TenseLoC: Tense Localization and Control in a Multilingual LLM

Ariun-Erdene Tumurchuluun[1,2], **Yusser Al Ghussin**[1,3],
David Mareček[2], **Josef van Genabith**[1,3], **Koel Dutta Chowdhury**[1]
[1]Saarland University, Saarland Informatics Campus
[2]Institute of Formal and Applied Linguistics, Charles University
[3] German Research Center for Artificial Intelligence (DFKI)
artu00001@stud.uni-saarland.de

Abstract

Multilingual language models excel across languages, yet how they internally encode grammatical tense remains largely unclear. We investigate how decoder-only transformers represent, transfer, and control tense across eight typologically diverse languages: English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai. We construct a synthetic tense-annotated dataset and combine probing, causal analysis, feature disentanglement, and model steering to LLaMA-3.1 8B. We show that tense emerges as a distinct signal from early layers and transfers most strongly within the same language family. Causal tracing reveals that attention outputs around layer 16 consistently carry cross-lingually transferable tense information. Leveraging sparse autoencoders in this subspace, we isolate and steer English tense-related features, improving target-tense prediction accuracy by up to 11% in a downstream cloze task[1] .

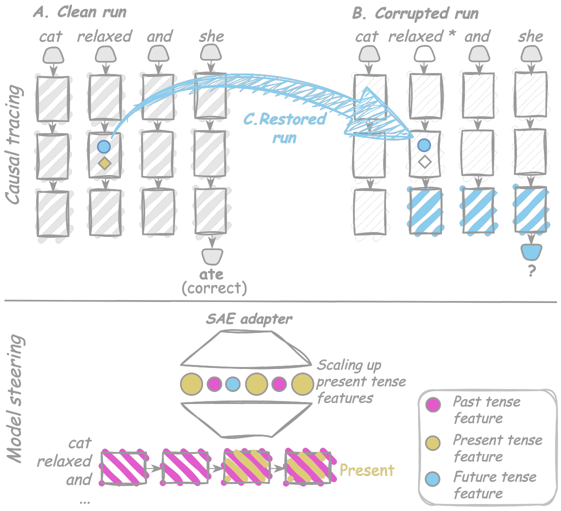


Figure 1: Main findings. Tense resides in a compact, causally active decoder subspace. **(Top)** Causal tracing shows that restoring a small projected subspace recovers tense probabilities across languages. **(Bottom)** SAE-based steering shows that scaling interpretable tense features in the residual stream shifts cloze completions toward the target tense, with minimal impact on other tenses at moderate scaling. These effects hold without temporal adverbials, indicating an internalized tense representation rather than surface-cue reliance.

1 Introduction

Recent transformer-based large language models (LLMs) learn high-dimensional contextual embeddings that yield state-of-the-art performance on multilingual tasks. However, these vectors conflate multiple linguistic features (Jawahar et al., 2019; Tenney et al., 2019; Belinkov, 2022), and as yet it remains unclear how these models represent tense internally. Grammatical tense, how languages mark past, present, and future, is fundamental to accurate human communication, reasoning and natural language processing (NLP) alike. Linguistic theories, from Reichenbach's tripartite model of event, reference, and speech time (Prior, 1967; Kamp, 1968) to later typological surveys, show that languages employ varied morphological and syntactic strategies, morphological inflections (e.g., “-ed”), auxiliaries (e.g., “will”), or adverbial cues (e.g., “Yesterday”), to situate events temporally.

Despite this foundational importance, the ways in which multilingual LLMs internally encode tense remain largely unchartered. Prior probing work (Li and Wisniewski, 2021) shows that morphological cues can predict tense in cross-lingual settings (i.e., French and Chinese), and large-scale studies report that multilingual encoders reliably encode morphological information including tense (Acs et al., 2023). Yet, these studies rely on correlation and cannot show whether the identified subspaces are functionally used by the model.

Along a related line of research, Sparse Autoencoders (SAEs) have been proposed to disentangle monosemantic features, hidden dimensions aligned with human-interpretable concepts (Tem-

[1] We release our data and code publicly at https://github.com/ariunerdenetum/tenseloc

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 243–264
November 8-9, 2025 ©2025 Association for Computational Linguistics

pleton et al., 2024; Gao et al., 2024). If successful, SAEs offer not only interpretability but also control, enabling researchers to steer model outputs by scaling these features (O'Brien et al., 2024; Härle et al., 2024). However, their application to grammatical tense remains under-explored. In particular, it is still unknown whether sparse tense features identified by SAEs, if they exist, are functionally necessary or sufficient for influencing model predictions.

In this work, we present a comprehensive analysis of **LLaMA-3.1 8B to examine how it encodes and uses grammatical tense across typologically diverse languages, and to determine whether these encodings are causally necessary, sufficient, and manipulable via sparse-feature interventions**. We show that targeted interventions on identified subspaces produce predictable changes in generation accuracy, providing a functional (not merely correlational) account of tense representation.

We combine probing, causal tracing, pre-trained SAEs, and targeted residual-stream interventions grounded in mechanistic interpretability to (i) locate tense subspaces, (ii) identify tense-carrying streams and layers (Section 3), (iii) disentangle human-interpretable, monosemantic tense features (Section 4), and (iv) test controllability via feature scaling (Section 5). We show our main findings in Figure 1.

Our contributions are fourfold:

1. We curate and release a multilingual, tense-annotated dataset of simple past, present, and future-tensed sentences in eight languages, with and without explicit temporal adverbials.

2. We show that linear tense signals are consistent throughout layers, generalize within language families in mid-layers, and that a causal bottleneck in attention-output around layer 16 (i.e., mid-layer) mediates functional use.

3. We extract monosemantic SAE features for each tense in English and check if those features are human interpretable, by validating their alignment with surface tense markers (e.g., "did", "will").

4. We manipulate the model's generation output via SAE features, showing that moderate scaling of target-tense features improves English cloze accuracy by up to 11% and transfers to German.

Language	UD Treebank
English (en)	UD_English-EWT $(*)$
German (de)	UD_German-GSD $(*)$
French (fr)	UD_French-GSD $(*)$
Italian (it)	UD_Italian-ISDT $(*)$
Spanish (es)	UD_Spanish-GSD $(*)$
Portuguese (pt)	UD_Portuguese-GSD $(**)$
Hindi (hi)	UD_Hindi-HDTB $(***)$
Thai (th)	UD_Thai-PUD $(***)$

Table 1: UD corpora and curation methods for eight languages. Inflection method is denoted by asterisk ("$*$"): $(*)$ - PatternLite; $(**)$ - mlconjug3; $(***)$ - custom rules.

2 Methods

2.1 Overview of our Approach

By combining probing, causal analysis, and feature disentanglement, we investigate how complex grammatical categories are represented in large multilingual transformers and establish a methodology for precise and interpretable control over temporal generation.

1. **Identification and isolation of tense representation:** We apply layer-wise probes and causal interventions to hidden activations to identify which layers and output streams carry tense signals and which are functionally necessary for tense prediction.

2. **Identifying human-interpretable tense features:** We apply pre-trained SAEs to these tense-bearing activations to disentangle tense to monosemantic features that align with human-readable tense markers (e.g., "did," "will"), and validate these features against probing and causal-tracing results.

3. **Steering tense generation:** We test whether SAE-derived features provide causal leverage by scaling them during inference. Through controlled interventions in the residual stream, we evaluate whether such scaling predictably steers tense generation in downstream cloze task.

2.2 Dataset

We build a controlled, multilingual, tense-annotated dataset from Universal Dependencies (UD) v2 (Consortium, 2021) and focus on languages that differ in morphological tense marking (curation in Table 1 and examples in Table 2). Dataset construction proceeds in two stages: (i) extraction of subject–verb–object clauses (SVO; SOV for Hindi; see Table 6 in Appendix; (ii) generation

Lang.	Tense	no_temp	with_temp
English	Past	We **lacked** sufficient information of an investigation.	*Yesterday*, we **lacked** sufficient information of an investigation.
	Present	We **lack** sufficient information of an investigation.	*Usually*, we **lack** sufficient information of an investigation.
	Future	We will **lack** sufficient information of an investigation.	*Tomorrow*, we will **lack** sufficient information of an investigation.
German	Past	Sie **hatte** eine Länge von Metern.	*Gestern*, **hatte** sie eine Länge von Metern.
	Present	Sie **hat** eine Länge von Metern.	*Normalerweise*, **hat** sie eine Länge von Metern.
	Future	Sie wird eine Länge von Metern **haben**.	*Morgen*, wird sie eine Länge von Metern **haben**.

Table 2: Synthetic sentence examples in past, present, and future tenses for two datasets ("no_temp" and "with_temrep"). Each sentence is generated via subject–verb–object extraction and verb inflection, producing three tense variants per sentence.

of three tense variants per sentence by automatically inflecting the main verb of the sentence using language-specific tools and rules.

Verb conjugation is performed with existing libraries and targeted rule sets: PatternLite (Smedt and Daelemans, 2012) for Romance and Germanic languages, mlconjug3 (Diao, 2023) for Portuguese (to capture irregular forms), and custom rule-based scripts for Hindi[2] and Thai[3] (see Table 1). Each example is annotated with `language`, `tense`, `sentence`, `main_verb`, and `verb_index`. The full corpus comprises 18,580 training and 4,646 test examples. To separate reliance on additional lexical temporal adverbials from internal verbal tense representations, we maintain two parallel splits: `no_temp` (i.e., temporal adverbials removed) and `with_temp`.

We select our target languages (Table 3) to cover typological diversity in tense marking (e.g., morphological inflection, auxiliaries, adverbials) and permit evaluation of within-family transfer. A per-language breakdown of tense-marking strategies and extraction configurations is provided in the Appendix C.

Language	Family	Writing system
English	Germanic	Latin
German	Germanic	Latin
French	Romance	Latin
Italian	Romance	Latin
Portuguese	Romance	Latin
Spanish	Romance	Latin
Hindi	Indo-Aryan	Devanagari
Thai	Kra-Dai	Thai script

Table 3: Target languages, families, and scripts. All languages are Indo-European except Thai.

2.3 Model

We use Meta LLaMA-3.1-8B (Meta, 2024), an autoregressive decoder-only transformer with byte-pair encoding. For each input sentence, we extract the hidden representation of the main-verb token (excluding auxiliaries) at every layer $\ell = \{0, \ldots, 32\}$. Model weights remain frozen for all experiments.

Sparse Autoencoders. We use two distinct pretrained SAEs for our analyses. (i) We employ LLaMA Scope (He et al., 2024) TopK-8x SAEs, which comprise 256 SAE components applied at each layer and stream (residual, attention, MLP), trained on the SlimPajama corpus (He et al., 2024). However, LLaMA Scope exhibits relatively high reconstruction loss, which restricts steering capabilities. Since it was trained on a primarily English dataset, we expect extreme sparse English features, which can limit interpretability and stability when applied cross-lingually[4].

