STEER: Sparse-Feature Tracing of Associative Knowledge Edits for Efficient Recovery

Anonymous ACL submission

Abstract

Human reasoning blends deliberate thinking with fast, intuitive associations, yet current language model training and evaluation largely ignore the associative dimension. We study what language models learn and sacrifice when fine-tuned on this intuitive behavior data. Using the multilingual Small World of Words, a psychological behavior dataset of speakers' mental associations, as a controlled probe, we fine-tune Llama-3.1-8B-Instruct on native-speaker word associations in four languages and examine what the model learns from this behavior data and at what price. We find that fine-tuning almost doubles the model's performance on human-like association generation, but reduces its general capability to 39–54%. We further identify a consistent mid-layer intervention band (L11–L14), where the SFT-induced representational shift is most useful for steering across all four languages. A steering vector at that band partially recovers the gain on association generation at minimal general capability cost, and sparse-autoencoder analysis reveals what separates the two: steering moves the feature basis in the same direction as fine-tuning (Pearson 0.78–0.86) but at a smaller magnitude, recovering register-shift features closely (72–91%) while leaving cultural-memory features only partially recovered (0–54%). Our work exposes the cost of full-parameter adaptation to culturally situated behavioral data and offers a transparent account of what a lightweight intervention can and cannot recover.

1 Introduction

Large language models are often evaluated on tasks that reward deliberate problem solving: solving mathematical questions, writing code, following instructions, or recalling facts (Grattafiori et al., 2024; Qwen et al., 2025; Hendrycks et al., 2021b; Chen et al., 2021). Human cognition, however, is not only deliberative. Much of it is associative (Collins and Loftus, 1975). When people see a word (e.g., smile) or object, related concepts are often triggered without conscious reasoning. These associations are intuitive System-1 thinking: sparse, automatic, and shaped by the environments and experiences in which people live (Kahneman, 2011).

Free word association offers a simple but powerful psychological tool to study this form of cognition (Rosenzweig, 1961). In this task, participants are given a cue word and asked to produce up to three words that come to mind (Wilson et al., 1988; Nelson et al., 2004). Aggregated across many speakers, these responses form a distribution over the mental representations that a group of speakers most readily associates with the cue. Large-scale multilingual and multicultural resources such as the Small World of Words (SWOW; Deyne et al., 2019) provide a rare window into how native speakers of different languages associate concepts differently, shaped by their lived experiences, environments, and cultural contexts. At the same time, these resources allow us to test whether language models can learn not only factual knowledge or reasoning skills, but also human-like associative behavior.

We treat associative learning as a basic but culturally and linguistically variable form of human capability. It tests whether a model can reproduce the distribution of concepts that native speakers intuitively bring to mind. Prior works have started to train language models to learn human-like associations (Liu et al., 2026; Dai et al., 2025), but they have not examined how such associative learning interacts with the model's internal representations and broader capabilities, and how the model captures variation across languages. To the best of our knowledge, this is the first study that examines associative learning, general capability, and internal representational change across multiple languages together.

This leads to our central question: *what does fine-tuning on culturally situated associations change inside a language model, and can we achieve a*

balance between human-like association generation and general capability by better understanding the model's internal learning behaviors? We study this question using SWOW as a controlled multilingual testbed across U.S. English, Mandarin Chinese, German, and Dutch. We fine-tune Llama-3.1-8B-Instruct (Grattafiori et al., 2024) on native-speaker associations, and conduct four linked analyses: (i) quantifying the impact of fine-tuning on general capability; (ii) localizing where the learned knowledge is stored layer by layer; (iii) constructing mean-difference steering vectors at those layers to recover the learned behavior without weight updates; and (iv) decomposing the learned representations with LlamaScope sparse autoencoders (He et al., 2024, SAE) to understand their feature properties.

Our results show that full-parameter fine-tuning improves association generation by $1.87\times$–$2.48\times$, but at the cost of reducing general capability to 39–54% capped retention, with mathematical reasoning and instruction following suffering most. Layer-by-layer analysis traces the representational shift from the baseline models to a consistent mid-layer band (L11–L14) in the fine-tuned models where SFT representations diverge most from vanilla. Importantly, we show that a single steering vector at that band can capture and transfer the learned associative ability, recovering 7–25% of the MASS@K gain while keeping capped capability retention at 84–95%. Interestingly, associative steering improves model performance on two general tasks, HumanEval and IFEval, across all four languages, suggesting that associative adaptation may be complementary to deliberate reasoning and instruction-following abilities when used appropriately.

Our contributions are: (1) the first end-to-end analysis of multilingual associative fine-tuning, spanning data, SFT, steering, and feature-level interpretation, and providing a transparent testbed for understanding how associative learning interacts with a language model's general capability; (2) evidence that the useful intervention direction for multilingual association learning emerges consistently in a shared mid-layer band, and that steering can partially recover the trade-off without weight updates; (3) a direct, feature-level comparison of steering and fine-tuning that quantifies exactly what a single steering direction can and cannot carry.[1]

[1]Anonymous code repo: https://anonymous.4open.science/r/swow-steering-artifact-FFBC.

2 Related Work

Word associations. Word association norms capture how people respond to a cue word with "the first word that comes to mind", providing a window into mental representations of semantic memory and cultural variations (Rosenzweig, 1961). Large-scale association norms (Deyne et al., 2019; Nelson et al., 2004) have been used to study value and gender bias (Ramezani and Xu, 2024; Du et al., 2019), multi-modal representation (De Deyne et al., 2021), commonsense knowledge (Liu et al., 2021, 2022), and cultural meanings (Lim et al., 2024).

Associative learning, cultural alignment, and fine-tuning trade-offs. Recent work has begun to use word associations to improve the cultural alignment of language models, showing that models can be trained to better reproduce human-like associations (Liu et al., 2026; Dai et al., 2025) and align with cultural specific tasks. This connects to a broader line of cultural NLP that asks whether language models capture culturally situated values and preferences (Liu et al., 2025; Durmus et al., 2024; Zhao et al., 2024). Yet the cost of such adaptation remains less clear. Cultural-alignment work is usually evaluated on culture-specific tasks, such as value alignment and cultural norms, but less often asks whether these improvements weaken the model's general capabilities. This matters because task-specific fine-tuning can improve target-task behavior while degrading general capability, especially when the task distribution is narrow (Kotha et al., 2024; Luo et al., 2025).

Steering and feature-level interpretation. Activation steering modifies behavior at inference time by adding directions to internal activations (Rimsky et al., 2024; Zou et al., 2023; Turner et al., 2024; Li et al., 2023) without fully updating model's weights. In this work, we explore whether activation steering can offer an alternative way to combine associative gains while preserving general capabilities. Layer-wise diagnostics such as the logit lens (nostalgebraist, 2020; Belrose et al., 2025) and analyses of layer-wise computation (Meng et al., 2022; Geva et al., 2023; Lad et al., 2024) are often used to localize where information is rewritten and where it becomes readable. Steering vectors control model outputs but are often entangled, offering limited feature-level interpretability. On the other hand, sparse autoencoders (Huben et al., 2024; Gao et al., 2025; Templeton et al., 2024) decompose activations into sparse feature bases

and have enabled feature-level analysis and editing (Marks et al., 2025), with public SAE suites such as Gemma Scope and LlamaScope (Lieberum et al., 2024; He et al., 2024).

3 Framework Overview

We address two questions: (1) what does the model gain from fine-tuning on SWOW and what does it cost in general capability (§ 4); and (2) can interpretability tools reveal where that knowledge is stored and whether a lighter-weight intervention can achieve a better trade-off (§ 5–7). We pursue the interpretability analysis through three interconnected components: localizing where models store the learned knowledge; constructing steering vectors to recover that knowledge without weight updates; and using sparse autoencoders to interpret the difference between steering and full fine-tuning at the feature level.

Datasets. We use word associations from the English, Mandarin Chinese, German, and Dutch versions of the Small World of Words project (De Deyne and Storms, 2008; Deyne et al., 2019; Li et al., 2024; Aeschbach et al., 2026). Each dataset contains cue words and, for each cue, a frequency-ranked distribution of free-association responses. We restrict the English data to U.S. respondents (SWOW-US). Table 1 reports split sizes. For cross-lingual SAE analysis, we additionally use a 215-cue overlap subset of basic-level concepts (colors, kinship, body parts, animals, etc.) whose translations are cues in every language. We identified candidate overlap cues with GPT-5.5; construction details are in Appendix D.

Lang.	Train	Valid	Test
EN-US	9,825	1,228	1,228
ZH	8,020	1,002	1,002
DE	4,701	588	588
NL	10,057	1,257	1,257
Aligned	–	–	215

Table 1: Dataset splits on four languages. The 215 aligned cues exist in all languages.

Models and Fine-tuning We use `Llama-3.1-8B-Instruct` (Grattafiori et al., 2024) as the base model. We use this setting because publicly available LlamaScope SAEs allow us to measure model behavior at the feature level. We train four language-specific frequency-aware SFT models: EN-SFT, ZH-SFT, DE-SFT, NL-SFT with full-parameter updates.[2] We adopt full-parameter SFT as a high-capacity adaptation setting because it gives the SWOW supervision maximal freedom to reshape the model, providing the clearest signal both for measuring the representational shift and for extracting the steering direction from the vanilla–SFT activation difference. Parameter-efficient alternatives such as LoRA (Hu et al., 2022) would answer a different question: whether a constrained low-rank update captures the same behavioral and feature-level changes, a question we leave to future work.

The **fine-tuning task is word association generation**: given a cue word, the model is instructed to generate an ordered list of associated words. The training objective is a frequency-aware objective that weights each output token by the empirical frequency mass of its parent association, $w_t = 1 + \lambda\, p_{\mathrm{gt}}(a(y_t))^{\gamma}$, and minimizes $\mathcal{L}_{\mathrm{freq}} = -Z^{-1}\sum_t w_t \log p_\theta(y_t \mid x, y_{<t})$ with $Z = \sum_t w_t$, λ=5, γ=0.5. Full derivation, training configuration, hyperparameters, per-language prompt templates, and convergence diagnostics are in Appendices A–C.

Task-specific Evaluation: Word Association Generation. We evaluate model performance on the task it is trained on to measure task-specific gains. We use the cues from the test sets (Table 1) to generate associations for each cue, and then compare the model outputs[3] against human associations in each language. We report MASS@K, Overlap@K, and Spearman. MASS@K measures the fraction of human-frequency probability mass recovered by the model's top-K outputs:

$$\text{MASS@}K = \frac{\sum_{a \in \widehat{A}_K \cap G_K} f(a)}{\sum_{a \in G_K} f(a)}. \qquad (1)$$

Here $\widehat{A}_K$ is the model top-K list and G_K is the gold top-K list ranked by human frequency. MASS@K directly aligns with the frequency-aware training objective. Overlap@K is the unweighted set-overlap fraction between top-K predicted and top-K gold associations, providing a frequency-agnostic complement to MASS@K. We set $K = 10$ to maintain consistency throughout the paper. Spearman is the rank correlation between predicted and gold response orderings over the union vocabulary. For steering experiments we

[2] A standard-SFT ablation without frequency weighting performs worse (Appendix G).

