TRACING LOW-LEVEL LINGUISTIC PROPERTIES THROUGH SPARSE REPRESENTATION TO BEHAVIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models (LMs) are built on human languages; therefore, the structural properties of languages are deeply embedded in a model's representations, affecting output-level behaviors. The relationship between these effects (as well as their nature) and the linguistic information remains largely unclear. We study this relationship through a combination of approaches across 40 low-level linguistic variables represented by 20,000 controlled contrast pairs across eight languages. In XGLM-564M, we compare supervised linear probes with features derived from a BatchTopK sparse autoencoder trained independently on multilingual natural text. Through evaluating these features across a set of metrics, performing ablation, and steering, which is then followed by next-token-level behavior evaluation, we find a set of six variables that show significant and consistent results across most tests while major subsets survive across different assessments. In this study, we aim to highlight the possibilities behind the broader intersection of computational linguistics and mechanistic interpretability, while showcasing how not just the features themselves, but the methods and approaches we use also produce significantly overlapping evidence and localization of representations.

1 INTRODUCTION

Language models encode substantial and extensive linguistic information, which is not enough to decode, as how the information is internally represented and whether it contributes to downstream computation requires further and detailed analysis. While supervised probes can recover morphological, syntactic, and other grammatical properties, their performance may rely on classifier capacity or correlations in the evaluation data (Conneau et al., 2018; Hewitt & Manning, 2019; Hewitt & Liang, 2019). Sparse autoencoders (SAEs) aim to decompose dense activations into a sparse learned dictionary of features as a complementary approach to supervised probes (Cunningham et al., 2023; Gao et al., 2024). However, as shown numerous times in this study, SAE features may cover a significantly narrower definition across lexical, morphological, language-specific, or tokenization patterns instead of the intended broad linguistic abstractions.

We examine these distinctions across 40 low-level linguistic variables covering morphology, grammatical categories, syntax, discourse, information structure, and orthography. We construct 20,000 controlled contrast pairs across the 40 variables and eight languages. We compare supervised probes with learned SAE features in XGLM-564M, then assess the features through semantic inspection, ablation and steering, and local next-token behavioral evaluation.

Supervised probes recover 26 out of 40 variables, while SAEs show robust results for 29 and 32 when the preservation of the overall held-out direction is considered. Across the two techniques, 23 variables are strongly recovered by both, indicating a heavy overlap between them. Ablation and steering identify 15 variables to contribute significantly to internal representations causally; however, only a subset of six produced false-discovery-rate-corrected and control-specific effects on local next-token preferences.

Our main contribution is a unified chain of methodologies connecting supervised probing, sparse autoencoders, semantic inspection, causal intervention, and behavioral evaluation, showcasing that the results yielded by these methods are only partially correlated. These findings help connect computational linguistics with mechanistic interpretability to separate linguistic information from model computation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 Related Work

2.1 Linguistic representations and probing in language models

A high volume of work establishes the recoverability of linguistic signals/information in neural language model (LM) representations utilizing classifiers and structural probes, including morphology, orthography, typography, syntax, and other grammatical properties (Conneau et al., 2018; Hewitt & Manning, 2019; Tenney et al., 2019). Expanding on this work, layer-based analyses suggest that different linguistic properties are recoverable/separable at different depths of the neural networks. Suggesting highly sparse and convoluted representations and behavioral effects across model layers (Tenney et al., 2019). However, probes have strict limitations in terms of identifying how information is represented internally and whether they contribute to behavioral effects. This limitation might depend on probe capacity, memorization, and correlations in the evaluation data (Hewitt & Liang, 2019; Belinkov, 2022). This motivates a shift from representational accessibility to methods that test the internal organization and computational role of linguistic information/structures.

2.2 Sparse autoencoders

Sparse autoencoders aim to decompose dense model activations into larger dictionaries, pushing individual neurons to represent individual latent features that activate sparsely, offering a means to transition from convoluted and distributed representations to more interpretable internal activations (Cunningham et al., 2023; Gao et al., 2024). Some variants of SAEs, such as BatchTopK, modify the sparsity constraint to improve the use of the dictionary without loss of sparse activations (Bussmann et al., 2024).

Exploring linguistic structures through SAEs has begun recently. Brinkmann et al. identify how sparse representations relate to grammatical concepts, such as number, gender, and tense across languages of wide topological diversity, and test them through causal interventions exploring how the manifestation of these concepts differs across languages (Brinkmann et al., 2025). Jing et al. utilize controlled linguistic contrasts to similarly discover SAE features and manipulate a subset of features to elicit behavioral effects including morphology, pragmatics, syntax, and semantics (Jing et al., 2025). The studies indicate that sparse feature dictionaries could be useful for extracting linguistic information. Statistical associations of latents with particular linguistic abstraction do not necessarily imply direct translation, as a feature may track a narrower definition across morphological forms, lexical correlates, punctuation patterns, or other surface realizations. That means more deep and structured research is needed to map exact mechanistic interpretability techniques and their precise interpretation in investigating linguistic features, including SAEs.

2.3 Causal interventions

The importance of causal methods has been discussed in previously mentioned literature, as it provides a stronger validation of whether linguistic representations contribute to model computation. To study mechanisms of syntactic agreement, causal mediation has been utilized (Finlayson et al., 2021), while naturalistic causal probing allows manipulation of morphosyntactic attributes such as grammatical number and gender (Amini et al., 2023). More recent work focuses on the combination of evaluation of linguistic tasks and internal interventions: CausalGym benchmarks causal interpretability methods across linguistics tasks (Arora et al., 2024), and Boguraev et al. use causal interpretability interventions on LMs to enhance linguistic capabilities through English filler-gap dependency constructions (Boguraev et al., 2025).

The aforementioned approaches establish that linguistic representations are causally manipulable, but the internal effects and output-level consequences are not always coherent. Related work also suggests that high-level behavior can be influenced by linguistic features: fine-tuning on data composed of systematically varying linguistic characteristics shifts downstream animal-welfare-related reasoning (Brazilek & Dunn, 2026). The present study combines the mentioned approaches into a single framework by investigating controlled linguistic contrasts through supervised-sparse feature recoverability and identification, natural-text interpretation, steering and ablation, and output-level white-box behavioral evaluation.

3 EXPERIMENTAL FRAMEWORK

3.1 CONTROLLED LINGUISTIC CONTRASTS

We begin by building a controlled features set of 40 low-level linguistic variables, based on the typological organization of the World Atlas of Language Structures (Dryer & Haspelmath, 2013) and adapted to the mechanistic interpretability experiments, while maintaining typological diversity and coverage. The final set was based on typologically motivated differences in morphology, discourse, orthography, information structures, grammatical categories, and syntax. The variables, among others, include voice, negation, modality, social deixis, reported speech, person marking, definiteness, agreement, causativity, case marking, mophosyntatic alignment, and information structure.

Along each variable, 500 contrast pairs, $(x_i^{\text{basis}}, x_i^{\text{changed}})$, where the *changed* sentence is modified to accommodate the properties of the target linguistic variable relative to the basis sentence while preserving the semantic content accurately. This yields a dataset of 20,000 contrast pairs, or, in total, 40,000 sentences. The dataset was split between training data, validation data, and test data, with a split of 1–400, 401–450, and 451–500, respectively.

The constructions are based on four language strategies filtered from a multilingual pool of English, Turkish, Japanese, Korean, Spanish, Russian, Arabic, and Mandarin Chinese to accommodate all features, as a number of variables cannot be adequately expressed by English. Variables marked as `EN` can be realized entirely by English contrasts. Variables marked as `EN+ctrl` require more tightly controlled templates, or often, lexical substitutions. `XL` variables are based on a selected non-English language when the corresponding linguistic distinction is missing or poorly defined in English, and `EN+XL` variables are English-accessible components that are more exhaustively realized in another language. Each cross-lingual variable utilizes at most one selected non-English language to reduce confounding caused by simultaneously varying language familiarity, corpus distribution, morphology, script, and model performance.

In addition to the train–validation–test split, the dataset is also characterized along `marker_family` and `lexical_domain`. These distinctions allow evaluation on data that differs across constructions and lexical contexts from those used to identify the sparse representations. Where applicable, a construction family might be withheld entirely from the training data. These controls are designed so representations that generalize across realizations of a linguistic contrast can be distinguished from those that predominantly memorize recurring words/templates - thus preventing overfitting.

The sparse autoencoder itself is trained on a separate multilingual natural-text corpus, while the controlled features dataset is reserved for supervised probing, identification of SAE features, and causal evaluation.

Table 1: Representative controlled linguistic contrast pairs from the feature dataset.

Variable	Basis	Changed
Voice / agent prominence	The intake clerk moved the appointment slot during the afternoon call.	The appointment slot was moved by the intake clerk during the afternoon call.
Evidentiality	I saw that the nurse updated the triage board before lunch.	The handover note says that the nurse updated the triage board before lunch.
Modality / mood	Nurse Patel is able to sterilize the dressing tray during the intake review.	Nurse Patel is required to sterilize the dressing tray during the intake review.
Quotation / reported speech	Dr. Chen reported, "The pharmacy team labeled the vaccine tray before noon."	Dr. Chen reported that the pharmacy team labeled the vaccine tray before noon.

3.2 MODEL AND ACTIVATION EXTRACTION

All final experiments are performed using `facebook/xglm-564M`, a multilingual autoregressive Transformer from the XGLM family (Lin et al., 2021). The model's residual-stream has a dimensionality of 1,024 and is assessed on the native tokenizer. Model parameters are frozen across all activation extractions, sparse-feature analyses, probing, and intervention experiments.

A standard forward pass is performed for each controlled sentence while collecting all hidden states. To identify the most relevant layers, layer-wise activations are retained for the supervised probing analysis testing, where each linguistic feature becomes recoverable. The two sequence-level representations considered are the residual activation at the final non-padding token and the mean residual activation over non-padding tokens.

The SAE and subsequent causal interventions are carried out at hidden-state index 12, which corresponds to the residual-stream after decoder layer 11. Separately from the layer-wise probe scan, this layer is fixed for the sparse-analysis stage. Causal effects are measured downstream at the final hidden layer so that interventions at the SAE site can be evaluated in accordance with their alterations of later representation of the same linguistic contrast. Mean-pooled activations provide the primary representation for the identification of SAE features, while final-token analyses are kept as supplementary information.

3.3 Supervised probe evaluation

Using supervised linear probes, we first evaluate the recoverability of the linguistic features directly from the model's residual activations. This analysis provides independent representational evidence, which can later be used as a comparative analysis baseline for the unsupervised SAE evidence.

Probes are trained for each variable and layer to distinguish between basis and changed sentences. Additionally, we analyze the representation of differences between pairs $\Delta h_i = h(x_i^{\text{changed}}) - h(x_i^{\text{basis}})$, which records the transformation-induced displacement of activations. Consistency across pair-difference directions indicates that different samples directly represent the linguistic features' characteristics through shared direction in the representation space, outperforming arbitrary classifiers.

