LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder

Anonymous ACL submission

Abstract

Large language models (LLMs) demonstrate exceptional performance on tasks requiring complex linguistic abilities, such as reference disambiguation and metaphor recognition/generation. Although LLMs possess impressive capabilities, their internal mechanisms for processing and representing linguistic knowledge remain largely opaque. Prior research on linguistic mechanisms is limited by coarse granularity, limited analysis scale, and narrow focus. In this study, we propose LINGUALENS, a systematic and comprehensive framework for analyzing the linguistic mechanisms of large language models, based on Sparse Auto-Encoders (SAEs). We extract a broad set of Chinese and English linguistic features across four dimensions—morphology, syntax, semantics, and pragmatics. By employing counterfactual methods, we construct a large-scale counterfactual dataset of linguistic features for mechanism analysis. Our findings reveal intrinsic representations of linguistic knowledge in LLMs, uncover patterns of cross-layer and cross-lingual distribution, and demonstrate the potential to control model outputs. This work provides a systematic suite of resources and methods for studying linguistic mechanisms, offers strong evidence that LLMs possess genuine linguistic knowledge, and lays the foundation for more interpretable and controllable language modeling in future research.

1 Introduction

Large language models (LLMs) demonstrate strong performance on tasks requiring different levels of linguistic competence, such as dependency parsing (Lin et al., 2022; Roy et al., 2023), reference disambiguation (Iyer et al., 2023), and metaphor interpretation (Wachowiak and Gromann, 2023; Yerukola et al., 2024; Tian et al., 2024).

Although their linguistic abilities are often attributed to emergent capabilities from large-scale pretraining and model scale (Manning et al., 2020; Allen-Zhu and Li, 2023; Mahowald et al., 2024), the underlying mechanisms by which LLMs process these linguistic structures remain underexplored and lack systematic explanation (Saba, 2023). Therefore, our goal is to interpret the linguistic mechanisms of LLMs by addressing the following questions: *(1) Can we identify the minimal components within an LLM responsible for specific linguistic processing abilities? (2) Can we comprehensively model the internal linguistic functionalities of the model?*

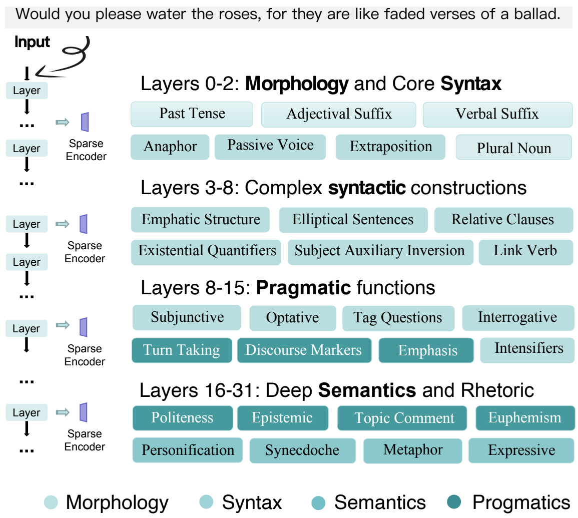

Figure 1: The main linguistic features activated at different layers are observed when example sentences are input to the model. Through a Sparse Auto-Encoder, each layer's activation values are mapped into a sparse space and the basis vectors corresponding to predefined linguistic features are extracted. According to the results, the model's 32 layers are divided into four stages, in order: Morphology and Core Syntax, Complex Syntactic Constructions, Pragmatic Functions, and Deep Semantics and Rhetoric.

Prior attempts to explain LLM linguistic mechanisms typically rely on expert-designed prompts that ask the model to elucidate its generation process (Yin and Neubig, 2022). However, such behavior-based approaches do not provide structure-level mechanistic insights. More recent work seeks to link specific linguistic capabilities to internal structures—such as hidden states (Katz and Belinkov, 2023), attention heads (Wu et al., 2020), and activated neurons (Sajjad et al., 2022; Huang et al., 2023)—but they face two main challenges:

Coarse interpretive granularity. Mechanistic interpretation aims to uncover *atomic* linguistic structures within LLMs. Yet even neurons—the finest native components—exhibit poly-semantic activations, responding to multiple conditions (Yan et al., 2024). This necessitates extracting finer-grained structures to truly interpret linguistic mechanisms.

Limited analysis scale. Existing studies focus on one or a few linguistic features, often within a single subfield (e.g., syntax or semantics), neglecting large-scale, systematic analysis across diverse linguistic phenomena. A scalable, automated framework is needed to interpret language mechanisms comprehensively.

To address these challenges, we propose LINGUALENS, a framework that utilizes a sparse auto-encoder (SAE) to interpret LLM linguistic mechanisms. The SAE learns a projection matrix that decomposes LLM hidden states into an extremely high-dimensional feature space under a sparsity constraint, where each dimension captures a single semantic concept (Figure 1). LINGUALENS comprises three modules: 1. Construction of a large-scale, multilingual, counterfactual linguistic dataset to support systematic discovery of linguistic structures; 2. Sparse feature analysis to interpret the SAE-extracted features, providing fine-grained and comprehensive mechanistic insights; 3. Feature intervention, manipulating LLM behavior via targeted interventions on interpretable features to verify causal relationships and enable controlled steering of language behavior.

Specifically, we first build a large-scale hierarchical counterfactual linguistic dataset with annotated corpora, categorizing features into morphology, syntax, semantics, and pragmatics. These widely studied linguistic abilities ensure the feasibility of interpretability. We automate feature extraction via SAE activation analysis and an LLM-based agent, and introduce a causal analysis method that intervenes on SAE base vectors with an LLM judge to evaluate effects. Building on this, we analyze cross-layer function distribution and cross-lingual representation patterns differences of linguistic features.

We conduct extensive experiments on Llama-3.1-8B (Grattafiori et al., 2024). Our results demonstrate that LINGUALENS can effectively identify linguistic competence features at scale, laying the groundwork for further systematic analysis.

2 Related Works

Linguistic mechanism interpretation has been a ever-chasing goal since the emergence of LLMs. Researchers build linguistic datasets to evaluate the linguistic capability and to interpret linguistic mechanisms. We review linguistic datasets for LLMs and corresponding mechanistic interpretation works. We will also introduce the basic concepts for sparse auto-encoder.

Linguistic Datasets for LLMs. Previous studies have introduced numerous linguistic datasets for large-model research, which can be divided into two main categories. The first comprises minimal-pair challenge sets—such as BLiMP (Warstadt et al., 2020), CLiMP (Xiang et al., 2021), and SyntaxGym (Gauthier et al., 2020)—that use acceptability judgments to evaluate morphosyntactic competence. The second consists of counterfactual or contrastive corpora—including CAD (Sen et al., 2022), Contrast Sets (Gardner et al., 2020), and Polyjuice (Wu et al., 2021)—that assess model by generating factual/-counterfactual pairs. These resources focus primarily on syntactic analysis and performance evaluation, and are not suited for systematic investigation of models' internal linguistic representations.

