	$\underline{.015}_{\pm.039}\uparrow$	$\underline{.008}_{\pm.045}\uparrow$	$\underline{.005}_{\pm.040}\uparrow$
`aya-expanse-8b`	Baseline (EN-prompt)	$\underline{\mathbf{.208}}_{\pm.042}$	$\underline{.226}_{\pm.139}$	$.249_{\pm.079}$	$.237_{\pm.112}$	$\underline{\mathbf{.248}}_{\pm.077}$
	+ vector steer (EN-prompt)	$\underline{\mathbf{.208}}_{\pm.042}$	$\underline{.226}_{\pm.139}$	$\underline{.251}_{\pm.076}\uparrow$	$\underline{.239}$fl$_{\pm.113}\uparrow$	$.247_{\pm.077}\downarrow$

Table 1: SEAHORSE Pearson r averaged over 6 quality dimensions (Comprehensible, Non-repetitive, Grammatical, Attributable, Main ideas, Concise) with standard deviation as subscripts (mean$_{\pm\text{std}}$), per language, using English prompts. Baseline uses σ=0 (no steering); steered uses the best σ selected on the validation split. **Bold** indicates the single highest value per language across the entire table; underline indicates the highest value per language within each model. Blue/↑ (red/↓) on steered rows indicates improvement (resp. degradation) relative to the baseline.

	de	es	ru	tr	vi
COMET					
Baseline	$.248_{\pm.087}$	$.266_{\pm.120}$	$.358_{\pm.130}$	$.380_{\pm.139}$	$.290_{\pm.116}$
+ vector steer	$\mathbf{.279}_{\pm.081}\uparrow$	$\mathbf{.318}_{\pm.184}\uparrow$	$.351_{\pm.155}\downarrow$	$.360_{\pm.144}\downarrow$	$.295_{\pm.136}\uparrow$
+ avgvec zero-shot	$.275_{\pm.084}\uparrow$	$.303_{\pm.146}\uparrow$	$\mathbf{.364}_{\pm.152}\uparrow$	$\mathbf{.392}_{\pm.165}\uparrow$	$\mathbf{.297}_{\pm.134}\uparrow$

Table 2: SEAHORSE Pearson r over 6 quality dimensions (Comprehensible, Non-repetitive, Grammatical, Attributable, Main ideas, Concise) with standard deviation as subscripts (mean$_{\pm\text{std}}$), per language for COMET (`wmt22-comet-da`). Baseline uses σ=0 (no steering); steered uses the best σ selected on the validation split. **Bold** indicates the highest value per language. Blue/↑ (red/↓) indicates improvement (resp. degradation) subscripts relative to the baseline.

duce meaningful signal. For Llama-3.2-3B, direct prompting degrades to near-zero or even negative correlations when prompting in English, with baseline values of -.004 for Spanish and -.001 for Vietnamese, likely due to the backbone's limited ability to follow instructions. Despite this weak baseline, steering recovers some signal across nearly all languages. With a larger model, direct prompting is the best-performing configuration overall, with aya-expanse-8b reaching the highest baseline correlations across four out of five languages. However, steering results in limited improvement over such baselines, producing marginal gains for Russian and Turkish, with a slight degradation for Vietnamese.

COMET shows strong but less consistent gains from language-specific steering. As discussed above, COMET baselines are substantially higher than LLM-based metrics across all languages. Language-specific steering improves German and Spanish substantially, but degrades Russian and Turkish, with a marginal gain for Vietnamese. Note that the multiligual variant, which we discuss in Section 5, shows a more consistent picture, improving all five languages and achieving the best results for Russian, Turkish, and Vietnamese.

Performance across languages does not follow resource levels. German consistently shows the lowest correlations across all configurations. Under Direct Prompting with aya-expanse-8b, differences across languages are small, suggesting that larger and more multilingual models may be more robust to language variation. Turkish shows notable variability: it is among the strongest languages under Direct Prompting with aya-expanse-8b, but weaker under GPTScore with Llama-3.2-3B, pos-

sibly due to its morphological complexity interacting poorly with token log-likelihood scoring. When steering is applied, degradations are rare and typically marginal, concentrated in Vietnamese and German for LLMs. The pattern is similar for COMET, where the two degradations under vector steering (Russian and Turkish) are reversed by the multilingual variant.

The effect of prompt language is model- and metric-dependent. Prompting in the source language (Appendix E) does not show a consistent advantage over English. Prompting in English and in the source language result in different baseline strengths depending on the metrics: source-language prompts typically produce stronger Direct Prompting baselines, while English prompts tend to result in higher GPTScore baselines. Steering is more consistently beneficial when prompting in English.

Performance is highly variable across dimensions Looking at Appendix A, baseline difficulty varies across dimensions. On average, Comprehensible is the easiest dimension, while Grammatical is the hardest one. One notable interaction concerns the metric: COMET substantially outperforms GPTScore and Direct Prompting on fluency-adjacent dimensions, but this advantage largely disappears for Grammatical and reverses for Attributable, where GPTScore outperforms COMET. Language-specific vector steering typically leads to larger gains on Comprehensible and Non-repetitive.

5 Do we need language-specific vectors?

Prior work in activation steering has found high level-concepts such as refusal or script to be broadly similar vectors across languages (Wang et al., 2025b; Shim et al., 2026; Lu et al., 2025). In this section, we investigate to what extent language-specific steering is needed for model-based evaluation metrics. We focus on GPTScore and COMET, as they show more consistent and interpretable behaviors than direct prompting. For simplicity, we experiment with prompting in English only.

To study the similarity of cross-language steering vectors, we measure the extent to which *Language X → English* directions are similar. As such, we compute the cosine similarity between our *Language X → English* vectors extracted from parallel data. Figure 2 shows the results of projecting language samples from the parallel data onto the principal components (PCA, left) and the cross-language similarity matrix (right) for GPTScore, for the Llama-3.2-3b model. We focus on the last model layer, as it is consistently identified as optimal on our development set across languages and settings. The PCA reveals that samples cluster by language, suggesting that the model's internal representations encode language identity at this layer. The similarity matrix, on the other hand, shows that the *Language X → English* steering vectors are relatively similar across languages.