[3] See Appendix E for parsing details.

add the recovery rate, which is the fraction of the SFT MASS@K gain recovered by steering:

$$\text{Recovery@10} = \frac{M^{10}_{\text{steer}} - M^{10}_{\text{van}}}{M^{10}_{\text{SFT}} - M^{10}_{\text{van}}}, \quad (2)$$

where M^{10}_s abbreviates MASS@10 for system s.

General Task Evaluation We evaluate a model's general task performance on benchmarks that are commonly adopted for evaluating general-purpose language models (Grattafiori et al., 2024). Specifically, we use MMLU (Hendrycks et al., 2021a), GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b), IFEval (Zhou et al., 2023), HumanEval (Chen et al., 2021), and TruthfulQA (Lin et al., 2022) in our work. We report per-benchmark scores and capped capability retention, $\text{CAPRET} = 100|\mathcal{B}|^{-1}\sum_{b\in\mathcal{B}}\min(s_{m,b}/s_{\text{vanilla},b}, 1)$, so gains above vanilla on one benchmark cannot offset losses on another.

4 What Is the General Capability Trade-off in Supervised Fine-tuning?

For each language and task, we compare model performance on the base model (Vanilla) and the fine-tuned model. Table 2 shows test-set results (vanilla and frequency-aware SFT) on word association generation. Frequency-aware SFT substantially improves multilingual word-association generation, achieving gains of 1.87×–2.48× across four languages compared to the Vanilla model. Table 3 presents general-capability benchmark results. In contrast, frequency-aware SFT substantially reduces general capabilities, particularly on mathematical reasoning (MATH, GSM8K), instruction following (IFEval), and code generation (HumanEval). Overall, these results highlight a clear trade-off: full SFT roughly doubles MASS@K but at the expense of more than halving general capability, suggesting a shift in how the model represents and prioritizes knowledge after fine-tuning.

This pattern is consistent with catastrophic forgetting under task-focused SFT (Kotha et al., 2024; Luo et al., 2025). Notably, degradation is most severe in English, the highest-resource language, so resistance to forgetting does not simply track pretraining coverage. Part of this collapse is also output-format mismatch: the SFT model carries its SWOW comma-list output habit into non-SWOW prompts, so the strict OpenCompass extractors miss many on-topic responses — HumanEval reads 0.00 everywhere because the model emits Python without markdown code fences. A format-tolerant rescore (Appendix I) lifts the SFT raw-score means by +5.0 to +6.1, foreshadowing the SAE finding (§7.2) that SFT silences the assistant "broadcaster" register that orchestrates instruction-following formats. We next ask where SFT changes the model, how much of that change steering can recover, and what the feature-level difference between the two operating points reveals.

Lang.	Model	Mass@10	Overlap@10	Spearman
EN	Vanilla	0.361	0.267	0.414
	SFT_{freq}	**0.676**	**0.515**	**0.583**
ZH	Vanilla	0.156	0.118	0.234
	SFT_{freq}	**0.387**	**0.284**	**0.407**
DE	Vanilla	0.257	0.188	0.342
	SFT_{freq}	**0.525**	**0.381**	**0.489**
NL	Vanilla	0.216	0.147	0.314
	SFT_{freq}	**0.527**	**0.394**	**0.511**

Table 2: Vanilla vs. SFT_{freq} on the test set.

Model	MMLU	MATH	GSM8K	IFEval	HumanEval	TruthQA	CapRet
Vanilla	70.43	11.28	79.83	66.43	25.61	0.29	100.0
EN SFT_{freq}	27.48	9.12	3.56	26.62	0.00	0.20	38.9
ZH SFT_{freq}	48.84	15.32	6.22	46.88	0.00	0.22	53.8
DE SFT_{freq}	40.10	7.68	6.82	38.73	0.00	0.21	43.9
NL SFT_{freq}	33.78	6.94	6.52	38.25	0.00	0.19	39.9

Table 3: General capability before and after frequency-aware SFT. CapRet is capped retention relative to vanilla, averaged over benchmarks.

5 Layer Localization

We identify candidate intervention layers with two diagnostics. Cosine divergence asks where SFT changes the model's internal representations most strongly relative to vanilla. The logit lens asks where the correct association distribution becomes readable from intermediate hidden states under a gold-candidate readout.

Cosine divergence. For each layer ℓ, we compute centered cosine similarity between vanilla and SFT residual-stream activations at the final prompt token, averaged over validation cues:

$$\text{CCS}_\ell = \mathbb{E}_x\left[\cos\left(\tilde{h}^{\text{van}}_\ell(x), \tilde{h}^{\text{SFT}}_\ell(x)\right)\right], \quad (3)$$

where $\tilde{h}^s_\ell(x) = h^s_\ell(x) - \bar{h}^s_\ell$ is the centered residual-stream activation for a model s. We use the corresponding cosine divergence, $\text{DIV}_\ell = 1 - \text{CCS}_\ell$, so larger values mark layers where SFT has the most updated representations relative to vanilla.

Frequency-aware logit lens. Following nostalgebraist (2020), we project each layer's hidden state through the unembedding W_U and restrict to the gold-association vocabulary for each cue. We take the top-K most probable candidate tokens as the layer's predicted associations and score the resulting distribution against the gold distribution using $\Delta\text{MASS@10} = \text{MASS@10}_{\text{SFT}} - \text{MASS@10}_{\text{vanilla}}$. Because this score is computed only over gold association candidates, it is a layerwise readout diagnostic rather than a free-generation metric and should not be compared directly to test-set MASS@K.

Result. Figure 1 presents the cosine-divergence results. Across all four languages, divergence peaks at L10 and remains elevated around the selected steering layers. Figure 2 shows the logit-lens results: the late layers achieve higher ΔMASS@10. The two diagnostics are complementary: divergence peaks at L10 in all languages, while logit-lens gain peaks later (EN L19, ZH L20, DE/NL L29). We treat candidate layers from both as inputs to a validation sweep over layers and steering strengths (§6).

6 Can Layer-Wise Steering Balance Fine-Tuning Task Performance and General Capability?

Full SFT improves association generation but destroys general capability. Activation steering offers a middle path: it modifies model behavior at inference time by adding a direction to the residual stream, leaving weights untouched. We hypothesize that the mean-difference vector between SFT and vanilla activations at the most informative layer can distill the learned association signal with minimal capability cost. We identify this layer through a sweep over several layers and steering strengths on the validation set (Appendix H), then evaluate the steered model on both association generation and general benchmarks.

Steering vector construction and sweep method. For each language and candidate layer ℓ we compute the mean-difference steering vector $\mathbf{v}_\ell = \mathbb{E}_{x\sim D_{\text{valid}}}[h^{\text{SFT}}_\ell(x) - h^{\text{vanilla}}_\ell(x)]$ (Rimsky et al., 2024) at the final prompt token, post-block residual stream, with chat-template prompts matching generation. At inference we run vanilla Llama and add $\alpha\mathbf{v}_\ell$ back into the original representation to the residual stream at layer ℓ via a persistent forward hook (i.e., $h^{\text{steered}}_\ell(x) = h^{\text{old}}_\ell(x) + \alpha\mathbf{v}_\ell$). We sweep candidate layers $\mathcal{L}$ spanning the cosine-divergence peak/band and the logit-lens peak, at $\alpha \in \{0, 0.1, 0.25, 0.5, 1.0, 1.5, 2.0\}$ on validation.

Results of layer and α sweeps on the validation set. The sweep identifies L14 for English and Chinese, L11 for German, and L12 for Dutch (full results in Appendix H). All selected layers fall near the cosine-divergence band rather than the logit-lens peak, confirming that intervention is most effective where SFT first reorients representations, not where associations are most linearly readable. For EN, ZH, and NL, α=0.5 is the validation-best cell; for DE, α=0.5 at L11 is within 0.015 of the α=1.0 peak. Figure 3 shows that $\alpha \geq 1.5$ collapses MASS@K to near zero in every language, indicating the model breaks before the vector stops helping. We adopt α=0.5 uniformly as a conservative but safe operating point.

Test-set results on word association generation. Table 4 reports the final test comparison. At α=0.5, activation steering improves over vanilla across the board in every language while leaving weights untouched. Recovery@10 ranges from 7% (DE) to 25% (EN): a single steering vector captures part of the association signal, but the gap to full SFT remains. The lower recovery for non-English languages reflects the limitation of a single-layer, single-direction intervention; more advanced approaches are left for future work.

General capability under steering. Table 5 shows that steering preserves substantially more general capability than full SFT. All four steered models retain 84%–95% of vanilla capability under capped retention, compared to 39–54% for the full SFT models (Table 3).[4] Notably, IFEval and HumanEval largely exceed vanilla under steering in all four languages, the same two benchmarks that collapse most severely under SFT.[5] A possible

[4] Per-benchmark profiles are in Appendix F.

[5] Applying a deterministic lenient parser identically to vanilla, SFT, and steered outputs (Appendix I) raises the SFT raw-score means by +5.0 to +6.1 points and lifts HumanEval to 24–34%, while leaving the steered rows essentially unchanged ($\Delta_{\text{mean}} \leq +3.5$); the qualitative SFT-vs-steered ordering is preserved, so the gap reported here is partly inflated by

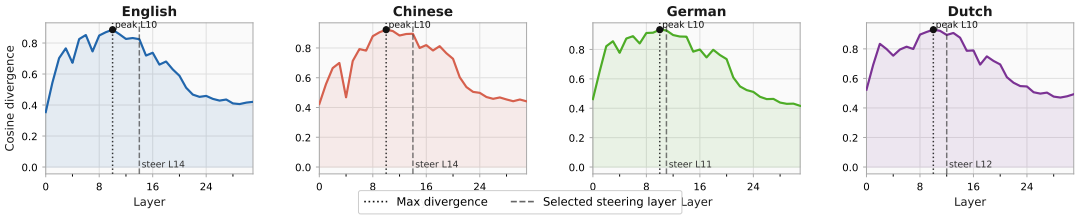

Figure 1: Cosine divergence ($1 - \mathrm{CCS}$) between vanilla and SFT activations across layers. Dotted lines mark divergence peaks; dashed lines mark selected steering layers, which lie near the high-divergence mid-layer band.

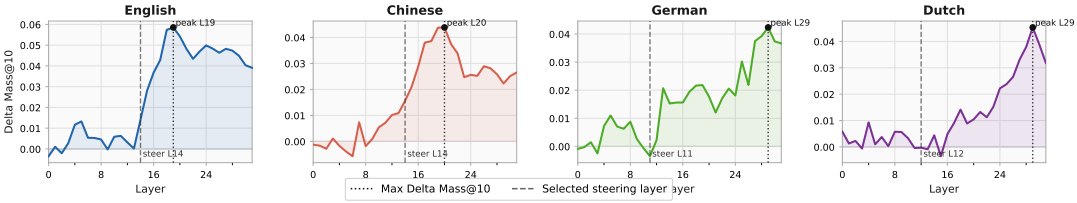

Figure 2: Frequency-aware logit-lens ΔMASS@10 (SFT $-$ vanilla) across layers. Dotted lines mark logit-lens peaks; dashed lines mark selected steering layers.