Probe results are assessed along a spectrum of dimensions. Activation recoverability shows whether basis and changed sentences are clearly separable from model activations. To measure how activation-based performance compares to surface-accessible baselines, we use text and artifact controls. To avoid overfitting, we use split generalization on held-out examples, including construction and lexical variation. To complement these measures, learned activation-direction viability is retained to accurately see whether a training activation estimated direction performs well on unseen pairs.

The supervised probe evaluation is repeated across three random seeds: 1, 2, and 42. The main cross-seed probe classification in later comparisons requires the following criteria to pass in all three seeds: activation-recovery, directional-consistency, and split-generalization.

3.4 Sparse autoencoder

We train a sparse autoencoder on hidden-state 12 activations from an independent multilingual natural-text corpus as complementary evidence to the supervised analysis, following recent work on sparse dictionaries as a tool to decompose language model activations (Cunningham et al., 2023; Gao et al., 2024). The relationships between the learned latents and the forty features are identified after training, as the SAE does not receive the controlled linguistic labels, thus allowing the model to not be forced to learn the variables.

The canonical SAE uses a $16\times$ BatchTopK architecture (Bussmann et al., 2024) with input dimensionality of 1,024 and a 16,384 dictionary of latent features. Training runs for four epochs using an average sparsity target of $k = 256$ active features per example. The architecture was chosen by analyzing both reconstruction performance and dictionary utilization.

The chosen model, which consists of four epochs and the BatchTopK architecture, achieves approximately 0.99863 validation explained variance and activates 16,372 out of 16,384 (99.9%) of its latent features across its entire training corpus at least once. The eight-epoch run slightly improved validation explained variance to approximately 0.99894; however, it substantially reduced the number of active features down to 15,764. Due to the significant trade-off from dictionary coverage against the small improvement in reconstruction, the four-epoch method was retained.

Native BatchTopK preserves only the strongest feature activations across the entire batch; therefore, whether `j` is kept as active for sentence `x` depends partly on other batch content. To enhance reproducibility and interpretability of downstream results, we use a fixed-threshold inference so that

regardless of unrelated batch content, a sentence always receives the same SAE representation. To preserve the learned encoder and decoder while making the activation independent of other batch compositions, the threshold is calibrated directly on the SAE's training activation distribution. All downstream natural-text inspection, feature identification, steering, and ablation experiments rely on this fixed inference structure.

3.5 SAE FEATURE IDENTIFICATION AND SEMANTIC INSPECTION

For each linguistic variable, the controlled basis and changed sentences are passed through the canonical SAE, and the feature-level pair difference is determined by $\Delta a_{if} = a_f(x_i^{\text{changed}}) - a_f(x_i^{\text{basis}})$, where a_f represents the activation of the f SAE feature.

Features are ranked along the training split by the consistency and magnitude of the differences between basis and changed activations - from which the candidate features are selected. Their significance is evaluated by paired sign-flip permutation tests and max statistic correction across the entire dictionary. Held-out test data are used to assess directional preservation and robustness, which allows us to classify SAE results along five tiers based on their performance: A, B1, B2, C and D. Only A and B1 show strong robustness, with A also denoting cross-variable specificity, while B2 and C preserve only held-out direction, and D fails to preserve it.

Statistical recovery is evaluated independently from semantic interpretation, where final candidates receive separate grades of A–D, from comparatively strong target alignment to weak or proxy interpretations. However, this grade does not filter for causal testing.

3.6 CAUSAL INTERVENTIONS

To test causal participation at hidden-state index 12, we follow the general causal-intervention tradition in mechanistic and linguistic model analysis of ablation and steering (Finlayson et al., 2021; Geiger et al., 2024; Arora et al., 2024; Brinkmann et al., 2025). Both are evaluated at the final hidden layer.

For ablation, the selected feature's decoded contribution is removed from the feature-high condition, the result of which is then measured along a linguistic direction computed from the training split. Each selected feature is compared with three selected SAE control features that have been chosen along activation magnitude, decoder norm, and firing rate.

For steering, the selected decoder direction is added to the feature-low condition along three different doses, $0.5\times$, $1\times$ (being the primary state), and $2\times$. Control vectors are scaled to match the target perturbation magnitude. Let Δ_v^{target} be the target intervention effect for variable v and $\Delta_v^{\text{control}}$ the mean effect for its 3 matched controls. We say that an intervention is *specific-positive* if

$$\Delta_v^{\text{target}} > 0, \qquad \Delta_v^{\text{target}} - \Delta_v^{\text{control}} > 0, \qquad q_v < 0.05,$$

where q_v is the Benjamini–Hochberg-adjusted value for the target-minus-control test. A variable can be chosen to be a part of the pre-behavioral analysis cohort when it satisfies a specific-positive ablation, a specific-positive steering as well as a monotonic steering dose response.

3.7 BEHAVIORAL EVALUATION

Behavioral evaluation is used to test whether the aforementioned interventions contribute to altering next-token preferences. For each held-out basis-changed pair, we look at the first divergent token with the shared prefix as a context provider. A behavioral effect is defined as the next-token logit margin's change between the feature-low and feature-high cases after steering with positive values indicating an increment in preference towards the feature-high case.

The mechanistic cohort is used for the primary analysis, while the all-40-variable analysis remains exploratory. Each variable is tested on 50 held-out test pairs with a $1\times$ dose as primary. The target effects are compared with the same matched controls; Benjamini–Hochberg correction (Benjamini & Hochberg, 1995) is applied to the mechanistic cohort.

4 RESULTS

4.1 LINGUISTIC INFORMATION IS BROADLY RECOVERABLE

Supervised probing discovered strong evidence for 26 out of the 40 linguistic features across all three random seeds. While sparse recovery was similarly - and slightly more - broad with 29 variables representing the strong SAE evidence tiers of A, B1 and B2, 32 preserving the overall held-out activation direction with the inclusion of tier C. The overlap is significant between the two methods, with twenty-three recovering strongly by both methods, six only through SAE analysis, three only by supervised probing, and eight by neither. This corresponds to an overlap of 0.775, a Jaccard similarity of 0.719, Cohen's $\kappa \approx 0.48$, and a significant relationship by Fisher's exact test ($p \approx 0.007$). Conclusively, the two methods detect significantly overlapping results of linguistic structures with sparse methods representing a broader set (Fig. 1).

4.2 SPARSE RECOVERY EXCEEDS SEMANTIC INTERPRETABILITY

Cleanly interpretable linguistic features were not yielded consistently by statistical SAE recovery. Semantic inspection assigned grades A, B, C, and D to 7, 13, 12, and 8 variables, respectively, while numerous statistically robust candidates tracked narrower definitions than the originally intended broad linguistic feature, including punctuation, typological distinctions, tokenization patterns, particular morphological forms, and language identity (Fig. 1). Robust samples from subsequent causal experiments did not coincide exhaustively with inspection grades.

Evidence summary across forty linguistic variables

Variables 1–20

ID	Variable	Probe	SAE	Insp.
01	Subject expression / pro-drop	R	D	C
02	basic_constituent_order	N	C	C
03	nominal_modifier_order	N	D	C
04	case_marking	R	A	B
05	morphosyntactic_alignment	R	B1	D
06	transitivity_valency	R	B2	C
07	voice_and_agent_prominence	R	B2	A
08	causativity_and_valency_change	R	B1	D
09	analytic_vs_synthetic_encoding	R	B2	C
10	morphological_segmentation_type	R	A	B
11	agreement_indexing_density	R	B2	C
12	optionality_vs_obligatoriness_of_marking	N	B2	D
13	redundancy_cumulative_exponence	N	A	B
14	definiteness_and_specificity	R	B2	C
15	number_marking	N	D	B
16	gender_noun_class	R	A	A
17	animacy_and_humanness	R	C	D
18	person_marking_and_person_hierarchy	R	B2	B
19	inclusive_exclusive_distinction	R	B2	C
20	pronoun_richness_and_reduction	R	A	A

Variables 21–40

ID	Variable	Probe	SAE	Insp.
21	possession_and_alienability	R	A	C
22	tense_prominence	R	D	C
23	aspect_and_event_structure	N	D	B
24	modality_and_mood	N	C	A
25	epistemic_modality	R	B2	A
26	evidentiality	N	D	B
27	mirativity_stance_and_affect_marking	R	A	C
28	negation_and_polarity_structure	R	B2	D
29	quantifier_scope_and_distributivity	R	B2	B
30	conditional_and_counterfactual_marking	R	A	B
31	subordination_and_embedding	N	A	B
32	quotation_and_reported_speech_structure	R	A	B
33	discourse_relation_marking	N	D	D
34	topic_comment_structure	R	B2	D
35	focus_and_given_new_marking	R	B2	B
36	genericity_and_kind_level_reference	R	B2	D
37	social_deixis_honorifics_status_encoding	N	B2	A
38	speech_act_force_and_request_directness	N	A	C
39	deixis_and_perspective_anchoring	N	B2	A
40	orthographic_and_tokenization_interface	N	D	B

Probe: R = robust, N = non-robust. SAE and inspection columns show categorical grades.

Figure 1: Evidence across all 40 linguistic variables, combining supervised-probe robustness, SAE evidence tier, and natural-text semantic inspection grade.

4.3 Ablation and steering reveal internal causal structure

Ablation and steering served as supplementary tests examining whether the identified sparse directions contribute to downstream representations. For ablation, 31 out of 40 variables showed a positive attenuation of the specific linguistic direction, 19 satisfied the specific-positive criterion, and 25 demonstrated more substantial effects for the target feature than the control variable. Similarly for steering, 31 variables produced positive effects, the target effects exceeded controls for 31, there was a monotonic response from 25 across doses, and a specific-positive effect for 23 features. Across the two analyses, 15 variables satisfied both mechanistic criteria. Throughout the analysis, representation cosine similarity remained substantially high, indicating that the effects were not caused by broad, distributed disruptions but rather localized changes (Fig. 2).

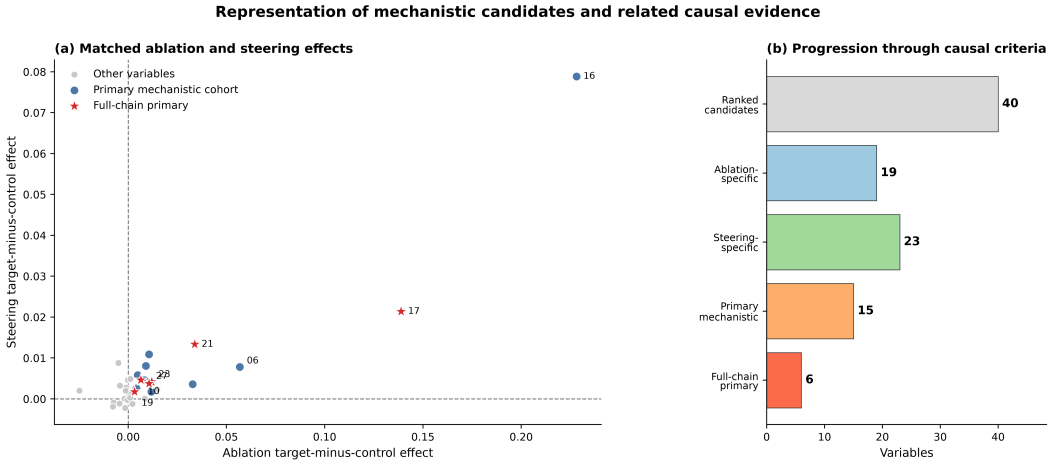

Figure 2: Causal intervention results. The figure summarizes matched ablation and steering effects across variables together with the progression to the 15-variable pre-behavior mechanistic cohort.