Motivated by this similarity, and in line with previous work (Lu et al., 2025), we explore the effect of taking a single steering vector across all languages. To do so, we average across language-specific vectors, and obtain our final steering vector as:

$$\bar{h}_v = \frac{1}{|S|} \sum_{s \in S} \left(\bar{h}_t^{(s)} - \bar{h}_s^{(s)} \right)$$

where s in a source language in our language set S and $|S| = 5$ in our setting. For GPTScore, we steer the last layer of the model, following previous observations; for COMET, similarly to the language-specific steering, we directly steer the pooling-space embedding directly. We set σ to -2.5 (the midpoint of our explored negative interval) for all languages, since the parameter is consistently negative across language.

Surprisingly, averaged zero-shot vector is at least as good as, and often better than, the language-specific tuned vector in nearly all configurations. Focusing on LLM-based metrics, Table 1 (focusing on the avgvec zero-shot rows) shows the results. For Llama-3.2-3B, improvements over the baseline are observed for all languages except Vietnamese (for which results are consistent with no intervention), and the averaged vector results in higher correlations than the language-specific tuned vector for all languages except Turkish. For aya-expanse-8b, improvements are consistent across all languages, always exceeding language-specific performance.

The effectiveness of a single multilingual vector is particularly interesting given we performed no parameter tuning and pre-selected an average negative σ, a setting that is comparatively simpler than the language-specific one. Since we do not search for hyperparameters and use a single language-independent vector, we can also scale to larger models with limited computational cost.

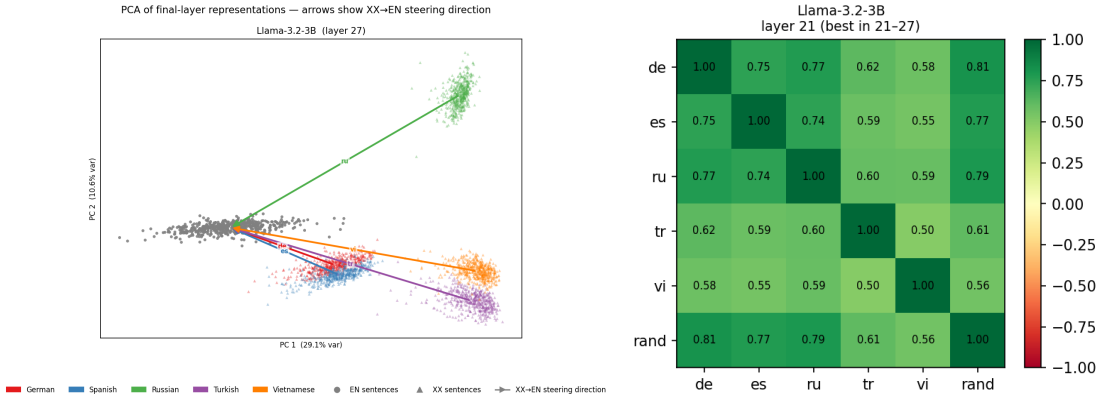

Figure 2: Projection of the instances from FLORES into the Principal Components (left) and cosine similarity across steering vectors (right) for Llama-3.2-3B.

Thus, we also explore whether our approach improves for larger models of the same family. After steering, GPTScore with Llama-3.1-8B achieves the best performance overall in Spanish, Russian, and Turkish, improving on non-trivial baselines; Vietnamese is the only language showing minor degradation, consistently with previous findings. aya-expanse-32b also shows improvement with respect to its baseline.

Improvements also stand for COMET, for which results are shown in Table 2. When using a cross-lingual vector (without parameter tuning), COMET results are remarkably strong. Despite requiring no language-specific tuning, it matches or exceeds the language-specific steered variant in four out of five languages, and achieves the best overall result for Russian ($r = .364$), Turkish ($r = .392$), and Vietnamese ($r = .297$). Also notice that, unlike the language-specific vector, the cross-lingual vector never degrades performance: for Russian and Turkish, where individual steering vectors hurt, the cross-lingual vector not only recovers but exceeds the baseline, suggesting that averaging across languages might have a natural regularizing effect that suppresses language-specific noise. Taken together, these results suggest that encoder-based metrics have strong potential when references are available, and can be steered effectively, making our approach generalizable beyond large decoder-based LLMs.

These results suggest that language-specific steering vectors might not be necessary: a single averaged vector, applied zero-shot with a fixed hyperparameter, might be sufficient to obtain consistent gains.

6 Conclusion

In this paper, we leverage the insight that LLMs internally use English as a pivot language (Wendler et al., 2024) to improve the reliability of neural metrics for multilingual summarization evaluation. We showed that test-time activation steering consistently improves metrics correlation with human judgments across a diverse set of languages, metrics, and backbone models, with no model retraining required. Our findings contribute to the growing body of work that leverages model-internal representations for practical downstream tasks (Marks and Tegmark, 2024), and extend it to the underexplored setting of automatic evaluation.

Notably, we find that language-specific steering is not necessary: a single averaged multilingual vector, applied zero-shot with a fixed hyperparameter, is sufficient to obtain consistent gains and often outperforms its language-specific counterpart.

We also apply our method to COMET: despite being a much smaller encoder-based model, steered COMET benefits strongly from our intervention and surpasses LLM-based evaluators in human agreement across multiple languages.

Several directions remain open for future work. Our approach assumes the availability of parallel data to derive steering vectors; moreover, future work should study how far the multilingual vector generalizes to unseen languages not included in the averaging set. Finally, while our work focuses on language steering, prior work has leveraged test-time interventions targeting specific evaluative dimensions such as fluency (Sheng et al., 2024); combining language- and dimension-specific interventions could yield further gains in multilingual evaluation.