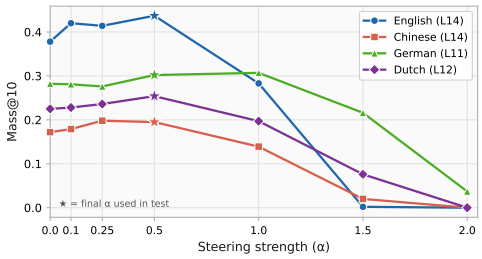

Figure 3: Validation MASS@10 as a function of steering strength α across candidate layers per language. We use the optimal $\alpha = 0.5$ uniformly for the final test.

Lang.	Model	L	α	Mass@10	Overlap@10	Spearman	Recov.
EN	Vanilla	—	—	0.361	0.267	0.414	—
	+ Steer	14	0.5	0.440	0.309	0.456	**25.1%**
	$\mathrm{SFT}_{\mathrm{freq}}$	—	—	0.676	0.515	0.583	100.0%
ZH	Vanilla	—	—	0.156	0.118	0.234	—
	+ Steer	14	0.5	0.187	0.138	0.289	13.4%
	$\mathrm{SFT}_{\mathrm{freq}}$	—	—	0.387	0.284	0.407	100.0%
DE	Vanilla	—	—	0.257	0.188	0.342	—
	+ Steer	11	0.5	0.277	0.197	0.345	7.5%
	$\mathrm{SFT}_{\mathrm{freq}}$	—	—	0.525	0.381	0.489	100.0%
NL	Vanilla	—	—	0.216	0.147	0.314	—
	+ Steer	12	0.5	0.243	0.161	0.327	8.7%
	$\mathrm{SFT}_{\mathrm{freq}}$	—	—	0.527	0.394	0.511	100.0%

Table 4: Final test-set comparison. Recovery@10 is the fraction of the SFT MASS@K gain over vanilla captured by steering at α=0.5.

explanation is that steering adds the new association capability while preserving the instruction-following register that full SFT suppresses; our SAE analysis (§7.2) supports this interpretation, though a full causal account is left for future work.

7 What Do SAEs Reveal About Fine-Tuning and Steering?

Sparse Autoencoders (SAEs) decompose a model's internal activations into a large set of interpretable *features* — directions in activation space that tend to fire on semantically related inputs (Templeton et al., 2024). We use LlamaScope SAEs (He et al., 2024) as a fixed measurement tool to ask: which features does SFT change, and does steering change the same ones?

We run all three model conditions (vanilla, SFT, α=0.5 steered) on each test cue at the language-specific steering layer, encode the residual-stream activations through the SAE (32,768 features each), and compute mean feature-shift vectors Δ^{SFT} and Δ^{ste} relative to vanilla. A positive shift means the feature became more active; a negative shift means it was suppressed. We then label the top-50 most amplified and top-50 most suppressed features by their top-12 activating cues, and group them into interpretable themes (via GPT-5.5).[6]

SFT's output-format collapse but is not explained by it.

[6]Full theme definitions, classification rules, and per-language cue evidence are in Appendix K.

Model	MMLU	MATH	GSM8K	IFEval	HumanEval	TruthQA	CapRet
Vanilla Llama-3.1-8B-Inst.	70.43	11.28	79.83	66.43	25.61	0.29	100.0
+ Steer EN (L14, α=0.5)	55.67	7.82	42.38	78.18	35.37	0.29	83.6
+ Steer ZH (L14, α=0.5)	64.37	18.60	67.93	80.70	46.34	0.28	95.4
+ Steer DE (L11, α=0.5)	38.69	15.38	55.04	76.14	45.73	0.28	86.8
+ Steer NL (L12, α=0.5)	55.93	17.10	64.75	80.10	45.12	0.26	91.9

Table 5: Steering results on general capability benchmarks. CapRet is the overall capped retention relative to vanilla, higher values are better. CapRet is 84–95% for steering, compared to 39–54% for full SFT (Table 3).

Lang.	Recov.@10	r (full)	r (215)	Overlap (full / 215)
EN	25.1%	0.855	0.790	119 / 83
ZH	13.4%	0.805	0.771	120 / 110
DE	7.5%	0.780	0.771	99 / 95
NL	8.7%	0.833	0.814	102 / 81

Table 6: Pearson correlation between SFT and steering feature-shift vectors (r), and overlap among top-200 positive features, on the full test set and on a 215-cue cross-lingual inventory. Recovery@10 is from Table 4.

Bin	EN	ZH	DE	NL
Positive (SFT amplifies)				
Embodied / concrete	12	10	11	14
Everyday / scene	7	7	8	9
Cultural memory	**8**	**17**	**11**	**12**
Evaluative / scalar	11	8	13	8
Mixed / fragmented	12	8	7	7
Negative (SFT suppresses)				
Assistant / instruction	19	13	5	4
Document / analysis	0	0	18	20
Abstract / method	9	9	0	0
Concrete / world attenuation	6	9	13	13
Affect / evaluative	11	8	6	7
Other / mixed	5	11	8	6

Table 7: Top-50 SFT-positive and SFT-negative SAE feature bins on the full test set. Embodied/everyday positives and broadcaster suppression appear in every language; broadcaster type and cultural-memory share vary by language.

7.1 Steering moves features in the same direction as SFT

The Pearson correlation between Δ^{SFT} and Δ^{ste} across all 32,768 features is 0.78–0.86 in every language (Table 6). Steering and SFT shift the feature basis in the same direction. The same pattern holds when we restrict to 215 basic-level concepts shared across all four languages (colors, body parts, animals, food), confirming it is not an artifact of each language's specific cue vocabulary. Figure 6 (Appendix J) visualizes this: most features lie near the identity line, showing that steering approximately reproduces the SFT direction.

7.2 What SFT amplifies and suppresses — and what steering recovers

Despite the high directional agreement, the *amount* of recovery differs sharply by feature type. To interpret which features are amplified or suppressed, we inspect the top-50 SFT-positive and top-50 SFT-negative features. We use ChatGPT-5.5 to (i) generate a label for each feature based on its top-12 activating cues, and (ii) group the labels into themes. We manually inspected these grouped themes and cues, and found them to be sensible. Table 7 reports the counts; three patterns emerge consistently across all four languages.

Broadcaster features (suppressed by SFT; recovered 72–91% by steering). These are features that fire on words like *instruction, provide, answer, output* — the language of a helpful assistant. Before SFT, the model is in "assistant mode" and these features are highly active; SFT on word associations silences them, moving the model into a free-association mode. We call them *broadcaster* features because they broadcast the "assistant mode" across the whole response. Steering recovers this suppression tightly (72–91%). Crucially, *which* broadcaster gets silenced differs by language and reflects each base model's pretraining bias, not speaker properties: EN suppresses chat scaffolding (19/50; *instruction, provide, output*); ZH suppresses an abstract reasoning register (9/50; 想法, 原理, 定理); DE suppresses a correction/exam register (18/50; *Korrektur, Klausur, Code*); NL suppresses a document-analysis register (20/50; *commentaar, samenvatting, analyse*). Same mechanism, different content.

Embodied / concrete features (amplified by SFT; recovered 35–48% by steering). These features fire on physical, sensory words — body parts, animals, food, household objects, nature. Together with everyday/scene features they occupy 19–23 of the top-50 SFT-positive slots in every language:

Lang.	Category	Feature	Top activating cues (English gloss)	$\Delta^{\text{SFT}} \rightarrow \Delta^{\text{ste}}$	Recov.
EN	Named places/culture	F24404	Robin Hood, Manhattan, Broadway, Methodist, New Orleans, Hawaii	$+4.04 \rightarrow +2.06$	51%
	Geography/culture	F20045	Swedish, Hebrew, Egypt, Cuba, Methodist, golf, arctic, volcano	$+2.96 \rightarrow +0.66$	22%
ZH	National/ethnic identity	F19960	中华, 党, 基层, 血缘, 汉族, 热血, 信念, 命, 仙 (China, party, grassroots, kinship, Han ethnicity, passion, belief, fate, immortal)	$+3.11 \rightarrow +1.69$	54%
	Slang/folk culture	F8331	杀马特, 油条, 麻子, 娘炮, 百家姓, 呻吟, 木板, 捣蛋 (subculture slang, fried dough, pockmarks, effeminate, Hundred Family Surnames, groan, plank, mischief)	$+3.79 \rightarrow +1.71$	45%
DE	Civic/intellectual canon	F17125	Goethe, Freiheit, Kunst, Rassismus, Islam, Wissenschaft, Selbstvertrauen (freedom, art, racism, science, self-confidence)	$+5.52 \rightarrow +2.21$	40%
	Historical figures/places	F16517	Putin, Hitler, Goethe, Einstein, Frankfurt, Vater, Bruder, Sohn (father, brother, son)	$+1.43 \rightarrow +0.00$	0%
NL	Flemish local culture	F9542	antwerpen, kerk, slasaus, hoer, trut, journaal, kaas, bier (Antwerp, church, sauce, slang terms, news, cheese, beer)	$+2.13 \rightarrow +0.79$	37%
	Belgian identity	F11213	vlaams, protestants, allochtoon, zuiders, jongere, wiskundig (Flemish, Protestant, immigrant, southern, youth, mathematical)	$+1.45 \rightarrow +0.00$	0%

Table 8: Representative cultural-memory SAE features per language. $\Delta^{\text{SFT}}, \Delta^{\text{ste}}$ are mean-activation differences relative to vanilla; "Recov." is $\Delta^{\text{ste}}/\Delta^{\text{SFT}}$. SFT amplifies community-specific features in every language; steering recovers 0–54%, with 0% on several DE and NL features.

free association is grounded in tangible everyday experience. Steering partially reproduces this (35–48%). Representative features per language are in Appendix L (Table 18).

Cultural memory features (amplified by SFT; recovered 0–54% by steering). These features are activated by community-specific content — named people, places, festivals, institutions, and slang. SFT amplifies them substantially; steering barely recovers them, with several features at 0% in DE and NL. ZH has the largest share (17/50; 中华, 党, 春节, 儒家, 微信); NL, DE, and EN follow with 12, 11, and 8 respectively. Table 8 gives representative examples with both deltas.

