DFKI-MLT at SemEval-2026 TASK 7: Steering Multilingual Models Towards Cultural Knowledge

Yusser Al Ghussin[1,2] **Daniil Gurgurov**[1,2] **Yasser Hamidullah**[1,2]
Josef van Genabith[1,2] **Cristina España-Bonet**[1,3] **Simon Ostermann**[1,2]

[1]German Research Center for Artificial Intelligence (DFKI GmbH),
[2]Saarland Informatics Campus, Saarbrücken, Germany
[3]Barcelona Supercomputing Center (BSC-CNS), Barcelona, Catalonia, Spain

Abstract

Large language models (LLMs) are increasingly used across diverse linguistic and cultural contexts, yet their cultural knowledge remains uneven across regions and languages. We present the **DFKI-MLT** system for **SemEval-2026 Task 7** on cultural awareness, where we apply *activation steering* to multilingual LLMs using language vectors extracted from parallel FLORES data. Our method performs inference-time adaptation by adding language-specific steering vectors to the residual stream at a selected transformer layer, without any parameter updates. We participated in both the short-answer (SAQ) and multiple-choice (MCQ) tracks; however, only our MCQ submission received an official score. In the official **MCQ** track, we achieved **86.96%** accuracy, ranking **7th out of 17** teams. To better understand system behavior, we conduct post-hoc analyses on the shared-task MCQ and SAQ settings. These analyses show that activation steering yields *modest* and *heterogeneous* improvements on cultural reasoning: gains are strongly *layer-sensitive*, vary substantially across language–region pairs (some configurations even degrade performance), and interact with prompt formulation (generic vs. culturally conditioned prompts). Our findings suggest that prompt design and activation steering should be jointly optimized for culturally aware multilingual inference. We release our code and experimental configurations at `https://github.com/Yusser96/SemEval-2026-Track7`.

1 Introduction

Large language models (LLMs) are increasingly deployed in multilingual settings, but strong multilingual performance does not necessarily imply strong *cultural* competence. Recent work shows that LLMs often underperform on culturally grounded reasoning and everyday cultural knowledge, especially for underrepresented regions and languages, even when they appear linguistically fluent (Myung et al., 2024; Romero et al., 2024). These concerns have motivated a growing body of research on *cultural awareness* and its evaluation in language models (Pawar et al., 2025). This challenge is central to SemEval-2026 Task 7 (Ousidhoum et al., 2026), which evaluates cultural knowledge and reasoning across diverse languages and cultures using BLEnD-style evaluation protocols (Myung et al., 2024).

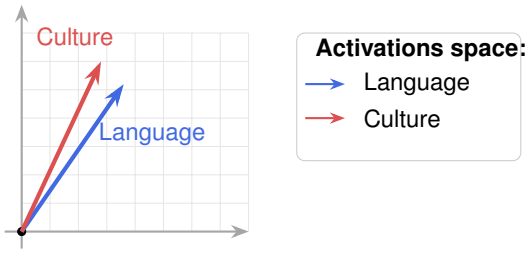

Figure 1: Motivation: if culture overlaps with language representations and language identity forms stable directions, then steering with language vectors may improve access to culturally relevant knowledge.

In this paper, we describe the **DFKI-MLT** submission to SemEval-2026 Task 7 (Ousidhoum et al., 2026; Ghosh et al., 2026). Prior work provides mechanistic evidence that multilingual LLMs encode cultural information in representations that overlap and interact with language-specific components (Namazifard and Poech, 2025), suggesting that intervening on *language-aligned directions* may also modulate culturally relevant behavior. Motivated by this, our system uses *activation steering*: instead of optimizing model parameters through fine-tuning, we modify internal activations at inference time using steering vectors (Rimsky et al., 2024). Concretely, we extract language steering vectors and inject them into the residual stream of multilingual LLMs during generation. We build on evidence that language identity is encoded as a stable direction in activation space (Marks and Tegmark, 2023), and hypothesize that

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2548–2563

steering along such directions can improve access to culturally relevant knowledge (Figure 1).

Our experiments across multiple multilingual instruction-tuned models, prompts and languages show that activation steering yields *modest* and *heterogeneous* effects on cultural reasoning: at best, we observe improvements of up to **+1.5%** absolute accuracy over the unsteered baseline on individual locales, but other configurations degrade performance, and gains do not generalize uniformly across language-region pairs. These results highlight both the appeal of steering as a lightweight inference-time intervention and its current limitations as a stand-alone solution to cultural alignment.

Beyond reporting shared-task performance, we aim to provide a detailed analysis of *when* and *why* using language vectors for activation steering can help cultural reasoning.

2 Task Background

SemEval-2026 Task 7 (Ousidhoum et al., 2026) evaluates the *cultural awareness* of language models and NLP systems across languages and regions. The task is based on the manually constructed BLEnD benchmark (Myung et al., 2024), which is designed specifically for evaluation and therefore does not provide training data. By withholding BLEnD from system training, the shared task aims to assess whether models can generalize to unseen cultural and linguistic contexts rather than memorizing benchmark content.

BLEnD currently covers multiple languages and cultures, and the shared task further expands coverage by adding additional language-culture pairs. Participants may compete in one or more tracks.

Track 1: Short Answer Questions (SAQ). In the SAQ track, systems answer short questions in the same language as the input question. The goal is to generate a culturally appropriate response while respecting linguistic and regional variation. Answers are evaluated against human-annotated BLEnD responses.

Track 2: Multiple-Choice Questions (MCQ). In the MCQ track, questions are provided in English, and each question includes four answer options representing different cultural perspectives (one option per country/region candidate, subject to the benchmark construction constraints). Systems must select the culturally appropriate option for the target region.

MCQ Example

Question: What sports do men like to watch the most in Ireland? A.baseball B.basketball C.cricket D.football
Gold label: D

Our participation. We participated in **both** Track 1 (SAQ) and Track 2 (MCQ). Our submission uses inference-time activation steering with language vectors extracted from multilingual parallel data, without model fine-tuning.

Evaluation metric. The official metric is **accuracy**, with evaluation designed to account for valid response variation. In the SAQ track, a generated answer is considered correct if it matches any acceptable human-annotated response for the same question. In the MCQ track, accuracy is computed based on whether the selected option matches the correct culturally appropriate choice.

3 System Overview

Our SemEval-2026 Task 7 submission uses *activation steering* as an inference-time intervention for culturally aware multilingual inference. Instead of fine-tuning model parameters, we intervene at inference time by adding a steering vector to the residual stream at a selected transformer layer.

The central hypothesis is that language identity is encoded as a direction in activation space (Marks and Tegmark, 2023) and that steering along this direction may modulate access to culturally relevant knowledge for a target language-region pair. We therefore construct language vectors from multilingual sentence representations and inject them during decoding.

The system has three components:

1. **Off-line Language vector extraction** from FLORES-based multilingual data;

2. **Inference-time activation steering** with tunable strength β;

3. **Development-time model / layer selection / steering strength** using the SemEval-2026 development phase.

In the final submission, we selected a single steering configuration based on development performance and applied it to both shared-task tracks.

3.1 Language Vector Extraction

We compute language vectors from FLORES (Team et al., 2022) sentences by averaging residual-stream activations and taking a difference of means similar to the approach used in AxBench (Wu et al., 2025). Let $h^{(l)}(x)$ denote the post-normalization residual-stream activation at layer l for input sentence x. For a target language ℓ, the language vector is defined as:

$$v_\ell^{(l)} = \frac{1}{|D_\ell|} \sum_{x \in D_\ell} h^{(l)}(x) - \frac{1}{|D_{\neg\ell}|} \sum_{x \in D_{\neg\ell}} h^{(l)}(x), \quad (1)$$

where D_ℓ is the set of sentences for the target language and $D_{\neg\ell}$ is the set of sentences from the remaining languages.

Activation extraction details: We use the **post-normalization residual stream** and compute the mean activation over **all tokens** in each sentence. Sentences are processed one at a time, and no additional prompt template is used during vector extraction (i.e., we feed the original FLORES sentence directly).