Limitations

Due to the relative lack of large-scale summarization evaluation datasets with human annotations, our study is limited by the potentially limited number of languages. For the specific dataset we employ (Clark et al., 2023), the additional constraint that this dataset only has binary annotations is a further limitation. Second, while we demonstrate the effectiveness of a single multilingual steering vector across the languages in our study, we do not evaluate its generalization to unseen languages i.e., languages not in the averaging set. Whether the vector transfers to typologically distant or lower-resourced languages remains an open question. Finally, steering representations toward an English pivot may introduce unwanted biases, e.g., cultural biases. We leave a systematic analysis of such biases to future work.

Ethical Statement

We acknowledge the usage of LLMs in the partial assistance of figure creation, code development, and in strengthening the coherence of parts of the manuscript.

References

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. *Preprint*, arXiv:2406.11717.

Niyati Bafna, Tianjian Li, Kenton Murray, David R. Mortensen, David Yarowsky, Hale Sirin, and Daniel Khashabi. 2025. The translation barrier hypothesis: Multilingual generation with large language models suffers from implicit translation failure. *Preprint*, arXiv:2506.22724.

Jeremy Barnes, Naiara Perez, Alba Bonet-Jover, and Begoña Altuna. 2025. Summarization metrics for spanish and basque: Do automatic scores and llm-judges correlate with humans? *Preprint*, arXiv:2503.17039.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Silvia Casola, Yang Janet Liu, Siyao Peng, Oliver Kraus, Albert Gatt, and Barbara Plank. 2025. References matter: Investigating the impact of reference set variation on summarization evaluation. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 274–291, Hanoi, Vietnam. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of text generation: A survey. *Preprint*, arXiv:2006.14799.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. 2023. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31822–31841, Vienna, Austria. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, and 37 others. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 96–120, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. METAL: Towards multilingual meta-evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2280–2298, Mexico City, Mexico. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Mateusz Krubiński and Pavel Pecina. 2022. From COMET to COMES – can summary evaluation benefit from translation evaluation? In *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems*, pages 21–31, Online. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Meng Lu, Ruochen Zhang, Carsten Eickhoff, and Ellie Pavlick. 2025. Paths not taken: Understanding and mending the multilingual factual recall pipeline. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15066–15096, Suzhou, China. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *Preprint*, arXiv:2310.06824.

Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. 2025. Beyond n-grams: Rethinking evaluation metrics and strategies for multilingual abstractive summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19019–19035, Vienna, Austria. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference*

on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Shuqian Sheng, Yi Xu, Tianhang Zhang, Zanwei Shen, Luoyi Fu, Jiaxin Ding, Lei Zhou, Xiaoying Gan, Xinbing Wang, and Chenghu Zhou. 2024. RepEval: Effective text evaluation with LLM representation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7019–7033, Miami, Florida, USA. Association for Computational Linguistics.

Ryan Soh-Eun Shim, Kwanghee Choi, Kalvin Chang, Ming-Hao Hsu, Florian Eichin, Zhizheng Wu, Alane Suhr, Michael A. Hedderich, David Harwath, David R. Mortensen, and Barbara Plank. 2026. Linear script representations in speech foundation models enable zero-shot transliteration. *Preprint*, arXiv:2601.02906.

Hannah Sterz, Fabian David Schmidt, Goran Glavaš, and Ivan Vulić. 2025. ReCoVeR the target language: Language steering without sacrificing task performance. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19390–19405, Suzhou, China. 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.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025a. Bridging the language gaps in large language models with inference-time cross-lingual intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5418–5433, Vienna, Austria. Association for Computational Linguistics.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schütze, and Barbara Plank. 2025b. Refusal direction is universal across safety-aligned languages. *Preprint*, arXiv:2505.17306.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Dimension-specific Results

Tables 3, 4, 5, 6, 7, and 8 present the results for the six evaluation dimensions in Seahorse when using an English prompt.

		de	es	ru	tr	vi
GPTScore						
`Llama-3.2-3B`	Baseline (EN-prompt)	.003	.319	.313	.099	.248
	+ vector steer (EN-prompt)	.025↑	.320↑	.315↑	.158↑	.244↓
	+ avgvec zero-shot (EN-prompt)	.023↑	.333↑	.331↑	.140↑	.258↑
`aya-expanse-8b`	Baseline (EN-prompt)	.029	.249	.353	.114	.305
	+ vector steer (EN-prompt)	.027↓	.255↑	.361↑	.118↑	.308↑
	+ avgvec zero-shot (EN-prompt)	.029↓	.256↑	.362↑	.118↑	.311↑
Direct Prompting						
`Llama-3.2-3B`	Baseline (EN-prompt)	.095	−.001	.055	.056	.044
	+ vector steer (EN-prompt)	.095↑	−.009↓	.058↑	.067↑	.049↑
	+ avgvec zero-shot (EN-prompt)	—	—	—	—	—
`aya-expanse-8b`	Baseline (EN-prompt)	.207	.335	.386	.477	.338
	+ vector steer (EN-prompt)	.206↓	.335↑	.385↓	.481↑	.337↓
	+ avgvec zero-shot (EN-prompt)	—	—	—	—	—
COMET						
`wmt22-comet-da`	Baseline	.297	.421	.578	.651	.425
	+ vector steer	**.361**↑	**.555**↑	.622↑	.626↓	**.454**↑
	+ avgvec zero-shot	.346↑	.489↑	**.627**↑	**.712**↑	.451↑

Table 3: SEAHORSE Pearson r for dimension **Comprehensible** per language, using English prompts. Baseline uses σ=0 (no steering); steered uses the best σ selected on the validation split; avgvec zero-shot uses a fixed vector averaged across all source languages. **Bold** indicates the highest value per language across the entire table; underline indicates the highest value per language within each model. Blue/↑ (red/↓) on steered rows indicates improvement (degradation) relative to the baseline.