To address these issues, (ii) we train multilingual SAEs of TopK-8x variants (expanding the hidden space by "factor 8") on Wikipedia text from seven languages: English, Spanish, French, Indonesian, Vietnamese, Chinese, and Japanese. Unlike LLaMA Scope's English-centric and highly sparse representations, our multilingual SAEs are designed to achieve lower reconstruction loss while producing sparse, language-agnostic features that enable more reliable cross-lingual comparison and steering within the same model architecture.

3 Identification and isolation of tense representation

We systematically probe how tense is encoded in LLaMA-3.1 8B, examining which layers and com-

[2] https://en.wikibooks.org/wiki/Hindi/Verbs

[3] https://en.wikipedia.org/wiki/Thai_language

[4] https://huggingface.co/Yusser/multilingual_llama3.1-8B_saes/tree/main

ponents represent tense and to what extent. To complement this, we use causal tracing to identify which layers are functionally responsible for carrying and applying tense signals during generation. In the body of the paper, we mainly focus on causal tracing, with additional detailed results on probing reported in Appendix D.

Causal Tracing of Tense Signals. As a preliminary experiment to ascertain if the tense representation is linearly decodable, we perform linear probing (Hewitt and Manning, 2019; Tenney et al., 2019; Chi et al., 2020) and find that tense representation resides throughout all the layers emerging from early layers and most robust in later layers (Figure 7 in Appendix D).

However, probing alone only shows where information is encoded and thus demonstrates correlation rather than causal influence; to address this, we test causality by intervening in intermediate activations to verify that the representations in question directly drive syntactic tense production. We adopt the causal tracing method introduced by Meng et al. (2022), implementing layerwise intervention and patching in our target model using the Pyvene library (Wu et al., 2024). In causal tracing, we care about the activations (i.e., hidden signals) as they travel through the network, which in our case, are tense signals.

Prompting. Each of our trials uses a one-shot prompt consisting of (i) a full sentence in the target tense and (ii) a truncated version of that sentence ending just before the verb:

Template

<partial-X-tense-ending-before-verb>

Example

Lily the cat relaxed on the mat and she ate an apple.
Lily the cat relaxed on the mat and she

The truncated sentence is fed to the model, forcing prediction of the verb and exposing how tense is internally represented. Since verbs may span multiple subtokens, we compute log-probabilities until the full sequence is generated.

Subspace Intervention. On this prompting setup, we apply a clean–corrupt–restore cycle at each transformer block to identify subspaces critical for tense encoding. We intervene across four activation streams S: attention output, MLP activation, MLP output, and post-residual block output. (i) In the **clean** step, we record the probability p_{gold} of the gold next token. (ii) In the **corrupt** step, Gaussian noise $\epsilon \sim \mathcal{N}(0, \delta^2 I)$ is injected into tense-bearing embeddings at layer 0. (iii) In the **restore** step, noisy activations at layer ℓ and stream S are overwritten with their clean counterparts.

We measure recovery as

$$\Delta p^{\ell,S}_{\text{restored}} = p^{\ell,S}_{\text{restored}} - p_{\text{corrupt}}$$

Averaging over prompt variants, noise seeds (Colas et al., 2018), and languages yields a recovery curve with Standard Error of the Mean (±SEM; Wooldridge, 2023) as a function of layer ℓ. Following Meng et al. (2022), we report the indirect effect, i.e., the change in output probability when a single state is restored. More details are in Appendix E.

Results. By corrupting and selectively restoring hidden-state activations, we observe that across the evaluated languages and tenses the **attention-output** stream shows a clear recovery peak around layer 16 (Figure 2); restoring the projected subspace at this layer yields a measurable increase in target-tense probability. This localizes mid-layers as **functionally necessary** for tense prediction, consistent with our preliminary probing results (Figure 7 Appendix D). Per-stream breakdowns are shown in Figure 12 in Appendix F.

Layer wise analysis indicates a processing progression: tense information emerges in the MLP activations near layer 15, is read by attention in layers 15-18, and then is propagated forward (Figure 13 in Appendix F). Recovery magnitudes in the MLP stream are smaller than in attention but indicate a measurable tense signal. This pattern is consistent with prior layer-wise intervention studies on other phenomena (e.g., factual knowledge in Meng et al. (2022)).

4 Identifying human-interpretable tense features.

Having identified critical layers ℓ^* for tense representation, we next ask whether features extracted by SAEs can be used to steer the model's outputs. Specifically, we test whether activating or inhibiting these features systematically shifts the predicted verb tense. This allows us to evaluate not only the interpretability of SAE features but also their causal influence on generation. To this end, we use SAEs to discover latent features in the model's hidden states that align with grammatical

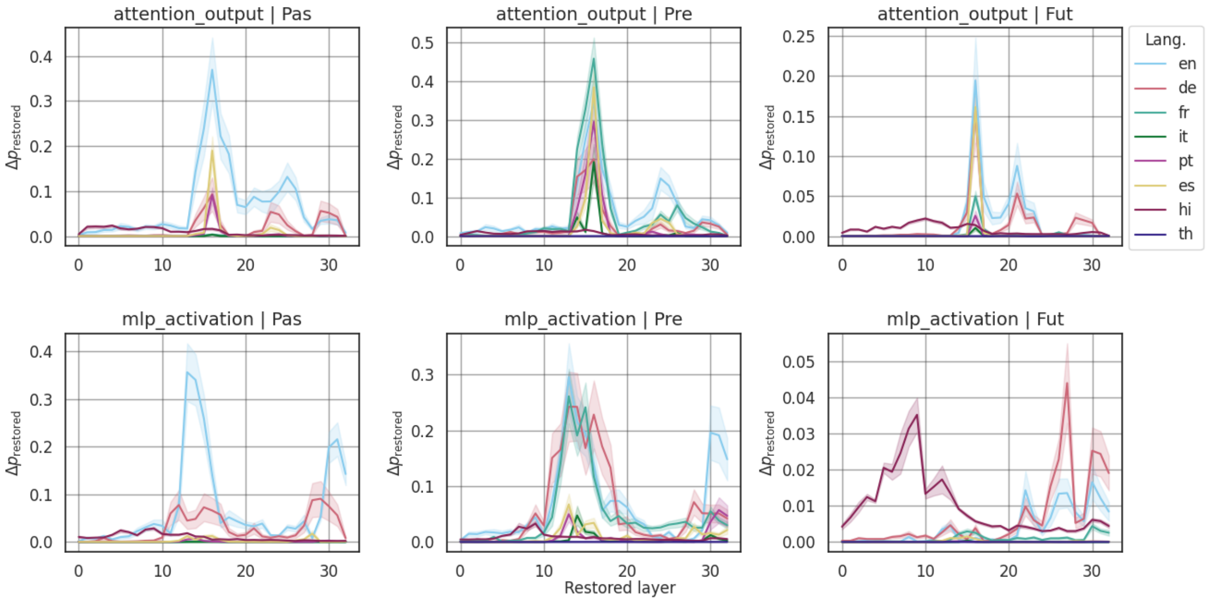


Figure 2: layerwise recovery curves $\Delta p_{\text{restored}}(\ell, S)$ in each language, faceted by stream and tense in attention output and MLP activation. High values indicate that restoring the corrupted token activations at that layer and stream most effectively recovers the correct verb-tense prediction.

tense in English. Our goals are twofold: (i) to validate that SAE-derived features are consistent with the probing and causal-tracing results (Figures 2 in Section 3 and 7 in Appendix D), and (ii) to identify monosemantic tense features that reliably map to tense labels and localize to the critical transformer layers ℓ^* identified earlier. Experiments are conducted on our curated datasets `no_temp` and `with_temp`.

Hidden-state extraction. Each sentence is fed individually through the original LLaMA model via the HookedTransformer interface in the sae_lens library (Bloom et al., 2024). We capture the hidden activations at the attention, MLP, and residual output streams for critical layers 15-31 (Figure 2 in Section 3). We extract the activation vector corresponding to the main verb token in each sentence.

SAE inference. We feed the hidden states to our trained multilingual SAEs and the LLaMA Scope SAEs (He et al., 2024) and obtain feature activations and corresponding decoder weights. We compute the reconstruction mean-squared error (MSE) on activations to identify which SAE best compresses the original signal with minimal loss (Figure 3). This helps us select the SAE whose low MSE guarantees fidelity to the model's internal representations (Shu et al., 2025; Engels et al., 2025).

However, since SAEs are trained with two loss functions for reconstruction and sparsity, there is a trade between having sparse monosemantic features and the steerabilty of the SAE (Bayat et al., 2025; Härle et al., 2024). Bayat et al. (2025) address this problem by adding a reconstruction error term to the SAE output while we propose training a new model that preforms better on reconstruction loss.

To assess how well each SAE isolates tense, we perform clustering on the encoder outputs at each layer and calculate the V-measure (v_ℓ) (Rosenberg and Hirschberg, 2007) against true tense labels[5]. We flatten the feature activations across all examples, use K-means with $k = 3$ (i.e., assumes 3 clusters and equal weight per class) for the past, present, and future tenses, and compute v_ℓ for each layer ℓ (Figure 4).

Extracting tense features at critical layers ℓ^*. After determining the optimal ℓ^*, we shortlist candidate features by intersecting two rankings. First, we rank each latent dimension based on the cosine similarity with static token embeddings from the model's unembedding matrix (He et al., 2024), which relies on the linear representation hypothesis (Nanda, 2023a; Bereska and Gavves, 2024). Second, we train *one-vs-rest linear probes* with LogisticRegression from scikit-learn on the encoder

[5] V-measure quantifies clustering quality as the harmonic mean of homogeneity (i.e., each cluster contains only members of a single class) and completeness (i.e., all members of a class are assigned to the same cluster).

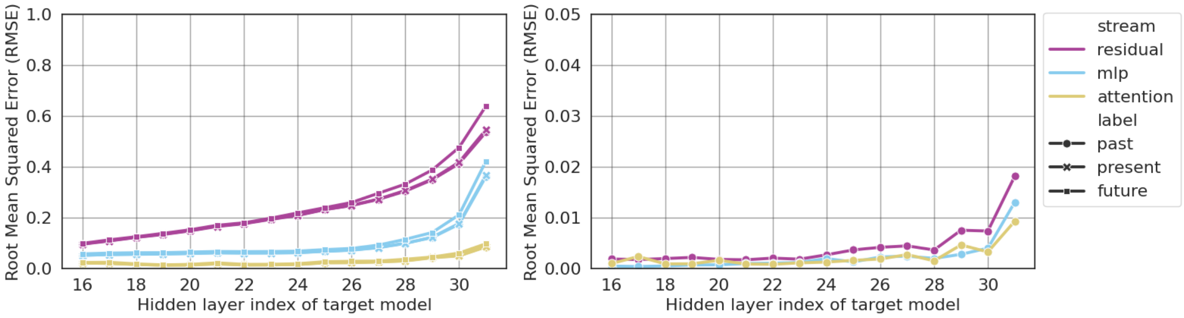


Figure 3: Layerwise MSE trends. **Left**: LLaMA Scope error grows. **Right**: Flat, near-zero error with Multilingual SAE.