FLORES mapping to shared-task language-region pairs: BLEnD targets language-region pairs (e.g., `ar-DZ`, `es-MX`), while FLORES (Team et al., 2022) provides language/script identifiers. We therefore define a mapping from shared-task pairs to FLORES language codes. For some cases where an exact regional mapping is unavailable in FLORES (e.g., multiple regions sharing the same language variety), we approximate using the closest available language-level FLORES code (e.g., a shared Spanish code for multiple Spanish-speaking regions). We provide the full mapping in Appendix A.

Data size and preprocessing: For each mapped language, we use the first **1,000** available FLORES dev sentences (Team et al., 2022) to compute the vector. We do not apply additional preprocessing beyond standard tokenization by the model tokenizer. A sample-size convergence study in Appendix B shows that the resulting DiffMean directions are already highly stable at substantially smaller sample sizes across the models we analyze.

3.2 Inference-Time Steering

During inference, we steer the hidden state at a selected transformer layer:

$$\tilde{h}^{(l)} = h^{(l)} + \beta \cdot v_\ell^{(l)}, \quad (2)$$

where $v_\ell^{(l)}$ is the language vector for the target language and β is a scalar steering strength.

We evaluated a small set of steering strengths $\beta \in \{1, 3, 5\}$ during development and find that $\beta = 1$ performs best for cultural steering in our setting. This value is used in the final submission.

3.3 Development-Time Model and Layer Selection

We perform model and layer selection during the SemEval development phase by evaluating a set of multilingual instruction-tuned LLMs and candidate steering layers. We tested older and newer models in different sizes that have proven to perform well in multilingual settings, including Qwen2.5-72B-Instruct and Qwen2.5-7B-Instruct (Team, 2024), Aya Expanse 8B and Aya Expanse 32B (Dang et al., 2024), and Qwen3-8B and Qwen3-32B (Team, 2025).

Based on development performance, we select **Qwen2.5-72B-Instruct** with steering applied at **Layer 26** for the final shared-task submission.

4 Experimental Setup

4.1 Decoding and Inference

We use greedy decoding (`temperature=0`) for both tracks to minimize confounding factors when evaluating activation steering. Since our method intervenes directly on internal representations, stochastic decoding (e.g., sampling with nonzero temperature) would introduce additional variance that can obscure whether performance changes are caused by the intervention or by decoding randomness. Deterministic decoding therefore allows a clearer attribution of gains or degradations to the steering configuration (layer and β), and improves reproducibility across layer sweeps and prompt comparisons.

Track 2 (MCQ). For each question, we prompt the model to choose one option from `A/B/C/D`. We score the four answer letters using their **output log-probabilities** and select the option with the highest log-probability. We generate at most **1 token**.

Track 1 (SAQ). We generate up to **32 tokens** to balance completeness and evaluation stability. Although SAQ targets concise answers, the required length varies across languages due to tokenization and morphology (e.g., multi-word expressions), and overly small limits risk truncating otherwise

correct answers. At the same time, longer generations increase the chance of irrelevant continuations that can hurt near-exact matching. To reduce formatting artifacts, we apply a lightweight normalization procedure to the generated text (Normalization details in Appendix C).

4.2 Prompting Strategy

We evaluate two prompt formulations for both tracks during analysis: a **generic prompt** and a **cultural prompt**. The official shared-task submission uses the **cultural prompt**.

Generic prompt. The generic prompt instructs the model to answer the question (or select one MCQ option) without explicitly mentioning the target region or language in the instruction text.

Generic prompt Template

Select exactly one option: A, B, C, or D.
Question: {question}
A. {option_a}
B. {option_b}
C. {option_c}
D. {option_d}
Answer (A/B/C/D):

Cultural prompt (official submission). The cultural prompt explicitly conditions the model on the target region and language (e.g., "for someone living in [region]" and "respond in [language]"). For SAQ, it additionally instructs the model to produce a concise answer without explanation.

Cultural prompt Template

You are answering a multiple-choice question for someone living in {Region}. Respond strictly in {Language} and select exactly one option: A, B, C, or D.
Question: {question}
A. {option_a}
B. {option_b}
C. {option_c}
D. {option_d}
Answer (A/B/C/D):

4.3 Hyperparameters

We select the steering strength from $\beta \in \{1, 3, 5\}$ on the SemEval-2026 development phase and use $\beta = 1$ in the final submission. We run layer sweeps to locate the best depth for steering. The steering layer (Layer 26) and backbone model (Qwen2.5-72B-Instruct) are also chosen based on development performance for the official results.

Track	Metric	Score	Rank
Track 1 (SAQ)	Acc.	N/A	– / 10
Track 2 (MCQ)	Acc.	86.96	7 / 17

Table 1: Official SemEval-2026 Task 7 results for our submission. The official submission used the **cultural prompt**. Our SAQ submission was not evaluated due to a corrupted/incorrect file and therefore has no official score.

Locale	Ours (%)	Our rank	Best (%)	Gap
es-EC	97.54	7	98.67	−1.13
en-GB	96.12	6	99.17	−3.05
es-MX	94.94	4	99.32	−4.38
ar-EG	94.84	2	91.03	+3.81
bg-BG	94.60	8	99.54	−4.94

Table 2: Top-5 language–region pairs (Track 2 MCQ) by our official accuracy. "Best" denotes the top-ranked system on the leaderboard (overall winner). Positive gap indicates our score exceeds the winner's per-locale score in the excerpt.

5 Results and Analysis

5.1 Official Shared-Task Results

Due to an incorrect/corrupted submission file, our Track 1 (SAQ) submission was not successfully evaluated by the organizers and therefore has no official score. We therefore report official leaderboard results only for Track 2 (MCQ).

Using the cultural prompt and activation steering, our Track 2 system achieved **86.96%** overall accuracy and ranked **7th** out of **17** teams (Table 1). The best-performing system on the leaderboard achieved **96.78%**, leaving a gap of **9.82** percentage points to our submission.

Table 2 lists five language-region pairs where our system performs best at the MCQ track. For each locale, we report our accuracy and our locale-specific rank (based on the official leaderboard). For ar-EG, we obtain **94.84%** while the overall winner system reports **91.03%**, meaning we outperform the winning system by **3.81** percentage points on this locale. In contrast, for bg-BG, we reach **94.60%** while the winner achieves **99.54%**, leaving a **4.94** percentage-point deficit. This heterogeneity aligns with our post-hoc analyses, which indicate that both steering and prompting effects are highly locale-dependent.

5.2 Post-hoc Analysis

To characterize system behavior beyond the single locked submission configuration, we ran post-hoc

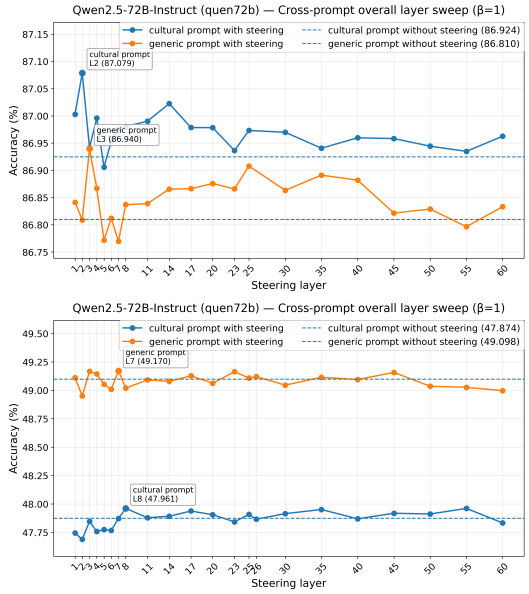

Figure 2: Post-hoc cross-prompt layer sweeps for Qwen2.5-72B-Instruct with $\beta = 1$ on **MCQ** (top) and **SAQ** (bottom). The official submission uses the **cultural prompt**.