		de	es	ru	tr	vi
GPTScore						
`Llama-3.2-3B`	Baseline (EN-prompt)	−.147	.068	.178	.095	.079
	+ vector steer (EN-prompt)	−.145↑	.067↓	.185↑	.094↓	.079↑
	+ avgvec zero-shot (EN-prompt)	−.139↑	.075↑	.191↑	.095↓	.075↓
`aya-expanse-8b`	Baseline (EN-prompt)	−.207	.033	.234	.168	.193
	+ vector steer (EN-prompt)	−.206↑	.038↑	.242↑	.176↑	.196↑
	+ avgvec zero-shot (EN-prompt)	−.199↑	.040↑	.245↑	.177↑	.200↑
Direct Prompting						
`Llama-3.2-3B`	Baseline (EN-prompt)	.088	.012	.057	.045	.059
	+ vector steer (EN-prompt)	.081↓	.017↑	.060↑	.043↓	.064↑
	+ avgvec zero-shot (EN-prompt)	—	—	—	—	—
`aya-expanse-8b`	Baseline (EN-prompt)	.210	.147	.173	.179	.180
	+ vector steer (EN-prompt)	.206↓	.146↓	.189↑	.195↑	.177↓
	+ avgvec zero-shot (EN-prompt)	—	—	—	—	—
COMET						
`wmt22-comet-da`	Baseline	.301	.338	.406	.414	.427
	+ vector steer	**.347**↑	**.513**↑	.468↑	.450↑	**.493**↑
	+ avgvec zero-shot	.345↑	.428↑	**.470**↑	**.468**↑	.488↑

Table 4: SEAHORSE Pearson r for dimension **Non-repetitive** per language, using English prompts. Baseline uses σ=0 (no steering); steered uses the best σ selected on the validation split; avgvec zero-shot uses a fixed vector averaged across all source languages. **Bold** indicates the highest value per language across the entire table; underline indicates the highest value per language within each model. Blue/↑ (red/↓) on steered rows indicates improvement (degradation) relative to the baseline.

		de	es	ru	tr	vi
GPTScore						
`Llama-3.2-3B`	Baseline (EN-prompt)	−.073	−.023	.190	.090	.124
	+ vector steer (EN-prompt)	−.074↓	−.023↑	.197↑	.090↓	.125↑
	+ avgvec zero-shot (EN-prompt)	−.070↑	−.033↓	**.198**↑	.092↑	.121↓
`aya-expanse-8b`	Baseline (EN-prompt)	−.101	.065	.140	.121	.171
	+ vector steer (EN-prompt)	−.101↑	**.068**↑	.148↑	.126↑	.174↑
	+ avgvec zero-shot (EN-prompt)	−.096↑	.066↑	.150↑	.128↑	.177↑
Direct Prompting						
`Llama-3.2-3B`	Baseline (EN-prompt)	.013	.010	.010	.023	−.038
	+ vector steer (EN-prompt)	.014↑	.016↑	.009↓	.025↑	−.031↑
	+ avgvec zero-shot (EN-prompt)	—	—	—	—	—
`aya-expanse-8b`	Baseline (EN-prompt)	**.159**	−.048	.139	.152	.129
	+ vector steer (EN-prompt)	**.159**↓	−.048↑	.139↓	.153↑	.126↓
	+ avgvec zero-shot (EN-prompt)	—	—	—	—	—
COMET						
`wmt22-comet-da`	Baseline	.090	.045	.197	.345	.245
	+ vector steer	.128↑	−.001↓	.170↓	.308↓	.237↓
	+ avgvec zero-shot	.119↑	.041↓	.183↓	**.352**↑	**.246**↑

Table 5: SEAHORSE Pearson r for dimension **Grammatical** per language, using English prompts. Baseline uses σ=0 (no steering); steered uses the best σ selected on the validation split; avgvec zero-shot uses a fixed vector averaged across all source languages. **Bold** indicates the highest value per language across the entire table; underline indicates the highest value per language within each model. Blue/↑ (red/↓) on steered rows indicates improvement (degradation) relative to the baseline.

		de	es	ru	tr	vi
GPTScore						
Llama-3.2-3B	Baseline (EN-prompt)	.279	.368	.334	.336	.278
	+ vector steer (EN-prompt)	.268↓	.371↑	**.342**↑	.334↓	.284↑
	+ avgvec zero-shot (EN-prompt)	.272↓	**.388**↑	.341↑	**.341**↑	.280↑
aya-expanse-8b	Baseline (EN-prompt)	.249	.101	.190	.197	.182
	+ vector steer (EN-prompt)	.254↑	.107↑	.196↑	.208↑	.186↑
	+ avgvec zero-shot (EN-prompt)	.258↑	.111↑	.198↑	.208↑	.189↑
Direct Prompting						
Llama-3.2-3B	Baseline (EN-prompt)	−.040	−.099	−.044	−.062	−.051
	+ vector steer (EN-prompt)	−.039↑	−.094↑	−.045↓	−.066↓	−.048↑
	+ avgvec zero-shot (EN-prompt)	—	—	—	—	—
aya-expanse-8b	Baseline (EN-prompt)	.293	.344	.266	.222	**.314**
	+ vector steer (EN-prompt)	**.295**↑	.345↑	.265↓	.219↓	**.315**↑
	+ avgvec zero-shot (EN-prompt)	—	—	—	—	—
COMET						
wmt22-comet-da	Baseline	.168	.219	.215	.204	.112
	+ vector steer	.226↑	.258↑	.202↓	.176↓	.123↑
	+ avgvec zero-shot	.209↑	.250↑	.211↓	.195↓	.121↑

Table 6: SEAHORSE Pearson r for dimension **Attributable** per language, using English prompts. Baseline uses σ=0 (no steering); steered uses the best σ selected on the validation split; avgvec zero-shot uses a fixed vector averaged across all source languages. **Bold** indicates the highest value per language across the entire table; underline indicates the highest value per language within each model. Blue/↑ (red/↓) on steered rows indicates improvement (degradation) relative to the baseline.

