

A Universal Vibe? Finding and Controlling Language-Agnostic Informal Register with SAEs

Uri Z. Kialy*
Ariel University
uri.kialy@ariel.ac.il

Avi Shtarkberg*
Ariel University
avishb1213@gmail.com

Ayal Klein
Ariel University
ayalk@ariel.ac.il

Abstract

While multilingual language models successfully transfer factual and syntactic knowledge across languages, it remains unclear whether they process culture-specific pragmatic registers, such as slang, as isolated language-specific memorizations or as unified, abstract concepts. We study this by probing the internal representations of Gemma-2-9B-IT using Sparse Autoencoders (SAEs) across three typologically diverse source languages: English, Hebrew, and Russian. To isolate pragmatic register processing from trivial lexical sensitivity, we introduce a novel dataset in which every target term is polysemous, appearing in both literal and informal contexts. We find that while much of the informal-register signal is distributed across language-specific features, a small but highly robust cross-linguistic core consistently emerges. This shared core forms a geometrically coherent “informal register subspace”, with greater geometric concentration at the deeper layer. Crucially, these shared representations are not merely correlational: activation steering with these features causally shifts output formality across all source languages and transfers zero-shot to six held-out languages spanning diverse language families and scripts. Together, these results provide the first mechanistic evidence that multilingual LLMs internalize informal register not just as surface-level heuristics, but as a portable, language-agnostic pragmatic abstraction.

1 Introduction

While large language models (LLMs) have largely mastered multilingual syntax and factual recall, human communication is fundamentally driven by social signaling. When a speaker uses informal register or slang, they are not merely conveying information; they are negotiating social distance, signaling cultural belonging, and expressing casual

intimacy. Because modern alignment pipelines overwhelmingly favor standard, highly edited text, LLMs suffer from a pervasive “register gap.” They frequently fail to appropriately process the dynamic, context-dependent language actually used in human interaction (Sun et al., 2024; Li et al., 2024), leading to miscalibrated responses and latent safety vulnerabilities where harmful intent is obfuscated by casual phrasing. Addressing this requires models calibrated not just for factual accuracy, but for deep sociolinguistic competence.

This register gap exposes a fundamental, unresolved question about the architecture of multilingual representation: when an LLM encounters informal register across diverse cultures, does it process each instance as an isolated, language-specific surface form? Or does it construct a unified, language-agnostic abstraction of “casual human interaction”? Prior mechanistic interpretability work has demonstrated that models build shared cross-lingual subspaces for facts and syntax (Conneau et al., 2020; Wendler et al., 2024), but whether they abstract *pragmatic* and *social* concepts remains entirely unknown. We ask: *Do LLMs represent informal register as a universal pragmatic concept, or as a disconnected collection of language-specific memorizations?*

To answer this, we probe the internal representations of Gemma-2-9B-IT using Sparse Autoencoders (SAEs). To isolate pragmatic register from the trivial confound of lexical bias (e.g., detecting informality merely due to the presence of specialized slang terms), we introduce a novel cross-lingual dataset across English, Russian, and Hebrew. Crucially, our dataset is strictly lexically controlled: every target term is polysemous, appearing in both its standard literal sense and its informal slang sense. This forces the model to resolve register from pragmatic context rather than mere vocabulary.

What we find fundamentally challenges the

* Equal contribution.

view of models as mere surface-pattern matchers. By isolating latent SAE features that activate differentially across slang and literal contexts, we identify a small, geometrically coherent “informal register subspace” in the model’s latent space, shared across all three typologically diverse source languages. Crucially, this abstraction is not merely correlational. Steering with this cross-linguistic feature set systematically and causally modulates output formality. Strikingly, this single sociolinguistic intervention transfers zero-shot to six typologically distant languages—including Japanese, Thai, Amharic, and Georgian—that were completely absent from our feature extraction pipeline. This zero-shot generalization provides compelling evidence that the model has internalized informal register not as a linguistic artifact, but as a shared, abstract pragmatic concept.

Our main contributions are threefold. **(i) Methodological:** We introduce a novel, lexically-controlled dataset of polysemous terms across English, Hebrew, and Russian designed to isolate pragmatic register from any vocabulary bias. **(ii) Observational:** We identify a geometrically coherent, language-agnostic “informal register subspace” within the SAE latent space, whose structure becomes more concentrated at greater model depth. **(iii) Causal:** We demonstrate that these shared latent features causally drive generation formality. The zero-shot transfer of this control to six typologically diverse, unseen languages provides the first mechanistic evidence for a cross-linguistic representation of pragmatic register.¹

2 Background

The Register Gap in NLP While large language models exhibit remarkable multilingual fluency, their training and alignment are overwhelmingly anchored in standard, edited text. This creates a critical “register gap” when models encounter the dynamic, culturally situated nature of informal language (Li et al., 2024). Prior multilingual research on register and style has largely focused on supervised formal-to-informal text transfer (Rao and Tetreault, 2018; Briakou et al., 2021b,a). Multilingual resources for *slang* remain exceptionally scarce; existing datasets are typically monolingual (Sun et al., 2024) or examine

variation within individual languages rather than across diverse typologies (Wuraola et al., 2024). Recent evaluations further show that even state-of-the-art models struggle to reliably deploy or comprehend informal register in context (Wu and Sun, 2025). Yet how multilingual models internally represent informal register—and whether these representations are shared across languages—remains largely unexplored.

Cross-Linguistic Model Interpretability A central debate in representation learning is whether multilingual models construct a shared, abstract “interlingua” or merely memorize parallel, language-specific surface patterns. Recent work has identified emergent cross-lingual alignment (Conneau et al., 2020) and demonstrated that semantically equivalent concepts occupy shared geometric subspaces (Wendler et al., 2024), often mediated by specialized multilingual neurons (Tang et al., 2024; Liu et al., 2024). However, these investigations have heavily indexed on factual knowledge and syntactic structure. Much less is known about whether models develop abstract, shared circuitry for *pragmatic register*—a dimension of language that is inherently social and culturally idiosyncratic. While prior work has successfully identified monolingual steering directions and style-sensitive neurons (Subramani et al., 2022; Rimsky et al., 2024; Lai et al., 2024), it remains an open question whether these mechanisms transfer cross-lingually. Our results bridge this gap, examining the balance between shared and language-specific latent features to determine if models possess a universal pragmatic architecture.

Sparse Autoencoders and Activation Steering

To untangle these complex pragmatic representations, we utilize Sparse Autoencoders (SAEs). By decomposing dense neural activations into an overcomplete, sparse basis, SAEs effectively mitigate polysemanticity, revealing human-interpretable features that standard neuron-level analyses obscure (Bricken et al., 2023; Templeton et al., 2024; Cunningham et al., 2023). Beyond serving as observational lenses, SAEs unlock *activation steering*: the ability to intervene on specific latent features at inference time to test their causal role in model behavior. This technique has successfully modulated model safety, mitigating social biases (Durmus et al., 2024) and controlling refusal behaviors (O’Brien et al., 2024; Arad et al., 2025).

¹Code available at https://github.com/UriKialy/universal_vibe.git.

Term	Usage	Example Sentence
<i>fire</i>	Literal	It says there was a big fire in London 300 years ago.
<i>fire</i>	Slang	That movie was straight fire.
<i>sick</i>	Literal	I think I better go home before I’m sick.
<i>sick</i>	Slang	My car’s got fresh 20’s, they look sick!

Table 1: Examples of polysemous target terms. The identical token forces register disambiguation from context alone.

We extend this causal framework to the sociolinguistic domain, using SAE-derived steering vectors to test whether a shared cross-linguistic “slang subspace” can predictably control output register across diverse languages.

3 Dataset

To evaluate cross-lingual register representations without the confound of lexical bias, we constructed a novel dataset across English, Hebrew, and Russian. To our knowledge, this is the first slang benchmark spanning more than two typologically distinct languages (Mei et al., 2024; Tiwari et al., 2025; Wuraola et al., 2024).