4.4 Behavioral influence is selective

Behavioral analysis was only performed on the previously described 15-variable cohort. Nine variables produced positive effects on the specific next-token preference, nine have monotonic responses across intervention doses, and ten produced larger target than control effects. Six satisfied the combined false-discovery-rate, control-specificity, and positive-effect criterion: aspect/event structure, animacy/humanness, morphological segmentation, mirativity/stance/affect, inclusive/exclusive distinction, and possession/alienability with respective grades from semantic inspection: B, D, B, C, C, and C. Conclusively, internal causal effects were more abundant than consistent output-level effects, while semantic interpretation showed little correlation with behavioral consequences (Fig. 3).

Table 2: Full-chain causal features satisfying the pre-behavior mechanistic criteria and the final control-specific, FDR-corrected behavioral criterion with monotonic behavioral dose response.

Variable	Feature	Grade	Δ_{beh}	q
Morphological segmentation type	15843	B	0.005927	0.002142
Animacy and humanness	6618	D	0.006903	0.002142
Inclusive/exclusive distinction	6062	C	0.000734	0.023322
Possession and alienability	15325	C	0.019592	0.002142
Aspect and event structure	6216	B	0.003034	0.009370
Mirativity, stance, and affect marking	1887	C	0.005211	0.002142

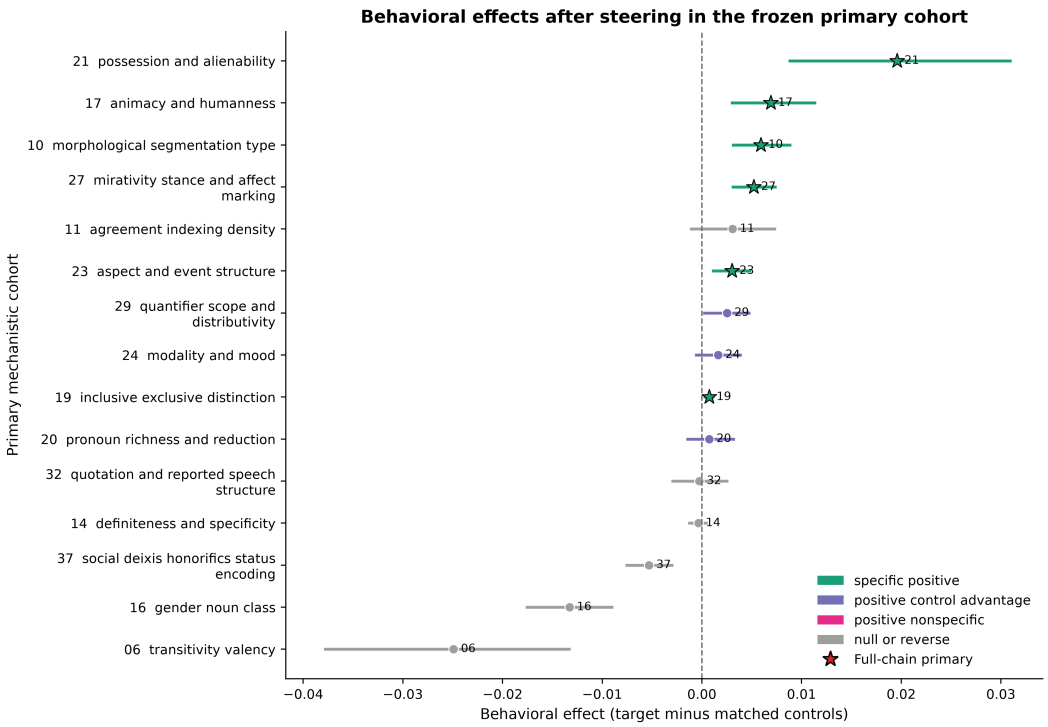

Figure 3: Behavioral target-minus-control effects for the frozen 15-variable mechanistic cohort under the primary steering dose. Stars indicate the six variables satisfying the full-chain behavioral criterion.

5 Discussion

5.1 Linguistic structure is represented partially and heterogeneously

Low-level linguistic features are broadly recoverable across XGLM's residual representations; however, sparse interpretations are oftentimes significantly narrower than the intended variable definitions, including language-specific surface patterns, suffixal or progressive morphology, kinship and possession terms, punctuation, or pronoun identity. Therefore, low-level linguistic structures appear to be represented through heterogeneous mixtures of surface-grounded, construction-specific, language-specific, and abstract features, complementing prior research that studies linguistic structures using SAEs while cautioning future approaches (Brinkmann et al., 2025; Jing et al., 2025).

The distinctions between probe and SAE results further showcase that caution is required when determining a variable to be "represented", as representational properties should be measured alongside. Linear probes examine the accessibility of information through a supervised classifier, while SAE analysis looks for a related direction in an independently learned sparse dictionary, of which neither method describes a feature's internal representation or its downstream effects wholly (Hewitt & Liang, 2019; Belinkov, 2022). Additionally, as we have shown, statistical recoverability and semantic interpretability can diverge significantly and should therefore be treated separately.

5.2 Participation of linguistic representations in computation

The results of the causal interventions indicate that a subset of the sparse linguistic features are not just decodable but also contribute to downstream computation. The 15-variable mechanistic cohort, derived from ablation and steering results, showcases internal mediation along probing and sparse autoencoder analysis. Nevertheless, of this cohort, only six variables satisfied the final behavioral criterion, suggesting that altering downstream representations is not necessarily enough to control broader or local output definitions.

Additionally, human interpretability proved to be unreliable in predicting the effectiveness of causal interventions, with several candidates of proxy-like semantic interpretations showing strong causal effects, and some semantically strong features failing to affect higher-level behavior. Therefore, these factors correlate less than expected and should be examined, treated, and interpreted along a broad set of rules with clear differentiation across each.

Our methodology aims to connect and extend previous approaches to causal interventions in linguistic representational environments, including supervised recoverability, causal interventions, sparse analysis, semantic inspection, and behavioral evaluation (Finlayson et al., 2021; Amini et al., 2023; Arora et al., 2024). It also complements recent work that uses SAEs for linguistic analysis (Brinkmann et al., 2025; Jing et al., 2025).

Additionally, the paper aims to point at the intersection of computational linguistics and mechanistic interpretability, highlighting the convoluted, low-level, yet computationally significant nature of it (Brinkmann et al., 2025; Jing et al., 2025). The highly convoluted nature of representations at this intersection arises from the fact that even a small set of features can yield narrower and broader interpretations, as well as significant overlaps and cross-variable effects. Low-level linguistic structures examined with white-box experiments might allow us to understand and steer complex model behavior efficiently by distinguishing how the construction of a language contributes to representations broadly and locally.

6 Limitations

The study considers a single model, XGLM-564M, and trains one SAE architecture at one residual-stream site, limiting claims about generality across models, layers, and architectures. Even though the controlled contrast pairs were designed to preserve meaning, synthetic pairs could be prone to retaining constructional, lexical, or other surface confounds. The multilingual nature of the datasets may reflect variation in tokenisation, pretraining exposure, or model proficiency as languages were chosen selectively. Due to human interpretation, semantic inspection is partly subjective. The performed causal interventions are aimed at specific sparse directions the necessity and sufficiency of which for the corresponding linguistic computation have not been established. Behavioral evaluation did not measure long-horizon generation, task-level behavior, or multi-turn behavior in deployment scenarios; it solely relied on the first divergence of next-token preferences. Additionally, the 40-variable list can only serve as a foundation, as languages have thousands of linguistic properties, a more exhaustive map of which might resolve some concerns raised in this paper.

7 Conclusion

Information from defined linguistic features is broadly recoverable from language model representations. Independently learned SAE features substantially overlap with information retained from supervised probes. However, numerous features show narrower, language-specific, or surface-grounded definitions instead of pointing towards their respective, abstract, and broadly defined linguistic features. Causal interventions show that a subset of the derived sparse directions contribute to downstream linguistic representation, and an even smaller subset reliably affects local next-token behavior. Recoverability, semantic interpretability, sparse isolation, internal causal contribution, and behavioral influence are correlated along limited boundaries and should be treated as such. Future work should test whether these claims persist across model scales, architectures, languages, layers, and the limitations identified above. Additionally, they should further explore the intersection of computational linguistics and mechanistic interpretability.

AI use statement

Generative AI tools were used to assist with the generation of the synthetic controlled-contrast dataset and with implementation/programming tasks. Smaller LLM judges were also used within the data-generation pipeline to assess, filter, and curate generated examples. We also used generative AI tools to summarize and report in the GitHub repository - some of which populate the appendices of this paper - to enhance reproducibility of the study. All generated data, code, and resulting analyses were reviewed by the author.

Reproducibility statement

The code, datasets, experimental configurations, and analysis artifacts required to reproduce the study will be available at a GitHub repository which will be released after the double-blind review period while implementation details and results are documented in the appendices. The repository includes the controlled linguistic contrasts and implementations of the linear probes, sparse-autoencoder, causal-interventions, and behavioral analyses; however, it does not include trained SAE weights.

References

Afra Amini, Tiago Pimentel, Clara Meister, and Ryan Cotterell. Naturalistic causal probing for morpho-syntax. *Transactions of the Association for Computational Linguistics*, 11:384–403, 2023. URL https://aclanthology.org/2023.tacl-1.23/.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://aclanthology.org/2024.acl-long.785/.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. URL https://aclanthology.org/2022.cl-1.7/.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Sasha Boguraev, Christopher Potts, and Kyle Mahowald. Causal interventions reveal shared structure across english filler–gap constructions. In *Proceedings of EMNLP*, 2025. URL https://aclanthology.org/2025.emnlp-main.1271/.

Jasmine Brazilek and Harper Dunn. Assert, don't describe: Linguistic features that shift LLM reasoning about animal welfare. *arXiv preprint arXiv:2606.26104*, 2026. URL https://arxiv.org/abs/2606.26104.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. Large language models share representations of latent grammatical concepts across typologically diverse languages. In *Proceedings of NAACL*, 2025. URL https://aclanthology.org/2025.naacl-long.312/.

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024. URL https://arxiv.org/abs/2412.06410.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2126–2136, 2018. URL https://aclanthology.org/P18-1198/.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023. URL https://arxiv.org/abs/2309.08600.

Matthew S. Dryer and Martin Haspelmath (eds.). *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL https://wals.info/.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 1828–1843, 2021. URL https://aclanthology.org/2021.acl-long.144/.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024. URL https://arxiv.org/abs/2406.04093.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743, 2019. URL https://aclanthology.org/D19-1275/.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of NAACL-HLT*, pp. 4129–4138, 2019. URL https://aclanthology.org/N19-1419/.