		de	es	ru	tr	vi
GPTScore						
`Llama-3.2-3B`	Baseline (EN-prompt)	.023	.267	.303	.290	.029
	+ vector steer (EN-prompt)	.021↓	.271↑	.308↑	.285↓	.021↓
	+ avgvec zero-shot (EN-prompt)	.020↓	.272↑	.308↑	.288↓	.022↓
`aya-expanse-8b`	Baseline (EN-prompt)	.073	.258	.284	.317	.136
	+ vector steer (EN-prompt)	.074↑	.261↑	.289↑	.324↑	.138↑
	+ avgvec zero-shot (EN-prompt)	.075↑	.263↑	.289↑	.325↑	.140↑
Direct Prompting						
`Llama-3.2-3B`	Baseline (EN-prompt)	.052	.013	−.030	−.039	−.013
	+ vector steer (EN-prompt)	.052↓	.025↑	−.024↑	−.035↑	−.006↑
	+ avgvec zero-shot (EN-prompt)	—	—	—	—	—
`aya-expanse-8b`	Baseline (EN-prompt)	.182	.303	.274	.241	.304
	+ vector steer (EN-prompt)	.182↑	.302↓	.272↓	.242↑	.303↓
	+ avgvec zero-shot (EN-prompt)	—	—	—	—	—
COMET						
`wmt22-comet-da`	Baseline	.324	.340	**.419**	**.378**	**.331**
	+ vector steer	.320↓	.331↓	.346↓	.335↓	.287↓
	+ avgvec zero-shot	**.325**↑	**.357**↑	.376↓	.346↓	.293↓

Table 7: SEAHORSE Pearson r for dimension **Main ideas** per language, using English prompts. Baseline uses σ=0 (no steering); steered uses the best σ selected on the validation split; avgvec zero-shot uses a fixed vector averaged across all source languages. **Bold** indicates the highest value per language across the entire table; underline indicates the highest value per language within each model. Blue/↑ (red/↓) on steered rows indicates improvement (degradation) relative to the baseline.

		de	es	ru	tr	vi
GPTScore						
`Llama-3.2-3B`	Baseline (EN-prompt)	.045	.279	.303	.303	.079
	+ vector steer (EN-prompt)	.035↓	.284↑	.312↑	.301↓	.075↓
	+ avgvec zero-shot (EN-prompt)	.037↓	**.290**↑	.314↑	**.306**↑	.075↓
`aya-expanse-8b`	Baseline (EN-prompt)	.058	.195	.237	.223	.102
	+ vector steer (EN-prompt)	.062↑	.199↑	.244↑	.235↑	.106↑
	+ avgvec zero-shot (EN-prompt)	.062↑	.202↑	.246↑	.236↑	.109↑
Direct Prompting						
`Llama-3.2-3B`	Baseline (EN-prompt)	.094	.042	.034	.018	−.008
	+ vector steer (EN-prompt)	.092↓	.036↓	.029↓	.016↓	−.001↑
	+ avgvec zero-shot (EN-prompt)	—	—	—	—	—
`aya-expanse-8b`	Baseline (EN-prompt)	.197	.274	.257	.153	.220
	+ vector steer (EN-prompt)	.201↑	.274↑	.257↑	.146↓	**.224**↑
	+ avgvec zero-shot (EN-prompt)	—	—	—	—	—
COMET						
`wmt22-comet-da`	Baseline	**.307**	.233	**.334**	.288	.201
	+ vector steer	.292↓	.253↑	.296↓	.267↓	.179↓
	+ avgvec zero-shot	.306↓	.254↑	.319↓	.277↓	.182↓

Table 8: SEAHORSE Pearson r for dimension **Concise** per language, using English prompts. Baseline uses σ=0 (no steering); steered uses the best σ selected on the validation split; avgvec zero-shot uses a fixed vector averaged across all source languages. **Bold** indicates the highest value per language across the entire table; underline indicates the highest value per language within each model. Blue/↑ (red/↓) on steered rows indicates improvement (degradation) relative to the baseline.

B Optimal Hyperparameters

		de		es		ru		tr		vi	
		layer	σ	layer	σ	layer	σ	layer	σ	layer	σ
GPTScore											
`Llama-3.2-3B`	EN-prompt	27	-2.0	27	$+2.0$	27	-1.0	27	-5.0	22	-2.0
	src-prompt	27	-2.0	27	$+1.0$	27	-1.0	27	-3.0	27	$+1.0$
`aya-expanse-8b`	EN-prompt	27	-5.0	31	-3.0	31	-2.0	31	-3.0	31	-2.0
	src-prompt	31	-3.0	31	-2.0	31	-1.0	31	-2.0	31	-1.0
Direct Prompting											
`Llama-3.2-3B`	EN-prompt	27	$+5.0$	27	-5.0	24	$+5.0$	25	$+5.0$	22	$+3.0$
	src-prompt	25	$+4.0$	27	-5.0	21	$+2.0$	23	$+4.0$	26	-4.0
`aya-expanse-8b`	EN-prompt	29	-5.0	31	$+5.0$	26	-5.0	26	-5.0	25	-5.0
	src-prompt	26	-5.0	26	-5.0	26	-2.0	29	-3.0	25	-5.0

Table 9: Best intervention layer and steering σ for LLM-based SEAHORSE evaluation. Layer and σ are selected jointly on the validation split by mean Pearson r across six quality dimensions. Avgvec rows use a fixed σ and the layer that maximised cross-lingual vector similarity.

	de	es	ru	tr	vi
COMET (wmt22-comet-da)					
vector steer	-5.0	-5.0	-2.0	$+5.0$	-2.0

Table 10: Best steering σ for COMET (`wmt22-comet-da`) SEAHORSE evaluation. COMET steers the pooling-space embedding directly (no per-layer selection); σ is chosen on the validation split by mean Pearson r across six quality dimensions.