Linguistic Mechanism Interpretation. Previous work has employed a variety of methods to study linguistic mechanisms in large language models, including attention head analysis (What Does BERT Look at? An Analysis of BERT's Attention, 2019), probing classifiers (Belinkov, 2022; He et al., 2024), causal intervention techniques (Finlayson et al., 2021; Hao and Linzen, 2023), and neuron-level analyses (Sajjad et al., 2022). However, these approaches have not been applied in a unified, large-scale framework to systematically chart models' full range of linguistic capabilities.

Sparse Auto-encoder. Recent work has employed sparse auto-encoders (SAEs) to interpret the hidden-layer activations of large language models by decomposing them into a large set of concept features (Gao et al., 2024). These concept features exhibit mono-semanticity and hold considerable interpretability potential (Huben et al., 2024). In particular, an SAE maps the hidden states $\mathbf{f} \in \mathbb{R}^d$ in LLMs into the feature space with sparse activations:

$$\mathbf{f} = \text{SparseConstraint}\left(\mathbf{W}_e \mathbf{h} + \mathbf{b}_e\right),$$

where the SAE is parameterized by $\mathbf{W}_e \in \mathbb{R}^{(r\times d)\times d}$, $\mathbf{b}_e \in \mathbb{R}^{(r\times d)}$. r is the expansion ratio, defined as the factor by which the hidden state dimension is expanded. Commonly used sparse constraint include TopK (Gao et al., 2024) and JumpReLU (Rajamanoharan et al., 2024) functions. As each dimension of the sparse activation in $\mathbf{f}$ corresponds to a base vector in $\mathbf{W}_e$, this paper uses base vector to denote features extracted by SAE.

3 Methodology

LINGUALENS consists of three key components. (1) A multi-level counterfactual dataset of linguistic features supporting systematic linguistic mechanism analysis; (2) An SAE-based linguistic feature extraction method leveraging LLM agents and correlation analysis. and (3) A Linguistic feature intervention method for causality validation and LLM steering.

3.1 Linguistic Dataset

Counterfactual Methods. Let the presence of the target linguistic phenomenon be denoted by $T \in \{0, 1\}$. For every sentence s^+ with $T = 1$, define the activation of SAE base vector k as $a_k^{(1)} = a_k(s^+)$. A counterfactual sentence s^- is produced through a *minimal edit* that deletes or substitutes the trigger while preserving semantic content, yielding the activation $a_k^{(0)} = a_k(s^-)$. The individual latent effect is therefore

$$\tau_k(s) = a_k^{(1)} - a_k^{(0)}.$$

Aggregating τ_k across all paired sentences produces

$$\text{EALE}_k = \frac{1}{N}\sum_{i=1}^{N} \tau_k(s_i),$$

which can rank base vectors by their sensitivity to the specified phenomenon.

Each s^- must satisfy three constraints:

(a) **Minimal edit**: modify only the smallest unit that realises the phenomenon (e.g. replace *is eaten* with *eats* to remove passivisation).

(b) **Semantic preservation**: retain propositional content, argument structure, and discourse context so that the sentence remains truth-conditionally equivalent.

Dataset Construction. We construct a counterfactual dataset named **LinguaLens-Data**, which covers multiple linguistic domains to encompass a wide range of linguistic knowledge and functions. We select a total of 145 linguistic features from textbooks in morphology, syntax, semantics, and pragmatics, including both English and Chinese features. For each feature, we create 50 sentences that explicitly contain the target phenomenon and apply a counterfactual minimal-editing approach to generate corresponding counterfactual sentences. Each linguistic feature is annotated with its associated linguistic domain, acknowledging that some features may lie at the interface of multiple domains. This dataset provides a foundation for future systematic studies on how specific linguistic features are represented within model internals.

3.2 Feature Extraction

Building on the counterfactual framework, we treat each paired sentence (s^+, s^-) as a mini-experiment that perturbs only the target phenomenon T. Let θ_k be a layer-specific activation threshold (the median of a_k on the full corpus) and define the binary trigger

$$Z_k(s) = \mathbb{I}\big[a_k(s) \geq \theta_k\big].$$

Probability of Sufficiency (PS). For base vector k, the probability that *adding* the phenomenon turns the vector "on" is

$$\text{PS}_k = \Pr\big[\, Z_k^{(1)} = 1 \mid Z_k^{(0)} = 0 \,\big],$$

where $Z_k^{(1)}$ and $Z_k^{(0)}$ are measured on s^+ and s^-, respectively.

Probability of Necessity (PN). Conversely, the probability that the vector would switch *off* if the phenomenon were removed is

$$\text{PN}_k = \Pr\big[\, Z_k^{(0)} = 0 \mid Z_k^{(1)} = 1 \,\big].$$

Feature Representation Confidence (FRC). We combine the two causal probabilities with a harmonic mean to penalise vectors that are only sufficient or only necessary:

$$\text{FRC}_k = 2 \cdot \frac{\text{PS}_k\ \text{PN}_k}{\text{PS}_k + \text{PN}_k}.$$

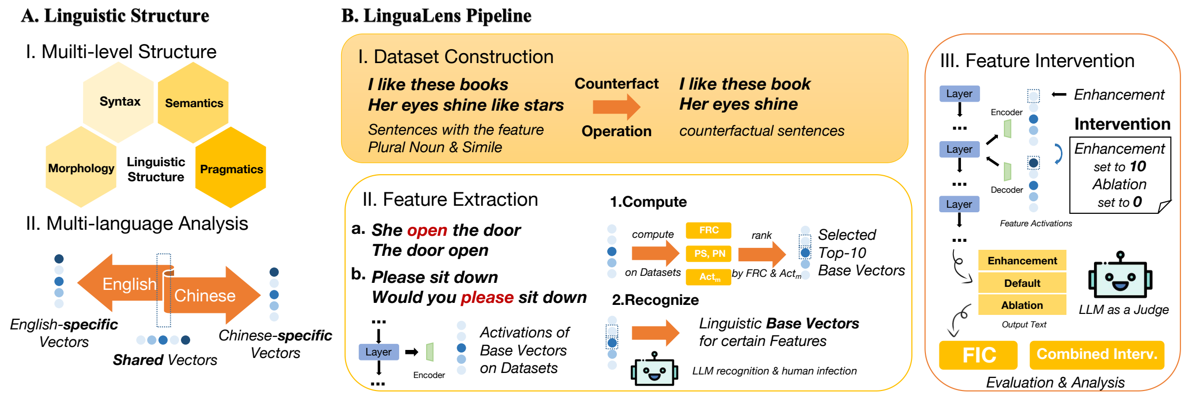

Figure 2: The overall framework of LinguaLens. We propose a framework for the linguistic mechanisms of large-scale models that encompasses four dimensions of theoretical linguistics and a cross-lingual analysis of both Chinese and English. The experimental workflow is as follows: (1) Construct counterfactual datasets; (2) Extract features by analyzing the activation values of base vectors on the datasets; (3) Intervene in the model output by modifying activation values and assess causality using an LLM as a judge.