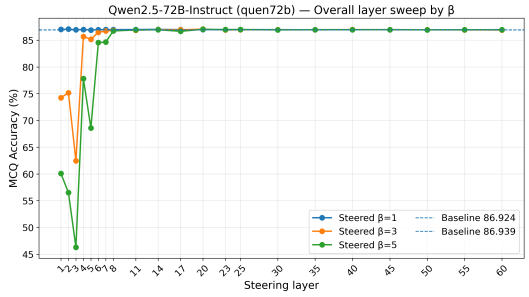

Figure 3: Post-hoc overall MCQ layer sweeps for Qwen2.5-72B-Instruct under different steering strengths ($\beta \in \{1, 3, 5\}$).

analyses on both MCQ and SAQ evaluation data across multiple models (Qwen2.5-72B/7B, Aya Expanse 8B/32B, Qwen3 8B/32B) using the same evaluation metrics provided by the SemEval-2026 organizers for each track. We observe:

(i) **strong layer sensitivity**: steering gains concentrate in a subset of layers while some layers degrade performance (e.g., the MCQ/SAQ cross-prompt layer sweeps in Figure 2); notably, for Qwen2.5-72B the best steering layer differs by both task and prompt (MCQ peaks at Layer 2 vs. 3, and SAQ peaks at Layer 8 vs. 7 for cultural vs. generic), illustrating that a single global layer choice is a compromise. Notably, the best post-hoc layer differs from the layer selected for the official submission due to differences in evaluation split.

(ii) β **sensitivity**: larger steering strengths are more prone to early-layer instability, whereas smaller strengths are generally more robust; in practice we found $\beta = 1$ to be the most reliable setting for Qwen2.5 models (Figure 3), while some Qwen3/Aya configurations tolerate stronger steering in post-hoc sweeps (Appendices D and E).

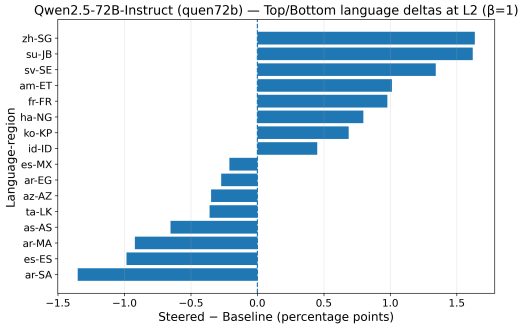

Figure 4: Top and bottom per-language MCQ accuracy changes (steered minus baseline, percentage points) using the cultural prompt.

(iii) **prompt-task interaction**: cultural prompting tends to be stronger for MCQ, where it conditions choice probabilities without changing output format, whereas the generic prompt is often better for SAQ across several models, likely because SAQ scoring depends on matching short surface forms and culturally conditioned prompts can induce verbose or stylistically marked responses (Figure 2). For example, for the SAQ item *"What is a popular snack at an amusement park in Azerbaijan?"*, the generic prompt yields a short candidate (*"Somsa/Samsa"*), while the cultural prompt produces a longer explanatory response (e.g., *"A popular snack at an amusement park in Azerbaijan is pakhlava, a sweet pastry..."*), which is more likely to fail evaluation even when broadly plausible. This aligns with findings that cultural prompting can be beneficial but is not uniformly effective across settings (Tao et al., 2024).

(iv) **model- and locale-dependent effects**: steering impacts vary substantially across language-region pairs (Figure 4), with some locales showing large gains and others degradations, and these patterns are not uniform across models, motivating model- and locale-aware steering policies in future work.

(v) **model- and β effects**: We do not observe a simple monotonic relationship between model parameter count or depth and the optimal steering strength β in our post-hoc sweeps. The preferred β appears model- and setting-dependent: across our evaluated models, $\beta = 1$ is the safest default, while

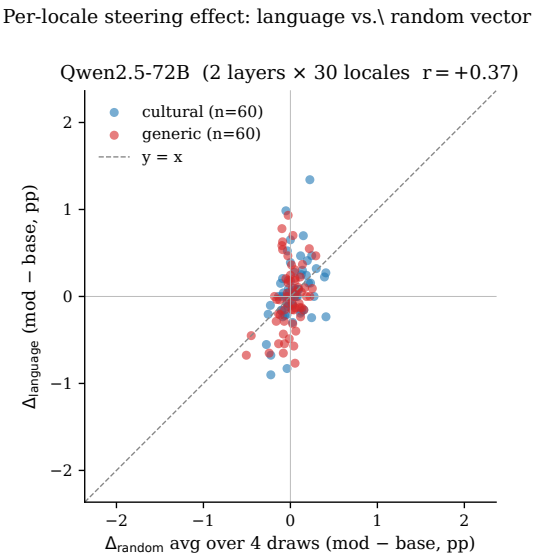

Figure 5: Per-locale steering effect for Qwen2.5-72B: Δ_{random} averaged over four Gaussian draws (x-axis) vs. Δ_{language} (y-axis), using the two dev-selected layers and both prompts. Each point is a (layer, prompt, locale) cell; n=60 per prompt. Random-vector effects concentrate near zero, while language-vector effects span a wider range and include negative outliers.

a few Qwen3/Aya configurations tolerate stronger steering in localized layers (Appendices D and E). We therefore caution against treating β as a function of scale alone, and recommend re-tuning it per model and prompt.

(vi) **random vs. language vector effects**: To check whether language-vector effects are distinguishable from generic activation perturbations, we compare them against L2-normalized Gaussian random vectors at the same layers with the same $\beta = 1$ intervention (Appendix F). For Qwen2.5-72B, random-vector effects remain concentrated near zero after averaging over four draws, while language-vector effects are somewhat more dispersed and include negative outliers (Figure 5). This suggests that random perturbations do not fully explain the language-vector effects, but the effects are also not reliably beneficial.

6 Discussion

Our post-hoc analyses indicate that activation steering for cultural MCQ/SAQ reasoning yields modest and highly context-dependent improvements rather than uniform gains. First, the steering effect is strongly layer-sensitive, with improvements concentrated in a subset of layers and other layers degrading performance. Second, per-config means stay under 0.5 pp on either track because most layers are neutral, and gains on one track do not predict gains on the other, so a single global steering layer (e.g., the Layer 26 used for the official submission) cannot be optimal for every (locale, track) pair. Third, prompt design interacts with steering in non-trivial ways: the cultural prompt used for the official submission and a simpler generic prompt produce different optimal steering layers and different per-language gains.

These findings indicate that prompt design and activation steering should be treated as a jointly optimized inference-time adaptation problem rather than independent components.

7 Limitations and Future Work

- **Official evaluation coverage.** Our Track 1 (SAQ) submission was not officially evaluated because of a corrupted file. All SAQ results are therefore *post-hoc* offline re-evaluations and not comparable to the official leaderboard. Future submissions should include stricter package validation before upload.

- **Scope of empirical comparison.** We analyze sensitivity to layer, β, prompt, model, and locale, but do not exhaustively compare against stronger prompt-only baselines, fine-tuning, or alternative steering methods such as CAA, ReFT, or SAE-based steering. Future work should benchmark DiffMean steering against these adaptation methods under matched compute and evaluation settings.

- **Language-derived vectors and cultural conflation.** Our vectors are derived from FLORES language-level data rather than culturally annotated or task-specific data. This conflates language identity with culture: several language–region pairs share the same FLORES code, so within-language regional variation is not captured. Future work should compare FLORES-based vectors with culture-specific and task-specific steering directions.

- **Single global steering configuration.** Our official submission uses one global (β, layer) pair, although post-hoc analyses show that locally optimal settings vary across models, prompts, layers, and locales. Future work should explore adaptive per-language, per-locale, or per-prompt steering policies.

Acknowledgments

This research was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project TRAILS (01IW24005).

References

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.

Debanjan Ghosh, Kai North, Ekaterina Kochmar, Mamoru Komachi, and Marcos Zampieri, editors. 2026. *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, San Diego, United States.

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

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Danial Namazifard and Lukas Galke Poech. 2025. Isolating culture neurons in multilingual large language models. 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 768–785.

Nedjma Ousidhoum, Junho Myung, Carla Perez-Almendros, Jiho Jin, Amr Keleg, Meriem Beloucif, Yi Zhou, Rodrigo Agerri, Vladimir Araujo, Naomi Baes, James Barry, Joanne Boisson, Nancy F. Chen, Christine de Kock, Aleksandra Edwards, Joseba Fernandez de Landa, Mohamed Fazli Imam, Huda Hakami, Shu-Kai Hsieh, and 11 others. 2026. SemEval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, 51(3):907–1004.