The behavioral Mass@K recovery (7–25%, §6) is consistent with these feature-level results. Taken together, the SAE analysis reveals that steering and full SFT are two operating points on the same feature axis. Both suppress the same assistant-mode broadcaster features and amplify the same embodied/everyday features. The key difference is magnitude and scope: steering moves each feature about half as far as SFT, and leaves cultural-memory features almost untouched. In practical terms, a steered model exits the assistant register and enters a free-association mode while retaining its general capabilities; a fully fine-tuned model does the same and additionally strengthens community-specific cultural knowledge, at the cost of more than halving its general capability. Which operating point is appropriate depends entirely on what the downstream task requires of the cultural-memory features — that is an empirical question about the task, not a claim we can make here. Whether multi-layer, per-cue-conditional, or higher-α steering could close the cultural-memory gap without the capability cost is left for future work.

8 Conclusion

We present the first systematic study tracing how a language model learns associative behavior using word associations across four languages, and show that this capability is captured mostly in the middle layers of Llama. While association steering only partially recovers the association generation gain, it largely preserves the model's general capability. Our sparse autoencoder results show that single-layer association steering recovers the embodied and concrete shifts reasonably well, but the deeper culture-specific features are captured far more fully by SFT, leaving future work to develop more advanced steering approaches.

Limitations

Single model family and scale. Our key focus is to explore how a language model can acquire associative learning and to design a framework for analyzing performance trade-offs and steering behaviors across different languages. Therefore, we focus on a single language model, Llama-3.1-8B-Instruct (Grattafiori et al., 2024). As a result, our findings are empirical and model-specific, including the best steering layers and the steering performance on general tasks. To what extent these findings apply to other architectures, such as Qwen (Qwen et al., 2025) and Mistral, or to other model scales remains open and requires further experimentation.

Single-direction, single-layer steering at fixed α. We test single mean-difference directions at individual layers with an optimal but fixed $\alpha = 0.5$. We show the capabilities and limitations of single-layer steering on both association tasks and general performance. We leave more complex steering approaches, such as multi-layer interventions, structured directions (e.g., low-rank directions), and percue conditional steering vectors, to future work, as they may close more of the gap to supervised fine-tuning.

SAE multilingual coverage. The LlamaScope L11/L12/L14 R-8x SAEs were trained on Llama-3.1-Base activations on SlimPajama (He et al., 2024), which is overwhelmingly English. As a result, the German, Dutch, and Chinese feature decompositions are produced by a feature dictionary biased toward English distributional patterns. We mitigate by (i) reporting per-language reconstruction quality and feature usage, (ii) framing the SAE as a fixed measurement device applied across languages rather than as a language-neutral concept dictionary, and (iii) being explicit that cross-lingual feature claims are conditional on this measurement asymmetry. In the future, a multilingually balanced SAE would strengthen the cross-cultural component.

SAE feature labels are interpretive. In the §7.2 SAE analysis we use commercial language models to generate the feature names and grouping to help us better understand the features. Feature labels in Tables are produced with LLM assistance from top activating cues and target outputs, then checked for coarse thematic consistency. Alternative labels are sometimes plausible. Recent work shows that SAE dictionaries may not be canonical and can vary across training choices (Leask et al., 2025; Paulo and Belrose, 2026); our claims concern relative feature-shift alignment between conditions, not the discovery of definitive concepts.

Cultural-memory features and contributor demographics. The cultural-memory content surfaced by SAE features reflects who contributed to each SWOW language version (e.g., the Dutch version leans Flemish; the English version is U.S.-coded). Cross-language comparisons should therefore be read as lexical and cultural salience in sampled communities, not as claims about national character or about all speakers of a language.

References

Samuel Aeschbach, Rui Mata, Kaidi Lõo, Simon De Deyne, and Dirk U Wulff. 2026. The small world of words german free-association norms. *arXiv preprint arXiv:2604.19620*.

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

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé, Jared Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 34 others. 2021. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

Allan M Collins and Elizabeth F Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407.

Xunlian Dai, Li Zhou, Benyou Wang, and Haizhou Li. 2025. From word to world: Evaluate and mitigate culture bias in LLMs via word association test. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24510–24526, Suzhou, China. Association for Computational Linguistics.

Simon De Deyne, Danielle Navarro, Guillem Collell, and Andrew Perfors. 2021. Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, 45.

Simon De Deyne and Gert Storms. 2008. Word associations: Norms for 1,424 dutch words in a continuous task. *Behavior research methods*, 40(1):198–205.

Simon De Deyne, Daniel J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "small world of words" english word association norms for over 12,000 cue words. *Behavior Research Methods*, 51:987 – 1006.

Yupei Du, Yuanbin Wu, and Man Lan. 2019. Exploring human gender stereotypes with word association test. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6133–6143.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

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.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. 2024. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *Preprint*, arXiv:2410.20526.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

Suhas Kotha, Jacob Springer, and Aditi Raghunathan. 2024. Understanding catastrophic forgetting in language models via implicit inference. In *International Conference on Learning Representations*, volume 2024, pages 24110–24139.

Vedang Lad, Wes Gurnee, and Max Tegmark. 2024. The remarkable robustness of LLMs: Stages of inference? In *ICML 2024 Workshop on Mechanistic Interpretability*.

Patrick Leask, Bart Bussmann, Michael T Pearce, Joseph Isaac Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. 2025. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations*.

Bing Li, Ziyi Ding, Simon De Deyne, and Qing Cai. 2024. A large-scale database of mandarin chinese word associations from the small world of words project. *Behavior Research Methods*, 57(1):34.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US. Association for Computational Linguistics.

Zheng Lim, Harry Stuart, Simon De Deyne, Terry Regier, Katerina Vylomova, Trevor Cohn, and Charles Kemp. 2024. A computational approach to identifying cultural keywords across languages. *Cognitive science*, 48:e13402.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Chunhua Liu, Trevor Cohn, Simon De Deyne, and Lea Frermann. 2022. WAX: A new dataset for word association eXplanations. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 106–120.

Chunhua Liu, Trevor Cohn, and Lea Frermann. 2021. Commonsense knowledge in word associations and ConceptNet. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 481–495, Online. Association for Computational Linguistics.

Chunhua Liu, Kabir Manandhar Shrestha, and Sukai Huang. 2026. Align: Word association learning for cultural alignment in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems 35*.

Douglas Nelson, Cathy Mcevoy, and Thomas Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, 36:402–7.

nostalgebraist. 2020. Interpreting GPT: The logit lens. LessWrong blog post.

Gonçalo Paulo and Nora Belrose. 2026. Sparse autoencoders trained on the same data learn different features. In *The Fourteenth International Conference on Learning Representations*.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 Technical Report. *arXiv preprint*. ArXiv:2412.15115 [cs].

Aida Ramezani and Yang Xu. 2024. Moral association graph: A cognitive model for moral inference. In *Annual Meeting of the Cognitive Science Society*.

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.

Mark R Rosenzweig. 1961. Comparisons among word-association responses in English, French, German, and Italian. *The American Journal of Psychology*, 74(3):347–360.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

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

Michael Wilson, George Kiss, and 1 others. 1988. Eat: The edinburgh associative corpus. *Oxford Text Archive Core Collection*.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *ArXiv*, abs/2311.07911.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Troy Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to ai transparency. *ArXiv*, abs/2310.01405.

A Frequency-Aware SFT: Configuration and Loss Derivation

Training setup. All four SFT models are fine-tuned from `meta-llama/Llama-3.1-8B-Instruct` with

Lang.	Steps	$\text{Train}_{\text{start}}$	$\text{Train}_{\text{end}}$	Eval_1	Eval_2	Wall (h)
EN	308	26.09	16.71	2.220	2.245	1.7
ZH	252	21.00	11.33	1.617	1.565	1.5
DE	148	17.85	11.80	1.672	1.665	0.9
NL	316	18.03	11.82	1.627	1.600	2.1

Table 9: SFT convergence. Train losses are frequency-weighted (comparable within-language only); Eval is ordinary NLL with weights stripped. Wall-clock on a single H100 node.

full-parameter updates using LLaMA-Factory with DeepSpeed ZeRO-3 on 2×H100. Hyperparameters: lr $1\mathrm{e}{-5}$, effective batch size 16 (per-device $8 \times$ grad-accum 2), max sequence length 2048, 2 epochs, AdamW (β_1=0.9, β_2=0.999), cosine LR with 3% warmup, bf16, random seed 42. Identical for all four languages and the standard-SFT ablation. Training YAMLs and the frequency-weighted-loss patch are released with the code.

Frequency-weighted loss derivation. Given a cue with associations $a_1, \ldots, a_n$ and observed response counts $f(a_i)$, the empirical association distribution is $p_{\text{gt}}(a_i) = f(a_i)/T$ with $T = \sum_j f(a_j)$. Each output token y_t is assigned a weight

$$w_t \;=\; 1 + \lambda\, p_{\text{gt}}(a(y_t))^{\gamma}, \quad \lambda = 5,\ \gamma = 0.5,$$

with separators and EOS at baseline weight 1. The loss is

$$\mathcal{L}_{\text{freq}} = -\frac{1}{Z} \sum_t w_t \log p_\theta(y_t \mid x, y_{<t}),$$

$$Z = \textstyle\sum_t w_t.$$

Normalising by Z keeps loss scale comparable across cues. Token weights are stripped at eval, so reported eval-loss is ordinary NLL and comparable across frequency-aware and standard-SFT settings.

B SFT Training Diagnostics

Table 9 reports training and eval losses at the end of each epoch. Training loss plateaus in every language and eval-loss is flat or decreasing, confirming clean convergence without overfitting. The small EN eval-loss rise (+0.025) is within run-to-run noise; the test-set MASS@K (Table 2) shows EN is the strongest performer and is not over-fit.

C Per-Language SFT Prompt Templates

All four languages share the same Alpaca-style conversational structure (system → user instruction + cue → assistant comma-separated list), with prompts translated and adapted per language. Each cue is presented in the user turn as `input`; the model is supervised to produce only a comma-separated list of associations (Chinese uses full-width comma ， , others ASCII comma). System and instruction text per language follows; gold outputs are real SWOW associations sorted by descending frequency.

English. System: "You are a sophisticated language model designed to explore word associations comprehensively."
Instruction: "Given a cue word, your task is to generate a comprehensive list of words associated with the cue word. Aim to cover as many relevant contexts, uses, and meanings as possible without repeating similar concepts. These words together provide a broad and insightful representation of all significant associations. Focus on revealing both common and unique aspects related to the cue word to ensure a balanced and thorough exploration of potential associations. Words should be distinct from each other. Your response shall only be the list of associated words. Do not generate words conditioned on the presence of other words but rather focus on the cue word itself."
Example. `input`: *elementary school*; `output`: *children, teacher, kids, reading, playground, education, grade school, learning, …*

Mandarin Chinese. System: "您是一款专为全面探索词语关联而设计的高级语言模型。"
Instruction: "给定一个提示词，你的任务是生成一个与该提示词相关联的全面词汇列表。目标是尽可能涵盖所有相关的语境、用法和含义，避免重复相似的概念。这些词共同提供对所有重要关联的广泛而深刻的表示。专注于揭示与提示词相关的常见和独特的方面，以确保对潜在关联进行平衡而彻底的探索。词语应彼此不同。你的回答只能是相关联的词语列表。不要生成受其他词语存在影响的词语，而是专注于提示词本身。"
Example. `input`: 谣言; `output`: 智者， 传播， 四起， 辟谣， 止于智者， 不攻自破， 真相， …

German. System: "Du bist ein anspruchsvolles Sprachmodell, das dafür entwickelt wurde, Wortassoziationen umfassend zu erkunden."