Yi Jing, Zijun Yao, Lingxu Ran, Hongzhu Guo, Xiaozhi Wang, Lei Hou, and Juanzi Li. Sparse autoencoder interprets linguistic features in large language models. *arXiv preprint arXiv:2502.20344*, 2025. URL https://arxiv.org/abs/2502.20344.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*, 2021. URL https://arxiv.org/abs/2112.10668.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019. URL https://aclanthology.org/P19-1452/.

A Linguistic Variable Inventory and Dataset Construction

A.1 Variable inventory and construction design

The controlled dataset contains 500 basis–changed pairs each for 40 variables (20,000 pairs and 40,000 sentences). Table 3 showcases the variable names, construction approaches, languages, and target contrast families.

The split is fixed by example identifier: 001–400 are used for training, 401–450 for validation, and 451–500 for testing. Each example also includes a marker family and a lexical domain. Across variables, the training domains are the following: healthcare, engineering, education, logistics, workplace operations, and community services. For validation: finance and agriculture; for testing: legal, scientific-research, and public-administration. Variable-specific marker families are listed in the YAML files under `forward_pass_runs/data/v1/feature_metadata_01_40/`.

The JSONL schema contains `id`, `variable_id`, `variable`, `approach`, `language`, `surface_type`, `contrast`, `split`, `pair`, `marker_family`, and `lexical_domain`. Along the above JSONL schema, deterministic validation and a Qwen-based judge were used during dataset auditing.

Table 3: Controlled linguistic-variable inventory. Contrast labels reproduce the released metadata.

ID	Variable	Approach	Language(s)	Target contrasts
1	subject_expression_pro_drop	XL	tr	explicit_subject_to_omitted_subject, lexical_np_subject_to_zero_subject, pronoun_subject_to_zero_subject, agreement_recoverable_subject_to_context_recoverable_subject
2	basic_constituent_order	EN+XL	en, ja	english_canonical_svo_diagnostic_control, english_marked_argument_order_control, japanese_canonical_sov_order, japanese_particle_preserved_scrambling, argument_order_shift_with_role_interpretation_preserved
3	nominal_modifier_order	EN+XL	en, es	english_adjective_noun_order_control, spanish_noun_adjective_order, english_prenominal_possessive_control, spanish_postnominal_de_possession, stacked_modifier_order_shift
4	case_marking	XL	ru	case_preserved_word_order_permutation, nominative_to_accusative_marking, nominative_to_dative_marking, canonical_order_to_case_preserving_scramble, role_reversal_with_overt_case_change

Continued on next page

Table 3 continued

ID	Variable	Approach	Language(s)	Target contrasts
5	morphosyntactic_alignment	XL	tr	turkish_case_role_alignment_approximation, intransitive_subject_to_transitive_agent, intransitive_subject_to_transitive_patient_control, zero_marked_subject_to_accusative_object, agent_patient_role_swap_with_case_marking
6	transitivity_valency	EN	en	intransitive_to_transitive, transitive_to_ditransitive, transitive_to_experiencer_frame, transitive_to_beneficiary_frame, transitive_to_instrument_frame, monovalent_to_bivalent_to_trivalent
7	voice_and_agent_prominence	EN	en	active_to_passive, active_to_by_phrase_passive, active_to_agentless_passive, active_to_middle_like, active_to_impersonal, agent_prominent_to_patient_prominent
8	causativity_and_valency_change	EN+XL	en, tr	noncausative_to_causative, direct_to_indirect_causation, permissive_to_coercive_causation, english_periphrastic_causative_control, turkish_morphological_causative, causer_added_to_event
9	analytic_vs_synthetic_encoding	EN+XL	en, tr	english_function_word_encoding_control, turkish_bound_morphology_encoding, analytic_relation_encoding_control, synthetic_relation_encoding_probe, multiword_relation_to_morphologically_packed_relation_parallel
10	morphological_segmentation_type	XL	tr	root_to_agglutinated_form, single_suffix_to_multi_suffix_chain, transparent_morpheme_chain_to_unsegmented_surface, case_suffix_chain_to_separate_gloss_control, possessive_case_number_suffix_stack
11	agreement_indexing_density	XL	es	subject_verb_person_number_indexing, determiner_noun_gender_number_agreement, noun_adjective_gender_number_agreement, determiner_noun_adjective_agreement_stack, overt_subject_to_verb_indexing_control, agreement_feature_shift_with_lexical_content_stable
12	optionality_vs_obligatoriness_of_marking	XL	zh	context_only_time_to_explicit_time_adverb, temporally_unmarked_to_temporal_adverb_marked, unmarked_event_to_optional_aspect_marker, aspect_marker_present_to_absent_with_context_stable, classifier_or_number_context_to_bare_nominal_control, optional_marker_present_to_marker_absent_with_context_stable
13	redundancy_cumulative_exponence	XL	ru	single_site_marking_to_multi_site_agreement, case_gender_number_bundle_inspection, noun_case_to_adjective_noun_case_agreement, subject_verb_person_number_agreement, cumulative_case_gender_number_exponence, redundant_agreement_stack_to_less_redundant_control
14	definiteness_and_specificity	EN	en	indefinite_to_definite, nonspecific_to_specific, new_reference_to_familiar_reference, generic_to_specific_individual, unique_reference_to_nonunique_reference, bare_plural_generic_to_definite_plural
15	number_marking	EN+XL	en, ar	english_singular_plural_control, arabic_singular_dual_plural_number, arabic_dual_to_plural_agreement, arabic_number_sensitive_modifier_agreement, cardinal_number_to_number_agreement, collective_or_singulative_number_control
16	gender_noun_class	XL	es	masculine_to_feminine_agreement, noun_gender_to_adjective_agreement, determiner_gender_shift, grammatical_gender_with_inanimate_nouns, natural_gender_to_grammatical_gender_control, gender_agreement_preserved_under_number_change
17	animacy_and_humanness	EN+ctrl	en	human_to_nonhuman_animate, animate_to_inanimate, sentient_to_nonsentient_living, living_to_nonliving, human_agent_to_artifact_agent_control, animal_patient_to_object_patient_control
18	person_marking_and_person_hierarchy	EN+XL	en, es	first_person_to_second_person, second_person_to_third_person, singular_person_to_plural_person, english_person_pronoun_control, spanish_person_number_verb_indexing, overt_pronoun_to_verb_indexing_control, speaker_or_addressee_to_third_person_reference
19	inclusive_exclusive_distinction	XL	zh	inclusive_we_to_exclusive_we_lexical, zanmen_to_women_contextual_contrast, speaker_addressee_included_to_excluded, group_membership_explicit_to_contextual_we, inclusive_like_phrase_to_exclusive_like_phrase
20	pronoun_richness_and_reduction	EN+XL	en, ja	english_overt_pronoun_baseline, japanese_overt_argument_to_zero_reference, repeated_pronoun_to_contextual_reference, first_person_overt_to_zero_subject, third_person_overt_to_zero_subject, pronoun_identity_recoverable_from_discourse
21	possession_and_alienability	EN+ctrl	en	alienable_relation_to_inalienable_relation_control, ownership_possession_to_kinship_relation, controllable_object_to_body_part_relation, temporary_possession_to_permanent_relation, legal_owner_to_associated_relation, possessor_control_high_to_possessor_control_low
22	tense_prominence	EN+XL	en, zh	english_past_present_future_tense_control, mandarin_temporal_adverb_time_reference, mandarin_contextual_time_inference, tense_marked_to_time_adverb_marked_parallel, explicit_time_anchor_to_contextual_time_anchor, morphological_tense_to_discourse_time_anchor_parallel
23	aspect_and_event_structure	EN+XL	en, ru	english_progressive_to_simple_aspect_control, russian_perfective_to_imperfective_aspect, habitual_to_single_event, completed_to_ongoing, iterative_to_punctual, aspect_pair_with_lexical_meaning_stable
24	modality_and_mood	EN	en	possibility_to_necessity, ability_to_permission, obligation_to_advice, must_to_should, may_to_can, indicative_to_imperative, permission_to_prohibition_control

Continued on next page

Table 3 continued

ID	Variable	Approach	Language(s)	Target contrasts
25	epistemic_modality	EN+ctrl	en	certain_to_probable, probable_to_possible, possible_to_doubtful, certainly_to_apparently, direct_claim_to_epistemically_hedged_claim, high_confidence_to_low_confidence
26	evidentiality	EN+XL	en, tr	english_source_marking_control, turkish_direct_past_to_reported_evidential, turkish_direct_to_inferred_evidential, reported_to_inferred_evidential, hearsay_source_to_inference_source, source_marker_present_to_absent_control
27	mirativity_stance_and_affect_marking	XL	ja	neutral_to_surprise_marker, neutral_to_unexpectedness_stance, plain_assertion_to_sentence_final_affect, direct_observation_to_mirative_reaction, neutral_description_to_evaluative_stance, speaker_attitude_absent_to_speaker_attitude_marked
28	negation_and_polarity_structure	EN	en	affirmation_to_simple_negation, simple_negation_to_denial, positive_claim_to_contradiction, wide_scope_negation_to_narrow_scope_negation, polarity_item_licensed_to_unlicensed_control, double_negation_to_single_negation
29	quantifier_scope_and_distributivity	EN	en	all_to_some, every_to_each, collective_to_distributive, none_to_some, most_to_all, subject_quantifier_scope_to_object_quantifier_scope
30	conditional_and_counterfactual_marking	EN	en	real_conditional_to_hypothetical, hypothetical_to_counterfactual, if_then_to_unmarked_dependency, conditional_to_concessive_conditional, future_open_to_past_counterfactual, necessary_condition_to_sufficient_condition_control
31	subordination_and_embedding	EN	en	main_clause_to_embedded_belief, main_clause_to_embedded_claim, single_embedding_to_recursive_embedding, reason_clause_to_evidence_clause, desire_embedding_to_belief_embedding, finite_to_nonfinite_embedding_control
32	quotation_and_reported_speech_structure	EN+XL	en, ja	english_direct_quote_to_indirect_report_control, english_punctuation_quote_to_reported_claim_control, japanese_quotative_reporting_structure, japanese_direct_quotation_to_indirect_report, embedded_claim_to_source_attribution, exact_words_to_paraphrased_report
33	discourse_relation_marking	EN+ctrl	en	cause_to_contrast, because_to_so_result, although_to_because, however_to_therefore, while_temporal_to_while_contrastive, since_causal_to_since_temporal, connective_present_to_connective_absent_control
34	topic_comment_structure	XL	ja	subject_marker_to_topic_marker, topic_fronting_to_neutral_order, old_information_topic_to_new_information_comment, contrastive_topic_to_plain_topic, topic_comment_to_subject_predicate_parallel, discourse_topic_present_to_topic_absent_control
35	focus_and_given_new_marking	EN+XL	en, ja	english_neutral_to_cleft_focus_control, english_contrastive_focus_control, japanese_focus_particle_marking, japanese_wa_topic_to_ga_focus_contrast, given_to_new_information_focus, focus_particle_to_plain_clause
36	genericity_and_kind_level_reference	EN	en	specific_individual_to_generic_kind, episodic_to_habitual, bare_plural_generic_to_definite_specific, norm_like_generalization_to_observed_event, kind_statement_to_member_statement, generic_to_statistical_generalization
37	social_deixis_honorifics_status_encoding	EN+XL	en, ko	english_deference_phrase_control, korean_plain_to_honorific_speech_level, informal_to_formal_request, equal_status_to_higher_status_addressee, referent_plain_to_referent_honorific, speaker_humility_to_neutral_description
38	speech_act_force_and_request_directness	EN+ctrl	en	assertion_to_question, statement_to_command, direct_command_to_indirect_request, request_to_warning, advice_to_obligation, promise_to_threat, explicit_performative_to_indirect_speech_act
39	deixis_and_perspective_anchoring	EN	en	here_to_there, now_to_then, this_to_that, speaker_centered_to_addressee_centered, first_person_to_third_person_perspective, proximal_to_distal_reference, today_to_that_day_context_control
40	orthographic_and_tokenization_interface	EN+XL	en, zh	english_capitalization_change_control, english_punctuation_present_to_absent_control, english_spacing_variant_control, mandarin_unsegmented_text_probe, mandarin_word_segmented_diagnostic_control, latin_punctuation_to_chinese_punctuation_control

B Supervised Probe Details

B.1 Probe families and representations

All XGLM hidden states are right-padded and cached for each controlled sentence. The two representations of the sequence are the residual activation of the last non-padding token and the average residual activation of non-padding tokens. The sentence probe is a logistic classifier of base vs changed sentences. The directional probe is based on the classification of the forward pair difference $h(x^{\text{changed}}) - h(x^{\text{basis}})$ against the exact reverse difference for the same pair. This construction tests the orientation of the controlled transformation.