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.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and 1 others. 2024. CULTURALBENCH: A human-verified benchmark for evaluating cultural knowledge in large language models. *ArXiv preprint*, abs/2410.02677.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

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

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025. Qwen3 technical report.

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. *ArXiv preprint*, abs/2501.17148.

Appendix

A FLORES Mapping for Shared-Task language-region Pairs

We map BLEnD language-region pairs to FLORES language/script identifiers to compute language vectors. In cases where FLORES does not provide a region-specific variety, we use the closest available language-level approximation (e.g., a shared Spanish FLORES code for multiple Spanish-speaking regions). Table 3 provides the complete mapping from BLEnD language-region pairs to FLORES language/script identifiers used to compute language vectors.

BLEnD locale	FLORES code	BLEnD locale	FLORES code	BLEnD locale	FLORES code	BLEnD locale	FLORES code
am-ET	amh_Ethi	ar-DZ	kab_Latn	ar-EG	arz_Arab	ar-MA	ary_Arab
ar-SA	ars_Arab	as-AS	asm_Beng	az-AZ	azj_Latn	bg-BG	bul_Cyrl
el-GR	ell_Grek	en-AU	eng_Latn	en-GB	eng_Latn	en-US	eng_Latn
es-EC	spa_Latn	es-ES	spa_Latn	es-MX	spa_Latn	eu-ES	eus_Latn
eu-PV	eus_Latn	fa-IR	pes_Arab	fr-FR	fra_Latn	ga-IE	gle_Latn
ha-NG	hau_Latn	id-ID	ind_Latn	ja-JP	jpn_Jpan	ko-KP	kor_Hang
ko-KR	kor_Hang	ms-SG	zsm_Latn	su-JB	jav_Latn	sv-SE	swe_Latn
ta-SG	tam_Taml	tl-PH	tgl_Latn	zh-CN	zho_Hans	zh-TW	zho_Hant
zh-SG	zsm_Latn	en-AS	asm_Beng	en-AZ	azj_Latn	en-BG	bul_Cyrl
en-CN	zho_Hans	en-DZ	kab_Latn	en-EG	arb_Latn	en-ES	spa_Latn
en-ET	amh_Ethi	en-FR	fra_Latn	en-GR	ell_Grek	en-ID	ind_Latn
en-IE	gle_Latn	en-IR	pes_Arab	en-JP	jpn_Jpan	en-KP	kor_Hang
en-KR	kor_Hang	en-LK	sin_Sinh	en-MA	arb_Latn	en-MX	spa_Latn
en-NG	hau_Latn	en-PH	tgl_Latn	en-PV	eus_Latn	en-SA	ars_Arab
en-SE	swe_Latn	en-SG	tam_Taml	en-TW	zho_Hant	en-EC	spa_Latn
en-JB	jav_Latn	ta-LK	sin_Sinh				

Table 3: Complete mapping from BLEnD language–region pairs to FLORES language/script identifiers used for language vector computation. Some mappings are approximations when an exact region-specific FLORES variety is unavailable.

B FLORES Sample-Size Convergence Study

We test whether the DiffMean language vectors used in §3.1 are sensitive to the number of FLORES sentences used for extraction. For each of the six post-hoc models (Qwen2.5-72/7B-Instruct, Qwen3-32/8B, and Aya-Expanse-32/8B) we re-estimate vectors for the same 28 FLORES languages at $N \in \{100, 200, \ldots, 1000\}$. For each language and layer, we compute $v_N^{(\ell,l)}$ from the first N FLORES dev sentences, using the same post-normalization residual-stream activations and token averaging as in §3.1. Since each N is a strict prefix of the $N{=}1000$ set, differences from v_{1000} isolate the effect of adding more sentences rather than changing the sample.

We measure convergence with cosine similarity, $\cos(v_N^{(\ell,l)}, v_{1000}^{(\ell,l)})$, for every language–layer–N cell. Figure 6 summarizes the results with per-language curves, averaged over layers, and a joint median with 25–75% IQR over all language–layer cells.

Across all six models, the joint median is already at least 0.99 at $N{=}100$ and reaches at least 0.999 by $N{=}500$; the IQR is essentially collapsed near 1.0 from about $N{=}300$ onward. Low-N outliers are model-dependent: the worst per-language layer means at $N{=}100$ range from about 0.96 for Aya models to about 0.86 for Qwen2.5-7B. Qwen2.5-72B does not show worse stability than smaller models, suggesting that greater depth does not require more FLORES data within the tested $[100, 1000]$ range.

These results indicate that using $1{,}000$ FLORES sentences is conservative for the six models studied here. Since the downstream intervention uses the unit-norm direction v with $\beta = 1$, a cosine similarity of 0.99 corresponds to a steering-direction change below $\arccos(0.99) \approx 8°$, smaller than the layer-to-layer variation observed in our steering sweeps. We therefore do not expect FLORES sample size to be a major source of instability in our reported results.

C Post-processing for SAQ

For SAQ evaluation, we normalize model outputs using simple string cleanup heuristics:

- truncate at <|end_of_text|> if present;
- keep only the first line;
- keep text before the first period;
- collapse repeated whitespace;
- remove quotation marks.

This post-processing is applied before matching generated answers to the set of human-annotated acceptable responses.

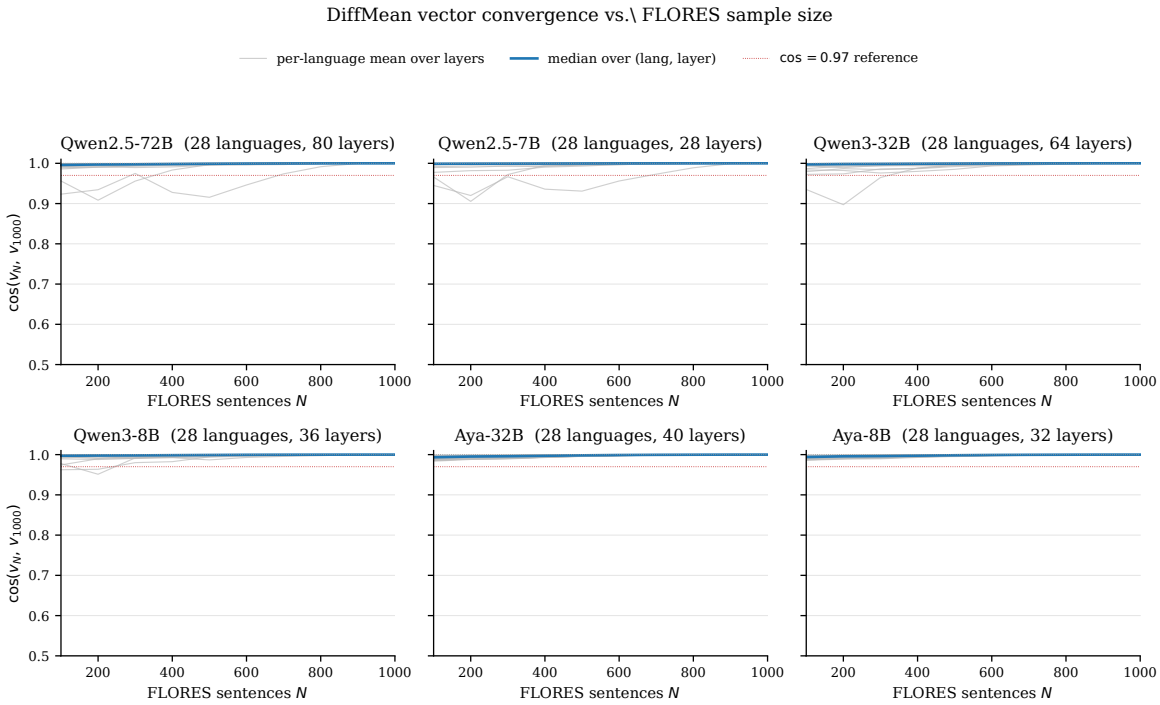

Figure 6: DiffMean vector convergence vs. FLORES sample size: $\cos(v_N, v_{1000})$ as a function of N for all six post-hoc models over 28 FLORES languages. Faint grey curves are individual languages, averaged over layers; the bold curve is the joint median over language–layer pairs, with the shaded 25–75% IQR. The dotted reference line marks $\cos = 0.97$, and the y-axis is zoomed to show sub-1 variation.