C Prompts

	English	German	Spanish	Russian	Turkish	Vietnamese
Compreh.	Generate a comprehensible summary for the following article: *[article]* Summary:	Erstelle eine verständliche Zusammenfassung des folgenden Artikels: *[article]* Zusammenfassung:	Genera un resumen comprensible del siguiente artículo: *[article]* Resumen:	Создайте понятное резюме следующей статьи: *[article]* Резюме:	Aşağıdaki makale için anlaşılır bir özet oluşturun: *[article]* Özet:	Tạo bản tóm tắt dễ hiểu cho bài viết sau: *[article]* Tóm tắt:
Non-rep.	Generate a summary without unnecessary repetition for the following article: *[article]* Summary:	Erstelle eine Zusammenfassung ohne unnötige Wiederholungen des folgenden Artikels: *[article]* Zusammenfassung:	Genera un resumen sin repeticiones innecesarias del siguiente artículo: *[article]* Resumen:	Создайте резюме без лишних повторений следующей статьи: *[article]* Резюме:	Aşağıdaki makale için gereksiz tekrar içermeyen bir özet oluşturun: *[article]* Özet:	Tạo bản tóm tắt không có thông tin lặp lại không cần thiết cho bài viết sau: *[article]* Tóm tắt:
Gramm.	Generate a grammatically correct summary for the following article: *[article]* Summary:	Erstelle eine grammatisch korrekte Zusammenfassung des folgenden Artikels: *[article]* Zusammenfassung:	Genera un resumen gramaticalmente correcto del siguiente artículo: *[article]* Resumen:	Создайте грамматически правильное резюме следующей статьи: *[article]* Резюме:	Aşağıdaki makale için dilbilgisel açıdan doğru bir özet oluşturun: *[article]* Özet:	Tạo bản tóm tắt đúng ngữ pháp cho bài viết sau: *[article]* Tóm tắt:
Attrib.	Generate a faithful summary for the following article: *[article]* Summary:	Erstelle eine quelltreue Zusammenfassung des folgenden Artikels: *[article]* Zusammenfassung:	Genera un resumen fiel al contenido del siguiente artículo: *[article]* Resumen:	Создайте достоверное резюме следующей статьи: *[article]* Резюме:	Aşağıdaki makale için kaynağa sadık bir özet oluşturun: *[article]* Özet:	Tạo bản tóm tắt trung thực với bài viết sau: *[article]* Tóm tắt:
Main ideas	Generate a summary capturing the main ideas of the following article: *[article]* Summary:	Erstelle eine Zusammenfassung mit den Hauptideen des folgenden Artikels: *[article]* Zusammenfassung:	Genera un resumen con las ideas principales del siguiente artículo: *[article]* Resumen:	Создайте резюме с основными идеями следующей статьи: *[article]* Резюме:	Aşağıdaki makalenin ana fikirlerini içeren bir özet oluşturun: *[article]* Özet:	Tạo bản tóm tắt nắm bắt ý chính của bài viết sau: *[article]* Tóm tắt:
Concise	Generate a concise summary for the following article: *[article]* Summary:	Erstelle eine prägnante Zusammenfassung des folgenden Artikels: *[article]* Zusammenfassung:	Genera un resumen conciso del siguiente artículo: *[article]* Resumen:	Создайте краткое резюме следующей статьи: *[article]* Резюме:	Aşağıdaki makale için kısa ve öz bir özet oluşturun: *[article]* Özet:	Tạo bản tóm tắt súc tích cho bài viết sau: *[article]* Tóm tắt:

Figure 3: GPTScore prompt templates per quality dimension and language. *[article]* is replaced by the source article; the second line (after the newline) is the continuation cue used for scoring.

	English	German	Spanish	Russian	Turkish	Vietnamese
Compreh.	Is the summary comprehensible?	Ist die Zusammenfassung verständlich?	¿Es el resumen comprensible?	Является ли резюме понятным?	Özet anlaşılır mı?	Bản tóm tắt có dễ hiểu không?
Non-rep.	Does the summary contain no unnecessary repetition?	Enthält die Zusammenfassung keine unnötigen Wiederholungen?	¿El resumen no contiene repeticiones innecesarias?	Не содержит ли резюме ненужных повторений?	Özet gereksiz tekrar içermiyor mu?	Bản tóm tắt có không chứa sự lặp lại không cần thiết không?
Gramm.	Is the summary grammatically correct?	Ist die Zusammenfassung grammatikalisch korrekt?	¿Es el resumen gramaticalmente correcto?	Является ли резюме грамматически правильным?	Özet dilbilgisi açısından doğru mu?	Bản tóm tắt có đúng ngữ pháp không?
Attrib.	Is the summary factually consistent with the article?	Ist die Zusammenfassung faktisch konsistent mit dem Artikel?	¿Es el resumen factualmente consistente con el artículo?	Является ли резюме фактически согласованным со статьёй?	Özet, makaleyle olgusal olarak tutarlı mı?	Bản tóm tắt có nhất quán về mặt thực tế với bài viết không?
Main ideas	Does the summary capture the main ideas of the article?	Erfasst die Zusammenfassung die Hauptideen des Artikels?	¿El resumen captura las ideas principales del artículo?	Отражает ли резюме основные идеи статьи?	Özet, makalenin ana fikirlerini yansıtıyor mu?	Bản tóm tắt có nắm bắt được các ý chính của bài viết không?
Concise	Is the summary concise?	Ist die Zusammenfassung prägnant?	¿Es el resumen conciso?	Является ли резюме кратким?	Özet kısa ve öz mu?	Bản tóm tắt có súc tích không?

Figure 4: Direct Prompting (G-eval) question per quality dimension and language.

Lang.	Task intro	Art. label	Sum. label	Answer cue
EN	Answer with 1 (yes) or 0 (no).	Article	Summary	Answer (1=yes, 0=no):
DE	Antworten Sie mit 1 (ja) oder 0 (nein).	Artikel	Zusammenfassung	Antwort (1=ja, 0=nein):
ES	Responda con 1 (sí) o 0 (no).	Artículo	Resumen	Respuesta (1=sí, 0=no):
RU	Ответьте 1 (да) или 0 (нет).	Статья	Резюме	Ответ (1=да, 0=нет):
TR	1 (evet) veya 0 (hayır) ile yanıtlayın.	Makale	Özet	Yanıt (1=evet, 0=hayır):
VI	Trả lời bằng 1 (có) hoặc 0 (không).	Bài viết	Bản tóm tắt	Câu trả lời (1=có, 0=không):

Figure 5: Fixed sentence-level labels used in Direct Prompting per language.