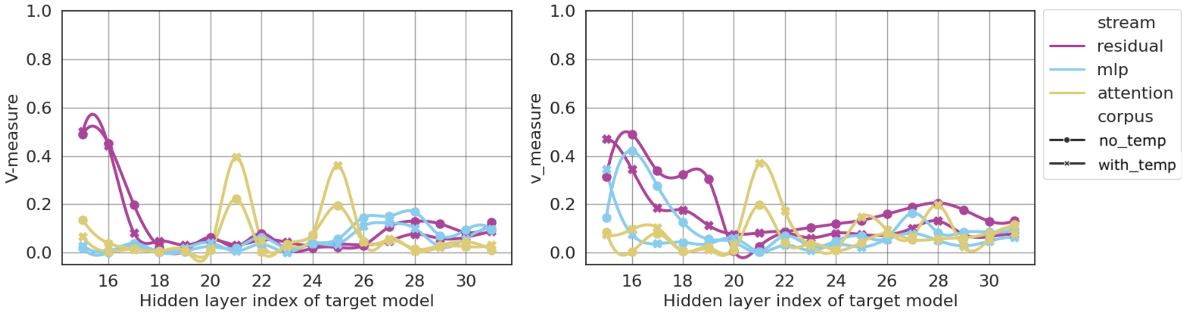


Figure 4: V-measure by layer for SAEs on corpora with/without temporal cues. Tense features are most distinct at layers 15–16.

outputs to predict temporal states.

Features are then ranked by their absolute probe weights to identify those that align with tense tokens and drive tense classification. However, this intersection might overlook weaker valid features or retain spuriously correlated ones if noise is agreed upon in both rankings. To address this, we conduct additional experiments in the steering experiment detailed in Section 5.

Results. Our multilingual SAE attains low mean-squared reconstruction error across layers (Figure 3), supporting $\ell^* \approx \{15, 16\}$ as the tense-critical layers where tense distinctions are most sharply encoded. This underscores the importance of intermediate layers for tense representation and aligns with our previous results in Figure 2 in Section 3. The intersected SAE feature set corresponds to human-readable tense markers (e.g., "did," "does," "will"; Figure 5), validating the interpretability of these features and their suitability for downstream steering. Additional visualizations appear in Figures 14 and 15 (Appendix G).

Visualizing high-dimensional SAE activations at ℓ^* via UMAP provides an intuitive snapshot of how the model's latent space isolates tense information (Figure 16 in Appendix G).

Baselines	
A	Original model, no adapter.
B1	LLaMA Scope SAE adapter applied at ℓ^*, with $\alpha = 1.0$ (i.e., no scaling).
B2	Our Multilingual SAE adapter applied at ℓ^*, with $\alpha = 1.0$ (i.e., no scaling).
Steering	
Excitation	Multiply each selected feature f by $\alpha > 1.0$ (positive intervention).
Inhibition	Multiply each selected feature f by $\alpha < 1.0$ (negative intervention).

Table 4: Definitions of baselines and feature-steering settings.

We find that in both SAE frameworks, future tense forms a distinct cluster while having more subtle distinctions between past and present tenses. This pattern suggests that SAEs capture a stronger, more uniform signal for the future tense than for the more subtle distinctions between past and present forms.

5 Steering Tense Generation

After identifying tense-sensitive features, we test whether these features can be used directly to control model behavior. Following an adapter-based steering paradigm Kissane et al. (2024b), we integrate SAE-derived "tense axes" into the residual stream and scale them during generation (McGrath

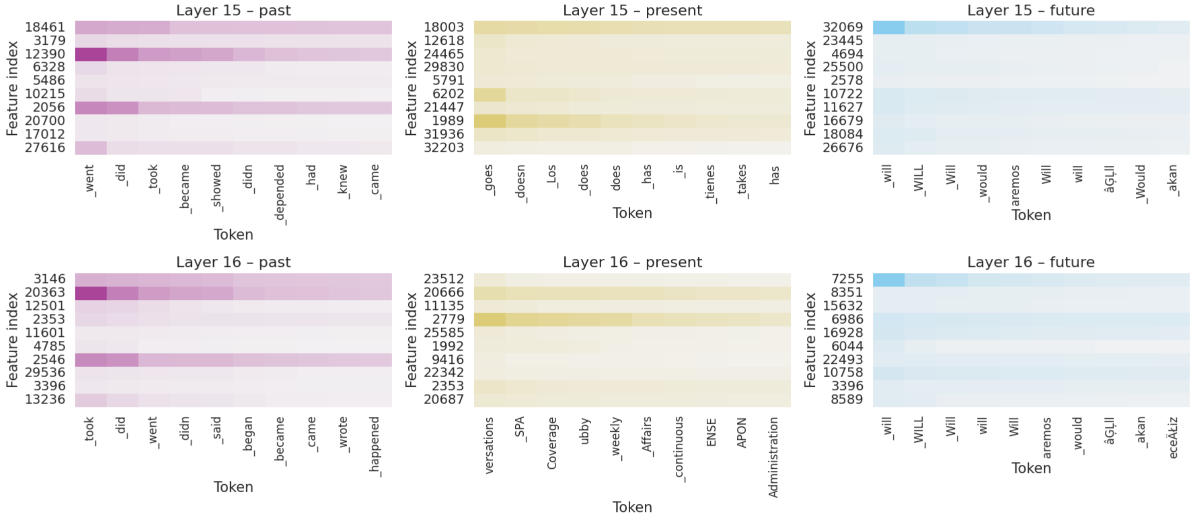


Figure 5: Token–feature heatmaps from the model's input embeddings at $\ell^* = 15$ (LLaMA Scope). Rows correspond to features, columns show the top ten tokens by cosine similarity.

.

et al., 2024; O'Brien et al., 2024; Härle et al., 2024). This mirrors prior steering work on syntactic and factual pathways and enables fine-grained tense control without retraining.

Specifically, we evaluate the causal impact of tense features by steering LLaMA-3.1 8B's hidden activations in a cloze task, which is explained below. Using feature scaling, we multiply selected latent dimensions by a factor α at critical layers ℓ^*, with $\alpha > 1$ for excitation and $\alpha < 1$ for inhibition on the cloze task.

Cloze task. We use a **cloze fill-in-the-blank** evaluation. Each prompt consists of a temporal cue (e.g., "Yesterday"), a sentence with a missing verb, three verb-form options (one per tense), and an answer placeholder. For instance:

Example

```
Yesterday, the dog ___ at the mailman.
A) barks
B) barked
C) will bark
Answer:
```

The model must output the correct option ("A", "B", or "C"). For each tense, we construct a balanced development set of 30 prompts and a test set of 500. Prompts pair diverse subjects/objects (e.g., "I", "we", "the mailman") with base verbs conjugated (including irregulars) into target tenses using PatternLite (Smedt and Daelemans, 2012); correct option order is randomized. We run the task in English and German and report accuracy.

Steering procedure. We compare three baselines and multiple steering configurations (Table 4).

SAE adapter combinations. We explore SAE adapters at individual layers (e.g., $\ell^* = 15$ then $\ell^* = 16$ separately), and both layers combined. For each combination, we apply the same feature set (Table 9 in Appendix H) and scaling α across all prompts in one run. This setup enables us to observe how scaling the SAE features, either individually or jointly across layers, affects the model's predictions in the downstream task.

Feature selection and scaling. Our SAE observation analysis (Section 4) yields a pool of features, but we need to ensure that these human-interpretable features are functional in the downstream task. Thus, we perform a grid search on the dev set to identify which features and α values work best for each tense. Specifically, for each candidate feature f in the combined pool, we run steering on the dev-prompts for each label, record the change in accuracy relative to baselines, and retain only those feature combinations that improve the target-label accuracy. This fine-grained search allows us to isolate the most effective features and scaling factors before the final test-set evaluation. The features determined in this fashion are listed in Table 9 in Appendix H.

Results. We evaluate cloze-task accuracy on the test set and find that moderate excitation ($\alpha = 5.0$) of tense features reliably enhances correct-tense

Language	Target tense	A	B1	B2	15			16			Both		
		-	1.0	1.0	2.0	5.0	0.1	2.0	5.0	0.1	2.0	5.0	0.1
English	Past	0.81	0.13	0.81	0.82	0.84	0.80	0.80	0.82	0.81	0.80	0.80	0.81
	Present	0.39	0.09	0.39	0.41	0.36	0.39	0.42	0.48	0.38	0.50	0.38	0.37
	Future	0.76	0.14	0.77	0.80	0.85	0.74	0.78	0.81	0.77	0.81	0.77	0.75
German	Past	0.63	-	0.63	0.64	0.68	0.63	0.65	0.65	0.67	0.62	0.63	0.63
	Present	0.60	-	0.61	0.62	0.66	0.64	0.71	0.66	0.72	0.60	0.60	0.57
	Future	0.44	-	0.43	0.43	0.44	0.42	0.38	0.43	0.36	0.44	0.45	0.45

Table 5: Test-set steering results. Baselines: A = original; B1 = LLaMA Scope SAEs (15–16); B2 = Multilingual SAEs (15–16). $F_{\ell^*} \in \{15, 16, \text{Both}\}$ is the Multilingual SAE hook layer(s); subheader shows α ($\alpha > 1$: excitation). (Highlighted) cells mark excitation that outperforms the baselines. In inhibition settings, lower accuracy indicates successful downward control.

predictions. Inhibition produces only small accuracy reductions, suggesting the model compensates for partial suppression via redundant or alternative features. A likely mechanistic explanation is that tense encoding is partially distributed and overlapping, so that inhibiting only a subset of target features may not have an effect as intended (McGrath et al., 2024). In line with McGrath et al. (2024), positive interventions are more effective than negative ones.

English tense features transfer to German for past and present (Table 5), but not for future tense. This finding suggests partial cross-lingual alignment and the presence of language-specific attention heads. We hypothesize that the observed English → German non-transfer primarily reflects distinct syntactic encodings (e.g., German verb-second and verb-final patterns) that alter where tense cues are represented across layers and components. The layer-wise causal differences reported in Figure 2, Section 3 align with this interpretation.

We do not rule out potential effects of tokenization or corpus frequency; confirming whether syntax alone explains the pattern will require targeted tests such as tokenization normalization, auxiliary alignment interventions, and controlled frequency experiments, which we leave for future work.

Moreover, since SAE features can partially overlap semantically, interventions on one tense may also influence others. We present these cross-label effects in Tables 10 and 11 in Appendix H.