D Post-hoc MCQ Analysis Plots

This appendix provides additional post-hoc analysis plots for all tested models.

D.1 Qwen2.5-72B-Instruct

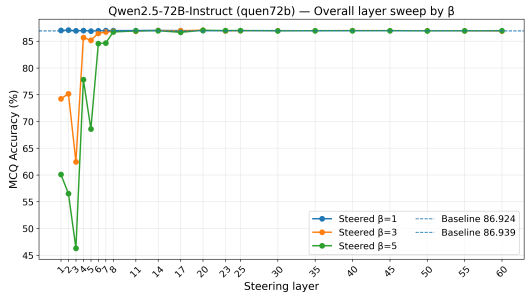

Figure 7: Post-hoc overall MCQ layer sweeps for Qwen2.5-72B-Instruct under different steering strengths ($\beta \in \{1, 3, 5\}$). Large steering strengths can substantially degrade performance in early layers, while $\beta = 1$ remains stable and yields the best overall trade-off in our experiments.

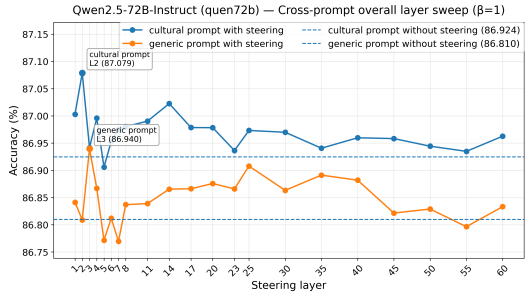

Figure 8: Post-hoc cross-prompt MCQ layer sweep for Qwen2.5-72B-Instruct with $\beta = 1$. The official submission uses the **cultural prompt**. Prompt choice affects both baseline accuracy and the optimal steering layer (here, Layer 2 for the cultural prompt and Layer 3 for the generic prompt).

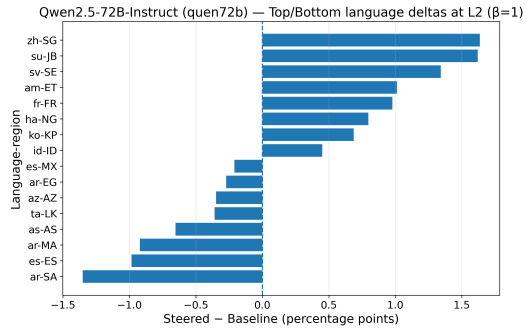

Figure 9: Top and bottom per-language MCQ accuracy changes (steered minus baseline, percentage points) for Qwen2.5-72B-Instruct at Layer 2 with $\beta = 1$ using the cultural prompt. Steering produces substantially different effects across language-region pairs, including both strong gains and degradations.

D.2 Qwen2.5-7B-Instruct

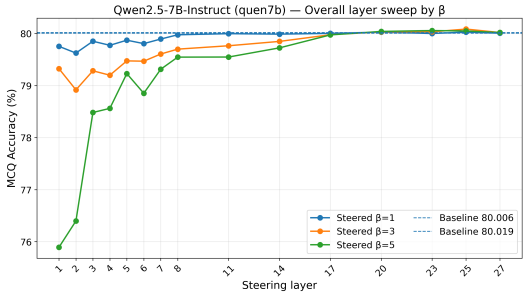

Figure 10: Post-hoc overall MCQ layer sweeps for Qwen2.5-7B-Instruct under different steering strengths ($\beta \in \{1, 3, 5\}$). Large steering strengths can substantially degrade performance in early layers, while $\beta = 1$ remains stable and yields the best overall trade-off in our experiments.

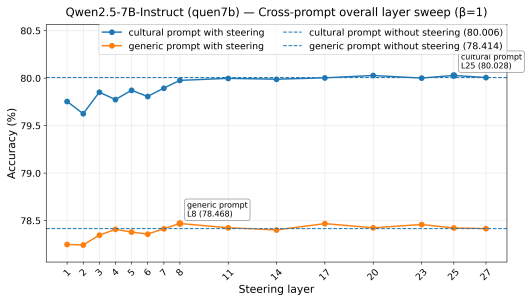

Figure 11: Post-hoc cross-prompt MCQ layer sweep for Qwen2.5-7B-Instruct with $\beta = 1$. The official submission uses the **cultural prompt**. Prompt choice affects both baseline accuracy and the optimal steering layer (here, Layer 25 for the cultural prompt and Layer 8 for the generic prompt).

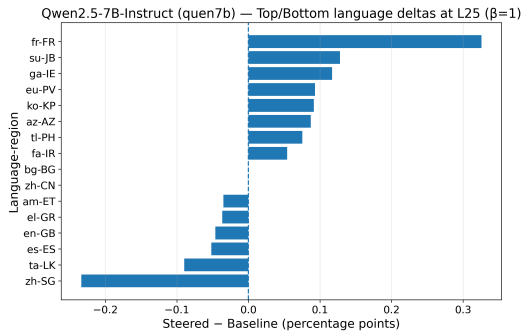

Figure 12: Top and bottom per-language MCQ accuracy changes (steered minus baseline, percentage points) for Qwen2.5-7B-Instruct at Layer 25 with $\beta = 1$ using the cultural prompt. Steering produces substantially different effects across language-region pairs, including both strong gains and degradations.

D.3 Aya Expanse 8B

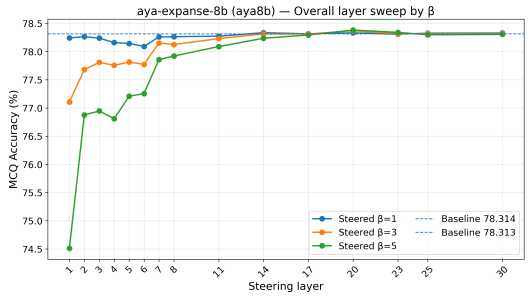

Figure 13: Post-hoc overall MCQ layer sweeps for Aya Expanse 8B under different steering strengths ($\beta \in \{1, 3, 5\}$). Large steering strengths can substantially degrade performance in early layers, while $\beta = 1$ remains stable and yields the best overall trade-off in our experiments.

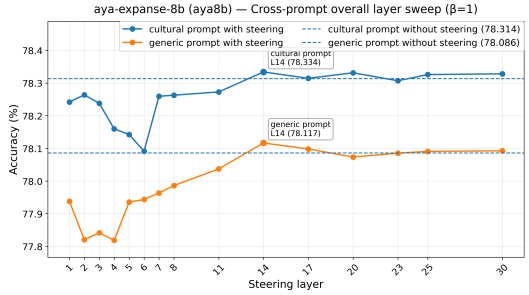

Figure 14: Post-hoc cross-prompt MCQ layer sweep for Aya Expanse 8B with $\beta = 1$. The official submission uses the **cultural prompt**. Prompt choice affects the baseline accuracy but, with this model, delivers the same optimal steering layer (Layer 14 for both the cultural and generic prompt).

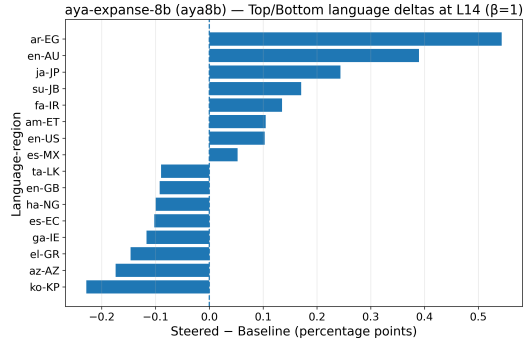

Figure 15: Top and bottom per-language MCQ accuracy changes (steered minus baseline, percentage points) for Aya Expanse 8B at Layer 14 with $\beta = 1$ using the cultural prompt. Steering produces substantially different effects across language-region pairs, including both strong gains and degradations.