Instruction: "Gegeben ein Reizwort, ist es deine Aufgabe, eine umfassende Liste von Wörtern zu generieren, die mit dem Reizwort assoziiert sind. Versuche, so viele relevante Kontexte, Verwendungsweisen und Bedeutungen wie möglich abzudecken, ohne ähnliche Konzepte zu wiederholen. Diese Wörter bieten zusammen eine breite und aufschlussreiche Darstellung aller wichtigen Assoziationen. Konzentriere dich darauf, sowohl häufige als auch einzigartige Aspekte im Zusammenhang mit dem Reizwort offenzulegen, um eine ausgewogene und gründliche Erkundung möglicher Assoziationen zu gewährleisten. Die Wörter sollten sich voneinander unterscheiden. Deine Antwort soll ausschließlich aus der Liste der assoziierten Wörter bestehen. Erzeuge keine Wörter, die durch das Vorhandensein anderer Wörter bedingt sind, sondern konzentriere dich auf das Reizwort selbst."
Example. `input:` *Amt*; `output:` *Behörde, Beamter, Verwaltung, Büro, Arbeitsamt, Bürokratie, Beamte, . . .*

Dutch. System: "Je bent een geavanceerd taalmodel dat is ontworpen om woordassociaties op een uitgebreide manier te verkennen."
Instruction: "Gegeven een sleutelwoord, is het jouw taak om een uitgebreide lijst van woorden te genereren die geassocieerd zijn met het sleutelwoord. Streef ernaar om zoveel mogelijk relevante contexten, gebruikswijzen en betekenissen te dekken zonder vergelijkbare concepten te herhalen. Deze woorden geven samen een breed en inzichtelijk beeld van alle belangrijke associaties. Richt je op het onthullen van zowel algemene als unieke aspecten die verband houden met het sleutelwoord om zo een evenwichtige en grondige verkenning van potentiële associaties te waarborgen. Woorden moeten van elkaar onderscheiden zijn. Je reactie mag alleen de lijst met geassocieerde woorden bevatten. Genereer geen woorden die afhankelijk zijn van de aanwezigheid van andere woorden, maar richt je op het sleutelwoord zelf."
Example. `input:` *sukkelaar*; `output:` *arm, medelijden, arme, dom, zielig, ocharme, bedelaar, duts, loser, . . .*

D Four-Language Overlap Cue Set

For cross-lingual SAE comparison we use a 215-cue subset shared across all four languages. We first identified basic-level concepts (colors, kinship, body parts, animals, food, household objects, basic verbs and adjectives) whose translations appeared likely to exist as cues in every language's test set. GPT-5.5 was used as an assistive tool to propose candidate translation mappings across English, Mandarin Chinese, German, and Dutch. These aligned cues are used only for cross-lingual SAE analysis (§7.1, "Aligned 215 cues" columns), not for training. The full list and per-language translation mappings are released with the code.

E Output Parser Details

A deterministic parser is applied uniformly across vanilla, SFT, and steered outputs in five stages: (1) strip whitespace and chat-template scaffolding; (2) split on language-appropriate separators (ASCII comma for EN/DE/NL, full-width comma ， for ZH); (3) remove numbered list prefixes (`1.`, `1)`, `(1)`, etc.); (4) remove cue echoes if the cue appears as the first item; (5) detect copy loops (any 3-gram repeated ≥ 4 times) and truncate at the first repeated 3-gram.

F Per-Benchmark General Capability

Figure 4 visualizes the per-benchmark profiles from Tables 3 and 5. The frequency-SFT polygons collapse inward most strongly on the format-sensitive axes (MATH, GSM8K, HumanEval), while the steered polygons closely retrace the vanilla footprint and extend past vanilla on IFEval and HumanEval in every language. The radar uses the strict OpenCompass scores; lenient-rescore equivalents (Appendix I) lift the SFT polygons partially toward steered without changing the qualitative ordering.

G Standard vs. Frequency-Aware SFT Ablation

Table 10 reports the standard-SFT ablation rows that motivate the frequency-aware loss. Frequency weighting consistently matches or slightly improves on standard SFT across all metrics in all four languages; effect sizes are small but uniform.

H Full Layer $\times$ α Validation Sweep

We sweep candidate layers $\mathcal{L}$ (per-language; including $L_{\text{logit-peak}}$) at $\alpha \in \{0, 0.1, 0.25, 0.5, 1.0, 1.5, 2.0\}$ on validation. Figure 5 shows the full MASS@K surface as a heatmap; Tables 11–14 report the per-language performance on the validation set. For EN, ZH, and NL, the chosen α=0.5 at the language-specific layer is the validation-best cell. For DE, the

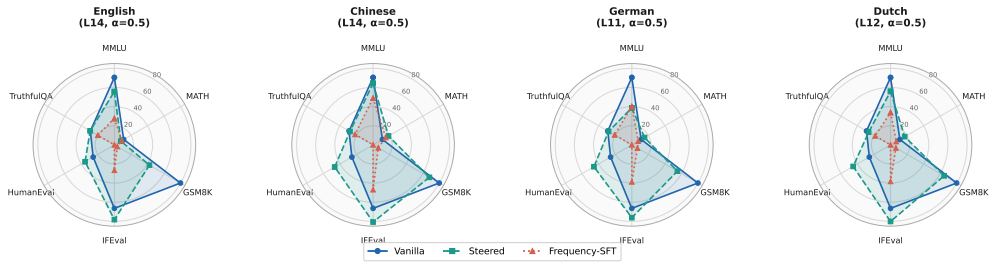

Figure 4: Per-benchmark general-capability comparison across vanilla, frequency-SFT, and α=0.5 steering for each language. SFT collapses inward on multiple axes; steering largely preserves the vanilla footprint.

Lang.	System	n	Mass@5	Mass@10	Mass@20	Overlap@10	Spearman
EN	Standard SFT	1228	0.677	0.670	0.644	0.508	0.581
	Frequency SFT	1228	**0.680**	**0.676**	**0.649**	**0.515**	**0.583**
ZH	Standard SFT	1002	0.386	0.373	0.345	0.272	**0.416**
	Frequency SFT	1002	**0.399**	**0.387**	**0.363**	**0.284**	0.407
DE	Standard SFT	588	0.525	0.518	0.505	0.372	0.487
	Frequency SFT	588	**0.532**	**0.525**	**0.505**	**0.381**	**0.489**
NL	Standard SFT	1257	0.508	0.513	0.507	0.384	0.510
	Frequency SFT	1257	**0.521**	**0.527**	**0.519**	**0.394**	**0.511**

Table 10: Standard SFT vs. frequency-aware SFT on the held-out test set (parser-applied scoring). Frequency weighting consistently matches or slightly outperforms standard SFT on MASS@K, Overlap@10, and Spearman in all four languages (one exception: ZH Spearman is fractionally higher for standard SFT, 0.416 vs 0.407).

validation-best cell is α=1.0 at L11 (0.302); we adopt α=0.5 at L11 (0.287, within 0.015 of the peak) for consistency with the uniform conservative setting across languages (capability-cost rationale in Appendix H.1). Bold marks the chosen (α, layer) cell.

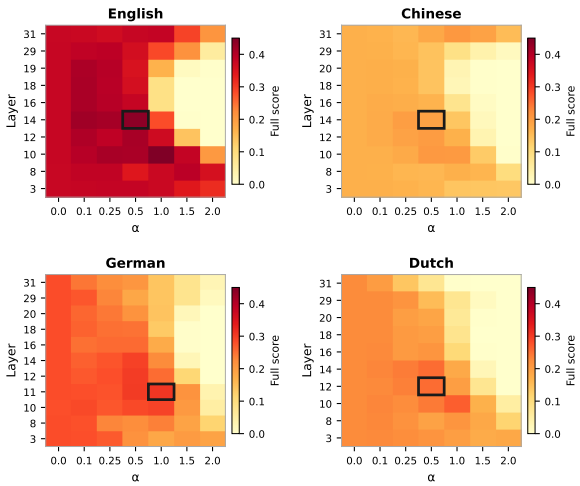

Figure 5: Validation MASS@K across layer $\times$ α per language.

Layer	α=0	0.1	0.25	0.5	1	1.5	2
L3	0.366	0.368	0.374	0.373	0.367	0.342	0.303
L8	0.366	0.370	0.375	0.348	0.379	0.364	0.332
L10	0.366	0.382	0.398	0.406	0.412	0.389	0.129
L12	0.366	0.392	0.403	0.437	0.401	0.093	0.000
L14	0.366	0.406	0.426	**0.440**	0.250	0.001	0.000
L16	0.366	0.397	0.408	0.397	0.064	0.000	0.000
L18	0.366	0.394	0.407	0.389	0.055	0.000	0.000
L19	0.366	0.395	0.406	0.383	0.108	0.008	0.001
L29	0.366	0.376	0.385	0.360	0.306	0.215	0.046
L31	0.366	0.364	0.360	0.368	0.354	0.265	0.167

Table 11: Validation Mass@10 sweep for EN. Bold marks the chosen (α, layer) cell.

H.1 General Capability at Higher Steering Strengths

For EN, ZH, and NL the chosen α=0.5 at the language-specific layer is already the validation-best cell (Tables 11, 12, 14), so the operating-point choice is uncontroversial. For DE, the validation-best cell is L11, α=1.0 at validation MASS@K = 0.302, 0.015 above the chosen L11, α=0.5 at 0.287 (Table 13). We adopt the lower-α cell because the small MASS@K gain at α=1.0 comes at a catastrophic capability cost: under the same six-benchmark protocol as Table 5, DE-Steer L11, α=1.0 collapses to **21.0**% CapRet (MMLU 0.00, MATH 0.00, GSM8K 1.44, IFEval 43.65, Hu-

Layer	α=0	0.1	0.25	0.5	1	1.5	2
L3	0.162	0.161	0.158	0.154	0.138	0.132	0.124
L8	0.162	0.161	0.161	0.157	0.159	0.138	0.100
L10	0.162	0.160	0.173	0.194	0.196	0.090	0.001
L12	0.162	0.169	0.188	0.191	0.168	0.021	0.000
L14	0.162	0.173	0.196	**0.200**	0.117	0.009	0.000
L16	0.162	0.168	0.176	0.174	0.067	0.004	0.000
L18	0.162	0.167	0.169	0.150	0.020	0.001	0.000
L20	0.162	0.167	0.168	0.152	0.024	0.001	0.000
L29	0.162	0.161	0.156	0.151	0.091	0.037	0.011
L31	0.162	0.161	0.155	0.161	0.163	0.132	0.085

Table 12: Validation Mass@10 sweep for ZH. Bold marks the chosen (α, layer) cell.