6 Conclusion

We present a four-phase diagnostic pipeline: probing, causal tracing, SAE disentanglement, and steering that links where tense information is linearly readable in latent representations to where it is functionally necessary and controllable. Linear probes show that LLaMA-3.1 8B (Meta, 2024) internally represents simple past, present, and future tenses in low-dimensional subspaces that are detectable across layers; with crosslingual transfer peaking in layers 20 to 30, suggesting a language agnostic encoding. Causal interventions (Meng et al., 2022) localize a functionally necessary subspace at around layers 15-16, primarily within the attention stream (with contributions from MLP activations and outputs), and restoring this small subspace recovers tense probability.

Applying SAEs (Kissane et al., 2024b; O'Brien et al., 2024; Härle et al., 2024) to activations at layers 15-16 yields monosemantic tense features that align with human-readable tense markers. Scaling these features in the residual stream systematically shifts cloze completions toward the target tense, improving correct-tense accuracy by up to 11% points with modest degradation. Crucially, the effect persists even without temporal adverbs (for example, "yesterday"), showing that the model internally encodes tense rather than relying on surface cues. English derived features transfer to German past and present but not future tense, suggesting that the model captures an abstract crosslingual temporal structure, though some future constructions may remain language specific or data limited.

To our knowledge, this is the first evidence in a multilingual LLM of a causally active, language agnostic tense subspace whose disentangled, interpretable features can steer generation. The finding holds across eight languages for simple tense forms, but broader generalization to richer aspectual patterns, other model families, and naturalistic contexts remains open. Future work should extend this framework to more complex temporal systems and finer grained circuit level analyses of cross-lingual temporal representation.

Limitations

This study operates in a controlled diagnostic setting that enables causal intervention but may limit generalization. Our experiments rely on automatically inflected sentences from UD treebanks, which simplify discourse context and may not mirror natural tense use. Rule-based inflections for Hindi and Thai add minor noise. We analyze only one decoder-family model and focus on basic tense forms—past, present, and future. While our interventions reveal clear mechanistic signals, we do not claim generalization to richer discourse contexts, morphologically complex or low-resource languages, other architectures, or compound aspectual tenses.

Future work should extend to human-annotated, naturalistic corpora with explicit tense labels, replicate analyses across architectures and tokenizers, and apply finer-grained causal probes and steering methods. Evaluating longer contexts and downstream tasks will further test whether the recovered features capture robust, generalizable temporal representations.

Acknowledgments

YG is supported by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project TRAILS (01IW24005). DM is supported by the project "Human-centred AI for a Sustainable and Adaptive Society" (reg. no.: CZ.02.01.01/00/23_025/0008691), co-funded by the European Union. KDC is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1102 Information Density and Linguistic Encoding.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Judit Acs, Endre Hamerlik, Roy Schwartz, Noah A. Smith, and Andras Kornai. 2023. Morphosyntactic probing of multilingual bert models. *Natural Language Engineering*, 30(4):753–792.

Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. 2025. Steering large language model activations in sparse spaces. *arXiv preprint arXiv:2503.00177*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety – a review. *Preprint*, arXiv:2404.14082.

Joseph Bloom, Curt Tigges, Anthony Duong, and David Chanin. 2024. Saelens. `https://github.com/jbloomAus/SAELens`.

Trenton Bricken and 1 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

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

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. 2018. How many random seeds? statistical power analysis in deep reinforcement learning experiments. *Preprint*, arXiv:1806.08295.

Bernard Comrie. 1985. *Tense*. Cambridge University Press, Cambridge.

Universal Dependencies Consortium. 2021. Universal dependencies v2. *Proceedings of the LREC 2020 Workshop on Universal Dependencies*.

Östen Dahl and Viveka Velupillai. 2011. Perfective/imperfective aspect. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*, chapter 65A. Max Planck Digital Library.

Sekou Diao. 2023. mlconjug3. *GitHub. Note: https://github.com/Ars-Linguistica/mlconjug3 Cited by*.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Joshua Engels, Logan Riggs, and Max Tegmark. 2025. Decomposing the dark matter of sparse autoencoders. *Preprint*, arXiv:2410.14670.

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

Ruben Härle, Felix Friedrich, Manuel Brack, Björn Deiseroth, Patrick Schramowski, and Kristian Kersting. 2024. Scar: Sparse conditioned autoencoders for concept detection and steering in llms. *arXiv preprint arXiv:2411.07122*.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, and 1 others. 2024. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Johan Anthony Willem Kamp. 1968. *Tense Logic and the Theory of Linear Order*. University of California, Los Angeles, CA, USA.

Connor Kissane, Robert Krzyzanowski, Joseph Isaac Bloom, Arthur Conmy, and Neel Nanda. 2024a. Interpreting attention layer outputs with sparse autoencoders. *Preprint*, arXiv:2406.17759.

Connor Kissane, robertzk, Arthur Conmy, and Neel Nanda. 2024b. Sparse autoencoders work on attention layer outputs. `https://www.lesswrong.com/posts/DtdzGwFh9dCfsekZZ/sparse-autoencoders-work-on-attention-layer-outputs`. AI Alignment forum post.

Bingzhi Li and Guillaume Wisniewski. 2021. Are neural networks extracting linguistic properties or memorizing training data? an observation with a multilingual probe for predicting tense. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3080–3089, Online. Association for Computational Linguistics.

Thomas McGrath, Daniel Balsam, Myra Deng, and Eric Ho. 2024. Understanding and Steering Llama 3 with Sparse Autoencoders. `https://www.goodfire.ai/papers/understanding-and-steering-llama-3`. Accessed: 2025-07-08.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *NeurIPS*.

Meta. 2024. Introducing llama 3.1: Our most capable models to date. `https://ai.meta.com/blog/meta-llama-3-1/`. Meta AI's Blog.

Eleni Metheniti, Tim Van De Cruys, and Nabil Hathout. 2022. About time: Do transformers learn temporal verbal aspect? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 88–101, Dublin, Ireland. Association for Computational Linguistics.

Neel Nanda. 2023a. 200 cop in mi: Techniques, tooling and automation. *Neel Nanda's Blog*. 30.

Neel Nanda. 2023b. 200 cop in mi: Techniques, tooling and automation. `https://www.lesswrong.com/posts/btasQF7wiCYPsr5qw/200-cop-in-mi-techniques-tooling-and-automation`. Neel Nanda's Blog.

Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Blake Bullwinkel, 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*.

Chris Olah. 2023. Superposition is not just neuron polysemanticity. Alignment Forum.

Terrence Parsons. 2002. Tense and aspect. Stanford Encyclopedia of Philosophy. `https://plato.stanford.edu/entries/tense-aspect/`.

Barbara H. Partee. 1973. Some structural analogies between tenses and pronouns in english. *Journal of Philosophy*, 70(18):601–609.

Arthur Prior. 1967. *Past, Present and Future*. Clarendon P., Oxford,.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic. Association for Computational Linguistics.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. 2025. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *Preprint*, arXiv:2503.05613.

Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *Journal of Machine Learning Research*, 13(66):2063–2067.

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

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *ICLR (Poster)*. OpenReview.net.

Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2012. Tempeval-3: Evaluating events, time expressions, and temporal relations. *Preprint*, arXiv:1206.5333.

Jeffrey M. Wooldridge. 2023. What is a standard error? (and how should we compute it?). *Journal of Econometrics*, 237(2, Part A):105517.

Zhengxuan Wu, Atticus Geiger, Aryaman Arora, Jing Huang, Zheng Wang, Noah Goodman, Christopher Manning, and Christopher Potts. 2024. pyvene: A library for understanding and improving PyTorch models via interventions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 158–165, Mexico City, Mexico. Association for Computational Linguistics.

Salah Yahiaoui and Iana Atanassova. 2023. TimeTank: A Corpus of Sentences Annotated with TimeInfo for Temporal Data. Dataset.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *ICLR*. OpenReview.net.

A Ethical Considerations

This work investigates how multilingual large language models represent and transfer grammatical tense across languages through causal and interpretability analyses. All experiments were conducted on open-weight models and publicly available datasets, including synthetically generated, tense-annotated corpora derived from existing treebanks. No human or private data were used. All model and data artifacts were used in full compliance with their respective licenses.

B Related work

This section addresses our study within interconnected research areas: linguistic theory of tense, interpretability methods used in our work, and current progress of linguistic analysis in mechanistic interpretability.

B.1 Tense in linguistics

Early linguistic work characterizes tense as the grammatical marking that locates an event in time. From *a syntactic perspective*, tense operates as a feature on a clause head that triggers morphological inflections (Partee, 1973). In contrast, *semantic frameworks* treat tense morphemes as operators that shift a reference time relative to the utterance time (Dahl and Velupillai, 2011). A further distinction arises between absolute tense, which ties events to the moment of speaking (e.g., simple past vs. present), and relative tense, which relates one event time to another (e.g., perfect or pluperfect) (Comrie, 1985).

B.2 Mechanistic interpretability

Superposition hypothesis. Superposition posits that internal vectors store more distinct features than their dimensionality by overlapping feature directions. Overlap causes crosstalk when recovering a single feature, because directions are not all orthogonal. This cost is acceptable when features are sparse (i.e., few active features per input) and when nonlinear readouts or learned decoders excite true signals and inhibit overlap (Bereska and Gavves, 2024; Olah, 2023).

Linear representation hypothesis. This hypothesis proposes that neural networks often depict high-level features as linear trajectories within the activation space (Bereska and Gavves, 2024). Linear representation can ease the comprehension and adjustment of neural network representations (Nanda, 2023b).

Relevant studies. Mechanistic interpretability has progressed through complementary observation and intervention methods. Linear and structural probes (Tenney et al., 2019; Hewitt and Manning, 2019; Jawahar et al., 2019) reveal that transformer layers encode syntactic and semantic categories. Multilingual probing of mBERT and XLM-R shows recoverable tense signals across dozens of languages (Acs et al., 2023; Li and Wisniewski, 2021). However, high-capacity probes risk spurious correlations and probing accuracy can be misleading (Hewitt and Manning, 2019; Belinkov, 2022). Consistent with Tenney et al. (2019) and Jawahar et al. (2019), we expect syntax is represented in early layers and higher-level abstractions in mid layers. Temporal semantics research—timeline inference and event ordering corpora (UzZaman et al., 2012; Yahiaoui and Atanassova, 2023)—and aspectual probes (Metheniti et al., 2022) target factual time relations rather than internal tense morphology. Causal tracing techniques (Abnar and Zuidema, 2020; Meng et al., 2022; Zhang and Nanda, 2024), have begun to link hidden activations to model behaviors, but have not yet been applied to tense. Finally, SAEs demonstrate that enforcing sparsity extracts monosemantic units for linguistic features (Bricken et al., 2023), offering a promising path to disentangle tense from other representations. Unlike prior work on encoders (e.g., mBERT probes), our work unifies these strands—probing, causal analysis, and SAE disentanglement—to fill the current gap in understanding and controlling tense in decoder-only multilingual transformers.