D Dataset

Data is released under a CC-BY 4.0 license, with which we comply fully.

System	DE	EN	ES	RU	TR	VI
T5-Base	0	236	0	0	0	0
T5-Base (250s)	0	241	0	0	0	0
T5-XXL	0	247	0	0	0	0
mT5-Small	231	239	234	222	218	234
mT5-Small (250s)	226	243	239	218	209	229
mT5-XXL	228	239	242	208	208	238
PaLM (1-shot)	241	250	239	241	207	223
PaLM (fine-tuned)	219	249	245	241	180	230
Reference	228	239	240	213	210	238
Total outputs	1,373	2,183	1,439	1,343	1,232	1,392

Table 11: Number of outputs samples, annotated with human judgments for the Seahorse validation set.

System	DE	EN	ES	RU	TR	VI
T5-Base	0	484	0	0	0	0
T5-Base (250s)	0	482	0	0	0	0
T5-XXL	0	476	0	0	0	0
mT5-Small	451	482	474	449	495	497
mT5-Small (250s)	459	484	477	427	488	490
mT5-XXL	443	481	479	450	485	492
PaLM (1-shot)	463	495	486	486	440	438
PaLM (fine-tuned)	434	487	473	485	409	483
Reference	435	484	465	439	486	494
Total outputs	2,685	4,355	2,854	2,736	2,803	2,894

Table 12: Number of outputs samples, annotated with human judgments for the Seahorse test set.

E Source prompting results

Table 13 presents the results when using a machine-translated prompt in the source language.

		de	es	ru	tr	vi
GPTScore						
`Llama-3.2-3B`	Baseline (src-prompt)	$\underline{.028}_{\pm.125}$	$.213_{\pm.143}$	$.273_{\pm.065}$	$.204_{\pm.103}$	$\underline{.144}_{\pm.093}$
	+ vector steer (src-prompt)	$.026_{\pm.127}\downarrow$	$\underline{.215}_{\pm.141}\uparrow$	$\underline{\mathbf{.278}}_{\pm.064}\uparrow$	$\underline{\mathbf{.212}}_{\pm.100}\uparrow$	$.143_{\pm.090}\downarrow$
`aya-expanse-8b`	Baseline (src-prompt)	$.004_{\pm.145}$	$.129_{\pm.079}$	$.195_{\pm.065}$	$.154_{\pm.069}$	$.158_{\pm.068}$
	+ vector steer (src-prompt)	$\underline{.006}_{\pm.144}\uparrow$	$\underline{.137}_{\pm.081}\uparrow$	$\underline{.200}_{\pm.065}\uparrow$	$\underline{.160}_{\pm.070}\uparrow$	$\underline{.161}_{\pm.067}\uparrow$
Direct Prompting						
`Llama-3.2-3B`	Baseline (src-prompt)	$.046_{\pm.039}$	$\underline{.029}_{\pm.038}$	$\underline{.056}_{\pm.035}$	$.081_{\pm.093}$	$\underline{.115}_{\pm.050}$
	+ vector steer (src-prompt)	$\underline{.051}_{\pm.043}\uparrow$	$.028_{\pm.025}\downarrow$	$\underline{.056}_{\pm.034}\uparrow$	$\underline{.083}_{\pm.094}\uparrow$	$.114_{\pm.053}\downarrow$
`aya-expanse-8b`	Baseline (src-prompt)	$\underline{\mathbf{.215}}_{\pm.083}$	$\underline{\mathbf{.244}}_{\pm.148}$	$\underline{.218}_{\pm.108}$	$\underline{.188}_{\pm.080}$	$.215_{\pm.056}$
	+ vector steer (src-prompt)	$.214_{\pm.085}\downarrow$	$\underline{\mathbf{.244}}_{\pm.148}\uparrow$	$\underline{.218}_{\pm.108}\downarrow$	$\underline{.189}_{\pm.079}\uparrow$	$\underline{\mathbf{.216}}_{\pm.056}\uparrow$

Table 13: SEAHORSE Pearson r averaged over 6 quality dimensions (Comprehensible, Non-repetitive, Grammatical, Attributable, Main ideas, Concise) with standard deviation as subscripts (mean$_{\pm\text{std}}$), per language, using source-language prompts. Baseline uses σ=0 (no steering); steered uses the best σ selected on the validation split. **Bold** indicates the single highest value per language across the entire table; underline indicates the highest value per language within each model. Blue/↑ (red/↓) on steered rows indicates improvement (degradation) relative to the baseline.

F AUC-ROC Results

Tables 14 and 15 present AUC-ROC results for the small models under English and source-language prompting respectively, corresponding to the Pearson r tables in the main paper (Table 1) and Appendix E. Table 16 shows AUC-ROC for the large-model avgvec zero-shot setting.