Layer	α=0	0.1	0.25	0.5	1	1.5	2
L3	0.267	0.267	0.262	0.213	0.173	0.157	0.109
L8	0.267	0.266	0.218	0.213	0.226	0.192	0.138
L10	0.269	0.274	0.261	0.270	0.283	0.239	0.041
L11	0.267	0.270	0.271	**0.287**	0.302	0.233	0.041
L12	0.267	0.258	0.272	0.281	0.252	0.083	0.000
L14	0.267	0.254	0.268	0.284	0.197	0.015	0.000
L16	0.267	0.242	0.254	0.243	0.135	0.002	0.000
L18	0.267	0.253	0.247	0.242	0.089	0.002	0.000
L20	0.267	0.249	0.241	0.214	0.094	0.012	0.000
L29	0.267	0.255	0.222	0.185	0.143	0.065	0.012
L31	0.267	0.232	0.214	0.208	0.141	0.075	0.021

Table 13: Validation Mass@10 sweep for DE. Bold marks the chosen (α, layer) cell. Validation-best cell is L11, α=1 (0.302); the chosen L11, α=0.5 (0.287) is within 0.015 and consistent with the uniform-α setting across languages.

manEval 1.22, TruthfulQA 0.16), below every full-SFT model (Table 3). DE-Steer L11, α=0.5 in contrast reaches 86.8% CapRet, as reported in Table 5. The DE comparison is therefore a concrete illustration of the broader pattern: across all four languages, $\alpha \geq 1.5$ at the chosen layer also collapses validation MASS@K itself to near zero (Tables 11–14, α=1.5 and α=2 columns), indicating the steering vector is breaking the model before it stops helping. A uniform α=0.5 also keeps the cross-lingual feature-shift comparison (§7) clean and removes per-language hyperparameter tuning as a confound. Taken together, α=0.5 sits inside the safe region for every language and preserves $\geq$ 84% CapRet (Table 5).

I General-Capability Output Parsing: OpenCompass Defaults and a Lenient Rescore

The general-capability numbers in Tables 3 and 5 use OpenCompass's default per-benchmark output extractors. For some benchmarks — in particular HumanEval — those extractors require a specific

Layer	α=0	0.1	0.25	0.5	1	1.5	2
L3	0.210	0.211	0.208	0.202	0.188	0.162	0.137
L8	0.209	0.215	0.207	0.181	0.188	0.156	0.120
L10	0.209	0.213	0.211	0.222	0.236	0.181	0.035
L12	0.209	0.213	0.228	**0.238**	0.187	0.055	0.000
L14	0.209	0.217	0.228	0.231	0.154	0.020	0.000
L16	0.209	0.211	0.200	0.203	0.105	0.002	0.000
L18	0.209	0.210	0.200	0.191	0.074	0.001	0.000
L20	0.209	0.212	0.190	0.180	0.064	0.001	0.000
L29	0.209	0.212	0.205	0.155	0.075	0.011	0.000
L31	0.209	0.192	0.135	0.071	0.002	0.000	0.000

Table 14: Validation Mass@10 sweep for NL. Bold marks the chosen (α, layer) cell.

output template (markdown code fences) that the frequency-aware SFT models often fail to produce, even when they generate an on-topic and otherwise valid response. To separate this output-format collapse from genuine capability loss, we re-score every prediction with a deterministic lenient parser applied identically to vanilla, SFT, and steered outputs. The lenient parser is guaranteed by construction to be at least as forgiving as the strict default: it never overrides an answer the strict extractor already accepted (per-item OR semantics), so lenient $\geq$ strict in every cell.

OpenCompass default extractors (the "strict" parsers). Taken directly from the OpenCompass v0.5.1 source code:

- **MMLU:** regex `(?i)ANSWER\s*:\s*([A-D])` applied to the raw response; if no match, the example is marked wrong.
- **GSM8K:** split on the next `Question:` marker, return the last number in the remaining text.
- **HumanEval:** extract the contents of the first "‘. . . "‘ markdown block; if no block is present, pass the raw text to the test harness.
- **MATH:** extract the contents of `\boxed{...}` and verify with `math_verify`/sympy.
- **IFEval:** run each per-prompt instruction check (e.g. "respond in JSON", "include no commas") on the raw response.
- **TruthfulQA:** we use OpenCompass's standard BLEU-based accuracy. For each model answer, the scorer compares its best BLEU match among the correct reference answers with its best BLEU match among the incorrect reference answers; the answer is marked correct when the former is higher.

Lenient parser The lenient parser differs from the OpenCompass defaults in five points:

1. **Conservative pre-cleaner.** Before any benchmark-specific extraction, every prediction is passed through a runaway-loop stripper that fires only when an n-gram of length 2–8 tokens repeats ≥ 5 times consecutively *and* the looped suffix covers $\geq 30\%$ of the prediction length. This catches collapses like "... Angelo and Melanie, Angelo and Melanie, Angelo and Melanie, ..." while leaving benign reasoning repetition untouched.

2. **MMLU fallbacks.** If the strict regex `ANSWER: [A-D]` does not match, the lenient parser tries (in order) `ANS: [A-D]`, "`the (correct) answer is X`", and "`option X`". It does not fall back to "any standalone A–D in text", which would introduce false positives from the reasoning chain.

3. **GSM8K answer-locality preference.** Loops are stripped first; the parser then prefers numbers attached to `the answer is`/`####`/`=` phrases, falling back to the last number only when no such anchor exists. This recovers cases where the correct numeric answer is followed by a non-numeric repetition tail.

4. **HumanEval template-tolerant code extraction.** When no markdown fence is present, the parser detects whether the model (i) re-emits the function signature un-indented (kept as-is, plus indented body), or (ii) continues the prompt's function body (kept until the first un-indented statement). Trailing pollution (`# test cases`, top-level `print(...)`, `if __name__`) is removed before the test harness runs the code. This eliminates the most common failure mode for SFT outputs, where a valid function body is followed by truncated test-case calls that break the entire module's syntax.

5. **Per-item OR semantics.** For every example, the strict extractor is consulted first; only if strict returns "wrong" does the lenient fallback run. This makes the lenient score a lower bound on the model's true capability under any post-hoc cleaner, and guarantees that the lenient column is never lower than the strict column.

MATH, IFEval, and TruthfulQA reuse the OpenCompass evaluators after the universal pre-cleaner; we found their default extraction was already robust enough that no additional lenient logic was needed.

Why this matters. Sampling the SFT prediction JSONs across MMLU/GSM8K/HumanEval, we find that comma-separated SWOW-style output (which the frequency-aware SFT loss strongly conditions on) is the dominant surface form on 79–90% of EN, DE, and NL items, with the remainder dominated by runaway repetition; ZH is the inverse, with $\sim 81\%$ runaway repetition on MMLU due to full-width punctuation patterns. The strict extractors are constructed around English-language assistant templates (`ANSWER:`, markdown fences) that the SFT model has largely unlearned. The lenient parser does not change the SFT model's behavior — it changes only what the scorer recognizes as an attempt at an answer.

Results. Table 15 reports the lenient rescore; Table 16 reports the per-cell gain over Tables 3 and 5. Three patterns emerge: (i) vanilla gains modestly ($+2.74$ mean), almost entirely from HumanEval ($+14.0$), confirming that even well-behaved models occasionally drop markdown fences; (ii) SFT rows gain uniformly across all four languages ($+5.0$ to $+6.1$ mean), with HumanEval contributing the largest single delta ($+23.8$ to $+34.2$); (iii) steered rows are essentially flat ($+0.42$ to $+3.49$ mean), with the largest gain (DE, $+3.49$) driven by MMLU letter-format fallbacks. The qualitative ordering of the three regimes is preserved: even under lenient scoring, vanilla and steered remain well above SFT on CapRet.

J SAE Feature-Delta Scatter: Per-Feature Alignment of SFT and Steering

Figure 6 visualizes the per-feature alignment between SFT and steering that underlies Table 6 and the recovery asymmetry in §7.2. Blue points (top-50 SFT-negative broadcaster features) cluster tightly along the identity line in every language — steering reproduces SFT's suppression at roughly full magnitude. Red points (top-50 SFT-positive pool and cultural-memory features) lie systematically below the diagonal — steering moves them in the same direction as SFT but at smaller magnitude. The visual gap between the diagonal cluster and the sub-diagonal scatter is the geometric basis

Model	MMLU	MATH	GSM8K	IFEval	HumanEval	TruthQA	CapRet
Vanilla Llama-3.1-8B-Inst.	70.81	11.28	81.80	66.43	39.63	0.44	100.0
EN SFT_{freq}	30.26	9.40	6.97	26.86	23.78	0.39	53.9
ZH SFT_{freq}	49.08	16.24	7.66	46.52	34.15	0.40	71.0
DE SFT_{freq}	40.87	8.84	10.08	38.37	30.49	0.40	62.3
NL SFT_{freq}	34.34	7.34	8.79	38.25	28.66	0.39	57.1
+ Steer EN (L14, α=0.5)	57.83	7.82	42.76	78.18	35.37	0.42	81.3
+ Steer ZH (L14, α=0.5)	65.21	18.60	69.83	80.94	46.34	0.46	96.2
+ Steer DE (L11, α=0.5)	54.23	15.38	59.74	76.02	46.34	0.47	91.6
+ Steer NL (L12, α=0.5)	60.57	17.10	66.79	80.10	45.12	0.43	94.2

Table 15: Lenient rescore of all general-capability benchmarks. Same prediction files as Tables 3 and 5; only the post-hoc extractor differs. CapRet is computed relative to the lenient vanilla row.

Model	MMLU	MATH	GSM8K	IFEval	HumanEval	TruthQA	Mean Δ
Vanilla Llama-3.1-8B-Inst.	+0.38	+0.00	+1.97	+0.00	**+14.02**	+0.00	+2.74
EN SFT_{freq}	+2.78	+0.28	+3.41	+0.00	**+23.78**	+0.00	+5.05
ZH SFT_{freq}	+0.24	+0.92	+1.44	+0.00	**+34.15**	+0.00	+6.13
DE SFT_{freq}	+0.77	+1.16	+3.26	+0.00	**+30.49**	+0.00	+5.95
NL SFT_{freq}	+0.54	+0.40	+2.27	+0.00	**+28.66**	+0.00	+5.31
+ Steer EN (L14, α=0.5)	+2.16	+0.00	+0.38	+0.00	+0.00	+0.00	+0.42
+ Steer ZH (L14, α=0.5)	+0.84	+0.00	+1.90	+0.12	+0.00	+0.00	+0.48
+ Steer DE (L11, α=0.5)	**+15.54**	+0.00	+4.70	+0.12	+0.61	+0.00	+3.49
+ Steer NL (L12, α=0.5)	+4.64	+0.00	+2.04	+0.12	+0.00	+0.00	+1.13

Table 16: Per-cell gain from lenient parsing (Table 15 minus Tables 3/5). HumanEval dominates the SFT gains: the model writes valid code on 24–34% of items but never inside the markdown fences the strict scorer requires.

for the 72–91% broadcaster recovery vs. 0–54% cultural-memory recovery reported in §7.2.