Our core design principle is strict lexical control. Prior to sentence collection, native speakers manually compiled a set of *polysemous* target terms, each carrying both a standard literal sense and a low-register informal sense (see Table 1). By evaluating register classification on sentences sharing the exact same target token, we ensure models cannot rely on word identity, out-of-vocabulary novelty, or term frequency; instead, they must resolve the register entirely from pragmatic context.

Instance-level labels (*Slang* vs. *Literal*) were assigned using GPT-4o-mini (OpenAI, 2024). While prior work establishes the broader LLM-as-a-judge paradigm for text evaluation (Zheng et al., 2023; Pangakis et al., 2023; Gilardi et al., 2023), the specific reliability of this framework for formal/informal register distinction is validated by our own human evaluation. Corroborated by author manual review, the automated labeling achieved a precision of ≥ 0.92 across all three languages (detailed in Appendix A). Detailed dataset statistics and slang-to-literal ratios are provided in Appendix A.²

²For access to the dataset, please contact the authors.

3.1 Corpus Composition

English. Comprises 2,835 labeled instances spanning 130 target terms (e.g., *fire*, *wet*, *sick*). Sentences were sourced from curated HuggingFace corpora and the naturalistic colloquial corpus introduced by Sun et al. (2024).

Hebrew. Contains 4,527 labeled instances across 18 target terms (e.g. *אש* – “fire/awesome”). Data was sourced from HuggingFace corpora (theatrical text, forums, lyrics). Hebrew’s non-concatenative morphology makes our strictly controlled token constraint particularly vital, as inflected surface forms of slang vary substantially.

Russian. Consists of 1,259 labeled instances covering 15 target terms (e.g., *пушка* – “gun/banger”). Sentences were organically sourced via API from VKontakte (VK) and Telegram, providing an ecologically valid representation of contemporary informal Russian.

Lexical diversity. Lexical concentration varies substantially across languages, with a broader target-term distribution in English and greater concentration in Hebrew and Russian (Appendix A).

4 Latent Feature Extraction and Analysis Setup

We use Gemma-2-9B-IT, leveraging its extensive suite of publicly available 131K-feature Sparse Autoencoders (Lieberum et al., 2024). Preliminary inspection identified substantial feature activity at Layer 9 and Layer 20, which serve as our mid- and deep-layer comparative baselines, respectively.³

4.1 SAE Feature Extraction

To isolate contextualized register representations from trivial lexical confounds, our extraction pipeline focuses strictly on the *target term* (the polysemous word used in either a slang or literal sense). For each sentence, we extract the model’s residual stream activations at the subword positions spanned by the target term and encode each activation via the pretrained SAE into a sparse feature vector:

$$\mathbf{f} = (f_1, \dots, f_{131072}), \quad f_i \geq 0,$$

where each dimension corresponds to one latent SAE feature. We apply this procedure indepen-

³A deeper layer (Layer 31) yielded negligible target-term activity passing our minimum threshold—likely because final layers heavily compress representations toward next-token prediction—and was therefore excluded from analysis.

dently across all three source languages to systematically evaluate the cross-lingual intersection of these target-term representations.

4.2 Feature Scoring and Selection

To identify slang-associated latent features, we compute the *differential activation frequency* for each feature i :

$$\Delta_i = P(f_i > 0 \mid \text{slang}) - P(f_i > 0 \mid \text{literal}). \quad (1)$$

We estimate these probabilities independently for each language and layer by pooling all target-term subword positions corpus-wide. Each subword position is treated as a separate observation and inherits the slang/literal label of its corresponding target-term instance; consequently, targets split into multiple subwords contribute multiple observations. Thus, Δ_i measures how much more frequently feature i activates at target-term positions in slang than in literal contexts. For subsequent analyses and experiments, we rank all 131,072 SAE features by Δ_i and retain the top- k features, using $k=100$ for defining the cross-linguistic overlap set (§4.3).

Separately, for the pooled feature-quality analysis in §5.1—where each feature is evaluated as a standalone binary classifier and rare features would yield unstable precision and recall estimates—we apply a *minimum activity filter* consisting of two criteria: a feature must activate on at least 5% of slang tokens and fire a minimum of 10 times across the combined dataset. These dual thresholds ensure that retained features represent a meaningful fraction of informal usage—rather than isolated idiosyncratic phrases—while accumulating sufficient absolute frequency for stable statistical estimation.

We stress that the two steps are independent and applied in this order: the per-language top- k rankings are computed on the unfiltered Δ_i ordering, and the activity filter is applied only to the pooled classifier analysis. Applying the filter before ranking would alter which features enter the top-100 lists, and hence the composition and size of the cross-linguistic core reported in §4.3.

4.3 Derived Feature Sets

Intersecting the top-100 ranked feature lists across the three languages, we identify a *cross-linguistic* feature core: features appearing in the top-100 lists of all three languages simultaneously—9 fea-

tures at Layer 9 and 10 at Layer 20. We also investigate high-salience bilingual feature sets for completeness (see details in Appendix J), but our primary focus is on the cross-linguistic core as the strongest candidate of shared informal register representation.

These set definitions play two roles in the paper. First, they support the observational analyses in Section 5, where we ask whether slang-related features exhibit shared cross-linguistic structure. Second, they seed the feature pools used later for activation steering (Section 6), where we test whether intervening on these feature sets causally shifts output register.

5 Observational Findings

We evaluate the extracted feature sets across four dimensions: individual discriminative quality, cross-lingual overlap, geometric coherence, and lexical readout. Across both analyzed layers, a consistent picture emerges: while the informal-register signal is distributed over multiple latent features, a small but non-trivial subset of these features is shared across languages, and this shared subset becomes more geometrically coherent at the deeper layer.

5.1 Individual Feature Quality and Distributed Signal

We begin by evaluating whether slang detection is dominated by a few especially strong latent features or instead distributed across a broader set. To do so, we treat each of the 131,072 SAE features as a binary classifier on the pooled cross-linguistic dataset (5,256 slang subword tokens; 7,697 literal subword tokens), using the rule *active* $\Rightarrow$ *slang*, and retaining only features that pass the minimum activity filter from Section 4.2. These counts are at the subword-token level rather than the instance level: the extraction pipeline (§4.1) records the residual-stream activation at every subword token spanned by a target term, so each labeled instance contributes one or more tokens depending on Gemma-2’s tokenizer. The ratio is close to 1.0 for English (977 tokens from 968 slang instances) but approaches 1.9 for Hebrew and Russian, whose target terms are frequently split into multiple subwords; instance-level counts are given in Appendix A, Table 3.

The resulting feature landscape shows a clear precision–recall trade-off at both layers. Some fea-

tures are relatively precise but fire on only a limited subset of slang tokens, whereas others provide broad coverage at lower precision. For example, high- Δ_i features such as 35440 (Layer 20, precision 0.849) and 38266 (Layer 9, precision 0.776) achieve relatively clean slang selectivity with moderate recall, while broad-coverage features such as 127994 (Layer 9) and 26379 (Layer 20) capture more than 85% of slang tokens at substantially lower precision. No single feature dominates all classifier metrics: Feature 93521 at Layer 20 is the only feature that appears simultaneously in the top-10 lists for precision, recall, and F1 (precision 0.966, recall 0.240, F1 0.384), but it leads none of them individually.

This distributed pattern motivates our focus on *feature sets* rather than individual features. The latent representation of informal usage is not concentrated in a single “slang neuron”, but spread across multiple partially overlapping sparse features.

5.2 Cross-Linguistic Feature Overlap

We next ask whether the most slang-associated features are largely language-specific or whether some are shared across languages. Intersecting the top-100 Δ_i -ranked features for English, Hebrew, and Russian at each layer yields a consistent three-tier structure: a small cross-linguistic core, a moderate bilingual set, and a larger language-specific remainder. The cross-linguistic core is small at both analyzed layers—9 features at Layer 9 and 10 at Layer 20—against 32 and 30 features respectively appearing in the top-100 lists of exactly two languages, and 209 and 210 appearing in exactly one. Thus, most slang-related latent features are not cross-linguistic, but the model consistently contains a small shared subset that survives across all three languages. Among language pairs, Hebrew–Russian shares more exclusive features than Hebrew–English at both layers (15 vs. 6 at Layer 9; 14 vs. 5 at Layer 20), possibly reflecting shared lower-resource status in the training distribution; English–Russian accounts for the remaining 11 at each layer.