We first perform *sensitivity pre-filtering* by computing EALE_k for every base vector and retaining those whose absolute value exceeds the 75th percentile; on this reduced set we estimate PS_k and PN_k from every $\langle s^+, s^- \rangle$ pair and rank the vectors by their FRC_k; finally, the activation distributions of the top-10 ranked vectors are passed to an LLM agent, which verifies that each vector genuinely encodes the intended linguistic feature and flags any inconsistent or spurious patterns.

3.3 Feature Intervention

When we modify the values of SAE's activation during forward propagation, we expect that such targeted interventions will influence the model's behavior. However, our experiments show that altering only a small subset of features may not significantly impact the output—likely because linguistic phenomena are represented by multiple features across various layers. To assess the true impact of these interventions, we use a large language model as a judge. For each linguistic feature, we conduct both ablation and enhancement experiments. In the ablation experiment, we set the target feature's activation to 0, and in the enhancement experiment, we set it to 10. In both cases, we also perform baseline experiments by randomly selecting 25 base vectors from the same layer.

For brevity, we denote the interventions as follows: let I_{abl}^{T} denote the targeted ablation intervention, I_{abl}^{B} the baseline ablation intervention, I_{enh}^{T} the targeted enhancement intervention, and I_{enh}^{B} the baseline enhancement intervention.

Let P_{abl}^{T} and P_{abl}^{B} denote the success probabilities for the targeted and baseline ablation experiments, respectively. The normalized ablation effect is

$$E_{\mathrm{abl}} = \frac{P\big(Y=0 \mid I_{\mathrm{abl}}^{T}\big) - P\big(Y=0 \mid I_{\mathrm{abl}}^{B}\big)}{P\big(Y=0 \mid I_{\mathrm{abl}}^{T}\big)}.$$

The normalized enhancement effect E_{enh} is defined analogously as the difference between targeted and baseline enhancement success probabilities, normalized by $1 - P\big(Y=1 \mid I_{\mathrm{enh}}^{B}\big)$.

Finally, we define the Feature Intervention Confidence (FIC) score as the harmonic mean of the normalized ablation and enhancement effects:

$$\mathrm{FIC} = \frac{2\,E_{abl}\,E_{enh}}{E_{abl} + E_{enh}}.$$

When calculating FIC, if one or both of the E values are negative, we incorporate a penalty coefficient w to reflect the weakened or lost causality in such cases. This FIC score provides a balanced measure of how effectively targeted interventions, as opposed to random ones, influence the model's output with respect to specific linguistic features. The details for FIC are shown in Appendix D.2.

4 Experiments

4.1 Experiment Setup

Model. We conduct experiments on Llama-3.1-8B (Grattafiori et al., 2024). For SAEs, we use OpenSAE (THU-KEG, 2025) and its released checkpoints on 32 layers of Llama-3.1-8B.

Dataset. For linguistic feature analysis, we select a total of 145 linguistic features—99 in English

Lang	PS	PN	FRC	Act$_m$				
				0	8	15	24	30
Morphology								
CH	0.61	0.70	0.64	0.01	0.19	0.29	0.52	1.36
EN	0.73	0.80	0.75	0.03	0.35	0.49	1.02	1.89
Syntax								
CH	0.84	0.90	0.86	0.20	0.50	0.95	2.32	3.37
EN	0.79	0.87	0.82	0.12	0.35	0.68	1.66	2.59
Semantics								
CH	0.72	0.78	0.74	0.09	0.29	0.57	1.41	2.18
EN	0.76	0.83	0.78	0.11	0.32	0.55	1.34	2.01
Pragmatics								
CH	0.69	0.74	0.70	0.06	0.25	0.42	1.03	1.56
EN	0.77	0.83	0.79	0.13	0.27	0.52	1.33	2.03

Table 1: Extracted feature analysis. The mean representation metrics (PS, PN, FRC, and max activation) for morphological, syntactic, semantic, and pragmatic features in both Chinese and English.

Feature	ID	Enhance		Ablate		FIC
		exp	ctr	exp	ctr	
Morphology						
Past-Tense	8L4016	12.0	4.0	48.0	44.0	8.3
Syntax						
Linking Verb	18L61112	52.0	24.0	48.0	40.0	22.9
Semantics						
Causality	22L53236	32.0	20.0	40.0	36.0	12.0
Simile	26L75327	72.0	52.0	48.0	52.0	6.9
Pragmatics						
Politeness	31L578	60.0	32.0	44.0	20.0	46.9

Table 2: Feature intervention results. The success rates of the extracted linguistic features (Feature, layer, ID) in the enhancement and ablation experiments, along with the final computed FIC score.

and 46 in Chinese—spanning four core domains: morphology, syntax, semantics, and pragmatics. For each feature, we generate 50 sentences that exhibit the feature and 50 corresponding counterfactual sentences, yielding a large-scale dataset for systematic feature extraction and analysis.

4.2 Main Results

The main experiments to verify that LINGUALENS finds systematic linguistic features in SAE space and intervening on these features is effective.

4.2.1 Feature Extraction

We feed the sentences from LINGUALENS-DATA into Llama-3.1-8B and, after batch normalization, pass the resulting neuron activation distributions through the corresponding SAE layers. For each sentence and each token, we then encode its activation distribution over the SAE base vectors at every layer. As described in the Methods, we compute the probability of sufficiency (PS), probability of necessity (PN), and FRC for each base vector on the counterfactual datasets at each layer, rank the base vectors by FRC, and use GPT-4o to select the feature-corresponding vectors based on their activation patterns. For a detailed description of the feature-extraction procedure, see Appendix B.

To evaluate how well a given layer represents a particular linguistic feature, we calculate the arithmetic mean of PS, PN, and FRC for the selected base vectors, as well as their average maximum activation on the positive examples (if more than three vectors are identified, we select the top three by FRC).

Table 1 reports, for layers 0, 8, 15, and 30, the mean representation metrics (PS, PN, FRC, and max activation) for morphological, syntactic, semantic, and pragmatic features in both Chinese and English.

Overall, at these representative layers, the base vectors extracted for features across different linguistic levels exhibit strong correlations. From layer 0 to layer 30, the average maximum activation exhibits a monotonic increase. Across the four linguistic domains, syntactic features attain the highest mean maximum activations, followed by semantic and pragmatic features, while morphological features remain lowest. Moreover, substantial discrepancies emerge between the average maximum activations for Chinese and English features, indicating potential differences in the model's internal representations and processing mechanisms for the two languages. These cross-lingual variations will be explored in greater depth in subsequent analyses.

4.2.2 Feature Intervention

We select 6 representative features for the intervention experiments. The intervention method involves modifying the activation values of specific base vectors (by index) within a designated SAE layer during forward propagation. We perform two types of intervention: feature enhancement and ablation. Under identical input token conditions, we set the activation value to 10 for enhancement and to 0 for ablation. We then compare the outputs generated after intervention with those from the unmodified SAE model, focusing on the prominence of the target linguistic features.