K SAE Feature Categorisation: Pipeline, Bin Definitions, and Cross-lingual Evidence

Annotation pipeline. For each language we selected the top-50 SFT-positive and top-50 SFT-negative features by Δ^{SFT} at the language-specific steering layer. GPT-5.5 produced an initial free-text label per feature from its top activating cues, top output words, decoder neighbours, and English translations where needed. The resulting labels for coarse consistency were merged into broader bins; we did not conduct independent native-speaker annotation for every language. Classification into the bins below was deterministic given the top-12 activating cues; ambiguous cases were assigned to the dominant signal, or "Other / mixed" when neither signal dominated. Per-feature classifications are released with the supplementary CSVs.

Positive bin definitions and exemplary cues.

Embodied / concrete. Cues are body parts, animals, plants, food, household objects, weather, or verbs of perception and motion. EN: raincoat, lighthouse, scarecrow, yoghurt, dairy. ZH: 春风, 柠檬, 面包, 兔子, 鲤鱼, 狮子, 大象, 鱼 *(spring breeze, lemon, bread, rabbit, carp, lion, elephant, fish)*. DE: Kopf, Herz, Blume, Kirche, Tasche, Tisch, Vogel, Zug *(head, heart, flower, church, bag, table, bird, train)*. NL: zool, handpalm, greppel, kluif, haarspeld, goot, inham, wand *(sole, palm, ditch, bone, hairpin, gutter, inlet, wall)*.

Everyday / scene. Cues are life, work, time, money, family, or social-role words without a dominant culturally specific signal. EN: book, buy, work, eat, sleep, sell, house, bed, children. ZH: 吃, 店, 床, 衣服, 打电话, 买, 工作, 写, 题目, 看 *(eat, shop, bed, clothes, call, buy, work, write, topic, look)*. DE: Euro, essen, Huhn, kaufen, Kartoffel, trinken, Autofahrer *(eat, chicken, buy, potato, drink, driver)*. NL: tijd, stoel, camera, krediet, wasmachine, vrouw, deur, koffie *(time, chair, credit, washing machine, woman, door, coffee)*.

Cultural memory. Cues include culturally specific named entities, festivals, institutions, slang, or person/place names. This bin captures community-specific long-tail content that steering cannot recover. EN: Robin Hood, Manhattan, Broadway,

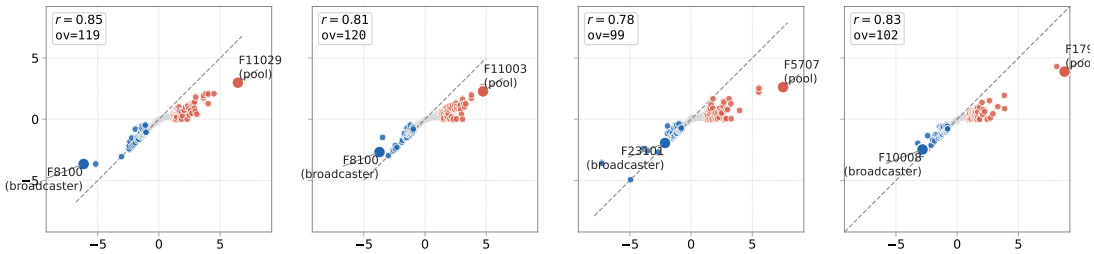

Figure 6: SAE feature-delta scatter on the full test set (one panel per language, at the language-specific steering layer). Each point is one of the 32,768 SAE features; the x-axis is Δ^{SFT} and the y-axis is Δ^{ste}. Red/blue points mark the top-50 SFT-positive/negative features respectively. Features near the identity line are recovered by steering at the same magnitude as SFT; features below the diagonal are under-recovered. The tight clustering of blue (broadcaster) points on the diagonal and the spread of red (pool and cultural-memory) points below it gives the visual basis for the recovery asymmetry reported in §7.2.

Methodist, New Orleans, Civil War, Hawaii, Cuba. ZH: 中华, 党, 基层, 血缘, 命, 仙, 热血, 汉族, 信念, 春节, 儒家, 道教, 微信, 杀马特 *(China, Party, grassroots, blood kinship, fate, immortal, hot-blooded, Han ethnicity, belief, Spring Festival, Confucianism, Taoism, WeChat, a Chinese youth subculture)*. DE: Goethe, Einstein, Putin, Hitler, Steuererklärung, Baden-Württemberg, Herr der Ringe *(tax return, Baden-Württemberg [German state], Lord of the Rings)*. NL: vlaams, antwerpen, unief, protestants, halfzus, kleermakerszit *(Flemish, Antwerp, university [student slang], Protestant, half-sister, cross-legged sit [Flemish])*.

Evaluative / scalar. Cues are adjectives, scalar terms, or affect words. EN: warranty, election, impossible, influential, cowardly, forbidden. ZH: 过失, 变换, 哀伤, 协作, 差别, 怪异, 意愿, 不正, 责备 *(fault, transform, sorrow, cooperation, difference, strange, willingness, wrongful, blame)*. DE: ziemlich, kaum, einverstanden, kümmern, mittelmäßig, großzügig *(quite, barely, agreed, care for, mediocre, generous)*. NL: levendig, liefelijk, onprettig, droef, hevig, vaag, slank, teder *(lively, lovely, unpleasant, sad, intense, vague, slim, tender)*.

Mixed / fragmented. Cue list does not concentrate in any of the above (function words, alphabet letters, mixed registers). Assigned only when no single signal dominates.

Negative bin definitions and exemplary cues.

Assistant / instruction. Top vanilla cues are task-framing instruction words suppressed by association-mode SFT. These are the primary broadcaster features recovered tightly by steering. EN: instruction, provide, shall, explore, investigation, via, minimize, language. ZH: 方法, 答案, 返回, 优化, 写, 开始, 题目, 关闭, 好 *(method, answer, return, optimize, write, start, topic, close, good)*. DE: Antwort, Aufzählung, unfreiwillig, namens, geistig, alt, anerkannt *(answer, enumeration, involuntary, named, intellectual, old, recognized)*. NL: antwoord, commentaar, samenvatting, proloog, vertalen, omgekeerd *(answer, commentary, summary, prologue, translate, reversed)*.

Document / analysis. Top vanilla cues are document-editing or structured-analysis register words. Concentrated in DE and NL; absent in EN and ZH. DE: Korrektur, Code, Klausur, Bearbeitung, Aufzählung, Ausführung, Übersicht *(correction, code, exam, editing, enumeration, execution, overview)*. NL: commentaar, samenvatting, structuur, analyse, vraag, voorwaarde, syntactisch, fouten *(commentary, summary, structure, analysis, question, condition, syntactic, errors)*.

Abstract / method. Top vanilla cues are abstract-reasoning or methodology words. Concentrated in ZH; absent in EN, DE, NL. ZH: 想法, 原理, 定理, 模式, 频率, 策略, 跳跃, 强调 *(idea, principle, theorem, pattern, frequency, strategy, jump, emphasize)*.

Concrete / world attenuation. Vanilla-active features for concrete nouns (animals, food, transport, geography) that are suppressed by SFT. EN: New Orleans, dice, camera, golf, butterfly, Cuba, badminton, elephant. ZH: 天气, 家具, 战争, 关爱, 甲虫, 快餐, 大象, 太阳, 海洋 *(weather, furniture, war, care, beetle, fast food, elephant, sun, ocean)*. DE: Fahrrad, Brot, Limonade, Kaffee, Trauma, Bier, Tasse, Zitrone, Kuh *(bicycle, bread, lemon-*

ade, coffee, trauma, beer, cup, lemon, cow). NL: vlaams, alsof, theater, dertien, water, bloed, vuurtoren, kaas *(Flemish, as if, theatre, thirteen, water, blood, lighthouse, cheese)*.

Affect / evaluative attenuation. Vanilla-active features for adjectives and emotional states suppressed by SFT. Weaker and less consistent than broadcaster suppression. EN: get rid of, shorter, shifty, impulsive, aloof, tactless, insincere. ZH: 原理, 客套, 两弹一星, 定理, 花卉, 轻视, 墨守成规, 骂 *(principle, politeness, Two Bombs One Satellite [nuclear program idiom], theorem, flowers, contempt, follow rules rigidly, scold)*. DE: lachen, geizig, luftig, etablieren, reparieren, erfrischend, hässlich *(laugh, stingy, airy, establish, repair, refreshing, ugly)*. NL: duidelijk, dus, vanbinnen, wel, overweldigend, verdrietig, taalvaardig *(clear, so/thus, within, well, overwhelming, sad, language-proficient)*.

Other / mixed. Cue list shows no concentrated signal, or splits evenly across two bins.

Bin	EN	ZH	DE	NL
Positive (SFT amplifies)				
Embodied / concrete	27	22	24	21
Everyday / scene	12	15	15	22
Cultural memory	0	0	0	0
Evaluative / scalar	7	9	8	4
Mixed / fragmented	4	4	3	3
Negative (SFT suppresses)				
Assistant / instruction	23	29	22	27
Document / analysis	0	0	0	0
Abstract / method	0	0	0	0
Concrete / world attenuation	13	12	23	17
Affect / evaluative	14	9	5	6
Other / mixed	0	0	0	0

Table 17: Bin counts on the four-language overlap setting (215 shared basic-level cues). The cultural-memory bin is zero by construction; the document/analysis and abstract-method broadcasters fold into the assistant/instruction column on this restricted inventory because the dominant suppressed cues on shared concepts are register words (*answer, language, question, write*) rather than community-specific editor or method registers. The shared embodied-positive and assistant-negative pattern reproduces in every language.

Overlap consistency check. Table 17 applies the same bin scheme to the 215-cue overlap inventory. Embodied + everyday positives total 34–43/50 in every language; broadcaster suppression dominates the negative space (22–29/50). The document/analysis (DE/NL) and abstract/method (ZH) broadcasters fold into assistant/instruction on this restricted inventory because their distinctive cues are absent from the shared vocabulary; the cultural-memory bin is zero by construction. The shared embodied-positive and register-suppression pattern survives a culture-neutral cue inventory.

L Representative Pool and Broadcaster Features

Table 18 lists the representative rank-1 pool feature (SFT-positive) and rank-1 broadcaster feature (SFT-negative) per language at the language-specific steering layer. Both classes appear consistently across all four languages and are jointly recovered by $\alpha = 0.5$ steering (Figure 6, Appendix J).

M Suggested Downstream Applications

A SWOW-steered model gives one additional controllable axis (pool/associate vs. broadcast/explain) without sacrificing general task ability. We expect it to perform well on (i) divergent and associative generation (brainstorming, copy, poetry, word puzzles), (ii) cognitively grounded NLP (vocabulary learning tools, distractor generation, lexical-screening prompts), (iii) cross-cultural and locale-aware persona generation, (iv) query expansion and synthetic concept-graph data in languages with poor lexical-resource coverage, and (v) lexical-network interpretability research that benefits from a controllable association-mode axis. For tasks that require the long-tail cultural-memory channel as the deliverable, full SFT remains the right tool (§7.2); the rest of this paper argues most downstream uses do not.