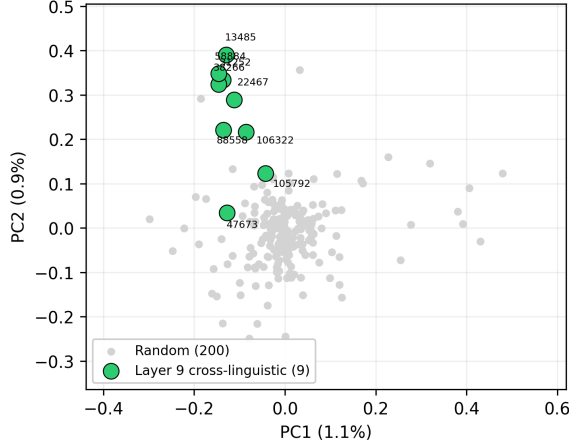
Statistical Significance of Overlap Counts. To test whether this overlap is greater than expected by chance, we perform a label-permutation test ($n = 10,000$). In each iteration, we independently shuffle slang/literal labels within each language, recompute feature rankings by Δ_i , and count the resulting triple overlap among the top-

100 lists. This permutation preserves the marginal activation frequencies and sparsity dynamics of every feature, isolating the target effect by destroying only the semantic mapping between activations and the slang–literal pragmatic labels. Under this null, the mean triple overlap is 0.74 (std = 0.90), and the maximum observed across all permutations is 7. The observed overlaps of 9 (Layer 9) and 10 (Layer 20) therefore lie outside the entire null distribution ($p < 10^{-4}$), supporting the conclusion that the cross-linguistic core reflects genuine shared structure tied to the slang–literal contrast rather than incidental feature co-occurrence. Taken together, these results suggest that the model represents informal usage through a mixed architecture: predominantly language-specific and bilingual features, plus a small but statistically robust cross-linguistic core.

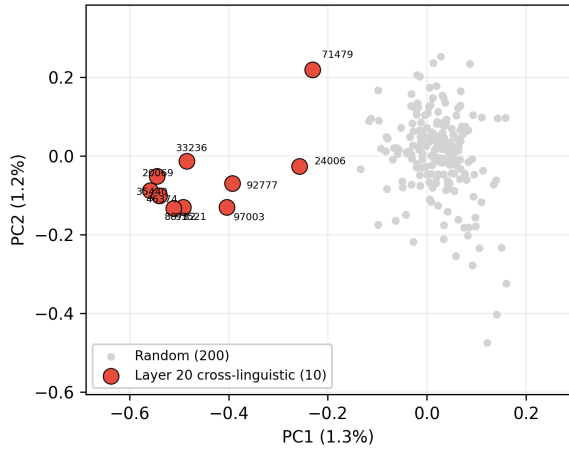
5.3 Geometric Coherence of the Shared Feature Set

Feature overlap alone does not imply that the shared features form a coherent latent subspace: in principle, the same features could recur across languages without occupying a meaningful region of decoder space. We therefore examine the geometry of the cross-linguistic feature core using the SAE decoder vectors $\{\mathbf{w}_i^{\text{dec}}\}_{i \in \mathcal{F}_{\text{cross}}}$. For each layer, we compute the mean pairwise cosine similarity among the decoder vectors of the cross-linguistic features, and compare it to the mean *absolute* cosine similarity between those features and 1,000 randomly sampled SAE features, averaged over five independent random draws. We take absolute values because the signed core-to-random mean is essentially zero at both layers (+0.0004 at Layer 9, −0.0008 at Layer 20); the magnitude of the alignment, not its sign, is the meaningful reference scale. The random sample itself is internally unstructured, with near-zero mean pairwise cosine similarity (0.0007 at Layer 9 and 0.0014 at Layer 20), indicating that any observed structure is not simply an artifact of the global SAE geometry. By contrast, the cross-linguistic features exhibit substantially elevated within-group similarity. To quantify this clustering, we define an *island score* as the ratio of the mean pairwise cosine similarity within the cross-linguistic set to its mean absolute cosine similarity with the random sample. A score substantially above $1 \times$ indicates a geometrically coherent cluster. At Layer 9, the mean within-group similarity is 0.065 against

a random reference of 0.0150, an island score of $4.3\times$. At Layer 20, coherence sharpens to 0.152 against 0.0195, a $7.8\times$ island score. Both values are stable under resampling (4.32–4.38 and 7.82–7.95 across five random seeds). Thus, the cross-linguistic features do not merely overlap by rank; they occupy a denser and more coherent region of decoder space, especially at the deeper layer (Figure 1). This geometric result strengthens the over-



(a) Layer 9 (island score $4.3\times$)



(b) Layer 20 (island score $7.8\times$)

Figure 1: Illustrative projection of cross-linguistic informal-register feature decoder vectors (colored) against randomly sampled SAE features (grey). Cross-linguistic features form a tighter grouping at Layer 20 than at Layer 9. The quantitative evidence comes from cosine-similarity analysis in the full decoder space; the 2D projection is shown only as a visual illustration, using an independent 200-feature random subsample.

lap analysis: the cross-linguistic core is not just statistically non-random, but also internally structured in a way consistent with a shared latent direction family.

5.4 Lexical Readout via Vocabulary Projection

As a post-hoc interpretability probe, we project each cross-linguistic feature’s decoder weight vector through the model’s *unembedding matrix* $W_U \in \mathbb{R}^{d \times |V|}$ —the model’s final linear map from a d -dimensional residual-stream vector to logits over the vocabulary V , and the output-side counterpart of the input embedding (tied to it in Gemma-2-9B-IT)—following Gur-Arieh et al. (2025), in order to inspect which vocabulary items each latent direction most promotes. Roughly half of the core features yield token lists clearly enriched for informal or register-marked expressions—including *dudes*, *vibes*, *swag*, and *dope*, alongside evaluative intensifiers and profanity—while the remainder promote tokens with no readily interpretable lexical structure. We therefore interpret this not as a standalone proof of feature meaning, but as a qualitative sanity check that a substantial subset of the identified cross-linguistic features have lexical readouts consistent with informal register. Examples for the two highest- Δ_i features are provided in Appendix C. Interestingly, features whose discriminative signal derives from Hebrew or Russian alone—including ones with no English slang—literal signal at all—often still project to English tokens, suggesting a systematic English bias in the *unembedding space* (the column space of W_U , spanned by the per-token output directions). This bias manifests at generation time as occasional language switches toward English, though these affect approximately 10% of steered completions overall (Appendix G). For this reason, we use vocabulary projection only as a supporting interpretability analysis: it is informative about the lexical tendencies of a latent direction, but it does not by itself establish the language-specificity or cross-linguistic generality of that feature. Those properties are instead supported by the overlap and steering analyses.

5.5 A Descriptive Depth Trend

Although our main results focus on the cross-linguistic feature core, the comparison between Layer 9 and Layer 20 reveals a consistent descriptive trend.

First, the model represents informal register far more sparsely at the deeper layer. Because features at Layer 20 are highly specialized, fewer of

them activate frequently enough to pass our 5% minimum activity threshold—the requirement that a feature fire on at least 5% of slang tokens—with 169 pooled features passing at Layer 20 against 892 at Layer 9. The same contrast appears at the token level: at Layer 9, slang targets trigger broader activation than literal targets, whereas at Layer 20 they engage fewer active features at lower total activation magnitude. Second, the cross-linguistic core identified at Layer 20 is more geometrically coherent than its Layer 9 counterpart, as reflected in its higher within-group cosine similarity and larger island score.