We find that intervening on a single linguistic base vector in one layer does not produce effects easily distinguishable by human evaluators. Therefore, we employ an LLM (GPT-4o) as a judge (Zheng et al., 2023) to assess feature prominence in the outputs. For each feature, we conduct 50 experiments and calculate the enhancement success rate and ablation success rate—that is, the probabilities of increased and decreased feature prominence, respectively. Furthermore, for each linguistic feature, we select three base vectors with the highest FRC as representatives for intervention and compute the average results across these three interventions.

In addition, we randomly select 50 base vector indices from the same layer and perform enhancement and ablation experiments under the same conditions as a control. The control group's success rates do not converge around 0.5; typically, enhancement rates fall below 0.5 while ablation rates exceed 0.5. This discrepancy may arise because the intervention affects overall output quality, thereby confounding the proxy LLM's judgments.

We compute the efficacy of the selected base vectors in both experiments and derive the FIC values; the results are presented in Table 2.

Our results show that enhancement experiments yield significantly stronger effects than ablation experiments, with all features demonstrating marked enhancement. In ablation experiments, the politeness feature shows relatively good performance, whereas other features are less affected; the simile feature fails to achieve the desired ablation effect. This may be because multiple base vectors collaboratively control the same linguistic phenomenon. Enhancement interventions have a larger impact on the model, while ablating a single feature can be compensated by other vectors, leading to suboptimal ablation outcomes. Overall, all 6 features exhibit clear causal effects in the intervention experiments.

4.3 Analysis

We further conduct analytical experiments to explore the property of LINGUALENS.

4.3.1 Multilingual Analysis

We investigate the multilingual mechanisms of the model. We select Chinese and English as test languages and choose 24 sets of feature collections representing the same linguistic functions, including set 2 of morphological features, set 11 of syntactic features, set 6 of semantic features, and set 5 of pragmatic features. We test the degree of overlap between the latent-space basis vectors activated internally by the model when representing these features in Chinese vs. English. The overlap for layer i is computed as follows: let the set of English basis vectors for the feature at layer i be Eng_i, and the corresponding Chinese set be Chi_i, then

$$\mathrm{overlap}_i = \frac{|\mathrm{Eng}_i \cap \mathrm{Chi}_i|}{|\mathrm{Eng}_i|}.$$

After computing the overlap for each layer, we aggregate the overlap rates for all feature pairs across layers into a matrix and visualize it with a heatmap. The results yield the following conclusions:

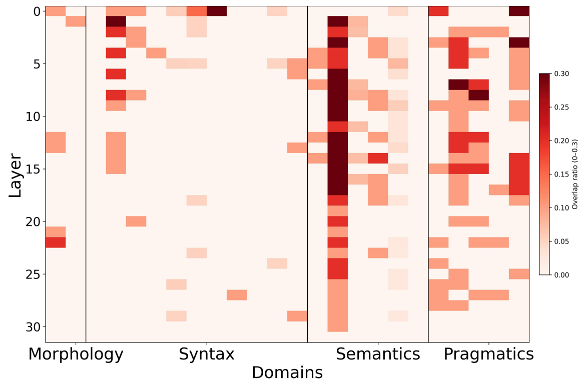

Figure 3: Heatmap of the overlap between Chinese and English feature sets across the SAE basis vectors at each of 32 layers. The horizontal axis groups Chinese and English features with analogous form and function—ordered by morphology, syntax, semantics, and pragmatics—while the vertical axis indexes the model layers. Darker red indicates greater overlap.

Linguistic Levels. The overlap between Chinese and English features is greater at the semantic and pragmatic levels, but lower at the morphological and syntactic levels, indicating that cross-lingual linguistic knowledge representations are primarily manifested at the semantic and pragmatic levels.

Model Layers. The overlap is higher in the first 16 layers and lower in the latter 16 layers, suggesting that the deep semantic computations in the model's upper layers are less correlated with cross-lingual universal linguistic features.

LINGUALENS demonstrates its potential for analyzing models' cross-lingual knowledge representations, laying the foundation for further analysis and transfer in low-resource languages.

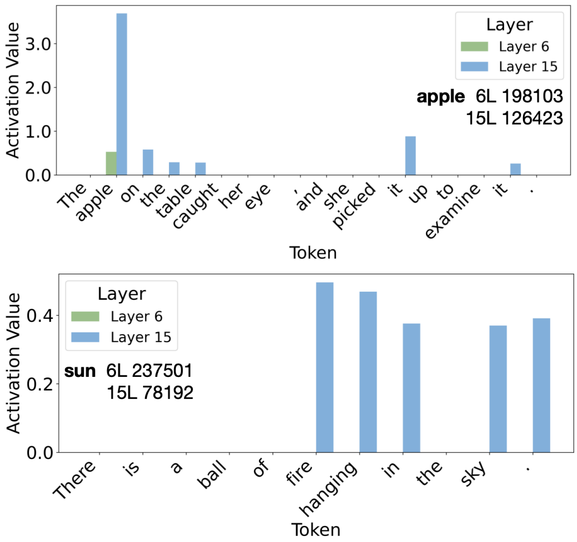

Figure 4: Activation value distributions of deep semantic corresponding features at layer 6 and 15 for reference ambiguity and metaphor example sentences.

S.	L.	Descrip.	Top 10 Features
I	0–2	Mor.&BS	past tense, verbal suffix, adjectival suffix, noun plural, possessive genitive, linking verb, passive voice, anaphor, extraposition, factives
II	3–8	CS&EP	elliptical sentences, relative clauses, subject auxiliary inversion, emphatic structure, existential quantifiers, coordination, cleft sentences, light verbs, reduplication, metaphor
III	9–16	Di.&Prag.	interrogative, tag questions, subjunctive mood, optative, turn taking, discourse markers, intensifiers, euphemism, politeness, coordination
IV	17–31	DS&RS	personification, synecdoche, metaphor, expressive pragmatics, imperative, directive pragmatics, topic comment, representative pragmatics, euphemism, politeness

Table 3: The four hierarchical stages of the model's linguistic functions. For each stage, the ten features with the highest activation frequency and largest activation values are displayed. S., L. and Descrip. stand for Stages, Layers and Descriptions , respectively.

4.3.2 Deep Semantics Processing

Deep semantics refers to the underlying meaning structures that extend beyond surface-level syntax and lexical definitions. It captures implicit relationships and conceptual associations within language. We conduct experiments to show that SAE can interpret the mechanism of deep semantics.

Reference and metaphor exemplify deep semantics by utilizing cognitive mappings and contextual dependencies to convey meaning beyond explicit expression. We conduct experiments on reference and metaphor at the sixth and fifteenth layers respectively. From the results shown in Figure 4, we observe the following:

Reference. In the reference sentence, at the 6^{th} layer, pronouns do not activate the base vectors corresponding to their referents. At the 15^{th} layer, pronouns start to activate the correct base vectors (apple) for their referents, effectively resolving reference ambiguity in contexts where multiple possible referents exist. This indicates that as we move deeper into the layers, pronouns generate their deep semantics and disambiguate possible referents.