C Tense Typologies

We survey the target languages in terms of family, script, word order, and tense marking strategies:

English (Indo-European, Germanic; Latin alphabet; SVO): English has a strong past/non-past distinction (Parsons, 2002). The simple past is marked by the suffix "-ed" (i.e., plus irregular forms), and the present is unmarked or marked by "-s" for a third person. Future time is typically expressed periphrastically using auxiliaries (e.g., "will", "going to") rather than an inflection (Parsons, 2002). Thus, English encodes tense morphologically for past and present but uses modal auxiliaries for future.

German (Indo-European, Germanic; Latin alphabet; Verb-Second order): German also marks tense morphologically. Present-tense verb forms (e.g., *geht* ("war")) contrast with a simple past (i.e., Präteritum) typically marked by suffixes or vowel ablaut (e.g., *ging* ("went")). German uses auxiliaries (e.g., "werden", "sein", "haben") to form periphrastic tenses, including the future and perfect. In subordinate clauses, it can use *wird gehen* ("will go") as a future. Overall, German has a two-way distinction (i.e., present vs. past) with optional future auxiliaries.

French (Indo-European, Romance; Latin alphabet; SVO): French has rich tense inflection on verbs. The present tense (e.g., *parle* ("speaks")) is marked, as is the simple past (i.e., passé simple, e.g. *parla*) and imperfect (e.g., *parlait*). The "passé composé" uses "avoir/être" + past participle to express past. French also has a true future suffix (e.g., "-ra", as in *parlera* ("will speak")) (Dryer and Haspelmath, 2013). Thus, tense is marked by a variety of suffixes and auxiliary constructions.

Italian (Indo-European, Romance; Latin alphabet; SVO): Italian, like other Romance languages, uses inflectional suffixes to mark tense. For example, "-ò" and "-ai" in *parlerò* ("I will speak") signal future tense, while "-ai" or "-i" mark past forms. The present tense is marked by suffixes on the verb stem (e.g., "-o", "-i", "-a", "-iamo", etc.). Compound tenses (i.e., passato prossimo) use "avere/essere" + participle for past reference. Thus, Italian distinguishes past, present, and future with a mix of suffixal and auxiliary marking.

Portuguese (Indo-European, Romance; Latin alphabet; SVO): Portuguese similarly marks tense on verbs. Present tense forms (e.g., *falo* ("speak")) contrast with a past preterite (e.g., *falei*) and a future suffix (e.g., *falarei*). There is also an imperfect (e.g., *falava*). The future tense can be formed analytically (i.e., using auxiliary "ir" + infinitive) or synthetically (i.e., "-rei" endings). Overall, Portuguese verb morphology encodes multiple tense distinctions.

Spanish (Indo-European, Romance; Latin alphabet; SVO): Spanish marks tense on verbs with multiple inflections. The simple past (i.e., preterite, e.g., *hablé* ("speak")) and imperfect (e.g., *hablaba*) are distinct suffixes, as are present (e.g., *hablo*) and future (e.g., *hablaré*) forms (Dryer and Haspelmath, 2013). The future tense is a suffix (i.e., usually "-ré") attached to the infinitive. Compound tenses use auxiliaries (i.e., "haber" + participle). Overall, Spanish has separate affixes for past, present, and future on the verb.

Hindi (Indo-European, Indo-Aryan; Devanagari script; SOV): Hindi's tenses are typically marked by verb inflections and auxiliaries. The simple present and past tenses are distinguished by different participial stems and agreement. For example, "-taa/-ti" suffixes for present continuous vs. "-yaa" participles for perfective past (e.g., *khaataa/khaatii* ("eating"), *khaayaa/khaayi* ("ate")). Hindi does not have a grammatical future inflection on the verb itself. Instead, periphrastic futures are formed with modal auxiliaries (e.g., *hoga* ("will be")) or with the verb *nikalnaa* ("to leave") implying future intent. Thus, Hindi effectively contrasts past vs. non-past, with future marked by particles or context.

Thai (Kra-Dai, Tai branch; Thai script; SVO): Thai is often described as a tenseless language. Thai verbs do not inflect for tense. Instead, time reference is conveyed by aspect markers and temporal adverbs. For example, particles such as *láew*

Language	NP Modifiers	VP Auxiliaries	PP Modifiers
en, de, fr, it, pt, es	det, amod, compound, poss, nummod	aux, aux:pass, compound:prt	det, amod, compound
hi	det, amod, compound, poss, nummod	aux, aux:pass, compound:prt	det, amod, compound
th	det, amod, compound, nummod	aux, aux:pass, compound:prt	det, amod, compound

Table 6: Simplified dependency-modifier configuration used for NP, VP, and PP extraction per language.

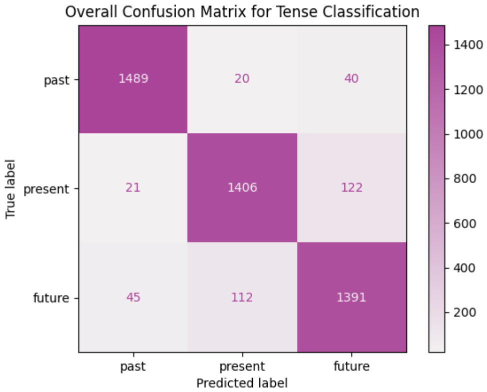
(“already”) or *jà* (“will”) and context words (e.g., “yesterday” or “tomorrow”) indicate past or future tense. Typologically, Thai lacks any inflectional future tense. It falls in the Southeast Asian area that does not mark future morphologically (Dryer and Haspelmath, 2013).

D Preliminary Linear Probing

We check if tense information is stored linearly by training classifiers on the model’s hidden states. We follow the probing framework of Hewitt and Manning (2019) for layer-wise analysis and the multilingual transfer evaluation of Chi et al. (2020). We conduct a series of experiments to assess internal tense representation after having observed strong diagonal accuracy from final layer (Figure 6). We utilize layerwise probes, where we train a separate probe for each layer on the dataset labeled as “no_temp” with a learning rate set at 1e-3.

$$\hat{y} = \mathrm{softmax}(W_\ell h_\ell(x)+b),\ \mathcal{L} = \mathrm{H}(\hat{y}, y)+\lambda\|W\|_1$$

where $y \in$ {past, present, future}, x is the main verb in the input and $\lambda \in \{0.01, 0.003, 0.001\}$.



	past	present	future
past	1489	20	40
present	21	1406	122
future	45	112	1391

Figure 6: Confusion matrix of classification performance at the final layer of Llama-3.1 8B. Rows are true tense labels, and columns are predicted labels. Strong diagonal values relative to the off-diagonal values confirm linear separability. It is measured on the main verb token (i.e., can be multiple tokens) embeddings.

D.1 Cross-lingual transfer

We adapt the layerwise paradigm to assess language-agnostic encoding by following the framework established by Chi et al. (2020) conducting two strategies: direct and hold-one-out transfer. This approach tests whether grammatical tense is encoded in a language-agnostic subspace or vary by language morphology. High transfer accuracy indicates a shared tense representation, while low accuracy suggests language-specific patterns. In the direct transfer approach, we train our model on one language and then test it on other languages, and in the hold-one-out method, we train the model on seven other languages while reserving one language for testing.

D.2 Results

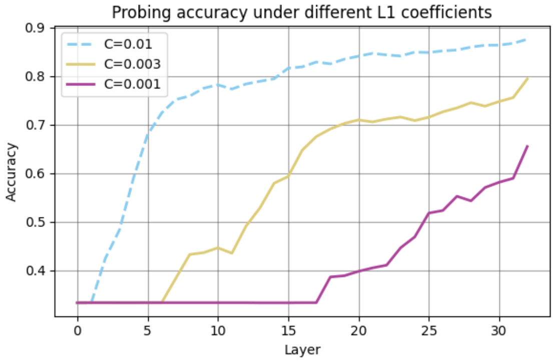


Figure 7: Probes trained with L1 regularization ($\lambda = 0.01, 0.003, 0.001$) show that tense is recoverable from early layers under weak regularization. Stronger penalties delay emergence to later layers, indicating that tense develops in early layers but strengthens in deeper ones, aligning with previous findings on syntactic feature emergence (Kissane et al., 2024a; Tenney et al., 2019). Early detection may also relate to morphology.

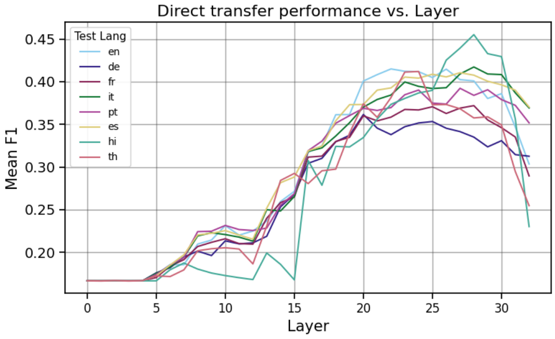


Figure 8: Direct-transfer performance across languages.