		de	es	ru	tr	vi
GPTScore						
Llama-3.2-3B	Baseline (EN-prompt)	$.483_{\pm.112}$	$.631_{\pm.080}$	$.698_{\pm.024}$	$.628_{\pm.069}$	$.604_{\pm.051}$
	+ vector steer (EN-prompt)	$.483_{\pm.111}\downarrow$	$.632_{\pm.082}\uparrow$	$\mathbf{.702}_{\pm.024}\uparrow$	$.626_{\pm.069}\downarrow$	$.603_{\pm.052}\downarrow$
aya-expanse-8b	Baseline (EN-prompt)	$.457_{\pm.135}$	$.597_{\pm.046}$	$.679_{\pm.055}$	$.625_{\pm.060}$	$.641_{\pm.063}$
	+ vector steer (EN-prompt)	$.459_{\pm.136}\uparrow$	$.600_{\pm.046}\uparrow$	$.684_{\pm.056}\uparrow$	$.631_{\pm.059}\uparrow$	$.643_{\pm.064}\uparrow$
Direct Prompting						
Llama-3.2-3B	Baseline (EN-prompt)	$.551_{\pm.044}$	$.497_{\pm.027}$	$.510_{\pm.034}$	$.499_{\pm.026}$	$.502_{\pm.035}$
	+ vector steer (EN-prompt)	$.550_{\pm.043}\downarrow$	$.498_{\pm.027}\uparrow$	$.511_{\pm.033}\uparrow$	$.499_{\pm.025}\uparrow$	$.505_{\pm.034}\uparrow$
aya-expanse-8b	Baseline (EN-prompt)	$.650_{\pm.064}$	$\mathbf{.669}_{\pm.098}$	$.676_{\pm.067}$	$\mathbf{.662}_{\pm.085}$	$.676_{\pm.055}$
	+ vector steer (EN-prompt)	$\mathbf{.651}_{\pm.063}\uparrow$	$\mathbf{.668}_{\pm.097}\downarrow$	$.674_{\pm.066}\downarrow$	$.661_{\pm.086}\downarrow$	$\mathbf{.678}_{\pm.055}\uparrow$

Table 14: SEAHORSE AUC-ROC averaged over 6 quality dimensions (Comprehensible, Non-repetitive, Grammatical, Attributable, Main ideas, Concise) with standard deviation as subscripts (mean$_{\pm\text{std}}$), per language, using English prompts. Baseline uses σ=0 (no steering); steered uses the best σ selected on the validation split. **Bold** indicates the single highest value per language across the entire table; underline indicates the highest value per language within each model. Blue/↑ (red/↓) on steered rows indicates improvement (degradation) relative to the baseline.

		de	es	ru	tr	vi
GPTScore						
Llama-3.2-3B	Baseline (src-prompt)	$.484_{\pm.108}$	$.631_{\pm.083}$	$.697_{\pm.026}$	$\mathbf{.631}_{\pm.065}$	$.607_{\pm.052}$
	+ vector steer (src-prompt)	$.482_{\pm.112}\downarrow$	$.633_{\pm.082}\uparrow$	$\mathbf{.701}_{\pm.027}\uparrow$	$\mathbf{.631}_{\pm.066}\downarrow$	$.607_{\pm.050}\downarrow$
aya-expanse-8b	Baseline (src-prompt)	$.451_{\pm.138}$	$.582_{\pm.041}$	$.651_{\pm.059}$	$.600_{\pm.060}$	$.627_{\pm.061}$
	+ vector steer (src-prompt)	$.453_{\pm.137}\uparrow$	$.588_{\pm.043}\uparrow$	$.655_{\pm.060}\uparrow$	$.604_{\pm.059}\uparrow$	$.629_{\pm.062}\uparrow$
Direct Prompting						
Llama-3.2-3B	Baseline (src-prompt)	$.533_{\pm.031}$	$.529_{\pm.032}$	$.550_{\pm.030}$	$.551_{\pm.069}$	$.575_{\pm.041}$
	+ vector steer (src-prompt)	$.539_{\pm.037}\uparrow$	$.529_{\pm.031}\uparrow$	$.551_{\pm.029}\uparrow$	$.552_{\pm.068}\uparrow$	$.575_{\pm.043}\downarrow$
aya-expanse-8b	Baseline (src-prompt)	$\mathbf{.621}_{\pm.030}$	$\mathbf{.656}_{\pm.102}$	$.631_{\pm.057}$	$.611_{\pm.047}$	$.631_{\pm.040}$
	+ vector steer (src-prompt)	$\mathbf{.621}_{\pm.029}\uparrow$	$\mathbf{.656}_{\pm.103}\downarrow$	$.630_{\pm.057}\downarrow$	$.611_{\pm.047}\uparrow$	$\mathbf{.632}_{\pm.038}\uparrow$

Table 15: SEAHORSE AUC-ROC averaged over 6 quality dimensions (Comprehensible, Non-repetitive, Grammatical, Attributable, Main ideas, Concise) with standard deviation as subscripts (mean$_{\pm\text{std}}$), per language, using source-language prompts. Baseline uses σ=0 (no steering); steered uses the best σ selected on the validation split. **Bold** indicates the single highest value per language across the entire table; underline indicates the highest value per language within each model. Blue/↑ (red/↓) on steered rows indicates improvement (degradation) relative to the baseline.

		de	es	ru	tr	vi
GPTScore						
Llama-3.1-8B-Instruct	Baseline (EN-prompt)	$.480_{\pm.120}$	$.643_{\pm.077}$	$.711_{\pm.028}$	$.659_{\pm.059}$	$.631_{\pm.051}$
	+ avgvec zero-shot (EN-prompt)	$\mathbf{.487}_{\pm.113}\uparrow$	$\mathbf{.649}_{\pm.084}\uparrow$	$\mathbf{.717}_{\pm.029}\uparrow$	$\mathbf{.660}_{\pm.058}\uparrow$	$.629_{\pm.051}\downarrow$
aya-expanse-32b	Baseline (EN-prompt)	$.448_{\pm.131}$	$.603_{\pm.039}$	$.680_{\pm.048}$	$.626_{\pm.061}$	$.642_{\pm.055}$
	+ avgvec zero-shot (EN-prompt)	$.451_{\pm.131}\uparrow$	$.607_{\pm.040}\uparrow$	$.683_{\pm.049}\uparrow$	$.631_{\pm.060}\uparrow$	$\mathbf{.645}_{\pm.056}\uparrow$

Table 16: SEAHORSE AUC-ROC averaged over 6 quality dimensions (Comprehensible, Non-repetitive, Grammatical, Attributable, Main ideas, Concise) with standard deviation as subscripts (mean$_{\pm\text{std}}$), per language, using English prompts. Baseline uses σ=0 (no steering); avgvec zero-shot uses a fixed steering vector averaged across all source languages with a pre-set σ. **Bold** indicates the single highest value per language across the entire table; underline indicates the highest value per language within each model. Blue/↑ (red/↓) on steered rows indicates improvement (degradation) relative to the baseline.