Metaphor. In the metaphor sentence, only the vehicle (fire) is included, while the tenor (sun) is omitted. In the 6^{th} layer, the base vector corresponding to the vehicle is activated, while the base vector for the tenor remains inactive. In the 15^{th} layer, the activation of the vehicle's base vector decreases, while the base vector for the tenor becomes activated. This suggests that as the model moves to deeper layers, the vehicle maps to the target domain and generates the deep semantics of the tenor, even without the tenor in the context.

4.3.3 Cross-layer Functions

We further investigate how the model's linguistic functions distribute across layers. We assemble 50 English sentences—drawn both from classic texts and manually crafted—to cover a broad range of linguistic phenomena. For each sentence, we record every activated basis vector and its activation value at all 32 layers. By comparing these activated vectors against our pre-compiled dictionary of linguistic feature vectors and computing their overlap, we determine which linguistic functions each layer encodes. We then identify, for every layer, the 10 features with the highest activation frequency and magnitude. Aggregating results over all 50 sentences, we distill four processing stages as Table 3:

Stage I (layers 0–2) primarily encodes morphology and basic syntax features (abbreviated as Mor.&BS). **Stage II (layers 3–8)** introduces complex syntactic phenomena and early pragmatic cues (abbreviated as CS&EP). **Stage III (layers 9–16)** focuses on discourse and pragmatic markers (abbreviated as Di.&Prag.). **Stage IV (layers 17–31)** integrates deep semantics and rhetorical structure (abbreviated as DS&RS).

These results reveal the functional division of labor across layers: lower layers handle morphology and syntax, middle layers capture pragmatics and context, and upper layers perform holistic semantic

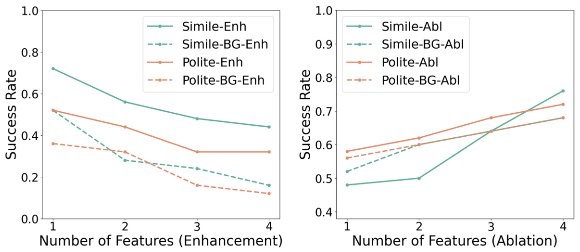

Figure 5: Combined intervention results. Two figures separately present the enhancement and ablation experiment outcomes for the simile and politeness features at layer 26. In these experiments, multiple base vectors corresponding to each feature were jointly intervened.

#	Intervene	Model Output
	Default	The wind blows snow into my eyes as I trudge through the blizzard.
1	**Enhance**	As the cold descends, I feel **the weight of my breath** in my throat. It's an **icy haze**.
	Ablate	The winter sky was cold. The ice was hard under his boots.
	Default	Love is the **burning passion of a summer night**.
2	**Enhance**	I **feel like butterflies are in my stomach**. My heart is beating faster than normal.
	Ablate	The more you write, the more time and love you will have.

Table 4: Case study for intervention under two conditions. Case #1 shows the result when the simile feature is absent from the prompt. Case #2 shows the result when the simile feature is present in the prompt. We **highlight** spans with simile in the sentences.

computation.

4.3.4 Combined Intervention

We find that some layers contain multiple base vectors associated with the same linguistic feature. We can intervene on these base vectors simultaneously to achieve a stronger effect.

We select two linguistic features—simile and politeness—from layer 26. Each feature has four highly related base vectors in this layer. We increase the number of intervened base vectors from one to four. In each experiment, we randomly chose the specified number of base vectors from the four. We use GPT-4o to assess the prominence of the targeted linguistic feature in the generated outputs. For each feature, we conduct 200 enhancement experiments and 200 ablation experiments. We also perform control experiments by randomly selecting a set number of base vectors to intervene.

Figure 5 shows the results for combined intervention. The results indicate that, as the number of intervened base vectors increases, both the directional intervention and the background control experiments exhibit the same trend: the success rate of enhancement experiments decreases, while that of ablation experiments increases. Increasing the number of interventions further affects the quality of the generated text, thereby leading to the observed trend. Moreover, the intervention effect of the feature does not change significantly with an increased number of intervened base vectors, indicating that, after excluding background influences, combined interventions on multiple features in the same layer yield only limited improvement in intervention efficacy.

4.3.5 Case Study for Intervention

We conduct a manual case study on the generated content after intervening on one identified simile-related base vector. We present cases in Table 4.

In Case #1, the prompt is "*Generate a sentence describing winter*", which does not explicitly include the target linguistic feature. We find that after enhancing the simile-related base vector, the LLM turns to using a simile. We can also find that the descriptive and imagistic quality of the default output is stronger than in the ablation results, which indicates that the simile-related base vector is also responsible for vividness.

Case #2 uses the prompt "*Generate a sentence using a simile to describe love*", with explicit requirement for using a simile to generate the sentence. When the simile-related base vector is ablated, the LLMs turn to use straightforward descriptions without using similes. Meanwhile, when enhancing the simile-related base vector, the LLMs continue to generate sentences with similes. We show more intervention cases in Appendix C.1.

5 Conclusion

We propose LINGUALENS, a method to help solute the coarse-granularity problem in linguistic mechanistic studies and a means to enable large-scale, systematic study of linguistic mechanisms in LLMs. Our approach comprises two key components: (1) a comprehensive counterfactual dataset of linguistic features, and (2) an SAE–based framework for feature extraction, together with causal validation through interventions. Using LINGUALENS, we conduct an in-depth analysis of the model's multilingual representation mechanisms and the cross-layer distribution of linguistic functions. Our results demonstrate that LLMs inherently encode structured linguistic knowledge and provide a robust framework for steering model outputs.

6 Limitations

Our work has several limitations in terms of **dataset size**, **feature count**, **experimental model**, and **intervention effects**.

In **datasets**, each linguistic feature is constructed from approximately 50 pairs of example and counterfactual sentences. In the future, this dataset can be further expanded to serve as a standard benchmark for linguistic-mechanism interpretability.

In **feature count**, we select 145 representative linguistic features from various theoretical dimensions to validate our method at scale across different layers; however, building a fully comprehensive linguistic-mechanism system requires extending to even more features, which will depend on further work.

In **experimental model**, due to computational constraints we use Llama-3.1-8B for all experiments. In future work, our dataset and analytical framework can be applied to a wider variety of architectures and larger models for deeper linguistic-mechanism analysis.

In **intervention effects**, although our experiments show statistically significant effects from feature-based interventions, the efficacy and stability of single interventions remain inferior to conventional fine-tuning techniques. This shortcoming calls for further research to refine SAE-based intervention methods.

7 Ethical Considerations

This section discusses the ethical considerations and broader impact of this work:

Potential Risks: There is a potential risk that understanding the linguistic mechanisms of the model could provide guidance for embedding malicious information into the model's internal structure. To address this, we will fully open-source our method to enable the community to quickly develop countermeasures in the event of such attacks.