Regularization and layer selection only use validation data. Reported test AUROC are supplemented with 1,000-sample bootstrap confidence intervals provided in the repository. Text controls include word- and character-level TF-IDF models and length/punctuation features.

B.2 Evidence ladder and cross-seed rule

The probe analysis uses five criteria: (1) activation recoverability, (2) comparison with surface-text controls, (3) held-out pair-direction consistency, (4) generalization across marker families and lexical domains, and (5) viability of the learned activation direction. The strict cross-seed probe classification used in the main paper requires levels 1, 3, and 4 to pass in each of seeds 1, 2, and 42. Learned-direction viability is reported separately and is not part of this binary definition. This rule produces 26 robust variables; 23 of these also receive strong SAE evidence.

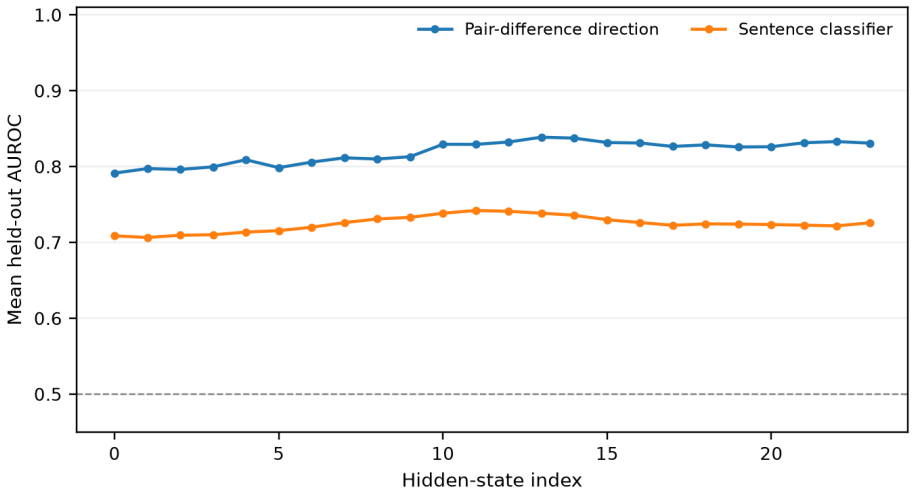

Figure 4: Mean held-out AUROC by hidden-state index, averaged over variables and seeds for mean-pooled sentence and pair-difference probes.

Table 4 gives the complete cross-seed summary. Seed-specific diagnostic figures already committed to the repository are located under `forward_pass_runs/v1/results_seed_{1,2,42}/visuals-tables/figures/`.

Table 4: Cross-seed supervised-probe summary. The pass count is the number of seeds (out of three) satisfying evidence levels 1, 3, and 4.

ID	Variable	Sent. AUROC	Layer	Δ AUROC	Passes	Status
1	Subject expression / pro-drop	0.950	12	1.000	3	robust_3of3
2	basic_constituent_order	0.593	8	0.622	0	no_core_pass_0of3
3	nominal_modifier_order	0.531	9	0.532	0	no_core_pass_0of3
4	case_marking	0.992	18	0.966	3	robust_3of3
5	morphosyntactic_alignment	0.821	8	0.968	3	robust_3of3
6	transitivity_valency	0.964	14	1.000	3	robust_3of3
7	voice_and_agent_prominence	0.860	0	0.980	3	robust_3of3
8	causativity_and_valency_change	0.857	13	0.960	3	robust_3of3
9	analytic_vs_synthetic_encoding	0.995	7	1.000	3	robust_3of3
10	morphological_segmentation_type	0.956	1	0.977	3	robust_3of3
11	agreement_indexing_density	0.754	0	0.732	3	robust_3of3
12	optionality_vs_obligatoriness_of_marking	0.628	12	0.664	0	no_core_pass_0of3
13	redundancy_cumulative_exponence	0.621	23	0.878	0	no_core_pass_0of3
14	definiteness_and_specificity	0.820	13	0.986	3	robust_3of3
15	number_marking	0.438	23	0.438	0	no_core_pass_0of3
16	gender_noun_class	0.867	0	0.988	3	robust_3of3
17	animacy_and_humanness	0.793	12	0.925	3	robust_3of3
18	person_marking_and_person_hierarchy	0.678	9	0.821	3	robust_3of3
19	inclusive_exclusive_distinction	0.744	11	0.910	3	robust_3of3
20	pronoun_richness_and_reduction	0.922	12	1.000	3	robust_3of3

Continued on next page

Table 4 continued

ID	Variable	Sent. AUROC	Layer	Δ AUROC	Passes	Status
21	possession_and_alienability	0.728	15	0.874	3	robust_3of3
22	tense_prominence	0.662	5	0.925	3	robust_3of3
23	aspect_and_event_structure	0.470	14	0.384	0	no_core_pass_0of3
24	modality_and_mood	0.638	21	0.744	0	no_core_pass_0of3
25	epistemic_modality	0.724	10	0.752	3	robust_3of3
26	evidentiality	0.483	6	0.702	0	no_core_pass_0of3
27	mirativity_stance_and_affect_marking	0.832	0	1.000	3	robust_3of3
28	negation_and_polarity_structure	0.827	4	0.960	3	robust_3of3
29	quantifier_scope_and_distributivity	0.860	1	0.974	3	robust_3of3
30	conditional_and_counterfactual_marking	0.707	15	0.928	3	robust_3of3
31	subordination_and_embedding	0.500	0	0.747	0	no_core_pass_0of3
32	quotation_and_reported_speech_structure	0.998	6	1.000	3	robust_3of3
33	discourse_relation_marking	0.459	6	0.086	0	no_core_pass_0of3
34	topic_comment_structure	0.886	7	1.000	3	robust_3of3
35	focus_and_given_new_marking	0.898	0	1.000	3	robust_3of3
36	genericity_and_kind_level_reference	0.806	12	0.947	3	robust_3of3
37	social_deixis_honorifics_status_encoding	0.554	3	0.750	0	no_core_pass_0of3
38	speech_act_force_and_request_directness	0.719	22	0.580	0	no_core_pass_0of3
39	deixis_and_perspective_anchoring	0.597	0	0.882	0	no_core_pass_0of3
40	orthographic_and_tokenization_interface	0.490	0	0.518	0	no_core_pass_0of3

C SAE Training and Model Selection

C.1 Training corpus and activation site

The independent SAE corpus contains 149,338 unique multilingual sentences after deterministic cleaning. The dataset retains linguistic metadata for audit purposes.

For XGLM-564M, activations are taken from hidden-state index 12 (the residual stream after decoder layer 11). Contexts are cut off at 192 tokens. Activation caching uses bfloat16 for the forward compute and float16 for storage. The training split contains 3,101,336 activation tokens, and the validation split 165,010, for a total of 3,266,346 unique tokens.

C.2 Canonical architecture and optimization

The canonical SAE is a $16\times$ BatchTopK model with input width 1,024, dictionary width 16,384, $k = 256$, auxiliary $k = 512$, auxiliary penalty $1/32$, decoder-bias preprocessing, and four epochs of training. Token batches contain 4,096 activations. Adam uses learning rate 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, no weight decay, 125 warm-up steps, decay over the final 20% of training, and gradient clipping at norm 1.0. The dead-feature window is 250 steps and the random seed is 42.

Initial JumpReLU runs collapsed despite robust reconstruction: the original $16\times$ run had validation explained variance approximately 0.9922 while 15,639 of 16,384 features were dead. BatchTopK substantially improved dictionary utilization. The four-epoch model attained validation MSE 0.004497, explained variance 0.998634, cosine similarity 0.994594, and validation mean $L_0 = 256$. Across the entire unique corpus, 16,372 features activated at least once. An eight-epoch diagnostic slightly improved explained variance to 0.998936 but reduced active features to 15,764; the four-epoch checkpoint was therefore selected.

C.3 Example-independent inference export

A global threshold of 0.11494007418131039 was calibrated from 758 training batches by tracking the minimum positive activation selected by BatchTopK with an EMA rate of 0.01. Downstream inference applies this threshold independently to every token. The export retains validation explained variance 0.998635, cosine similarity 0.994597, and mean $L_0 = 255.9135$.

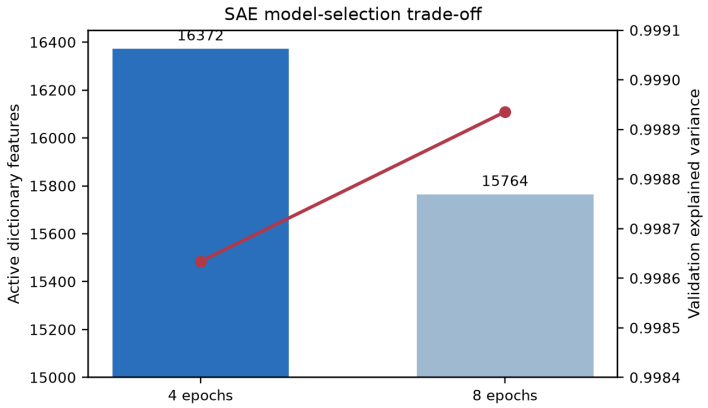

Figure 5: SAE model-selection trade-off. Four additional epochs give only a small reconstruction gain while materially reducing dictionary coverage.