D.4 Aya Expanse 32B

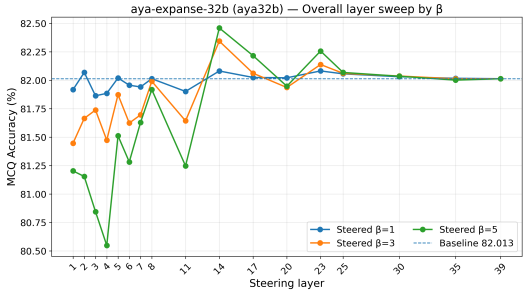

Figure 16: Post-hoc overall MCQ layer sweeps for Aya Expanse 32B under different steering strengths ($\beta \in \{1, 3, 5\}$). Large steering strengths can substantially degrade performance in early layers, while improving or meeting performance in mid to late layers. In comparison, $\beta = 1$ remains stable across layers although $\beta = 5$ yields the best overall Acc in our experiments.

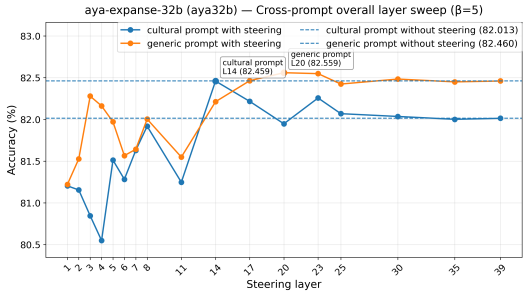

Figure 17: Post-hoc cross-prompt MCQ layer sweep for Aya Expanse 32B with $\beta = 5$. The official submission uses the **cultural prompt**. Prompt choice affects both baseline accuracy and the optimal steering layer (here, Layer 14 for the cultural prompt and Layer 20 for the generic prompt).

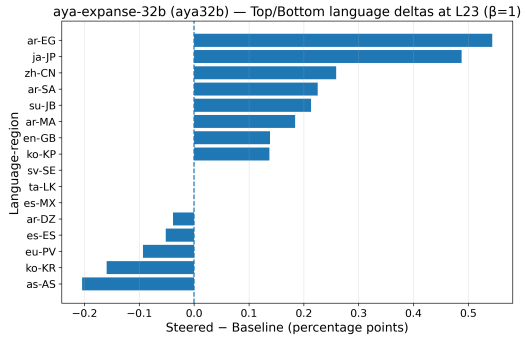

Figure 18: Top and bottom per-language MCQ accuracy changes (steered minus baseline, percentage points) for Aya Expanse 32B at Layer 23 with $\beta = 1$ using the cultural prompt. Steering produces substantially different effects across language-region pairs, including both strong gains and degradations.

D.5 Qwen3-8B

Figure 19: Post-hoc overall MCQ layer sweeps for Qwen3-8B under different steering strengths ($\beta \in \{1, 3, 5\}$). Large steering strengths can substantially degrade performance in early layers, while $\beta = 1$ remains stable and yields the best overall trade-off in our experiments.

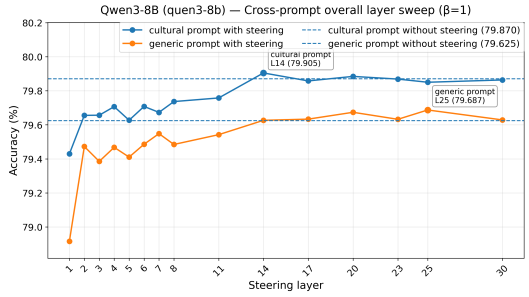

Figure 20: Post-hoc cross-prompt MCQ layer sweep for Qwen3-8B with $\beta = 1$. The official submission uses the **cultural prompt**. Prompt choice affects both baseline accuracy and the optimal steering layer (here, Layer 14 for the cultural prompt and Layer 25 for the generic prompt).

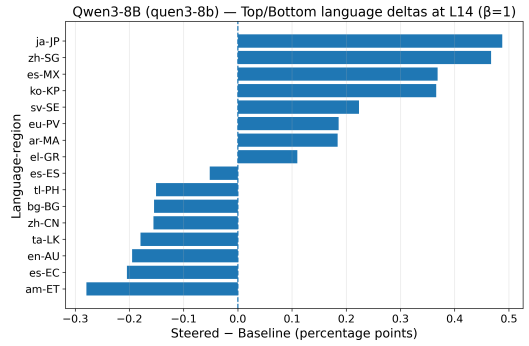

Figure 21: Top and bottom per-language MCQ accuracy changes (steered minus baseline, percentage points) for Qwen3-8B at Layer 14 with $\beta = 1$ using the cultural prompt. Steering produces substantially different effects across language-region pairs, including both strong gains and degradations.

D.6 Qwen3-32B

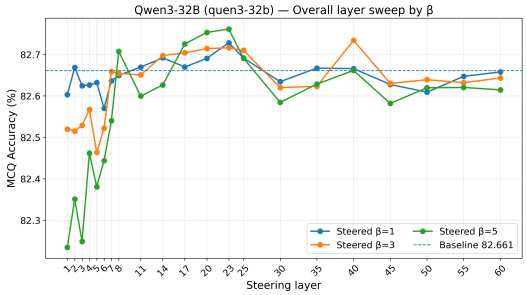

Figure 22: Post-hoc overall MCQ layer sweeps for Qwen3-32B under different steering strengths ($\beta \in \{1, 3, 5\}$). Large steering strengths can substantially degrade performance in early layers, while $\beta = 1$ remains stable although $\beta = 5$ yields the best overall Acc in our experiments.

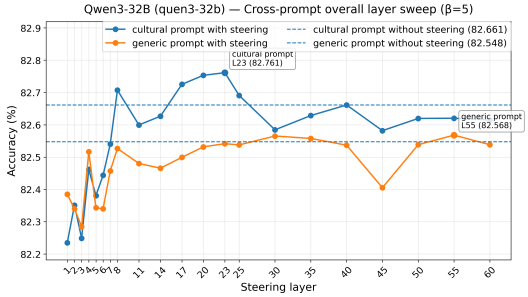

Figure 23: Post-hoc cross-prompt MCQ layer sweep for Qwen3-32B with $\beta = 5$. The official submission uses the **cultural prompt**. Prompt choice affects both baseline accuracy and the optimal steering layer (here, Layer 23 for the cultural prompt and Layer 55 for the generic prompt).

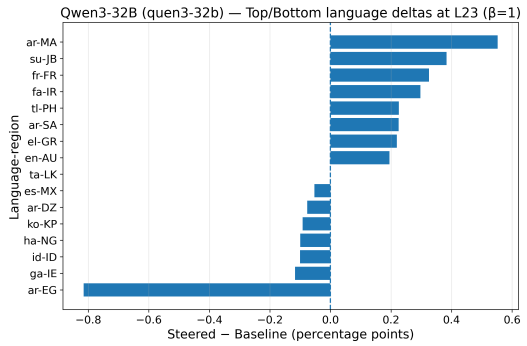

Figure 24: Top and bottom per-language MCQ accuracy changes (steered minus baseline, percentage points) for Qwen3-32B at Layer 23 with $\beta = 1$ using the cultural prompt. Steering produces substantially different effects across language-region pairs, including both strong gains and degradations.

E Post-hoc SAQ Analysis Plots

This appendix provides additional post-hoc analysis plots for all tested models.

E.1 Qwen2.5-72B-Instruct

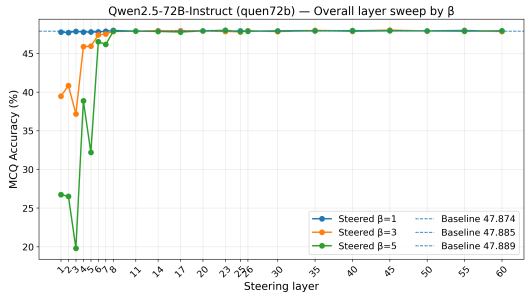

Figure 25: Post-hoc overall SAQ layer sweeps for Qwen2.5-72B-Instruct under different steering strengths ($\beta \in \{1, 3, 5\}$). Large steering strengths can substantially degrade performance in early layers, while $\beta = 1$ remains stable and yields the best overall trade-off in our experiments.