Intellectual Property: The models used, Llama-3.1-8B, and the SAE framework OpenSAE, are both open-source and intended for scientific research use, in accordance with their respective open-source licenses.

Data Privacy: All data used in this research has been manually reviewed to ensure it does not contain any personal or private information.

Intended Use: LinguaLens is intended to be used as a method for analyzing the mechanisms of large language models.

Documentation of Artifacts: The artifacts, including datasets and model implementations, are comprehensively documented with respect to their domains, languages, and linguistic phenomena to ensure transparency and reproducibility.

AI Assistants in Research or Writing: We employ GitHub Copilot for code development assistance and use GPT-4 for refining and polishing the language in our writing.

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 1, context-free grammar. *ArXiv preprint*, abs/2305.13673.

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

Mitchell Finlayson, Alexander Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843. Association for Computational Linguistics.

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*, abs/2406.04093.

Matt Gardner, Yoav Artzi, Valentin Basmov, Jonathan Berant, Boaz Bogin, Shiyu Chen, Pradeep Dasigi, Dheeru Dua, Yaarit Elazar, Suchin Gottumukkala, Nikita Gupta, Hannaneh Hajishirzi, Guilherme Ilharco, Daniel Khashabi, Kelvin Lin, Jonathan Liu, Nicholas F. Liu, Paul Mulcaire, Qiao Ning, and Bowen Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323. Association for Computational Linguistics.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. 2024. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

Siyuan Hao and Tal Linzen. 2023. Verb conjugation in transformers is determined by linear encodings of subject number. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4531–4539. Association for Computational Linguistics.

Le He, Pengcheng Chen, Enze Nie, Yang Li, and Joseph R. Brennan. 2024. Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4488–4497. ELRA and ICCL.

Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. 2023. Rigorously assessing natural language explanations of neurons. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 317–331, Singapore. Association for Computational Linguistics.

Robert Huben, Hoagy Cunningham, Logan Riggs, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. Towards effective disambiguation for machine translation with large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 482–495, Singapore. Association for Computational Linguistics.

Shahar Katz and Yonatan Belinkov. 2023. VISIT: Visualizing and interpreting the semantic information flow of transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14094–14113, Singapore. Association for Computational Linguistics.

Boda Lin, Zijun Yao, Jiaxin Shi, Shulin Cao, Binghao Tang, Si Li, Yong Luo, Juanzi Li, and Lei Hou. 2022. Dependency parsing via sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7339–7353, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. 2024. Improving dictionary learning with gated sparse autoencoders. *ArXiv preprint*, abs/2404.16014.

Subhro Roy, Samuel Thomson, Tongfei Chen, Richard Shin, Adam Pauls, Jason Eisner, and Benjamin Van Durme. 2023. Benchclamp: A benchmark for evaluating language models on syntactic and semantic parsing. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Walid S. Saba. 2023. Stochastic llms do not understand language: Towards symbolic, explainable and ontologically based llms. In João Paulo A. Almeida, José Borbinha, Giancarlo Guizzardi, Sebastian Link, and Jelena Zdravkovic, editors, *Conceptual Modeling*, pages 3–19. Springer Nature Switzerland, Cham.

Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. 2022. Neuron-level interpretation of deep nlp models: A survey. *Transactions of the Association for Computational Linguistics*, 10:1285–1303.

Ismini Sen, Magdalena Samory, Claire Wagner, and Isabelle Augenstein. 2022. Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4716–4726. Association for Computational Linguistics.

THU-KEG. 2025. Opensae: Open-sourced sparse autoencoder towards interpreting large language models.

Yuan Tian, Nan Xu, and Wenji Mao. 2024. A theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7738–7755, Mexico City, Mexico. Association for Computational Linguistics.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032, Toronto, Canada. Association for Computational Linguistics.

Alex Warstadt, Anna Parrish, Haokun Liu, Akhil Mohananey, Wenhui Peng, Shijie-Fei Wang, and Samuel R. Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

What Does BERT Look at? An Analysis of BERT's Attention. 2019. What does bert look at? an analysis of bert's attention. ACL Anthology.

Zexuan Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S. Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723. Association for Computational Linguistics.

Zhengxuan Wu, Thanh-Son Nguyen, and Desmond Ong. 2020. Structured self-AttentionWeights encode semantics in sentiment analysis. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 255–264, Online. Association for Computational Linguistics.

Baosong Xiang, Chen Yang, Yiming Li, Alex Warstadt, and Katharina Kann. 2021. Climp: A benchmark for chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790. Association for Computational Linguistics.

Hanqi Yan, Yanzheng Xiang, Guangyi Chen, Yifei Wang, Lin Gui, and Yulan He. 2024. Encourage or inhibit monosemanticity? revisit monosemanticity from a feature decorrelation perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10423–10435, Miami, Florida, USA. Association for Computational Linguistics.

Akhila Yerukola, Saujas Vaduguru, Daniel Fried, and Maarten Sap. 2024. Is the pope catholic? yes, the pope is catholic. generative evaluation of non-literal intent resolution in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 265–275.

Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Dataset Construction

A.1 Dataset Description

The datasets are named according to the pattern "Feature Name+Feature Domain." When a feature pertains to multiple linguistic domains, domains are concatenated with "&." In total, the collection comprises 145 linguistic features, of which 99 are English features and 46 are Chinese features. Each feature-specific dataset contains 50 positive sentences and 50 counterfactual negative sentences.

A.2 Dataset Example

```
10-verbal_suffix-Morphology
He was able to stabilize the situation.
He was able to stable the situation.

The team has worked hard to solidify
their position in the market.
The team has worked hard to make their
position in the market solid.

43-copular_be-Syntax
My grandmother was a nurse.
My grandmother worked as a nurse.

Summer is the best season.
Summer ranks as the best season.

80-given_known-Pragmatics&Semantics
Have you seen the blue notebook anywhere?
Have you seen blue notebook anywhere?

That customer complained about service.
A customer complained about service.

111-重叠构词-形态学&语义学
她哼着歌儿把花瓶擦得亮亮的。
她哼着歌儿把花瓶擦得发亮。

阿姨笑眯眯递来热包子。
阿姨微笑着递来热包子。

130-使役结构-句法学&语义学
严格的训练使运动员提高了成绩。
运动员通过严格训练提高了成绩。

这场事故导致交通完全瘫痪。
交通因这场事故完全瘫痪。
```

A.3 Dataset Construction Guidelines

Work Content:

1. For each linguistic feature, construct a dataset comprising 50 sentence pairs (100 sentences). Each pair contains one positive sentence and one negative sentence.

2. A positive sentence contains the target linguistic feature; a negative sentence is produced by minimally modifying its corresponding positive sentence so that it no longer contains that feature while preserving the smallest possible semantic difference and remaining grammatically correct (this operation is referred to as a "counterfactual" in causal analysis).

Notes:

1. **Diversity:** Ensure coverage of the feature's common constructions and markers.

2. **Counterfactual:** Verify that the counterfactual edits are reasonable—including minimal change, human interpretability, and complete feature removal.