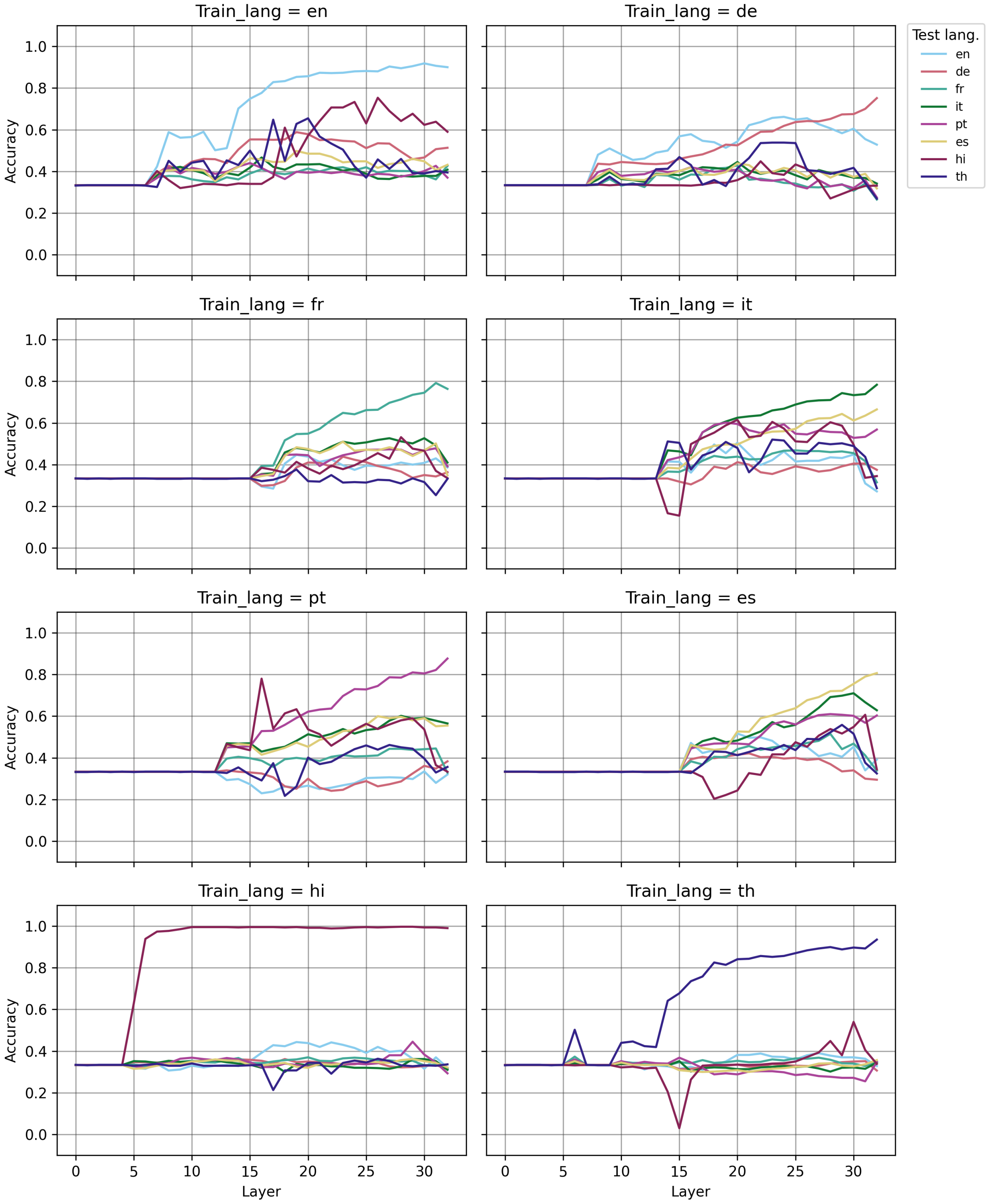


Figure 9: Direct-transfer accuracy by layer. Each subplot shows, for a fixed train language, the probe's accuracy on all test languages at each layer. Languages within the same family transfer more effectively to one another, with peak transfer performance in the mid-to-late layers. Romance languages exhibit strong within-group transfer, although French yields the weakest performance among them. Hindi and Thai show poor cross-transfer from most other languages, indicating distinct tense encoding, likely attributable to their divergent typology, writing systems, and language families. English and German nonetheless transfer moderately well into Hindi and Thai, possibly because auxiliary constructions in Hindi and future-tense markers in Thai partially align with Germanic and Romance patterns.

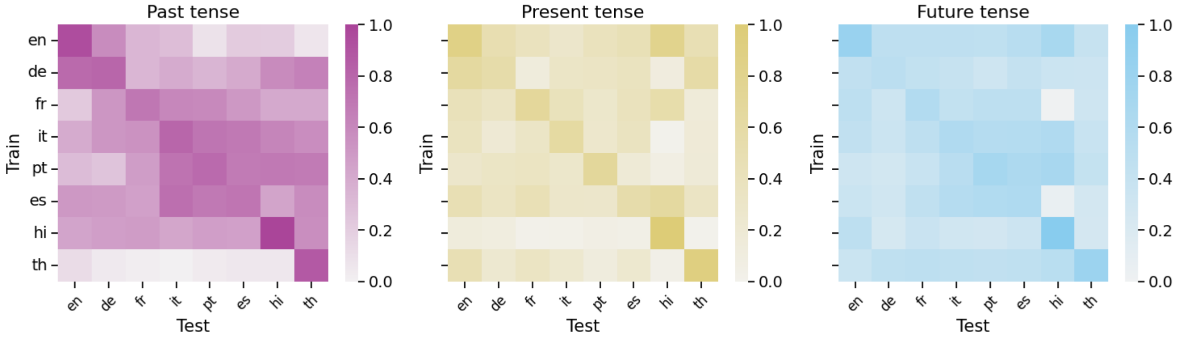


Figure 10: Direct transfer performance broken down into tense at layer 25, where the transfer performance peaks. Transfer between languages within the same family is noticeable, while self-transfer is also distinguishable.

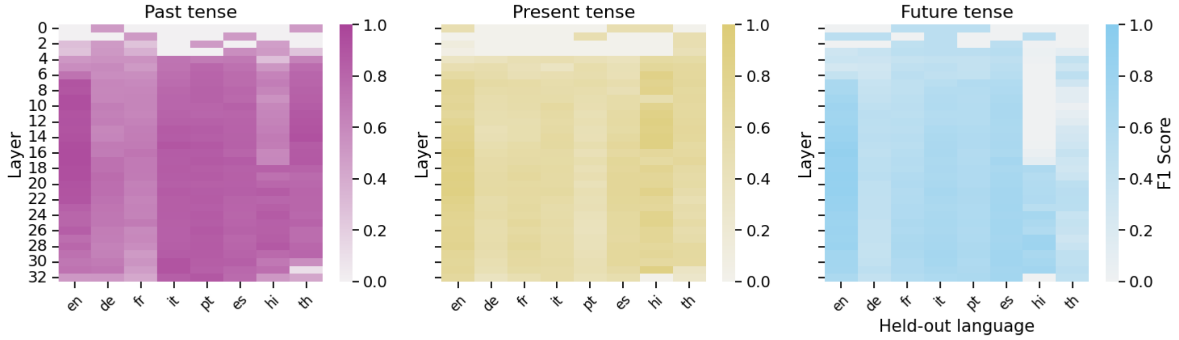


Figure 11: Hold-one-out transfer probing performance across layers and languages, broken down by tense. A single hyperplane trained on all languages except the held-out target still separates tense above chance for most languages, signifying language-agnostic features—past tense yields the highest hold-out F1-macro across all held-out languages, with English highest and German lowest. In the Romance group, only the past tense remains robust; present and future collapse toward chance, especially for French. Hindi's past/future peaks in late layers; present emerges earlier. Thai's past-tense transfer peaks mid-layers; present/future remain near chance.

Language	Past	Present	Future
English	Lily the cat **relaxed** on the mat and she **ate** an apple.	Lily the cat **relaxes** on the mat and she **eats** an apple.	Lily the cat **will relax** on the mat and she **will eat** an apple.
German	Lily die Katze **entspannte sich** auf der Matte und sie **aß** einen Apfel.	Lily die Katze **entspannt sich** auf der Matte und sie **isst** einen Apfel.	Lily die Katze **wird sich entspannen** auf der Matte und sie **wird** einen Apfel **essen**.

Table 7: Semantically minimal, tense-varying template example in English and German.

E Causal Tracing

E.1 Prompt design

We construct semantically minimal sentence frames that differ only in verbal inflection (i.e., past, present, or future) across eight languages.

Few-shot. We create prompts with two identical full-tense sentences separated by a distractor of alternate tense. We inject noise in the verb positions of the first and last sentences to assess whether causal tracing method can flip the generated tense.

Template

<full-X-tense-sentence>
<full-Y-tense-sentence>
<partial-X-tense-ending-before-verb>

Example

Lily the cat relaxed on the mat and she ate an apple.
Lily the cat relaxes on the mat and she eats an apple.
Lily the cat relaxed on the mat and she

Original generation: `ate`. After noise injection: `eats`.

One-shot. To confirm cross-language validity, we generate five variants per tense by varying subjects (e.g., “I,” “Aki the dog”), verbs, and objects while preserving argument structure. English templates were manually drafted, translated using Google Translate, and validated through back-translation. Table 7 shows representative templates.

Template

<full-X-tense-sentence>
<partial-X-tense-ending-before-verb>

Example

Lily the cat relaxed on the mat and she ate an apple.
Lily the cat relaxed on the mat and she

Original generation: `ate`. After noise injection: `is`.

E.2 Experimental setup

1. **Prompts.** Five prompts per tense and language, varying subject/pronoun and verb-object lexemes.
2. **Noise Seeds.** $M_{noise} = 5$, seeds to ensure independent Gaussian draws for reproducibility.
3. **Window Size.** $Window = 3$, restoring layer ℓ activations at some token positions with its previous and next layers.
4. **Streams.** Four sub-components per layer: attention output, MLP activation, MLP output, block output.

Restoration positions In the few-shot prompt experiment, we perform restoration on all token positions. Based on the results, we decided to focus on critical token positions where restoration is most effective in the one-shot experiment (Table 8).

Position	Description
`<|begin_of_text|> (pos 0)`	The very first token embedding.
Pre-verb	The token immediately preceding the first main-verb subtoken.
Tense-bearing subtokens	All subtokens of the auxiliary + main-verb.
Final token	The last token in the “partial …ending” line.

Table 8: Critical token positions.

E.3 Evaluation metrics

We interpret higher $\Delta p_{restored}$ values as more substantial evidence that a given layer and stream are critical for tense generation. We report means with Standard Error of Mean (SEM) across different seeds. The Standard Error of the Mean (SEM) quantifies the precision with which we have estimated the true mean of $\Delta p_{restored}$ across noise-seed replicates. Formally, if $\{x_i\}_{i=1}^{M}$ are the $\Delta p_{restored}$ values for M independent seeds, and $\overline{x} = \frac{1}{M}\Sigma_i x_i$ with sample standard deviation $\sqrt{\frac{1}{M-1}\Sigma_i \left(x_i - \overline{x}\right)^2}$, then