D Sparse Feature Evidence and Inspection

D.1 Feature-wide association and corrected null

For each variable and each of 16,384 features, the primary normalized paired effect is

$$e_f = \frac{\text{mean}_i\, \Delta a_{if}}{\sqrt{\text{mean}_i\, \Delta a_{if}^2}}, \qquad \Delta a_{if} = a_f(x_i^{\text{changed}}) - a_f(x_i^{\text{basis}}).$$

Candidate ranking uses the training split only. The top 25 absolute effects are retained separately for mean-pooled and final-token representations, producing 2,000 initial candidate rows. A 2,000-permutation paired sign-flip test records the maximum absolute effect across all 16,384 features on each permutation. The empirical max-statistic p value therefore controls feature search within each variable and pooling analysis; its minimum possible value is $1/2001$.

Candidate specificity is assessed by ranking the target variable's effect for a feature against its effects over all 40 variables, together with the ratio, margin, and standardized difference relative to non-target variables.

D.2 Evidence tiers and semantic inspection

Tier A requires held-out direction, subgroup robustness across construction and domain, and cross-variable selectivity. B1 is variable shared, but maintains construction robustness. B2 keeps the held-out direction with incomplete subgroup robustness C only keeps the aggregate held-out direction D loses the held-out direction. The resulting counts are 11 A, 2 B1, 16 B2, 3 C and 8 D.

Statistical evidence and semantic interpretation are treated as separate measures. For each causally tested feature the repository records the natural multilingual top activations, the number of unique top tokens and languages, and an inspection rationale. Inspection grades A-D correspond decreasingly to the natural activation pattern to the intended variable. Alternative candidates from the set of already surviving ones were inspected if the original statistical winner was a close proxy. Feature selection was performed prior to consideration of causal outcomes. Table 5 reports both the statistical winner and the feature ultimately taken to causal testing.

Table 5: Complete sparse-feature evidence and inspection summary. The statistically selected feature and the causally tested feature can differ after alternative-candidate inspection.

ID	Variable	SAE f	Tier	Train	Val.	Test	Spec.	Causal f	Grade	Inspection
1	Subject expression / pro-drop	8431	D	0.869	0.101	-0.248	1	8041	C	mixed_subject_expression_proxy

Continued on next page

Table 5 continued

ID	Variable	SAE f	Tier	Train	Val.	Test	Spec.	Causal f	Grade	Inspection
2	basic_constituent_order	9791	C	0.823	0.807	0.033	1	932	C	clause_order_proxy
3	nominal_modifier_order	4426	D	-0.566	-0.076	0.199	1	4426	C	prenominal_possessive_modifier
4	case_marking	13269	A	0.511	0.390	0.585	1	13269	B	case_bearing_nominals
5	morphosyntactic_alignment	7247	B1	-0.695	-0.693	-0.665	10	7247	D	shared_punctuation_proxy
6	transitivity_valency	5382	B2	0.939	0.802	0.075	1	164	C	argument_structure_morphology_proxy
7	voice_and_agent_prominence	8235	B2	0.855	0.921	0.623	1	8235	A	voice_passive_participle_structure
8	causativity_and_valency_change	8182	B1	0.894	0.934	0.783	2	8182	D	causativity_association_uninterpretable
9	analytic_vs_synthetic_encoding	6579	B2	-0.502	-0.561	-0.480	1	144	C	analytic_synthetic_morphology_proxy
10	morphological_segmentation_type	6125	A	-0.708	-0.531	-0.456	1	15843	B	turkish_suffixal_morphology
11	agreement_indexing_density	758	B2	0.826	0.818	0.673	1	758	C	agreement_inflection_proxy
12	optionality_vs_obligatoriness_of_marking	9607	B2	0.404	0.519	0.538	1	9607	D	optionality_uninterpretable
13	redundancy_cumulative_exponence	9749	A	-0.681	-0.836	-0.775	1	4814	B	cumulative_inflection_morphology
14	definiteness_and_specificity	13819	B2	-0.508	-0.483	-0.528	1	12995	C	indefinite_one_proxy
15	number_marking	8113	D	0.762	0.204	-0.045	2	8685	B	plural_number_morphology
16	gender_noun_class	5205	A	0.897	0.908	0.909	1	5205	A	grammatical_gender_feminine_morphology
17	animacy_and_humanness	3104	C	0.499	0.573	0.035	2	6618	D	animacy_lexical_proxy
18	person_marking_and_person_hierarchy	13235	B2	0.378	0.208	0.270	1	13235	B	first_person_plural_marking
19	inclusive_exclusive_distinction	6062	B2	-0.854	-0.494	-0.023	1	6062	C	first_person_pronoun_proxy
20	pronoun_richness_and_reduction	7551	A	-0.823	-0.800	-0.810	1	6944	A	crosslingual_third_person_pronouns
21	possession_and_alienability	9576	A	0.491	0.516	0.508	1	15325	C	kinship_possession_proxy
22	tense_prominence	1385	D	-0.396	-0.138	0.000	2	1385	C	temporal_reference_lexemes
23	aspect_and_event_structure	6216	D	-0.705	-0.259	0.291	1	6216	B	progressive_aspect_ing
24	modality_and_mood	3941	C	0.483	0.208	0.141	1	13197	A	deontic_modality_markers
25	epistemic_modality	15221	B2	0.770	0.566	0.590	1	15221	A	epistemic_uncertainty_markers
26	evidentiality	11375	D	0.876	0.028	-0.743	1	6900	B	turkish_indirect_evidential_mis
27	mirativity_stance_and_affect_marking	1887	A	0.975	0.855	0.952	1	1887	C	mirativity_japanese_entangled
28	negation_and_polarity_structure	4382	B2	0.604	0.282	0.386	1	4382	D	negation_association_uninterpretable
29	quantifier_scope_and_distributivity	8551	B2	0.700	0.475	0.313	1	8551	B	quantified_noun_phrases
30	conditional_and_counterfactual_marking	15951	A	0.718	0.179	0.304	1	11810	B	counterfactual_conditional_morphology
31	subordination_and_embedding	7643	A	0.927	0.911	0.649	1	11461	B	complementizer_embedding
32	quotation_and_reported_speech_structure	1754	A	-0.830	-0.859	-0.904	1	1754	B	quotation_mark_surface_control
33	discourse_relation_marking	9728	D	-0.881	-0.112	0.886	3	9728	D	discourse_relation_uninterpretable
34	topic_comment_structure	10896	B2	-0.680	-0.290	-0.609	1	1302	D	topic_comment_uninterpretable
35	focus_and_given_new_marking	7136	B2	0.699	0.419	0.102	1	7136	B	japanese_korean_focus_particles
36	genericity_and_kind_level_reference	5836	B2	0.372	0.233	0.345	2	1858	D	genericity_quantifier_proxy
37	social_deixis_honorifics_status_encoding	2081	B2	0.820	0.422	0.372	1	2081	A	korean_formal_honorific_morphology
38	speech_act_force_and_request_directness	3396	A	0.440	0.057	0.400	1	14406	C	speech_act_exclamation_surface
39	deixis_and_perspective_anchoring	14441	B2	-0.702	-0.316	-0.270	1	14441	A	crosslingual_temporal_deixis
40	orthographic_and_tokenization_interface	3132	D	-0.418	-0.141	0.246	3	3132	B	orthographic_punctuation_interface

E Ablation Results

At every token in hidden-state index 12, ablation subtracts the selected feature's decoded contribution, $a_f W_{\mathrm{dec},f}$, after undoing the activation scale. The downstream score is the projection onto a linguistic pair-difference direction estimated on the training split and evaluated at XGLM's final hidden layer. Positive attenuation means that ablation reduced the held-out distinction.

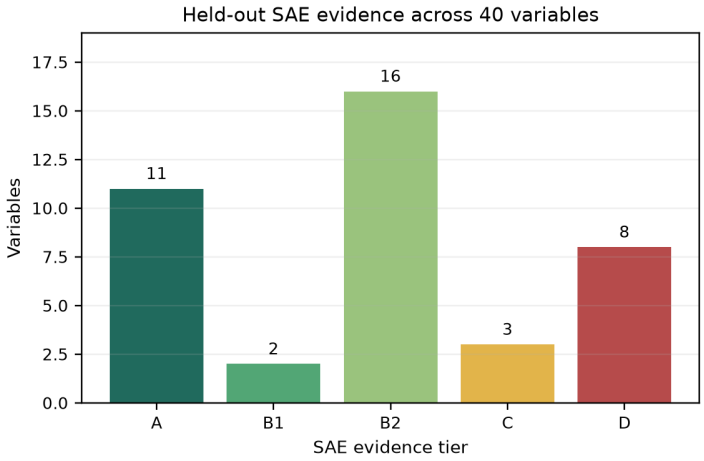

Figure 6: Distribution of the 40 variables across SAE evidence tiers.

Each target is compared with three SAE features matched using training-only activation magnitude, firing rate, and decoder norm. Pair bootstrap inference uses 2,000 resamples and seed 42; target-versus-control p values are adjusted over the 40 variables using Benjamini–Hochberg. Nineteen variables satisfy the specific-positive ablation rule. The mean cosine similarity between intervened and baseline final representations remains close to one, supporting a localized rather than globally destructive intervention.

Table 6 provides all 40 variable-level results.

Table 6: Complete ablation results. Positive attenuation denotes reduction of the held-out downstream linguistic distinction.

ID	Variable	f	Side	Atten.	Target–ctrl.	q	Cos.	Category
1	Subject expression / pro-drop	8041	changed	0.00453	0.00849	0.002	0.99998	specific_positive
2	basic_constituent_order	932	changed	0.00134	0.00109	0.344	1.00000	positive_control_advantage
3	nominal_modifier_order	4426	basis	0.00012	0.00010	0.645	1.00000	positive_control_advantage
4	case_marking	13269	changed	-0.00220	-0.00005	0.993	1.00000	null_or_reverse
5	morphosyntactic_alignment	7247	basis	-0.00773	-0.00780	0.002	0.99999	null_or_reverse
6	transitivity_valency	164	changed	0.05631	0.05683	0.002	0.99977	specific_positive
7	voice_and_agent_prominence	8235	changed	0.00318	0.00050	0.680	1.00000	positive_control_advantage
8	causativity_and_valency_change	8182	changed	0.01733	0.01472	0.214	0.99876	positive_control_advantage
9	analytic_vs_synthetic_encoding	144	basis	0.00171	0.00212	0.020	1.00000	specific_positive
10	morphological_segmentation_type	15843	changed	0.00626	0.00645	0.002	1.00000	specific_positive
11	agreement_indexing_density	758	changed	0.03090	0.03275	0.002	0.99997	specific_positive
12	optionality_vs_obligatoriness_of_marking	9607	changed	-0.00171	-0.00155	0.002	1.00000	null_or_reverse
13	redundancy_cumulative_exponence	4814	changed	-0.00257	-0.00740	0.002	0.99999	null_or_reverse
14	definiteness_and_specificity	12995	basis	0.00428	0.00418	0.002	1.00000	specific_positive
15	number_marking	8685	changed	0.00078	-0.02484	0.002	1.00000	positive_nonspecific
16	gender_noun_class	5205	changed	0.21767	0.22811	0.002	0.99989	specific_positive
17	animacy_and_humanness	6618	changed	0.13385	0.13888	0.002	0.99995	specific_positive
18	person_marking_and_person_hierarchy	13235	changed	-0.00019	-0.00191	0.002	1.00000	null_or_reverse
19	inclusive_exclusive_distinction	6062	basis	0.00111	0.00323	0.002	1.00000	specific_positive
20	pronoun_richness_and_reduction	6944	basis	0.00485	0.01171	0.002	1.00000	specific_positive
21	possession_and_alienability	15325	changed	0.03118	0.03382	0.002	0.99998	specific_positive
22	tense_prominence	1385	basis	0.00000	-0.00020	0.406	1.00000	null_or_reverse
23	aspect_and_event_structure	6216	basis	0.00209	0.01193	0.002	1.00000	specific_positive
24	modality_and_mood	13197	changed	0.00833	0.01062	0.002	1.00000	specific_positive
25	epistemic_modality	15221	changed	0.00361	-0.00420	0.003	1.00000	positive_nonspecific
26	evidentiality	6900	changed	0.00390	0.00111	0.425	1.00000	positive_control_advantage