Taken together, these observations suggest a depth-dependent reorganization of informal-register information: the signal is more broadly distributed at Layer 9, while at Layer 20 it is represented by a sparser set of features occupying a more geometrically coherent region of decoder space. We treat this as a descriptive empirical trend rather than a strong mechanistic claim.

6 Causal Experiments

We next test whether the cross-linguistic informal-register SAE features identified above can causally control register during generation, rather than merely tracking it as a correlational artifact. We use the Layer 20 cross-linguistic core as our primary intervention set and report complementary Layer 9 experiments in Appendix H. Crucially, testing these interventions across linguistic boundaries allows us to assess the extent to which these features encode a cross-linguistic, abstract representation of pragmatic register.

6.1 Experimental Setup

Steering Mechanism For each feature set $\mathcal{F}$ (e.g., cross-linguistic features), we construct a steering vector $\mathbf{v}_{\mathcal{F}}$ by averaging the SAE decoder weight vectors of its constituent features:

$$\mathbf{v}_{\mathcal{F}} = \frac{1}{|\mathcal{F}|} \sum_{i \in \mathcal{F}} \mathbf{w}_i^{\text{dec}} \quad (2)$$

where $\mathbf{w}_i^{\text{dec}}$ is the decoder weight vector for feature i . The resulting vector is L2-normalized: $\mathbf{v}_{\mathcal{F}} \leftarrow \mathbf{v}_{\mathcal{F}} / \|\mathbf{v}_{\mathcal{F}}\|_2$. During generation, we inject the steering vector into the post-block residual stream at layer L via a forward hook applied at every generated token position t :

$$\mathbf{h}'_{L,t} = \mathbf{h}_{L,t} + \alpha \cdot \mathbf{v}_{\mathcal{F}} \quad (3)$$

where $\mathbf{h}_{L,t}$ is the residual stream activation at layer L and token position t , α is the steering coefficient, and $\mathbf{h}'_{L,t}$ is the modified activation. Negative α steers toward formal language; positive α steers toward informal language. We test 6 coefficients ranging from -150 to 100 . We construct steering vectors from the cross-linguistic feature set: the 10-feature core at Layer 20 for the primary experiment (§6.2), and the 9-feature core at Layer 9 (Appendix H). Results for bilingual-exclusive feature sets are reported in Appendix J.

Prompts and Languages We use 20 open-ended sentence prompts (e.g., “I think that...”) across nine languages (see Appendix E for the full prompt list), yielding $\sim 1,080$ sentence completions. The prompts were translated from English into Hebrew, Russian, and German by native speakers, with translations validated by GPT-4o, and into the remaining five languages by the same model directly. Our evaluation spans three *source* languages whose data contributed to feature extraction (English, Hebrew, and Russian) and six *zero-shot generalization* languages entirely absent from our feature extraction data: German, Japanese, Hindi, Thai, Georgian, and Amharic. The six zero-shot languages were chosen to maximize typological diversity, spanning six different language families (Germanic, Japonic, Indo-Aryan, Kra-Dai, Kartvelian, and Semitic) and six distinct script traditions.

Evaluation Formality is evaluated on a continuous scale $[0, 1]$ using GPT-4o-mini as an automated judge, where 1 represents highly formal and 0 highly informal language. Steering effectiveness is measured via Pearson correlation between α and formality score, computed over all individual completions (i.e., 120 per language per feature set). To validate the automated formality scores, we manually rated 30 completions per language for four languages: the three source languages (English, Hebrew, and Russian) and one zero-shot language (German). For each language, the completions were scored by one fluent speaker of that language, drawn from the authors or their fluent acquaintances. GPT-4o-mini scores correlated with the human ratings at $r = 0.62$, indicating moderate agreement. Given the inherently subjective nature of formality judgments, we treat the automated scores for the remaining five zero-shot languages as reasonable proxies for human judgment.

Target Language	r (L20)
English	-0.506
Hebrew	-0.398
Russian	-0.505
German	-0.458
Japanese	-0.508
Hindi	-0.425
Thai	-0.499
Georgian	-0.328
Amharic	-0.437

Table 2: Summary of steering results — Pearson r correlation between α (steering strength) and generated text formality, computed over the operating range $\alpha \in [-150, 100]$. The cross-linguistic feature set effectively controls the register of generated texts at all source languages, and this transfers to six zero-shot languages absent from the feature extraction pipeline. Results of steering in Layer 9 are reported in Appendix H.

6.2 Results

Source Language Control and Bilingual Transfer As detailed in Table 2, applying the cross-lingual steering vectors to the three source languages produces clear, statistically significant negative correlations between the steering coefficient α and output formality (r between -0.40 and -0.51 , all $p < 10^{-5}$). This confirms that the shared feature set causally modulates register within its derived languages.

As an intermediate test of cross-lingual abstraction, we evaluate *bilingual-exclusive* feature sets—defined from two source languages and applied exclusively to the held-out third. These feature sets also produce significant monotonic decreases in formality (Appendix J), establishing that causal register control does not require the target language itself to contribute directly to feature extraction.

Zero-Shot Transfer to Unseen Languages

Building on this abstraction, we test whether this causal control transfers zero-shot to six languages entirely absent from feature identification. Despite their typological distance from the source-language set, every tested language shows a significant decrease in formality as α increases, with the effect concentrated in the informal direction (Figure 2).

The strongest zero-shot transfer occurs in Japanese and Thai, while even Georgian and Amharic exhibit moderate but highly significant negative correlations ($p < 10^{-3}$). Across the zero-shot set, mean formality decreases substan-

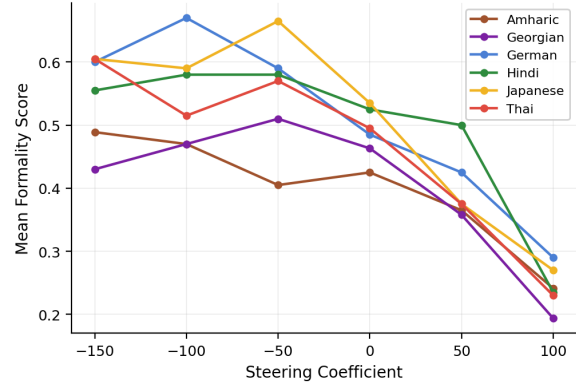


Figure 2: Per-language formality scores under cross-lingual feature steering for all six zero-shot transfer languages. All zero-shot languages exhibit a formality decrease as steering coefficient α increases, with the sharpest changes in the informal direction.

tially (from 0.548 at $\alpha = -150$ to 0.245 at $\alpha = 100$), demonstrating that the cross-lingual features successfully modulate register in typologically diverse, zero-shot settings.

Complementary Layer 9 experiments reproduce the same qualitative steering pattern across all three source languages, with correlations somewhat larger in magnitude than those observed at Layer 20 (Appendix H).⁴ Furthermore, comparison against a random feature baseline confirms that these formality shifts are specific to the identified cross-lingual subspace rather than a generic property of residual-stream perturbation (Appendix F).

Native Slang, Not Translation Beyond aggregate formality scores, we test whether steered completions remain coherent and in the target language. Gemma-2-9B-IT self-perplexity remains low and stable across all nine languages throughout $\alpha \in [-150, 100]$, then rises sharply beyond $\alpha = 100$ (Figure 3), motivating this interval as our operating range. External reference models (Qwen2.5-7B-Instruct, and DictaLM 2.0 for Hebrew) show the same pattern (Appendix I). Automatic language identification further shows that approximately 90% of completions remain in the target language, with most switches directed toward English (Appendix G).