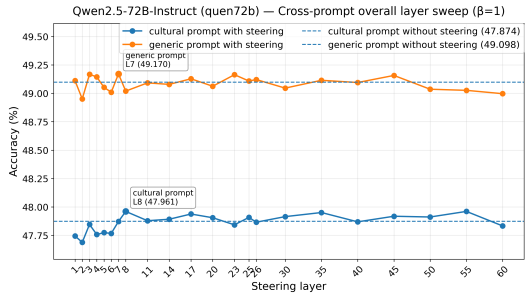

Figure 26: Post-hoc cross-prompt SAQ layer sweep for Qwen2.5-72B-Instruct with $\beta = 1$. The official submission uses the **cultural prompt**. Prompt choice affects both baseline accuracy and the optimal steering layer (here, Layer 8 for the cultural prompt and Layer 7 for the generic prompt).

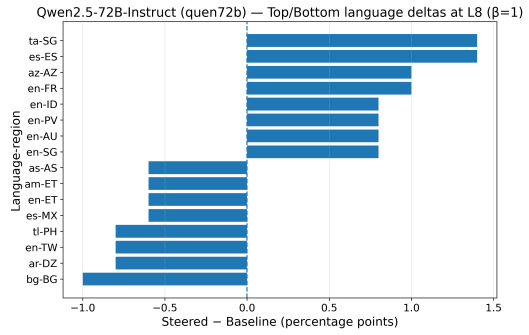

Figure 27: Top and bottom per-language SAQ accuracy changes (steered minus baseline, percentage points) for Qwen2.5-72B-Instruct at Layer 8 with $\beta = 1$ using the cultural prompt. Steering produces substantially different effects across language-region pairs, including both strong gains and degradations.

E.2 Qwen2.5-7B-Instruct

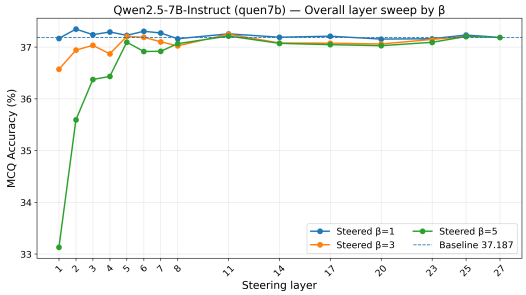

Figure 28: Post-hoc overall SAQ layer sweeps for Qwen2.5-7B-Instruct under different steering strengths ($\beta \in \{1, 3, 5\}$). Large steering strengths can substantially degrade performance in early layers, while $\beta = 1$ remains stable and yields the best overall trade-off in our experiments.

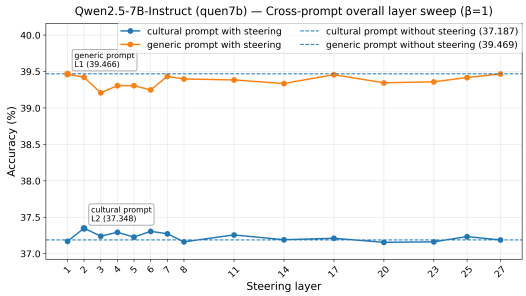

Figure 29: Post-hoc cross-prompt SAQ layer sweep for Qwen2.5-7B-Instruct with $\beta = 1$. The official submission uses the **cultural prompt**. Prompt choice affects both baseline accuracy and the optimal steering layer (here, Layer 2 for the cultural prompt and Layer 1 for the generic prompt).

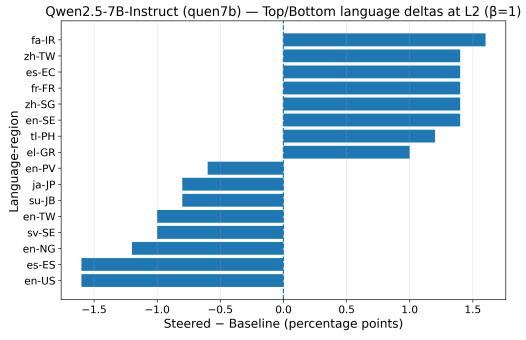

Figure 30: Top and bottom per-language SAQ accuracy changes (steered minus baseline, percentage points) for Qwen2.5-7B-Instruct at Layer 2 with $\beta = 1$ using the cultural prompt. Steering produces substantially different effects across language-region pairs, including both strong gains and degradations.

E.3 Aya Expanse 8B

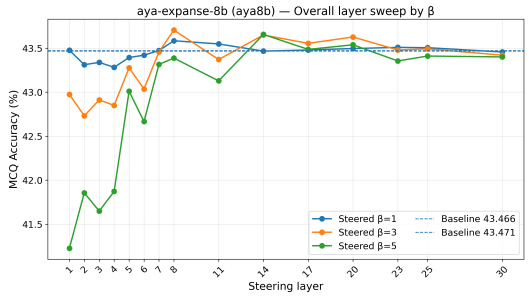

Figure 31: Post-hoc overall SAQ layer sweeps for Aya Expanse 8B under different steering strengths ($\beta \in \{1, 3, 5\}$). Large steering strengths can substantially degrade performance in early layers, while $\beta = 1$ remains stable and yields the best overall trade-off in our experiments.

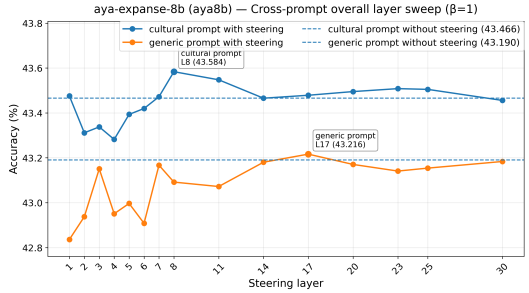

Figure 32: Post-hoc cross-prompt SAQ layer sweep for Aya Expanse 8B with $\beta = 1$. The official submission uses the **cultural prompt**. Prompt choice affects both baseline accuracy and the optimal steering layer (here, Layer 8 for the cultural prompt and Layer 17 for the generic prompt).

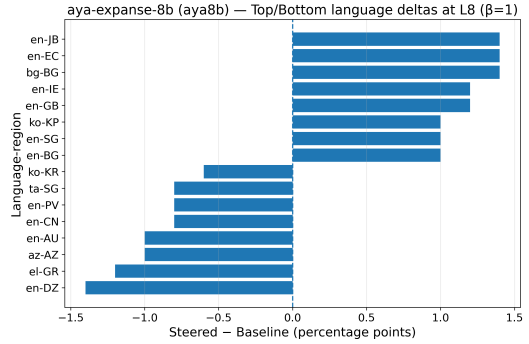

Figure 33: Top and bottom per-language SAQ accuracy changes (steered minus baseline, percentage points) for Aya Expanse 8B at Layer 8 with $\beta = 1$ using the cultural prompt. Steering produces substantially different effects across language-region pairs, including both strong gains and degradations.

E.4 Aya Expanse 32B

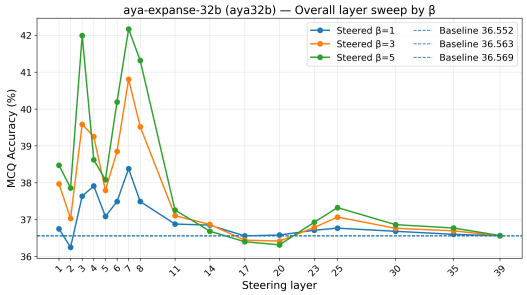

Figure 34: Post-hoc overall SAQ layer sweeps for Aya Expanse 32B under different steering strengths ($\beta \in \{1, 3, 5\}$). Large steering strengths can substantially improve performance in early layers and $\beta = 5$ yields the best overall Acc in our experiments.