N Per-Bin Representative Features (Full Test)

Tables 19 and 20 show the rank-1 SFT-positive and SFT-negative SAE feature per bin per language on the full test set, with both Δ^{SFT} and Δ^{ste} so per-feature alignment vs. recovery gap is directly visible. Bins follow Appendix K; "—" indicates the bin had no feature in the top-50 for that language. The complete top-50 feature tables per language per setting (full-test and overlap) with both deltas, full evidence columns, and steered-direction counterparts are released as supplementary material.

O Compute and Reproducibility

Compute. SFT: $\sim 4 \times 16 = 64$ H100-hours across the four languages (per-language wall-clock

Lang.	Role	Δ^{SFT}	Δ^{ste}	Feature	Top activating cues (*italic* = EN translation)
EN	Free-association pool	$+6.45$	$+2.98$	F11029	cuddle, luxurious, silverware, swimsuit, celebration, sandals, moisturizer, heroism
EN	Assistant broadcaster	-6.15	-3.66	F8100	instruction, provide, shall, explore, investigation, via, minimize
ZH	Free-association pool	$+4.74$	$+2.27$	F11003	媒体, 店, 艺术, 新, 液, 冰, 食, 生产, 路, 热 *(media, shop, art, new, liquid, ice, food, production, road, heat)*
ZH	Abstract-method broadcaster	-3.72	-2.68	F8100	想法, 原理, 定理, 模式, 频率, 预算, 录像, 避免 *(idea, principle, theorem, pattern, frequency, budget, video, avoid)*
DE	Free-association pool	$+7.51$	$+2.62$	F5707	Kopf, Herz, Blume, Kirche, Tasche, Tisch, Vogel, Zug, Schlange, Musik *(head, heart, flower, church, bag, table, bird, train, snake, music)*
DE	Exam/correction broadcaster	-2.13	-1.94	F23101	Korrektur, Code, Klausur, Bearbeitung, Aufzählung, Ausführung, Übersicht *(correction, code, exam, editing, enumeration, execution, overview)*
NL	Free-association pool	$+8.77$	$+3.89$	F17972	tijd, stoel, camera, krediet, wasmachine, vrouw, deur, koffie, chocolade, suiker *(time, chair, camera, credit, washing machine, woman, door, coffee, chocolate, sugar)*
NL	Document-analysis broadcaster	-2.83	-2.47	F10008	commentaar, samenvatting, structuur, analyse, vraag, voorwaarde, syntactisch, fouten *(commentary, summary, structure, analysis, question, condition, syntactic, errors)*

Table 18: Representative pool and broadcaster SAE features per language at the language-specific steering layer (full-test setting). Broadcaster suppression is recovered tightly (DE F23101: 91%; NL F10008: 87%; ZH F8100: 72%), while pool-feature amplification is only partially recovered (EN: ∼46%; NL: ∼44%; DE: ∼35%; ZH: ∼48%).

in Table 9). Validation steering sweeps: ∼240 H100-hours total. SAE encoding and analysis: negligible. OpenCompass benchmark evaluation: ∼2–8 H100-hours per system.

Code and artifacts. Code (LLaMA-Factory frequency-weighted-loss patch, training YAMLs, layer-localization diagnostics, steering scripts, SAE encoding and comparison), model checkpoints (vanilla, standard SFT, frequency-aware SFT), per-layer steering vectors, SAE feature activations and reviewed feature annotations, the four-language 215-cue aligned cue list, and full per-language validation-sweep CSVs are released anonymously at https://anonymous.4open.science/r/swow-steering-artifact-FFBC. Random seed 42 throughout.

Bin	Lang.	Feature	Δ^{SFT}	Δ^{ste}	Top activating cues (full test; *italic* = EN translation)
Embodied / concrete	EN	F23208	+3.65	—	raincoat, lighthouse, scarecrow, yoghurt, dairy, washing
	ZH	F29065	+3.07	—	春风, 柠檬, 面包, 兔子, 鲤鱼, 狮子, 大象, 鱼 *(spring breeze, lemon, bread, rabbit, carp, lion, elephant, fish)*
	DE	F5707	+7.51	+2.62	Kopf, Herz, Blume, Kirche, Tasche, Tisch, Vogel, Zug *(head, heart, flower, church, bag, table, bird, train)*
	NL	F5548	+3.32	—	zool, handpalm, greppel, kluif, haarspeld, goot, inham, wand *(sole, palm, ditch, bone, hairpin, gutter, inlet, wall)*
Everyday / scene	EN	F10874	+4.02	—	book, buy, work, eat, sleep, sell, house, bed, children
	ZH	F10874	+1.91	—	吃, 店, 床, 衣服, 打电话, 买, 工作, 写, 题目, 看 *(eat, shop, bed, clothes, call, buy, work, write, topic, look)*
	DE	F16534	+5.54	+2.53	Euro, essen, Huhn, kaufen, Kartoffel, trinken, Autofahrer *(eat, chicken, buy, potato, drink, driver)*
	NL	F17972	+8.77	+3.89	tijd, stoel, camera, krediet, wasmachine, vrouw, deur, koffie *(time, chair, credit, washing machine, woman, door, coffee)*
Cultural memory	EN	F24404	+4.04	—	Robin Hood, Manhattan, Broadway, soprano, badminton, Methodist, New Orleans
	ZH	F19960	+3.06	—	中华, 党, 基层, 血缘, 命, 仙, 热血, 汉族, 信念 *(China, Party, grassroots, blood kinship, fate, immortal, hot-blooded, Han ethnicity, belief)*
	DE	F17125	+5.52	+2.21	Herz, Kunst, Freiheit, Selbstvertrauen, Rassismus, Goethe, Wissenschaft, Islam *(heart, art, freedom, self-confidence, racism, Goethe, science, Islam)*
	NL	F11213	+1.45	—	protestants, jongere, vlaams, allochtoon, wiskundig, zuiders *(Protestant, youngster, Flemish, immigrant, mathematical, southern)*
Evaluative / scalar	EN	F12389	+2.69	—	warranty, election, impossible, influential, cowardly, forbidden
	ZH	F24736	+3.79	—	过失, 变换, 哀伤, 协作, 差别, 怪异, 意愿, 不正, 责备 *(fault, transform, sorrow, cooperation, difference, strange, willingness, wrongful, blame)*
	DE	F20326	+3.25	+1.26	ziemlich, kaum, einverstanden, kümmern, mittelmäßig, großzügig *(quite, barely, agreed, care for, mediocre, generous)*
	NL	F29792	+1.53	—	levendig, liefelijk, onprettig, droef, hevig, vaag, slank, teder *(lively, lovely, unpleasant, sad, intense, vague, slim, tender)*

Table 19: Rank-1 SFT-positive feature per bin per language on the full test set. "—" for Δ^{ste} indicates the feature is in SFT's top-50 but not in steering's top-50 at the same rank (full per-feature steering deltas in supplementary CSVs). Embodied + everyday features dominate every language; cultural-memory features carry community-specific named entities (*Robin Hood*, 中华, Goethe, vlaams). English translations provided for ZH, DE, and NL cues.

Bin	Lang.	Feature	Δ^{SFT}	Δ^{ste}	Top vanilla cues (full test; *italic* = EN translation)
Assistant / instruction	EN	F8100	-6.15	-3.66	instruction, provide, shall, explore, investigation, minimize, language
	ZH	F8023	-1.90	—	方法, 答案, 返回, 优化, 写, 开始, 题目, 关闭, 好 *(method, answer, return, optimize, write, start, topic, close)*
	DE	F7731	-3.89	-2.41	Antwort, Aufzählung, unfreiwillig, namens, geistig, alt, anerkannt *(answer, enumeration, involuntary, named, intellectual, old, recognized)*
	NL	F25454	-1.99	—	antwoord, commentaar, samenvatting, proloog, vertalen, omgekeerd *(answer, commentary, summary, prologue, translate, reversed)*
Document / analysis	EN	—	—	—	—
	ZH	—	—	—	—
	DE	F23101	-2.13	-1.94	Korrektur, Code, Klausur, Bearbeitung, Aufzählung, Ausführung, Übersicht *(correction, code, exam, editing, enumeration, execution, overview)*
	NL	F10008	-2.83	-2.46	commentaar, samenvatting, structuur, analyse, vraag, voorwaarde, syntactisch *(commentary, summary, structure, analysis, question, condition, syntactic)*
Abstract / method	EN	—	—	—	—
	ZH	F8100	-3.72	-2.68	想法, 原理, 定理, 模式, 频率, 策略, 跳跃, 强调 *(idea, principle, theorem, pattern, frequency, strategy, jump, emphasize)*
	DE	—	—	—	—
	NL	—	—	—	—
Concrete / world attenuation	EN	F28607	-2.10	—	New Orleans, dice, camera, golf, butterfly, Cuba, badminton, elephant
	ZH	F11029	-3.46	—	天气, 家具, 战争, 关爱, 甲虫, 快餐, 大象, 太阳, 海洋 *(weather, furniture, war, care, beetle, fast food, elephant, sun, ocean)*
	DE	F7549	-7.30	-3.59	Fahrrad, Brot, Limonade, Kaffee, Trauma, Bier, Tasse, Zitrone, Kuh *(bicycle, bread, lemonade, coffee, trauma, beer, cup, lemon, cow)*
	NL	F7513	-3.21	—	vlaams, alsof, theater, dertien, water, bloed, vuurtoren, kaas *(Flemish, as if, theatre, thirteen, water, blood, lighthouse, cheese)*
Affect / evaluative attenuation	EN	F9918	-3.06	—	get rid of, shorter, shifty, impulsive, aloof, tactless, insincere
	ZH	F7150	-2.42	—	原理, 客套, 两弹一星, 定理, 花卉, 轻视, 墨守成规, 骂 *(principle, politeness, Two Bombs One Satellite, theorem, flowers, contempt, follow rules rigidly, scold)*
	DE	F29172	-1.72	-1.10	lachen, geizig, luftig, etablieren, reparieren, erfrischend, hässlich *(laugh, stingy, airy, establish, repair, refreshing, ugly)*
	NL	F29387	-1.83	—	duidelijk, dus, vanbinnen, wel, overweldigend, verdrietig, taalvaardig *(clear, so/thus, within, well, overwhelming, sad, language-proficient)*

Table 20: Rank-1 SFT-negative feature per bin per language on the full test set. The document/analysis broadcaster is unique to DE and NL; the abstract/method broadcaster is unique to ZH; the assistant broadcaster appears in all four but is largest in EN ($\Delta^{\text{SFT}} = -6.15$). Steering recovers broadcaster suppression closely where it engages with the same feature. English translations provided for ZH, DE, and NL cues.