Qualitative manual inspection further shows

⁴This contrasts with §5, where the Layer 20 cross-linguistic features are more geometrically coherent than those at Layer 9. This dissociation suggests that geometric organization and steering efficacy capture distinct aspects of the representation rather than necessarily increasing together.

that informal steering can be realized through language-specific slang rather than direct translations of English expressions. For the shared prompt stem “In my opinion, the movie was...”, strong formal steering ($\alpha = -150$) yields measured, explanatory prose in every language, whereas strong informal steering ($\alpha = +100$) produces casual completions with language-specific informal expressions: English *so good*, Russian *крутым* (“awesome”) and *бомба* (“a blast”), Hebrew *ממש כיף* (“really fun”). These distinct realizations are consistent with steering a shared pragmatic representation rather than reproducing a fixed English lexical pattern. Steering is similarly effective on English stems that begin in an informal register. Formal steering still strips out the casual tone—“Yo, that party was...” becomes “a lot of fun. I met some interesting people”—while informal steering intensifies it further, as in “...**LIT** last night!”

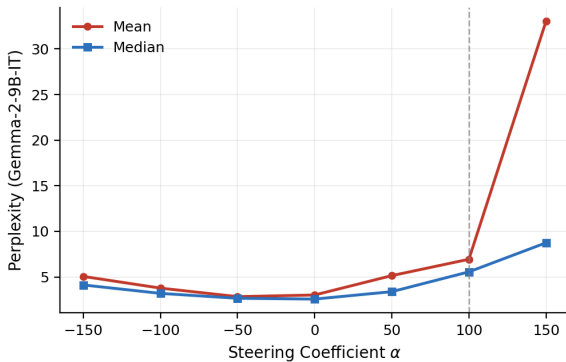


Figure 3: Mean and median perplexity (Gemma-2-9B-IT self-perplexity) across all nine languages as a function of steering coefficient α . Perplexity is stable across $[-150, 100]$ and rises sharply beyond $\alpha = 100$, motivating the chosen operating range.

7 Discussion

What Does the Shared Subspace Represent?

Our findings may reflect a representation broader than slang alone. The 9–10 cross-linguistic features we identify may partly capture a broader social-pragmatic function of informal language—for example, in-group signaling through linguistic deviation—that recurs across languages (Greenberg, 1963; Eble, 1996; Mattiello, 2008). This interpretation is consistent with the transfer of steering to languages absent from feature identification, and predicts that related phenomena such as code-switching, taboo language, or in-group jargon may engage overlapping cross-linguistic fea-

tures. In this sense, the “slang island” may provide a window into a broader pragmatic representational space.

At the same time, such a representation should not be expected to separate cleanly from semantics. Our contrastive design controls for target-word identity by comparing literal and slang uses of the same lexical item, but these uses can still differ in meaning, topic, affect, and social stance. Indeed, strict semantic–pragmatic disentanglement seems to be conceptually ill-posed: register is itself realized partly through semantic and affective choices. We therefore interpret the identified subspace as strongly associated with informal register rather than as a fully content-abstract register representation. Its cross-linguistic recurrence, causal steerability, and register-marked lexical readout provide converging evidence for this interpretation, while more tightly controlled contrasts could further characterize what information the subspace contains.

From Features to Circuits. The identified SAE features also provide a natural starting point for circuit-level analysis (Elhage et al., 2021). Activation patching (Wang et al., 2023) and emerging SAE feature-circuit methods (Marks et al., 2025) could test whether these features participate in a shared causal pathway for register processing across languages, extending the present feature-level analysis toward a mechanistic account of pragmatic control.

8 Conclusion

We investigated whether multilingual LLMs represent informal register through language-specific features alone or also through shared latent structure. Using Sparse Autoencoders in Gemma-2-9B-IT, we identified a small, geometrically coherent cross-linguistic feature core shared across English, Hebrew, and Russian. Steering this feature set causally modulates output formality and transfers zero-shot to six languages absent from feature identification. Together, these findings provide evidence that multilingual LLMs encode informal register through a partly shared, abstract pragmatic representation rather than solely through language-specific surface patterns.

9 Limitations

While our findings provide mechanistic evidence for cross-lingual pragmatic representations, sev-

eral limitations contextualize our results. First, our layer selection was constrained by the availability of pre-trained SAEs; a comprehensive layer sweep could reveal a more precise depth profile of slang representation. Furthermore, our experiments are conducted exclusively on Gemma-2-9B-IT; whether this geometrically coherent “informal register island” generalizes across other model architectures or scale regimes remains an open question.

Second, we operationalize informal register primarily through slang-sensitivity (the slang–literal contrast) evaluated at the target-term level. Sociolinguistically, slang (ephemeral, in-group vocabulary) and informal register (a broad, stable communicative mode) are distinct, and register shifts often manifest as sequence-level syntactic relaxations rather than isolated lexical choices. However, our causal steering experiments demonstrate that intervening on these token-derived features systematically controls broader generation formality. Furthermore, the steering results extend beyond the source languages and are not reducible to target-word identity alone, supporting a broader pragmatic interpretation of these features.

Finally, our evaluation methodology and the model’s own latent geometry exhibit an English-centric bias. Vocabulary projection of the universal features heavily favors English terms, and steered generation occasionally induces language switching toward English, particularly at stronger steering coefficients ($\alpha > 100$). This suggests that the “language-agnostic” subspace may function partially as an English-anchored hub rather than a perfectly symmetric abstract interlingua. Additionally, relying on GPT-4o-mini as an automated formality judge for typologically distant languages (e.g., Amharic, Thai) risks projecting Western-aligned definitions of formality onto cross-cultural pragmatics. Future work incorporating comprehensive native-speaker evaluations will be necessary to fully disentangle universal pragmatics from pretraining data imbalances.

Acknowledgments

We used OpenAI’s ChatGPT during manuscript preparation to assist with language editing, stylistic refinement, and the organization and phrasing of portions of the Introduction, Discussion, Conclusion, and other manuscript text. All generated suggestions were critically reviewed and revised

by the authors, who take full responsibility for the final content.

References

- Dana Arad, Aaron Mueller, and Yonatan Belinkov. 2025. SAEs are good for steering – if you select the right features. *arXiv preprint arXiv:2505.20063*.
- Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021a. [Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1321–1336, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021b. [Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, and 1 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.
- Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034.
- Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.
- Esin Durmus, Alex Tamkin, Jack Clark, Jerry Wei, Jonathan Marcus, Joshua Batson, Kunal Handa, Liane Lovitt, Meg Tong, Miles McCain, and 1 others. 2024. Evaluating feature steering: A case study in mitigating social biases. *Anthropic Research*.
- Connie Eble. 1996. *Slang & Sociability: In-group Language Among College Students*. University of North Carolina Press.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal

- Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2021/framework/index.html>.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*.
- Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press, Cambridge, MA.
- Yoav Gur-Arieh, Roy Mayan, Chen Agassy, Atticus Geiger, and Mor Geva. 2025. [Enhancing automated interpretability with output-centric feature descriptions](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5757–5778, Vienna, Austria. Association for Computational Linguistics.
- Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-specific neurons for steering LLMs in text style transfer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443.
- Jiahuan Li, Yiqing Cao, Shujian Huang, and Jiajun Chen. 2024. Formality is favored: Unraveling the learning preferences of large language models on data with conflicting knowledge. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5307–5320.
- Tom Lieberum, Senthoran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*.
- Weize Liu, Yinlong Xu, Hongxia Xu, Jintai Chen, Xuming Hu, and Jian Wu. 2024. [Unraveling Babel: Exploring multilingual activation patterns of LLMs and their applications](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11855–11881, Miami, Florida, USA. Association for Computational Linguistics.
- 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. In *ICLR*.
- Elisa Mattiello. 2008. *An Introduction to English Slang: A Description of its Morphology and Semantics*. Polimetrica.
- Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, and Xueqi Cheng. 2024. SLANG: New concept comprehension of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.
- Kyle O’Brien, David Majercak, Xavier Fernandes, Richard Edgar, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangdeh. 2024. Steering language model refusal with sparse autoencoders. *arXiv preprint arXiv:2411.11296*.
- OpenAI. 2024. [GPT-4o system card](#). *Preprint*, arXiv:2410.21276.
- Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative AI requires validation. *arXiv preprint arXiv:2306.00176*.
- Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.
- Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of NAACL-HLT*, pages 129–140.
- Nina Rimskey, 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*, pages 15504–15522.
- Shaltiel Shmidman, Avi Shmidman, Amir DN Cohen, and Moshe Koppel. 2024. [Adapting LLMs to Hebrew: Unveiling DictaLM 2.0 with enhanced vocabulary and instruction capabilities](#). *Preprint*, arXiv:2407.07080.
- Nishant Subramani, Nivedita Suresh, and Matthew E Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581.
- Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. 2024. Toward informal language processing: Knowledge of slang in large language models. In *North American Chapter of the Association for Computational Linguistics*.
- Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. [Language-specific neurons: The key to multilingual capabilities in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.
- Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, and 1 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Tanmay Tiwari, Vibhu Gupta, Manikandan Ravikiran, and Rohit Saluja. 2025. HiSlang-4.9k: A benchmark dataset for Hindi slang detection and identification. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 464–472, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. [Interpretability in the wild: a circuit for indirect object identification in GPT-2 small](#). In *The Eleventh International Conference on Learning Representations (ICLR)*.