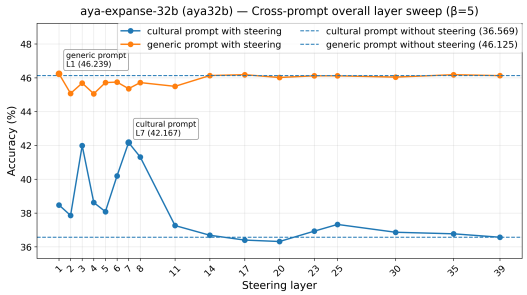

Figure 35: Post-hoc cross-prompt SAQ layer sweep for Aya Expanse 32B with $\beta = 5$. The official submission uses the **cultural prompt**. Prompt choice affects both baseline accuracy and the optimal steering layer (here, Layer 7 for the cultural prompt and Layer 1 for the generic prompt).

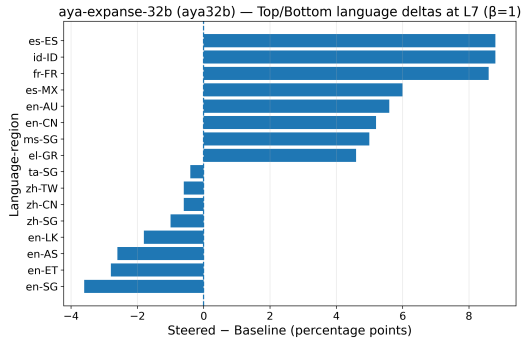

Figure 36: Top and bottom per-language SAQ accuracy changes (steered minus baseline, percentage points) for Aya Expanse 32B at Layer 7 with $\beta = 1$ using the cultural prompt. Steering produces substantially different effects across language-region pairs, including both strong gains and degradations.

E.5 Qwen3-8B

Figure 37: Post-hoc overall SAQ layer sweeps for Qwen3-8B under different steering strengths ($\beta \in \{1, 3, 5\}$). Large steering strengths can substantially degrade performance in early layers, while $\beta = 1$ remains stable and yields the best overall trade-off in our experiments.

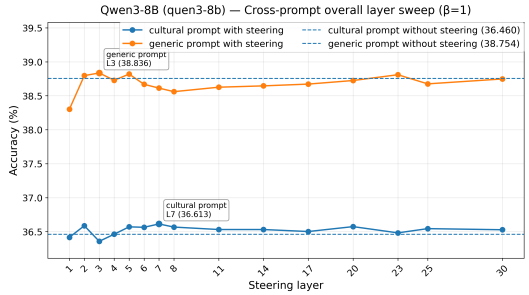

Figure 38: Post-hoc cross-prompt SAQ layer sweep for Qwen3-8B with $\beta = 1$. The official submission uses the **cultural prompt**. Prompt choice affects both baseline accuracy and the optimal steering layer (here, Layer 3 for the cultural prompt and Layer 7 for the generic prompt).

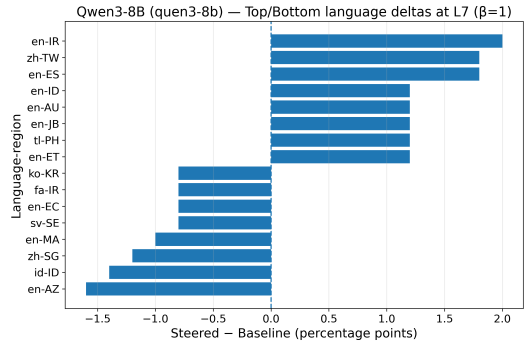

Figure 39: Top and bottom per-language SAQ accuracy changes (steered minus baseline, percentage points) for Qwen3-8B at Layer 7 with $\beta = 1$ using the cultural prompt. Steering produces substantially different effects across language-region pairs, including both strong gains and degradations.

E.6 Qwen3-32B

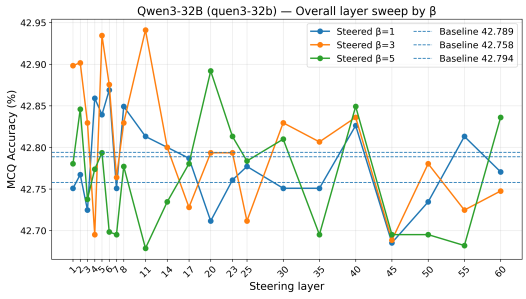

Figure 40: Post-hoc overall SAQ layer sweeps for Qwen3-32B under different steering strengths ($\beta \in \{1, 3, 5\}$). Steering strengths show unstable performance across all layers, while $\beta = 5$ yields the best overall Acc in our experiments.

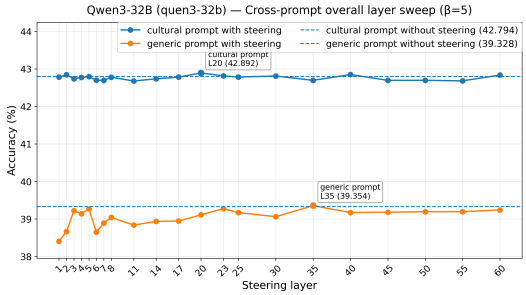

Figure 41: Post-hoc cross-prompt SAQ layer sweep for Qwen3-32B with $\beta = 5$. The official submission uses the **cultural prompt**. Prompt choice affects both baseline accuracy and the optimal steering layer (here, Layer 20 for the cultural prompt and Layer 35 for the generic prompt).

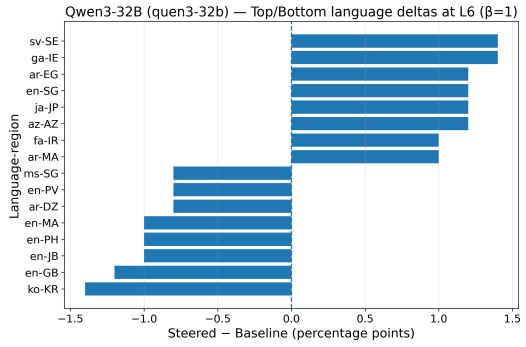

Figure 42: Top and bottom per-language SAQ accuracy changes (steered minus baseline, percentage points) for Qwen3-32B at Layer 6 with $\beta = 1$ using the cultural prompt. Steering produces substantially different effects across language-region pairs, including both strong gains and degradations.

F Random- vs. Language-Vector Steering

This appendix summarizes the random-vector control used to test whether language-vector steering effects reflect language-specific structure or generic activation perturbations.

Setup. For each model, layer, prompt, and locale, we compare two unit-norm steering directions applied with the same intervention

$$\tilde{h}^{(l)} = h^{(l)} + \beta u^{(l)}, \qquad \beta = 1.$$

The language direction is the FLORES DiffMean vector from §3.1,

$$v_\ell^{(l)} = \frac{1}{|D_\ell|} \sum_{x \in D_\ell} h^{(l)}(x) - \frac{1}{|D_{\neg\ell}|} \sum_{x \in D_{\neg\ell}} h^{(l)}(x),$$

computed from token-mean post-normalization residual activations and then L2-normalized. The random control samples

$$r_{\ell,k}^{(l)} \sim \mathcal{N}(0, I_d), \qquad \hat{r}_{\ell,k}^{(l)} = r_{\ell,k}^{(l)} / \|r_{\ell,k}^{(l)}\|_2,$$

with the same dimensionality and norm as the language vector. Within each draw k, the same random vector is reused for all locales sharing the corresponding FLORES language code, mirroring the language-vector setup.

Bootstrap. For the random baseline, we run four independent Gaussian draws and compute

$$\Delta_{\text{random},k} = \text{acc}_{\text{mod},k} - \text{acc}_{\text{base},k}.$$

We then average $\Delta_{\text{random},k}$ over draws and compare it to the single language-vector delta,

$$\Delta_{\text{language}} = \text{acc}_{\text{lang}} - \text{acc}_{\text{base}}.$$

Both deltas are reported in percentage points. The comparison is matched at the same model, layer, prompt, locale, and steering strength.

Interpretation. The Qwen2.5-72B comparison in Figure 5 shows that averaged random-vector effects remain concentrated near zero, while language-vector effects are somewhat more dispersed. This supports the main-text conclusion that language-vector steering is not simply equivalent to a generic Gaussian perturbation, but also does not yield reliably positive gains at the tested layers.