$$SEM = \frac{s}{\sqrt{M}}. \tag{1}$$

F Layer-wise Recovery Analysis

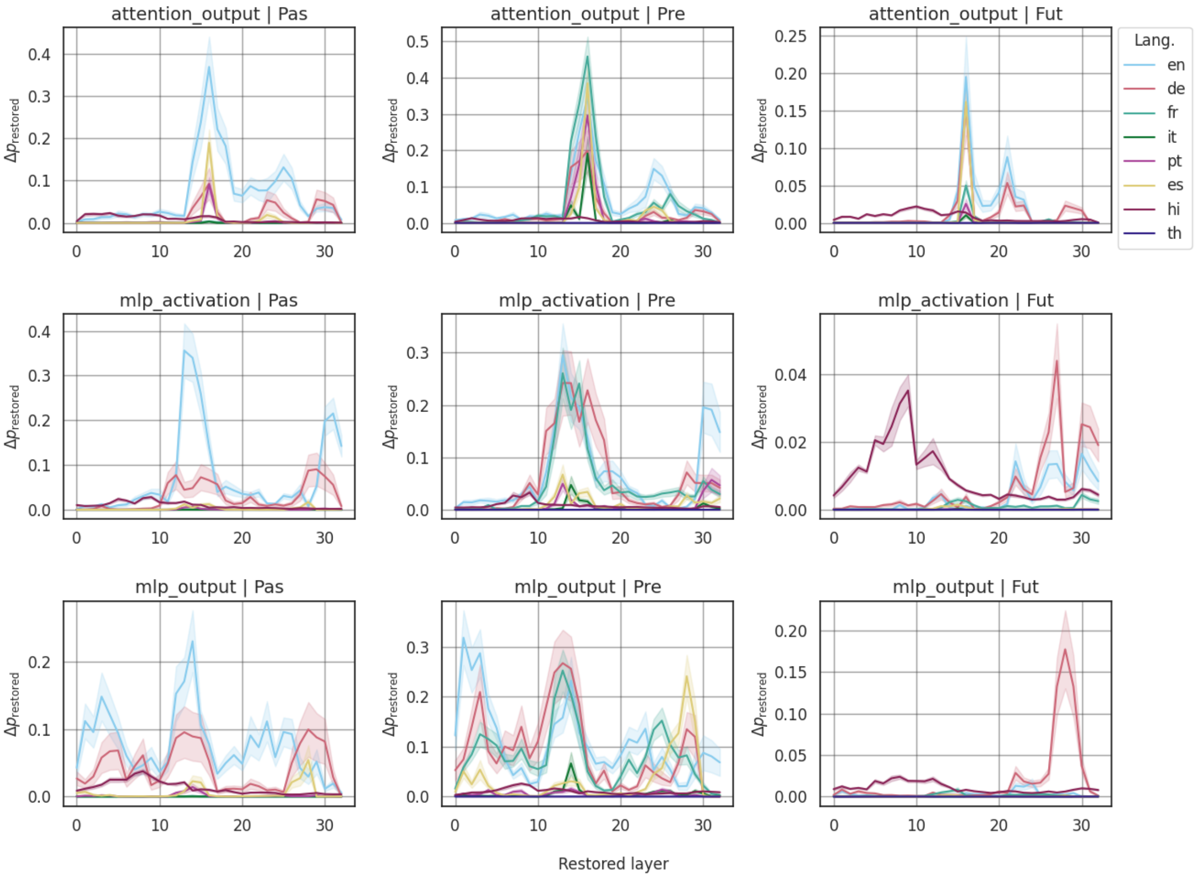


Figure 12: layerwise recovery curves $\Delta p_{\text{restored}}(\ell, S)$ in each language, faceted by stream and tense. High values indicate that restoring the corrupted token activations at that layer and stream most effectively recovers the correct verb-tense prediction.

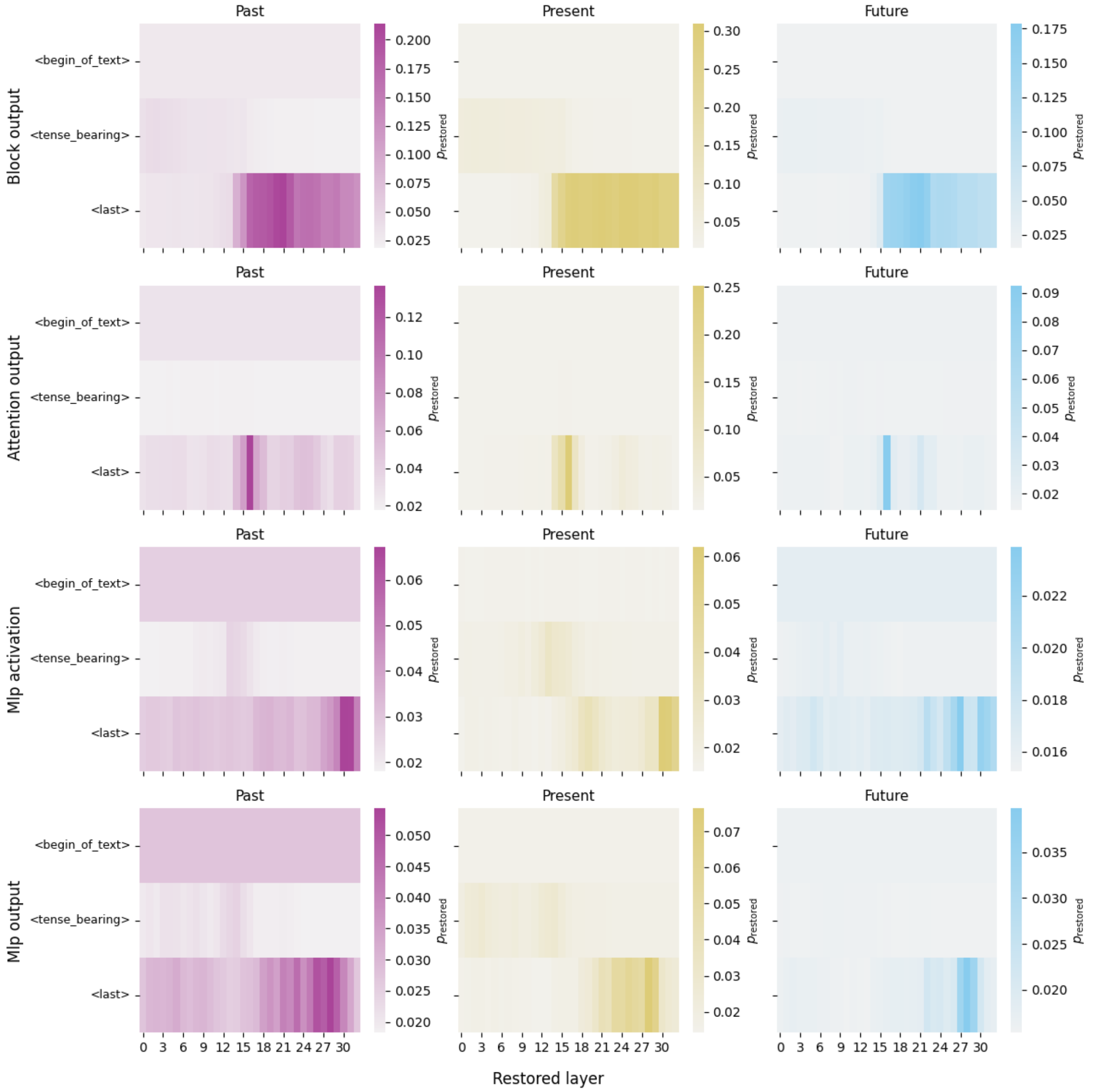
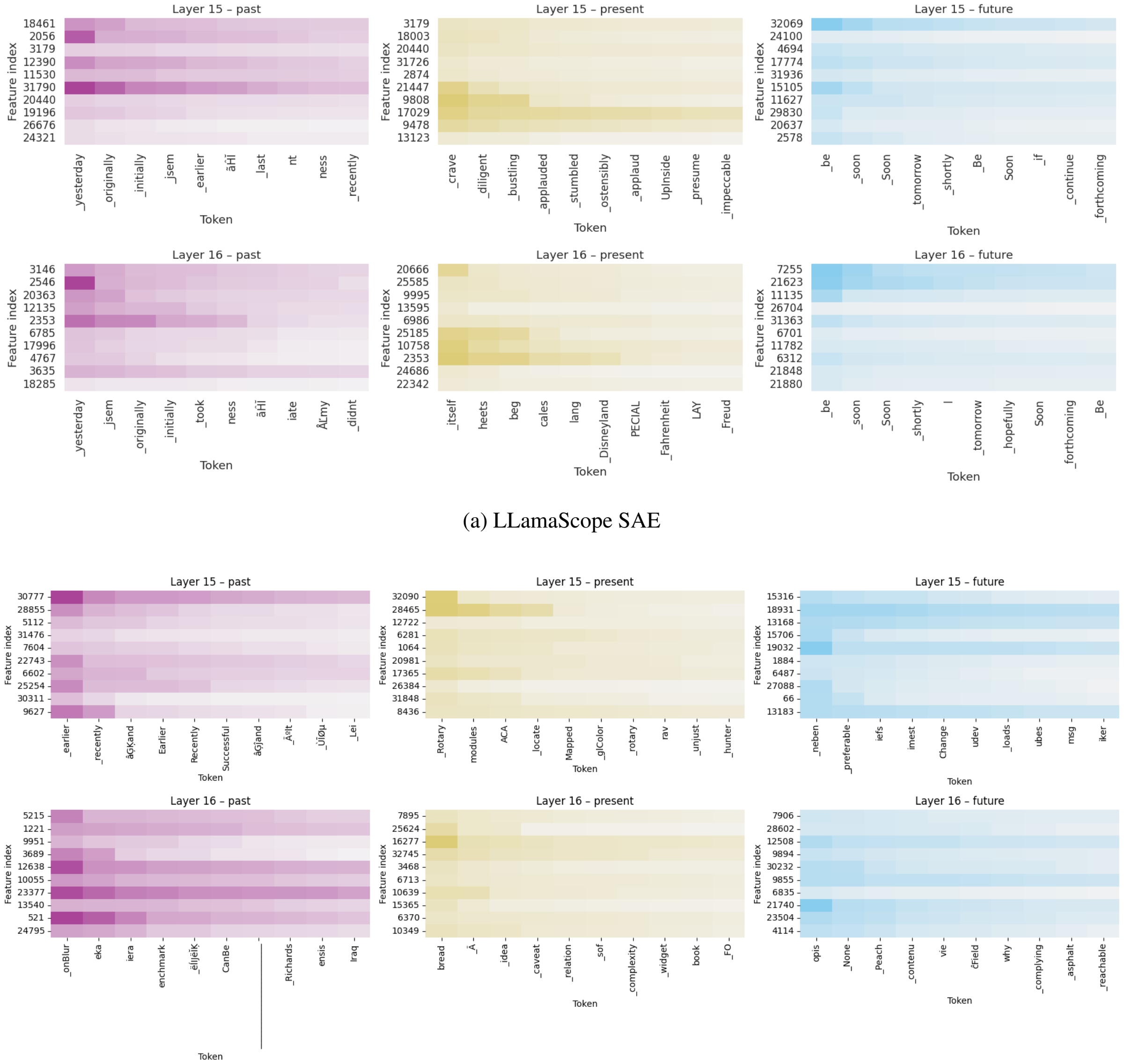


Figure 13: Causal analysis for each tense, averaging across language results.