Continued on next page

Table 6 continued

ID	Variable	f	Side	Atten.	Target–ctrl.	q	Cos.	Category
27	mirativity_stance_and_affect_marking	1887	changed	0.00416	0.01058	0.002	1.00000	specific_positive
28	negation_and_polarity_structure	4382	changed	-0.00037	-0.00104	0.211	1.00000	null_or_reverse
29	quantifier_scope_and_distributivity	8551	changed	0.00582	0.00830	0.002	1.00000	specific_positive
30	conditional_and_counterfactual_marking	11810	changed	0.00003	0.00020	0.406	1.00000	positive_control_advantage
31	subordination_and_embedding	11461	changed	0.00056	-0.00502	0.006	0.99998	positive_nonspecific
32	quotation_and_reported_speech_structure	1754	basis	0.00754	0.00902	0.002	0.99999	specific_positive
33	discourse_relation_marking	9728	basis	-0.00201	-0.00050	0.746	0.99999	null_or_reverse
34	topic_comment_structure	1302	changed	0.00251	-0.00443	0.002	1.00000	positive_nonspecific
35	focus_and_given_new_marking	7136	changed	-0.00002	-0.00163	0.002	1.00000	null_or_reverse
36	genericity_and_kind_level_reference	1858	changed	0.00215	0.00216	0.002	1.00000	specific_positive
37	social_deixis_honorifics_status_encoding	2081	changed	0.00488	0.00477	0.002	1.00000	specific_positive
38	speech_act_force_and_request_directness	14406	changed	0.00058	0.00101	0.003	1.00000	specific_positive
39	deixis_and_perspective_anchoring	14441	basis	0.00080	-0.00124	0.225	1.00000	positive_nonspecific
40	orthographic_and_tokenization_interface	3132	basis	0.00011	-0.00040	0.102	1.00000	positive_nonspecific

F STEERING RESULTS

Steering adds the selected decoder direction to the feature-low member of each held-out pair at hidden-state index 12. The coefficient is the absolute training-set activation gap multiplied by dose. Control coefficients are decoder-norm adjusted so that the residual-space L_2 perturbation exactly matches the target intervention. Doses $0.5\times$, $1\times$, and $2\times$ are evaluated, with $1\times$ fixed as the primary inferential condition before held-out analysis. The same final-layer linguistic direction and the same three training-only controls as in ablation are used.

23 variables satisfy the specific-positive steering rule, 25 have a monotonic target response across the three doses, and 31 show a target effect larger than the matched-control mean.

Table 7 gives the complete dose-response and primary-test summary.

Table 7: Complete steering results. Doses scale the training-set target activation gap; the $1\times$ condition is primary.

ID	Variable	f	$0.5\times$	$1\times$	$2\times$	Mono.	Target–ctrl.	q	Category
1	Subject expression / pro-drop	8041	0.00049	0.00055	-0.00080	–	0.00007	0.978	positive_control_advantage
2	basic_constituent_order	932	0.00128	0.00183	0.00452	✓	0.00052	0.316	positive_control_advantage
3	nominal_modifier_order	4426	0.00001	0.00081	0.00243	✓	-0.00059	0.593	positive_nonspecific
4	case_marking	13269	-0.00278	-0.00114	0.00021	✓	0.00003	0.978	null_or_reverse
5	morphosyntactic_alignment	7247	0.00008	-0.00480	-0.00309	–	-0.00191	0.236	null_or_reverse
6	transitivity_valency	164	0.00065	0.00405	0.01255	✓	0.00781	0.002	specific_positive
7	voice_and_agent_prominence	8235	0.00113	0.00229	0.00598	✓	0.00106	0.095	positive_control_advantage
8	causativity_and_valency_change	8182	-0.00263	-0.00093	-0.00795	–	0.00234	0.182	null_or_reverse
9	analytic_vs_synthetic_encoding	144	0.00038	-0.00021	0.00114	–	-0.00121	0.277	null_or_reverse
10	morphological_segmentation_type	15843	-0.00072	0.00408	0.00811	✓	0.00460	0.002	specific_positive
11	agreement_indexing_density	758	0.00110	0.00218	0.00340	✓	0.00360	0.002	specific_positive
12	optionality_vs_obligatoriness_of_marking	9607	-0.00006	0.00068	-0.00031	–	0.00015	0.513	positive_control_advantage
13	redundancy_cumulative_exponence	4814	-0.00059	-0.00141	-0.00401	–	-0.00087	0.004	null_or_reverse
14	definiteness_and_specificity	12995	0.00153	0.00244	0.00368	✓	0.00257	0.002	specific_positive
15	number_marking	8685	0.00100	0.00388	0.00910	✓	0.00200	0.020	specific_positive
16	gender_noun_class	5205	0.03564	0.07863	0.17849	✓	0.07886	0.002	specific_positive
17	animacy_and_humanness	6618	0.00833	0.02174	0.04662	✓	0.02138	0.002	specific_positive
18	person_marking_and_person_hierarchy	13235	0.00047	0.00069	0.00109	✓	0.00005	0.978	positive_control_advantage
19	inclusive_exclusive_distinction	6062	0.00194	0.00244	0.00301	✓	0.00177	0.002	specific_positive

Continued on next page

Table 7 continued

ID	Variable	f	$0.5\times$	$1\times$	$2\times$	Mono.	Target–ctrl.	q	Category
20	pronoun_richness_and_reduction	6944	0.00094	0.00253	0.00567	✓	0.00177	0.004	specific_positive
21	possession_and_alienability	15325	0.00683	0.01420	0.02439	✓	0.01334	0.002	specific_positive
22	tense_prominence	1385	0.00191	0.00445	0.00936	✓	0.00468	0.002	specific_positive
23	aspect_and_event_structure	6216	0.00139	0.00190	0.00394	✓	0.00418	0.002	specific_positive
24	modality_and_mood	13197	0.00265	0.00712	0.01669	✓	0.01091	0.002	specific_positive
25	epistemic_modality	15221	0.00156	0.00663	0.01383	✓	0.00323	0.002	specific_positive
26	evidentiality	6900	0.00237	0.00138	0.00197	–	0.00490	0.002	specific_positive
27	mirativity_stance_and_affect_marking	1887	0.00303	0.00626	0.01327	✓	0.00378	0.002	specific_positive
28	negation_and_polarity_structure	4382	-0.00196	0.00070	-0.00018	–	0.00270	0.002	specific_positive
29	quantifier_scope_and_distributivity	8551	0.00126	0.00347	0.00648	✓	0.00456	0.002	specific_positive
30	conditional_and_counterfactual_marking	11810	-0.00099	0.00062	0.00287	✓	-0.00025	0.593	positive_nonspecific
31	subordination_and_embedding	11461	0.00452	0.00877	0.01630	✓	0.00880	0.002	specific_positive
32	quotation_and_reported_speech_structure	1754	0.00271	0.00658	0.01115	✓	0.00808	0.002	specific_positive
33	discourse_relation_marking	9728	0.00024	0.00041	-0.00113	–	0.00071	0.040	specific_positive
34	topic_comment_structure	1302	0.00178	0.00030	0.00140	–	-0.00112	0.020	positive_nonspecific
35	focus_and_given_new_marking	7136	0.00024	-0.00157	-0.00096	–	-0.00219	0.002	null_or_reverse
36	genericity_and_kind_level_reference	1858	-0.00034	-0.00055	0.00040	–	-0.00107	0.178	null_or_reverse
37	social_deixis_honorifics_status_encoding	2081	0.00480	0.00579	0.01548	✓	0.00579	0.002	specific_positive
38	speech_act_force_and_request_directness	14406	0.00016	-0.00038	-0.00107	–	0.00032	0.499	null_or_reverse
39	deixis_and_perspective_anchoring	14441	0.00096	0.00014	0.00502	–	0.00203	0.027	specific_positive
40	orthographic_and_tokenization_interface	3132	-0.00076	-0.00167	-0.00152	–	-0.00025	0.513	null_or_reverse

G Behavioral Evaluation

The primary behavioral family comprises variable IDs 6, 10, 11, 14, 16, 17, 19, 20, 21, 23, 24, 27, 29, 32, and 37. For each of 50 held-out test pairs per variable, the tokenized members are compared at their first divergent token. Their longest shared prefix is used as context; when divergence occurs at token 1, the XGLM beginning-of-sequence token supplies the common context. The outcome is the change under steering in the logit margin between the feature-high and feature-low continuation tokens.

The target and the same three matched controls are evaluated at $0.5\times$, $1\times$, and $2\times$; $1\times$ is primary. Bootstrap inference uses 2,000 resamples. Benjamini–Hochberg correction is applied to the frozen 15-variable family; the separate all-40 correction remains exploratory.

Nine cohort variables have positive target effects, ten outperform matched controls, and six satisfy the combined positive, control-specific, FDR-corrected criterion. All six also show monotonic dose response. They are morphological segmentation type (feature 15843), animacy and humanness (6618), inclusive/exclusive distinction (6062), possession and alienability (15325), aspect and event structure (6216), and mirativity/stance/affect marking (1887).

Table 8 reports the complete primary-family results.

Table 8: Behavioral results for the frozen 15-variable mechanistic cohort.