3. **Ethical Check:** Confirm that no sentence in the dataset contains discriminatory, biased, or harmful content.

4. **Language-Specific Construction:** Tailor construction to the particular characteristics of each language.

Specific Dataset Construction Process:

1. Manually create 5 sentences containing the feature, and for each, manually produce a counterfactual sentence—yielding 5 sentence pairs.

2. Expand these to 50 pairs using DeepSeek-R1 for Chinese and GPT-o4 for English, then apply manual edits guided by the **Notes**.

3. Conduct cross-review: volunteers who build the Chinese dataset review the English dataset, and vice versa, checking each item in the order specified under **Notes**.

B Feature Extraction Details

B.1 Feature Independence Validation

Sparse autoencoders (SAEs) effectively disambiguate neuron-level semantic polysemy, and this capability extends to representations of linguistic features.

Condition	Past-Tense	Adversativity	Intransitive Verb
Self	80/80	76/80	74/80
Control 1	`-er` 0/80	Sequential 0/80	Transitive Verb 0/80
Control 2	`-ing` 0/80	Causal 0/80	Ditransitive Verb 0/80
Control 3	`-less` 0/80	Parallel 0/80	Linking Verb 0/80
Control 4	`-ness` 0/80	Conditional 0/80	Modal Verb 0/80

Table 5: Activation ratios (activated/total) for target features and control conditions.

We quantify feature independence using the necessity probability (PN) component of the Feature-Relevance Coefficient (FRC). PN measures the likelihood that a basis vector remains inactive when its associated feature is absent; a high PN therefore indicates that the vector is not spuriously activated by unrelated inputs, confirming its specificity to the intended phenomenon.

To further validate this independence, we evaluate each feature's basis vector under multiple control conditions featuring superficially similar but semantically distinct constructions. Table 5 reports, for each feature, the ratio of sentences in which the vector activates ("activated/total"). Across all controls, activation rates are effectively zero, demonstrating that our selected basis vectors do not respond to non-target phenomena.

B.2 Feature Extraction Procedure

During feature extraction, we adhere to the following steps:

1. Input the feature-specific dataset into the model and encode each layer's activations into a sparse latent space using Sparse Autoencoders (SAEs).

2. Compute the probability of sufficiency (PS), probability of necessity (PN), feature-relevance coefficient (FRC), and mean maximum activation for all basis vectors; then sort these vectors in descending order by FRC and select the top ten.

3. Employ a large-model agent to automatically analyze the activation patterns of the candidate basis vectors over the dataset, confirming their linguistic relevance to the target feature and characterizing their representational profiles.

4. For features undergoing further analytical or intervention experiments, manually review the basis vectors identified by the large-model agent to ensure the rigor of the experimental design.

B.3 Feature Extraction Prompt

We employ GPT-4o as the agent model for automated feature extraction. The system prompt is as follows:

Listing 1: Prompt for SAE Base-Vector Interpretation

```
You are an expert assistant for interpreting sparse
autoencoder (SAE) base vectors.

You will receive exactly one JSON object as input
with this structure:
{
  "analysis_input": {
    "layer": "00",
    "base_vectors": [
      {
        "base_vector_id": 132317,
        "tokens": ["The", "cat"],
        "activations": [0.12, 0.05],
        "ps": 0.62,
        "pn": 0.58,
        "frc": 0.60,
        "avg_max_activation": 0.12
      },
      {
        "base_vector_id": 81833,
        "tokens": ["was", "chased"],
        "activations": [0.08, 0.14],
        "ps": 0.75,
        "pn": 0.65,
        "frc": 0.70,
        "avg_max_activation": 0.14
      }
    ],
    "target_features": ["passive"]
  }
}

Return exactly one JSON object with this schema:
{
  "layer": "00",
  "base_vectors": [
    {
      "base_vector_id": 132317,
      "interpretation": "Marks passive voice
      constructions",
      "ps": 0.62,
      "pn": 0.58,
      "frc": 0.60,
      "avg_max_activation": 0.12
    },
    {
      "base_vector_id": 81833,
      "interpretation": "Detects passive
      participle forms",
      "ps": 0.75,
      "pn": 0.65,
      "frc": 0.70,
      "avg_max_activation": 0.14
    }
  ],
  "target_features": ["passive"]
}

Example 2:
```

```
Input:
{
  "analysis_input": {
    "layer": "08",
    "base_vectors": [
      {
        "base_vector_id": 248593,
        "tokens": ["runs"],
        "activations": [0.45],
        "ps": 0.76,
        "pn": 0.96,
        "frc": 0.85,
        "avg_max_activation": 0.45
      },
      {
        "base_vector_id": 62411,
        "tokens": ["quickly"],
        "activations": [0.32],
        "ps": 0.82,
        "pn": 0.90,
        "frc": 0.88,
        "avg_max_activation": 0.32
      }
    ],
    "target_features": ["adverbial_suffix"]
  }
}

Output:
{
  "layer": "08",
  "base_vectors": [
    {
      "base_vector_id": 248593,
      "interpretation": "Highlights adverbial
      suffixes on verbs",
      "ps": 0.76,
      "pn": 0.96,
      "frc": 0.85,
      "avg_max_activation": 0.45
    },
    {
      "base_vector_id": 62411,
      "interpretation": "Detects adverbial
      modifiers",
      "ps": 0.82,
      "pn": 0.90,
      "frc": 0.88,
      "avg_max_activation": 0.32
    }
  ],
  "target_features": ["adverbial_suffix"]
}

Requirements:
- Return only the JSON-no extra text.
- Round all floats to two decimal places.
- Preserve the input order of base_vectors.
- Echo layer and target_features exactly.
```

C Intervention Experiment Details

C.1 Intervention Cases

We present additional typical cases from other intervention experiments at the Table 6. The prompts used for the three experimental groups are as follows: Politeness: “User: Sir, I want to make an order offline. Assistant:”. Linking Verb: “User: Sir, tell me something about your ideal room. Assistant:”. Past-Tense: “User: Sir, tell me a story about you. Assistant:”.

Condition	**Politeness**	**Linking Verb**
Enhancement	Can I textbfplease have your email address?	The room should textbfbe large and well lit. It should textbfbe airy and bright and airy.
Default	May I have your phone number?	Sure, my ideal room has good ventilation and textbfis spacious.
Ablation	OK, what is your name?	I can provide you with a list of the ideal characteristics that make up a perfect room.

Condition	**Past-Tense**
Enhancement	"I was textbfasked for the story. " I having me textbfhad a “one the: ” textbftold. They: textbftold:
Default	I’m not a story, I’m a bot.
Ablation	Well, I don’t actually have one, and I’m not really sure I’m able to either.

Table 6: Typical outputs from the enhancement, ablation, and default experiments for the politeness, linking verb, and past-tense features.