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

Siyang Wu and Zhewei Sun. 2025. How do language models generate slang: A systematic comparison between human and machine-generated slang usages. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6536–6559, Suzhou, China. Association for Computational Linguistics.

Ifeoluwa Wuraola, Nina Dethlefs, and Daniel Marciniak. 2024. Understanding slang with LLMs: Modelling cross-cultural nuances through paraphrasing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15525–15531, Miami, Florida, USA. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

A Dataset Statistics

	English	Hebrew	Russian	Total
Labeled instances	2,835	4,527	1,259	8,621
Slang instances	968	1,776	538	3,282
Literal instances	1,857	2,751	721	5,329
Unique target terms	130	18	15	—

Table 3: Dataset statistics per language. Counts are labeled (sentence, target-term) instances: a sentence containing two target terms contributes two instances. Ten English instances carry an inconclusive judge label and are excluded from both classes, so the slang and literal rows sum to 8,611 rather than 8,621.

This section provides a detailed breakdown of the multilingual dataset used in our experiments. Table 3 summarizes the distribution of slang and literal instances across the three source languages: English, Hebrew, and Russian. We collect sentences via two complementary methods. First, we sample from existing slang corpora such as OpenSubtitles-Slang (Sun et al., 2024), retaining only sentences where a target term from our human-curated list appears and is confirmed as slang usage by an LLM judge. Second, we scrape social media platforms and open HuggingFace datasets, applying the same automated validation to confirm both that the sentence contains slang and that the slang term belongs to our curated list. Importantly, all sentences are drawn from naturally occurring human-produced text no synthetic data was used in constructing the dataset.

The dataset comprises 8,621 labeled instances. The vast majority of literal counterparts were sourced from the same corpora as the slang samples for each language, maintaining stylistic consistency within each language.

Validation of Automated Slang/Literal Labels

Portions of the dataset arrive with pre-existing labels, while the remainder were labeled by GPT-4o-mini. To validate the latter, we manually annotated a balanced sample per language and evaluated the automated labels against it. In English ($n=50$; 25 slang / 25 literal) and Russian ($n=50$; 25 slang / 25 literal), GPT-4o-mini achieves precision 0.96 and recall 1.00 on the slang class, with a single literal sentence misclassified as slang and all slang sentences classified correctly. In Hebrew ($n=100$; 50 slang / 50 literal), it reaches precision 0.92 and recall 0.98. These results indicate that the slang/literal distinction is robust and reliably iden-

tifiable by the model. Moreover, our contrastive setup—which probes the activation delta between slang and literal senses of the same word—targets a dominant, recurrent latent direction and is therefore stable under minor label noise, as the aggregate signal remains anchored by the consistent majority of correctly identified pragmatic contrasts.

Lexical Diversity. To assess whether slang instances are dominated by a small subset of target terms, we compute the proportion of slang instances accounted for by the top-5 most frequent target words per language. The top-5 types account for 22.6% of English slang instances (219/968, out of 130 target terms), 58% of Russian slang instances (312/538, 15 terms), and 78.5% of Hebrew slang instances (1,400/1,776, 18 terms), reflecting the larger and more varied English target-term inventory relative to Hebrew and Russian. Top ranked target-terms for all three languages are provided in Table 4.

	English	Hebrew	Russian
1	off (5.7%)	שם (26.6%)	бомба (12.6%)
2	man (5.7%)	מח (19%)	жестъ (11.9%)
3	hit (4.6%)	דפוק (14.1%)	пушка (11.9%)
4	check (3.7%)	רצח (11.1%)	треш (11.0%)
5	bad (2.9%)	פצצה (07.7%)	огонь (10.6%)

Table 4: Top-5 most frequent slang target terms per language, with each term’s share of that language’s total slang instances.

B Top Discriminative SAE Features by Language

This section lists the cross-linguistic and per-language discriminative features identified in Section 5.2. Cross-linguistic features appear in the top-100 discriminative set for all three source languages; per-language features are each language’s top-10 ranked by Δ_i (slang minus literal activation rate). The two layers share no cross-linguistic features in common.

Rank	Feature	Avg Δ_i
1	35440	+33.6%
2	93521	+31.6%
3	97003	+21.2%
4	33236	+19.9%
5	88782	+14.8%
6	92777	+11.0%
7	20069	+10.6%
8	71479	+10.1%
9	24006	+9.6%
10	46374	+9.2%

Table 5: Cross-linguistic features (Layer 20), ranked by average Δ_i .

Rank	Feature	Slang %	Literal %	Δ_i
<i>English</i>				
1	35440	59.9	7.8	+52.1
2	33236	45.9	2.1	+43.7
3	93521	40.7	1.0	+39.7
4	7697	36.0	17.3	+18.7
5	108864	25.8	7.2	+18.6
6	21876	20.1	2.4	+17.7
7	28165	40.4	23.1	+17.3
8	40141	42.0	25.0	+17.0
9	88782	17.3	0.3	+17.0
10	103326	18.6	2.4	+16.3
<i>Hebrew</i>				
1	115163	51.7	29.6	+22.1
2	35440	24.0	2.8	+21.2
3	74327	71.1	56.6	+14.5
4	93521	13.4	0.3	+13.0
5	97003	16.7	3.9	+12.8
6	30218	45.6	33.2	+12.4
7	7163	13.1	0.7	+12.4
8	8408	17.0	5.7	+11.3
9	104622	23.5	13.0	+10.5
10	86719	48.0	38.0	+10.0
<i>Russian</i>				
1	93521	43.0	1.0	+42.0
2	97003	39.8	1.7	+38.1
3	115163	46.8	8.9	+37.9
4	45410	37.3	7.8	+29.5
5	35440	28.8	1.3	+27.5
6	83575	23.8	2.4	+21.4
7	92777	21.3	0.3	+21.0
8	88782	21.5	0.6	+20.9
9	24006	27.5	8.4	+19.1
10	3598	21.1	2.2	+19.0

Table 6: Top 10 discriminative features per language (Layer 20).

Rank	Feature	Avg Δ_i
1	38266	+32.5%
2	58884	+32.1%
3	13485	+22.5%
4	22467	+22.4%
5	47673	+18.3%
6	106322	+17.5%
7	31252	+15.2%
8	105792	+15.0%
9	88558	+14.0%

Table 7: Cross-linguistic features (Layer 9), ranked by average Δ_i .