ID	Variable	f	Target	Target–ctrl.	q	Mono.	BOS	Category
6	transitivity_valency	164	-0.017288	-0.024911	0.002	–	0	null_or_reverse
10	morphological_segmentation_type	15843	0.005894	0.005927	0.002	✓	0	specific_positive
11	agreement_indexing_density	758	-0.004354	0.003068	0.194	–	42	null_or_reverse
14	definiteness_and_specificity	12995	-0.000631	-0.000344	0.525	–	20	null_or_reverse
16	gender_noun_class	5205	-0.015715	-0.013284	0.002	–	50	null_or_reverse
17	animacy_and_humanness	6618	0.005830	0.006903	0.002	✓	0	specific_positive
19	inclusive_exclusive_distinction	6062	0.000689	0.000734	0.023	✓	0	specific_positive
20	pronoun_richness_and_reduction	6944	0.002054	0.000733	0.588	✓	5	positive_control_advantage
21	possession_and_alienability	15325	0.017484	0.019592	0.002	✓	0	specific_positive

Continued on next page

Table 8 continued

ID	Variable	f	Target	Target–ctrl.	q	Mono.	BOS	Category
23	aspect_and_event_structure	6216	0.000505	0.003034	0.009	✓	0	specific_positive
24	modality_and_mood	13197	0.003382	0.001627	0.194	✓	0	positive_control_advantage
27	mirativity_stance_and_affect_marking	1887	0.009712	0.005211	0.002	✓	0	specific_positive
29	quantifier_scope_and_distributivity	8551	0.002139	0.002527	0.063	✓	8	positive_control_advantage
32	quotation_and_reported_speech_structure	1754	-0.004094	-0.000252	0.877	–	0	null_or_reverse
37	social_deixis_honorifics_status_encoding	2081	-0.005310	-0.005307	0.002	–	9	null_or_reverse

H Reproducibility Details

The public repository consists of four analysis blocks. The controlled datasets, metadata, activation/probe scripts, three seed-specific result trees, and existing probe figures are available under `forward_pass_runs/`. `SAE/` contains the natural training corpus, canonical SAE implementation and configuration, post-canonical feature evidence, and natural-text inspection artifacts. All intervention scripts, manifests, controls, pair-level summaries, and final causal evidence are located in `causal_steps/`. The ultimate synthesis is captured in the root-level `Final Conclusions.md`.

The released files include the full 20,000 pair controlled dataset, the 149,338 sentence SAE corpus, exact SAE and intervention configurations, seeds, analysis code, and summary outputs. We do not commit the trained SAE weight tensor, XGLM activation caches, and some large pair-level intermediate arrays. So, to reproduce it, one needs to re-cache XGLM hidden-state-12 activations and re-train/export the SAE before running the post-canonical and causal stages. The canonical SAE configuration is in `SAE/canonical_sae/sae_canonical/training_config.yaml` and the fixed-threshold export parameters are in `sae_inference_config.json`. The three JSON manifests in `causal_steps/causal_interventions/` keep track of the exact causal settings.

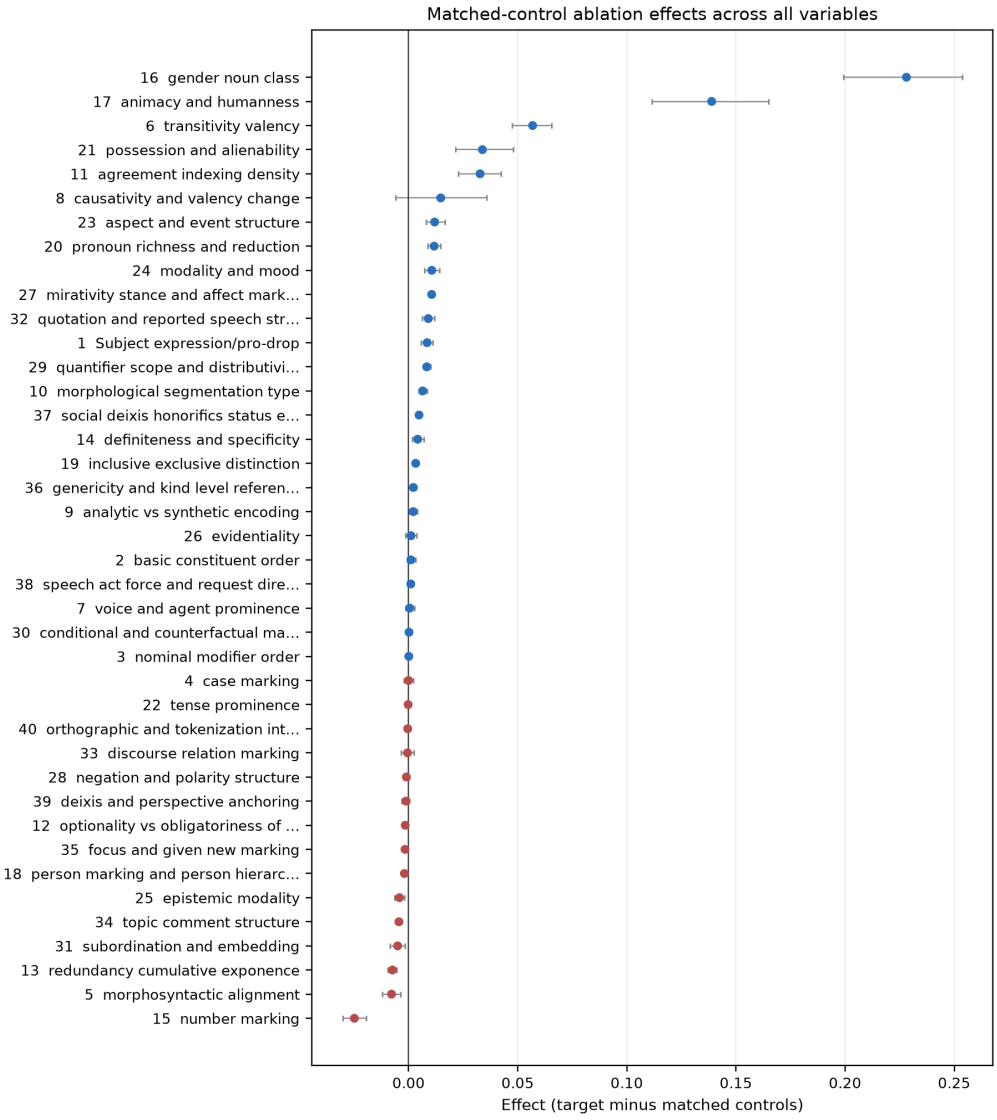

Figure 7: Target-minus-matched-control ablation effects with 95% bootstrap confidence intervals for all variables.

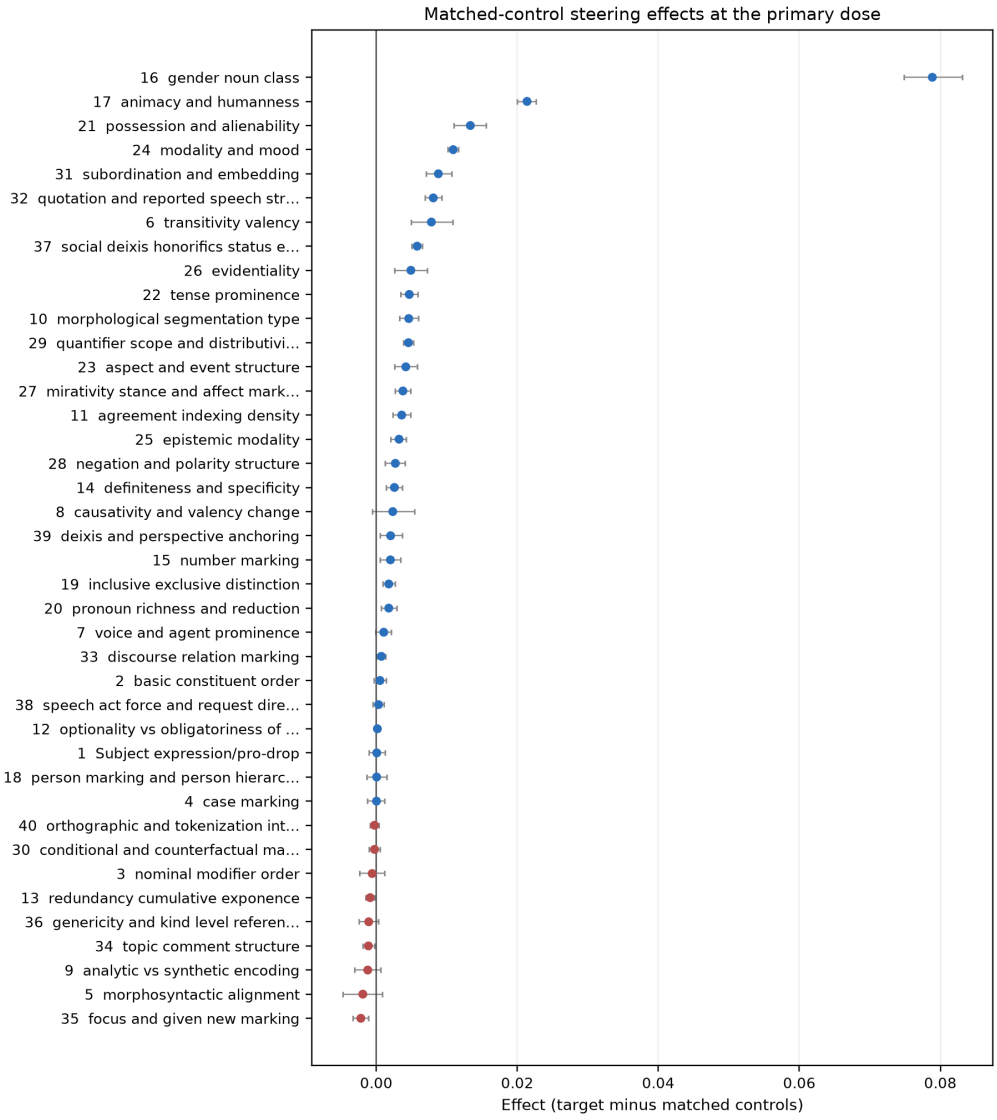

Figure 8: Target-minus-matched-control steering effects at the primary $1\times$ dose, with 95% bootstrap confidence intervals.

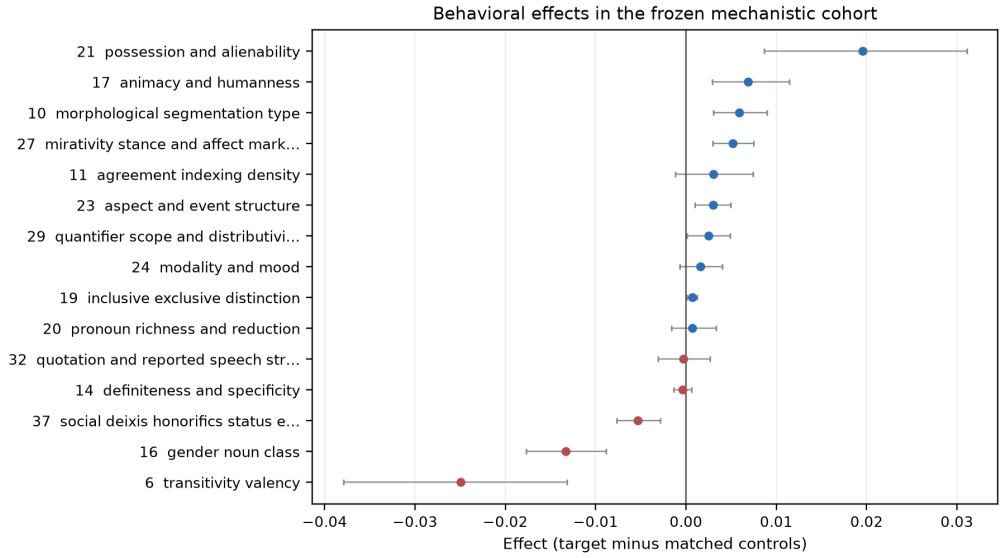

Figure 9: Target-minus-matched-control behavioral effects with 95% bootstrap confidence intervals for the frozen mechanistic cohort.