G Layer-specific Analysis

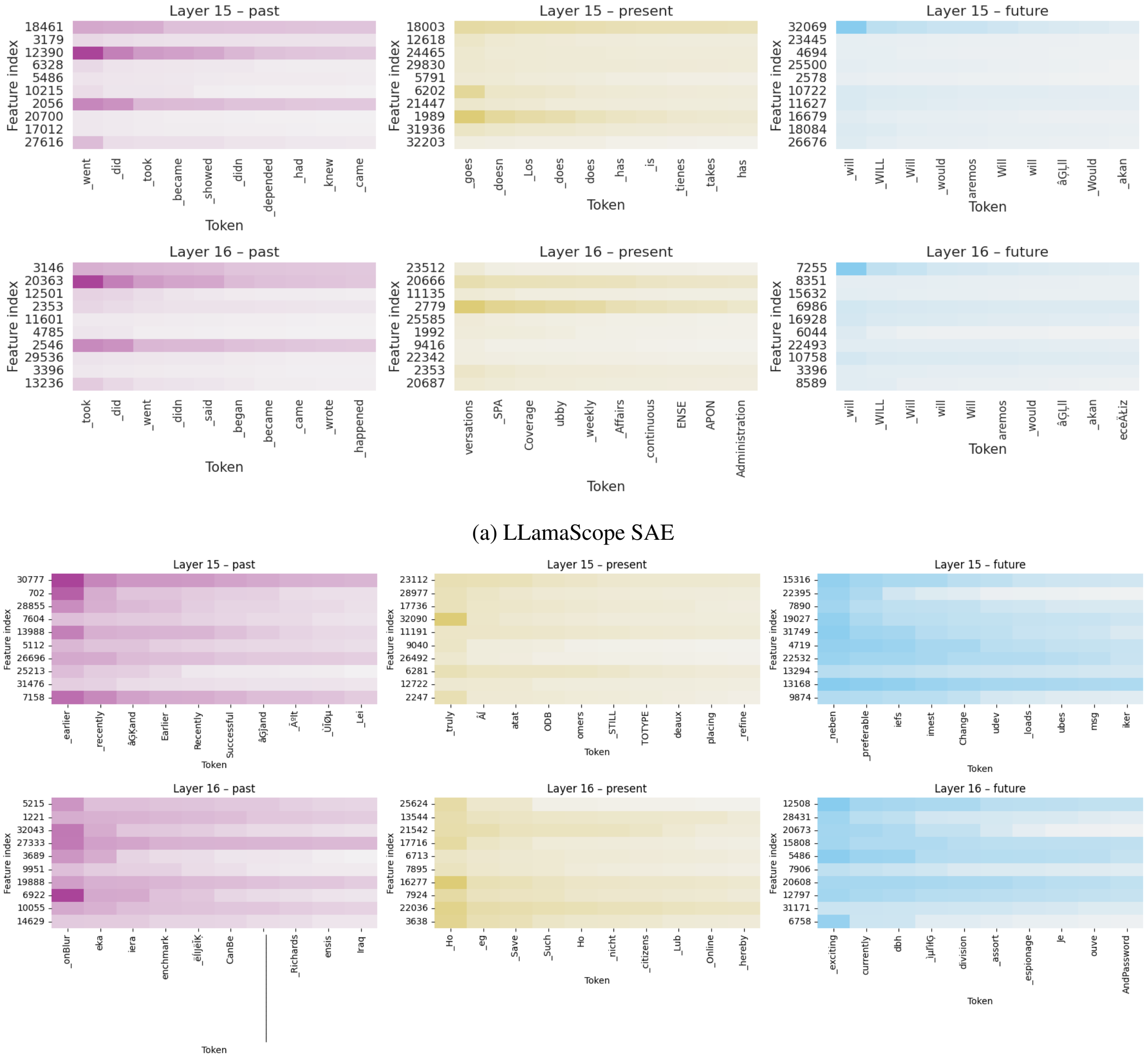


(a) LLamaScope SAE

(b) Multilingual SAE

Figure 14: Token–feature heatmaps at layers 15–16 (ℓ^*) for LLaMA Scope and Multilingual SAEs. Each heatmap shows cosine similarity between SAE-derived feature vectors (from the decoder's tense-encoding subspace) and the model's **output embeddings**. Rows are features; columns list the top ten tokens by similarity. LLaMA Scope features show clear past cues (e.g., "yesterday," "earlier") and future cues (e.g., "tomorrow," "soon"), while Multilingual SAE features align more weakly. A corresponding visualization using the model's input embeddings is shown in Figure 15.

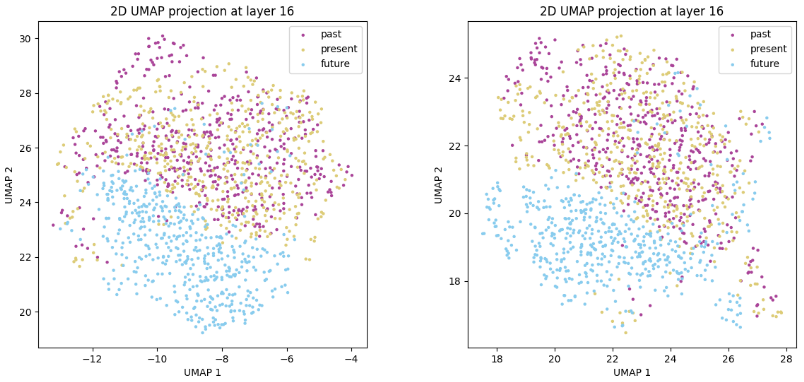
H Model Steering



(a) LLamaScope SAE

(b) Multilingual SAE

Figure 15: Token–feature heatmaps using model’s input embedding matrix at layers 15 and 16 for LLaMA Scope and multilingual SAEs.



(a) LLamaScope SAE

(b) Multilingual SAE

Figure 16: 2D UMAP of SAE activations at layer 16 for both Multilingual and LLaMA Scope frameworks. "Future" examples form a tight, distinct cluster, while "past" and "present" intermingle, reflecting stronger, more consistent signals for future tense due to the invariant token "will." By contrast, past tense relies on irregular forms or the "-ed" suffix, and present alternates between the bare verb and "-s," producing overlapping activations. This pattern highlights that steering future tense is more straightforward, whereas disentangling past versus present remains challenging due to subtle morphological distinctions and semantic overlap.

Layer	Past		Present		Future	
	Feature	α	Feature	α	Feature	α
15	15316	10.0	5112	4.0	702	1.5
	23112	7.0	7890	8.0	5112	1.5
	28855	10.0	15706	6.0	7890	3.0
	30777	9.0	26492	10.0	12722	8.0
			30777	7.0	15316	7.0
					15706	2.0
					23112	5.0
					26492	1.5
					28855	2.0
					30777	1.5
					32090	1.5
16	1221	8.0	3638	4.0	1221	3.0
	3638	7.0	5215	6.0	3638	2.0
			7895	9.0	3689	5.0
			9951	7.0	5215	8.0
			23504	8.0	6922	1.5
			25624	3.0	7895	9.0
					9951	1.5
					12508	1.5
					17716	1.5
					23504	1.5
					25624	4.0
					28602	2.0
					32043	7.0

Table 9: Tense features identified from multilingual SAEs at layers 15 and 16. For each target tense, we report feature indices and their optimal scaling factor α on the dev set (30 prompts per tense). Higher α indicates a weaker baseline signal requiring stronger scaling, while lower α reflects robust intrinsic tense encoding. Both tense-specific and tense-agnostic features are included.

Setting	F_{ℓ^*}	α	Past features			Present features			Future features		
			Pas	**Pre**	**Fut**	**Pas**	**Pre**	**Fut**	**Pas**	**Pre**	**Fut**
Baseline	A	—	0.81	0.39	0.76	0.81	0.39	0.76	0.81	0.39	0.76
	B1	1.0	0.13	0.09	0.14	0.13	0.09	0.14	0.13	0.09	0.14
	B2	1.0	0.81	0.39	0.77	0.81	0.39	0.77	0.81	0.39	0.77
$\alpha > 1$	15	2.0	0.82	0.40	0.80	0.80	0.41	0.78	0.81	0.40	0.79
		5.0	0.84	0.36	0.85	0.77	0.42	0.81	0.83	0.38	0.87
	16	2.0	0.81	0.40	0.78	0.80	0.42	0.77	0.80	0.41	0.78
		5.0	0.82	0.35	0.81	0.77	0.48	0.78	0.77	0.45	0.79
	Both	2.0	0.82	0.39	0.81	0.79	0.42	0.78	0.80	0.41	0.81
		5.0	0.80	0.36	0.84	0.72	0.50	0.81	0.76	0.48	0.82
$\alpha < 1$	15	0.1	0.80	0.40	0.74	0.82	0.39	0.76	0.80	0.39	0.75
		0.0	0.79	0.40	0.74	0.82	0.39	0.76	0.80	0.39	0.75
	16	0.1	0.81	0.39	0.76	0.81	0.38	0.77	0.81	0.38	0.77
		0.0	0.81	0.39	0.76	0.81	0.38	0.78	0.81	0.38	0.77
	Both	0.1	0.78	0.40	0.73	0.81	0.37	0.76	0.81	0.40	0.75
		0.0	0.78	0.40	0.73	0.81	0.37	0.76	0.81	0.39	0.75

Table 10: Model steering results on English test set. Baseline A: Original model; Baseline B1: LLaMA Scope SAEs at layers 15, 16; Baseline B2: Multilingual SAEs at layers 15, 16; F_{ℓ^*} denotes the layer(s) where SAE adaptors are applied during inference, and α is the scaling factor. Feature columns report accuracy when these features are scaled. (Highlighted) cells mark excitation that outperforms the baselines. In inhibition settings, lower accuracy indicates successful downward control.

Setting	F_{ℓ^*}	α	Past features			Present features			Future features		
			Pas	**Pre**	**Fut**	**Pas**	**Pre**	**Fut**	**Pas**	**Pre**	**Fut**
Baseline	A	—	0.63	0.60	0.44	0.63	0.60	0.44	0.63	0.60	0.44
	B	1.0	0.63	0.61	0.43	0.63	0.61	0.43	0.63	0.61	0.43
$\alpha > 1$	15	2.0	0.64	0.63	0.44	0.63	0.62	0.44	0.64	0.64	0.43
		5.0	0.68	0.65	0.42	0.61	0.66	0.42	0.66	0.64	0.44
	16	2.0	0.63	0.62	0.43	0.62	0.64	0.42	0.61	0.63	0.42
		5.0	0.65	0.65	0.42	0.57	0.71	0.40	0.58	0.68	0.38
	Both	2.0	0.65	0.65	0.44	0.62	0.66	0.42	0.63	0.66	0.43
		5.0	0.67	0.59	0.44	0.53	0.72	0.38	0.56	0.69	0.36
$\alpha < 1$	15	0.1	0.62	0.59	0.44	0.64	0.60	0.44	0.63	0.58	0.44
		0.0	0.62	0.58	0.44	0.64	0.59	0.44	0.63	0.58	0.44
	16	0.1	0.63	0.59	0.44	0.64	0.60	0.45	0.64	0.59	0.45
		0.0	0.63	0.59	0.44	0.64	0.59	0.45	0.64	0.59	0.45
	Both	0.1	0.63	0.58	0.44	0.64	0.57	0.46	0.64	0.57	0.45
		0.0	0.63	0.58	0.44	0.65	0.57	0.46	0.64	0.57	0.45

Table 11: Model steering results on German test set using the tense features found in English dataset. Baseline A: Original model; Baseline B: Multilingual SAEs at layers 15, 16; F_{ℓ^*} indicates the layer indices where SAE adaptors are hooked to the model during inference. α is the scaling factor. Feature columns report accuracy after scaling. (Highlighted) cells mark excitation that outperforms the baselines. In inhibition settings, lower accuracy indicates successful downward control.