Rank	Feature	Slang %	Literal %	Δ_i
<i>English</i>				
1	58884	38.7	2.0	+36.7
2	31547	39.9	4.8	+35.1
3	13485	38.8	3.7	+35.1
4	38266	38.1	3.7	+34.3
5	82164	38.6	4.3	+34.3
6	65823	31.0	1.1	+29.9
7	99647	33.1	3.9	+29.2
8	22467	31.6	4.0	+27.6
9	127209	53.8	26.8	+27.1
10	124105	27.5	0.6	+26.9
<i>Hebrew</i>				
1	38266	29.6	7.9	+21.7
2	119607	76.1	54.7	+21.4
3	22467	22.8	2.8	+20.1
4	64480	69.0	49.9	+19.1
5	58884	29.2	10.6	+18.5
6	8819	41.0	23.2	+17.8
7	13273	22.0	5.1	+16.9
8	86686	31.6	15.1	+16.5
9	25081	16.9	0.4	+16.5
10	39125	49.1	33.0	+16.0
<i>Russian</i>				
1	101504	57.8	15.2	+42.6
2	38266	49.1	7.7	+41.5
3	58884	52.8	11.7	+41.1
4	1855	49.8	10.4	+39.4
5	120652	36.3	1.1	+35.1
6	106322	35.6	2.7	+32.9
7	129047	36.4	5.0	+31.3
8	47673	36.8	5.8	+30.9
9	60050	41.8	12.9	+28.8
10	12375	39.3	12.4	+27.0

Table 8: Top 10 discriminative features per language (Layer 9).

C Vocabulary Projection: Top Promoted Tokens per Cross-Linguistic Feature

Table 9 lists the top promoted vocabulary tokens for the two highest- Δ_i cross-linguistic features (35440 and 93521), as well as for their averaged vector. Following Gur-Arieh et al. (2025), we pass each SAE decoder weight vector through the model’s final layer norm and project it through the

unembedding matrix; tokens are ranked by projection weight. The recurrence of informal expressions such as *dudes*, *vibes*, *swag*, and *funky* across the individual features and their average provides qualitative lexical support for interpreting these directions as associated with informal register.

Rank	Feat. 35440	Feat. 93521	Avg (35440+93521)
1	dudes	emojis	dudes
2	guys	vibes	vibes
3	funky	swag	badass
4	fellas	kaarangay	funky
5	dude	coolest	guys
6	homie	dope	vibe
7	rappers	vibe	dude
8	badass	awesome	swag
9	streetwear	swagger	swagger
10	dope	vibing	coolest

Table 9: Top 10 vocabulary tokens promoted by the two highest- Δ_i cross-linguistic features (35440, 93521) and their averaged vector. All promoted tokens reflect informal register semantics consistently across individual features and their combination.

This English-dominant pattern is not confined to the cross-linguistic core. Feature 115163, which is discriminative in Hebrew ($\Delta_i = +22.1\%$) and Russian ($\Delta_i = +37.9\%$) but carries essentially no slang–literal signal in English ($\Delta_i = -0.0\%$), promotes *stuff*, *stupid*, *crazy*, and *gonna*. The Russian-specific Feature 3598 ($\Delta_i = +19.0\%$ in Russian; outside the top-100 for both English and Hebrew) likewise promotes *amazing*, *fantastic*, *good*, and *VERY* among its highest-ranked interpretable tokens. That features carrying no English discriminative signal still read out through English vocabulary points to a systematic English bias in the model’s unembedding matrix, rather than to any limitation of the cross-linguistic features.

D GPT-4o Formality Evaluation Prompt

We use GPT-4o-mini as an automated formality judge to evaluate steered model outputs. Texts are scored in batches of 20 with temperature 0 for reproducibility. The system and user prompts are as follows:

System Prompt.

You are a formality scorer. Return only a JSON array of numbers between 0 and 1.

User Prompt.

Rate the formality of each text below on a scale from 0 to 1.

- 0 = very informal, casual, slang, colloquial
- 1 = very formal, professional, academic

Return ONLY a JSON array of numbers, nothing else. Example: [0.2, 0.8, 0.5]

Texts:

1. "<text_1>"
2. "<text_2>"
- ...

Scores are clipped to [0, 1]. Failed batches are retried up to 3 times. To validate the automated scores, 30 completions per language (120 total) were independently scored by human annotators, yielding a Pearson correlation of $r = 0.62$ between GPT-4o-mini and human judgments.

D.1 Per-Language Human–LLM Agreement

Per-language Pearson r between GPT-4o-mini and human ratings is 0.768 (English), 0.803 (German), 0.474 (Hebrew), and 0.368 (Russian), with $n=30$ per language and a pooled r of 0.626. Higher-resource Latin-script languages show strong agreement, while Hebrew and Russian show moderate agreement, broadly in line with published human–human inter-annotator agreement on formality (≈ 0.5 – 0.7 ; Pavlick and Tetreault, 2016). Mean scores are well-calibrated across all languages ($|\text{GPT mean} - \text{human mean}| \leq 0.10$), indicating that the lower correlations reflect unbiased per-item noise rather than systematic bias. Human inter-annotator agreement on the same items is $r=0.814$ (English) and $r=0.615$ (Hebrew), placing GPT-4o-mini’s per-language agreement within the range of human–human variability. Judge noise is therefore non-trivial but does not threaten the rank-level trends our steering analysis relies on.

D.2 Prompt Sensitivity of the Formality Judge

To assess sensitivity to the judge prompt, we re-scored the 120 human-rated completions under two rephrasings (V2: “evaluate level of formality”; V3: “score the register, 0 = colloquial / 1 = formal written”). Inter-prompt agreement is high (V1–V2 $r=0.771$, V1–V3 $r=0.773$, V2–V3 $r=0.944$; all $p < 10^{-24}$); per-language correlations span 0.69–0.99 across all 12 prompt×language cells, and 89% of texts vary by less than 0.20 on the 0–1 scale across prompts. Prompt reformulation thus introduces minor calibration shifts but reliably preserves the underlying linear relationships. These results indicate that the

steering evaluation is relatively insensitive to modest reformulations of the judge prompt.

E Steering Evaluation Prompts

Table 10 lists the English sentence prompts used for activation steering evaluation. Each prompt is an incomplete sentence that the model completes under varying steering coefficients. Prompts include both neutral stems (1–15) and informally-primed stems (16–20) to test steering effectiveness across starting registers. Hebrew, Russian, and German were translated by native speakers, and the remaining five languages by GPT-4o; each language has 20 prompts. Full lists are available in the supplementary code.

#	Prompt
1	I think that
2	She believes the new restaurant downtown
3	In my opinion, the movie was
4	They feel like the music
5	The exam results were
6	The presentation was
7	The game last night was
8	The outfit she wore looked
9	The book I finished was
10	The car he bought seems
11	I feel that the concert was
12	The coffee from that shop tastes
13	The new phone features are
14	The vacation we took was
15	He thought his explanation was
16	Yo, that party was
17	Bro, check out this
18	That beat is so
19	Dude, the game last night was
20	Man, this food is

Table 10: English steering prompts. Prompts 1–15 are register-neutral; 16–20 are informally primed. Hebrew, Russian, and German prompts follow the same design.

F Random Feature Steering Ablation

To rule out the possibility that arbitrary directions in the SAE latent space produce comparable formality shifts—or that steering-induced output degradation is misclassified as informality by the automated judge—we conducted a random feature ablation. We sampled 5 independent sets of 10 random features from the full 131,072-feature dictionary (matching the size of the cross-linguistic feature set), constructed normalized steering vectors by averaging their decoder weights (identical to Eq. 2), and applied them to English and German (zero-shot) at $\alpha \in \{-150, 0, 150\}$. Note that the main experiments use $\alpha \in [-150, 100]$ as the

coherence-safe operating range (Appendix I); we deliberately extend to $\alpha = 150$ here to stress-test both the random vectors and the automated judge under conditions at the boundary of coherent generation. Each condition produced 20 completions, scored by GPT-4o-mini using the same protocol described in Appendix D.

As shown in Table 11, cross-linguistic universal features produce strong negative correlations between α and formality (English: $r = -0.791$, $p < 10^{-13}$; German: $r = -0.743$, $p < 10^{-10}$), while random feature vectors yield near-zero correlations (English: mean $r = 0.006 \pm 0.133$; German: mean $r = -0.060 \pm 0.077$). One-sample t -tests comparing $|r_{\text{random}}|$ against $|r_{\text{universal}}|$ confirm the difference is highly significant (English: $t = -35.58$, $p < 10^{-5}$; German: $t = -26.41$, $p < 10^{-4}$).

Crucially, this ablation serves a dual purpose. First, it establishes that the formality modulation produced by cross-linguistic features is specific to the identified feature subspace rather than a generic property of high-magnitude residual-stream perturbations: mean formality under random steering remains flat across the full range (English: $0.434 \rightarrow 0.419 \rightarrow 0.435$; German: $0.470 \rightarrow 0.488 \rightarrow 0.439$), whereas the universal vector drives it from 0.615 to 0.165 in English and from 0.660 to 0.205 in German. Second, it validates the automated judge: since random steering at $\alpha = \pm 150$ —including beyond the main operating range—produces coherent completions that receive baseline formality scores (~ 0.4 – 0.5), the low formality scores assigned to universal-feature-steered outputs reflect genuine register shifts rather than an artifact of the judge equating degraded text with informality.

Language	Vector Type	Pearson r	p -value
English	Universal	-0.791	5.6×10^{-14}
English	Random (mean $\pm$ std)	0.006 ± 0.133	—
German	Universal	-0.743	1.1×10^{-11}
German	Random (mean $\pm$ std)	-0.060 ± 0.077	—

Table 11: Pearson correlations between steering coefficient α and formality score for universal vs. random feature vectors (5 independent draws of 10 random features each). Coefficients tested: $\alpha \in \{-150, 0, 150\}$.

G Language Preservation Under Steering

To verify that activation steering modulates formality rather than switching the output language,

we applied automatic language identification to all steered completions across nine languages and six coefficients ($\alpha \in \{-150, -100, -50, 0, 50, 100\}$). Overall, approximately 90% of completions remained in the target language. Preservation was highest at negative (formal-direction) coefficients and decreased at positive coefficients, consistent with the model’s English-dominant informal register exerting a pull at high α . The observed language switches were directed predominantly toward English, with only isolated switches to other non-target languages. Hebrew, Russian, and Japanese all exceeded 95% preservation; the remaining languages stayed above 80%. These results confirm that the formality shifts reported in Section 6 reflect genuine within-language register changes. We note that approximately 90% target-language retention is itself a robust result given the difficulty of the setting: steering is applied zero-shot to typologically distant, low-resource languages (e.g., Amharic, Georgian) never used to derive the steering features. Moreover, the $\sim 10\%$ of switches are not random but overwhelmingly default to English, consistent with the English bias in the unembedding space noted in §5.4—rather than evidence that the cross-linguistic register features themselves are unstable.

H Additional Steering Information

Table 12 reports Layer 9 cross-linguistic feature steering correlations, mirroring the Layer 20 results in the main text, and Figure 4 shows the corresponding mean formality curves by steering coefficient. Figure 5 shows the mean formality scores for the three source languages at Layer 20.

Language	r	p
English	-0.623	$< 10^{-14}$
Hebrew	-0.637	$< 10^{-14}$
Russian	-0.631	$< 10^{-14}$

Table 12: Layer 9: Pearson correlation between steering coefficient and formality score for source languages.

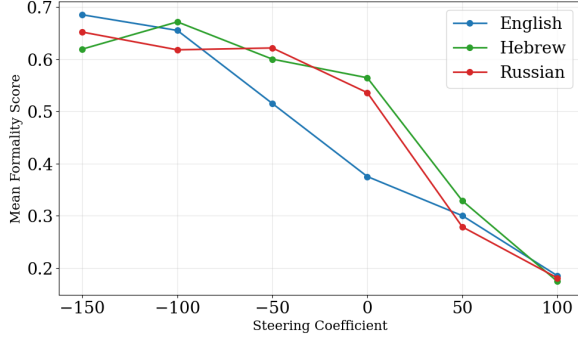


Figure 4: Layer 9: Mean formality score by steering coefficient for each source language.

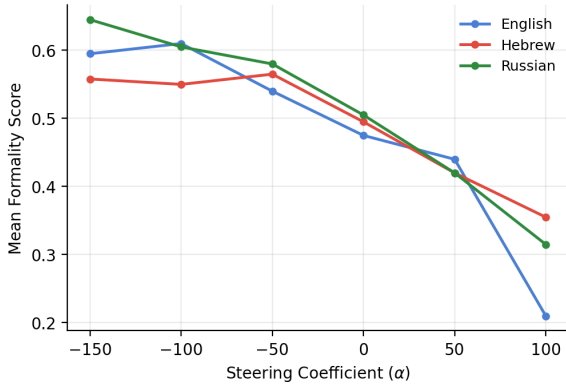


Figure 5: Mean formality score as a function of steering coefficient α for the three source languages (English, Hebrew, Russian) using universal cross-linguistic features at Layer 20.

I Additional Perplexity Analysis

To independently validate that steered outputs remain coherent within our chosen operating range (Section 6.2, Figure 3), we additionally compute perplexity using external reference models: Qwen2.5-7B-Instruct (Team, 2024) for eight languages and DictaLM 2.0 Instruct (Shmidman et al., 2024) for Hebrew, due to limited Hebrew tokenizer coverage in Qwen2.5. Perplexity is computed on the completion tokens only, conditioned on the prompt; completions with perplexity above 10,000 (degenerate or empty generations) are excluded from the reported medians. As shown in Figure 6, median perplexity under these external models remains stable across the full $[-150, 100]$ range for all nine languages, with values staying below 25 for most languages and within $2\times$ of baseline for $|\alpha| \leq 50$. This cross-model consistency confirms that the formality shifts reported in Section 6 reflect genuine register changes rather than degenerate outputs.

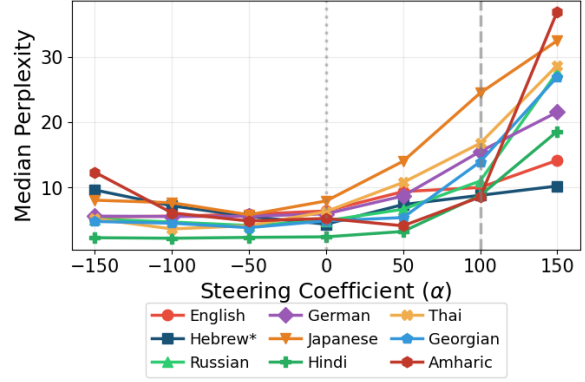


Figure 6: Median perplexity by steering coefficient α for all nine languages evaluated with independent reference models (Qwen2.5-7B-Instruct for eight languages; DictaLM 2.0 for Hebrew). The stable perplexity within $[-150, 100]$ independently corroborates the coherence window identified via Gemma self-perplexity (Figure 3).

J Bilingual-Exclusive Feature Steering

As a stricter test of cross-linguistic abstraction, we evaluate *bilingual-exclusive* feature sets. For each target language, we first identify features that appear in the top-20 Δ_i lists of the other two languages, and then remove any feature that also appears in the target language’s own top-20 list. Thus, when steering English, we use only features shared by Hebrew and Russian but absent from English’s top-20 set; analogous constructions are used for Hebrew and Russian.

This design makes the transfer test more demanding than the universal-feature setting: the steering features are not only derived without using the target language as a source, but are explicitly constrained to exclude the target language’s own most salient slang-associated features. If such feature sets still support causal control in the held-out language, this strengthens the claim that the model contains partially language-abstract mechanisms for informal-register processing, rather than relying solely on language-specific latent features.

Despite this exclusion, bilingual-exclusive features still yield strong negative correlations between steering coefficient α and output formality in the held-out language. At Layer 9, the resulting correlations are $r = -0.701$ for English (steered by Hebrew+Russian features), $r = -0.675$ for Hebrew (English+Russian), and $r = -0.628$ for Russian (Hebrew+English). At Layer 20, Hebrew+Russian features achieve $r = -0.724$ for English, English+Russian features achieve

$r = -0.768$ for Hebrew, and Hebrew+English features achieve $r = -0.770$ for Russian, all statistically significant at $p < 10^{-19}$. These results provide an additional robustness check on the main transfer findings, showing that causal register control can arise even from feature sets that are shared across languages but exclude the target language itself—suggesting that informal register signal extends well beyond the strict trilingual core.