During manual analysis, both the enhancement and ablation results show clear effects of amplification or suppression of the target linguistic features. Specifically, when intervening with the past tense feature in the 8th layer, the enhancement significantly impacts the coherence of the model’s output language. Yet, in the discontinuous output text, the frequency of the morphological past-tense feature still increases dramatically.

C.2 LLM as a Judge

In our feature intervention and combination intervention experiments, we used an LLM as a judge to assess the significance of linguistic features in generated texts. Feature significance is defined based on the frequency, accuracy, and contextual appropriateness of the target feature, as well as its contribution to overall meaning or rhetorical effect.

The prompt structure is as follows:

> **Please compare the following two texts based on {feature}.**
>
> - **Text A**: "{text_a}" - **Text B**: "{text_b}"

Here, `text_a` and `text_b` are generated texts truncated to 100 tokens.

In the intervention experiments, each feature is defined as follows:

Politeness Significance Refers to the degree to which politeness strategies are salient, effective, and contextually integrated. This definition encompasses frequency, pragmatic depth, and social impact in shaping interpersonal rapport, mitigating face threats, and reinforcing cooperative intent.

Past Tense Verb Significance Refers to the degree to which past tense verbs are salient, accurate, and contextually integrated. It includes frequency, morphological consistency, and the rhetorical or narrative impact on establishing a coherent sense of time and providing historical context.

Causality Significance Refers to the degree to which cause-and-effect relationships are clearly indicated, logically structured, and contextually coherent. This includes the frequency and precision of causal connectives (e.g., *because, therefore, thus*) and the depth of reasoning to explain how conditions lead to outcomes.

Linking Verb Structure Significance Refers to the degree to which linking verbs (e.g., *be, become, seem, appear*) are salient, accurate, and contextually integrated. It emphasizes frequency, morphological correctness, semantic clarity, and effectiveness in conveying states, characteristics, or identities.

Simile Significance Refers to the degree to which similes (e.g., comparisons using *like* or *as*) are salient, creative, and contextually integrated. This definition encompasses frequency, imagery richness, and the rhetorical impact on clarity, vividness, and reader engagement.

D Metric Calculation

D.1 Feature Representation Confidence (FRC)

In our feature analysis experiments, we introduce two key causal probabilities that serve as the basis for computing the Feature Representation Confidence (FRC).

The Feature Representation Confidence (FRC) is computed as the harmonic mean of PN and PS: $FRC = \frac{2\,PN\,PS}{PN+PS}$. The harmonic mean is chosen because it ensures that FRC remains low if either PN or PS is low, thereby providing a balanced measure that only yields a high score when both necessity and sufficiency are strong. This approach allows us to robustly quantify the ability of the SAE latent space's base vectors to represent the targeted linguistic features.

D.2 Feature Intervention Confidence (FIC)

In our methodology, the Feature Intervention Confidence (FIC) score is computed as the harmonic mean of the normalized ablation effect E_{abl} and the normalized enhancement effect E_{enh}:

$$FIC = \frac{2\,E_{abl}\,E_{enh}}{E_{abl} + E_{enh}}.$$

This formulation ensures that FIC is high only when both the ablation and enhancement interventions yield strong effects.

In practice, however, it is possible that one or both of these effects are negative, indicating that an intervention produces an effect opposite to the intended direction. Moreover, even if only one effect is significant while the other is near zero, the feature may still exhibit causal influence. Simply setting an effect that is near zero or negative to 0 would result in an FIC score of 0, which does not adequately capture the underlying causality.

To address this, we introduce a penalty coefficient w to adjust for negative or near-zero effects. Specifically, we define the penalized effect E' for each intervention as follows:

$$E' = \begin{cases} E, & \text{if } E \geq 0, \\ w \cdot |E|, & \text{if } E < 0. \end{cases}$$

Here, w is empirically set to 0.5. Thus, if one of the normalized effects (either E_{abl} or E_{enh}) is negative, we compute its penalized value as 0.5 times its absolute value rather than setting it directly to 0. This approach ensures that even when one of the effects is weak or slightly negative, the FIC score does not vanish entirely, preserving the indication of causality.

Accordingly, the FIC score is then computed as:

$$FIC = \frac{2\,E'_{abl}\,E'_{enh}}{E'_{abl} + E'_{enh}}.$$

In our experiments (see Table2), only the metaphor feature shows a slightly negative ablation effect, while the enhancement and ablation effects for the other features are positive. The introduction of the penalty coefficient w effectively moderates the impact of the negative effect for the metaphor feature, resulting in a more balanced and meaningful FIC score.

This penalty mechanism is crucial because even when only one of the interventions (ablation or enhancement) shows a significant effect, it still provides evidence of the feature's causal role. By incorporating w, we ensure that such cases are not misrepresented by an FIC score of 0, thus offering a more robust measure of the overall causal strength.

E Linguistic Structure

E.1 Linguistics Levels

Morphology The study of the internal structure of words—how roots, prefixes, suffixes, and inflectional endings combine to create different word forms and convey grammatical information such as tense, number, or case.

Syntax The study of how words are arranged into larger units—phrases, clauses, and sentences—and the rules that govern their permissible order and hierarchical relationships within a language.

Semantics The field that investigates meaning at the level of words, phrases, and sentences: how linguistic expressions map to concepts, objects, events, or states of affairs in the world, and how compositional principles let smaller meanings combine into larger ones.

Pragmatics The study of how context and communicative intentions shape meaning in real-world use—how speakers choose utterances to achieve goals, how listeners infer implied or indirect meaning, and how factors like shared knowledge, discourse history, and social norms influence interpretation.

E.2 Linguistic Feature List

past_tense Morphology & Semantics — verb form that locates an event before speech time.

noun_plural Morphology — form marking more than one noun referent.

agentive_suffix Morphology — suffix creating nouns for the doer of an action.

negation_prefix Morphology — prefix that reverses or denies the base meaning.

degree_prefix Morphology — prefix intensifying or scaling the base concept.

temporal_prefix Morphology — prefix adding time relations such as "pre-" or "post-".

quantitative_prefix Morphology — prefix conveying amount or number.

spatial_or_directional_prefix Morphology — prefix indicating place or direction.

nominal_suffix Morphology — suffix that turns a base into a noun.

verbal_suffix Morphology — suffix that turns a base into a verb.

adjectival_suffix Morphology — suffix that turns a base into an adjective.

adverbial_suffix Morphology — suffix that turns a base into an adverb.

possessive_form Morphology & Syntax — morphological marking of ownership or relation.

third_person_singular Morphology & Syntax — verb agreement form for he/she/it.

past_participle Morphology & Syntax — verb form used in perfect aspect or passive voice.

present_participle Morphology & Syntax — "-ing" form used for progressives or gerunds.

comparative Morphology & Semantics — form showing a higher degree of a property.

superlative Morphology & Semantics — form showing the highest degree of a property.

past_tense_irregular Morphology — past form that does not end in "-ed".

past_participle_irregular Morphology — irregular past participle form.

intransitive_verb Syntax — verb that takes no direct object.

transitive_verb Syntax — verb that requires a direct object.

linking_verb Syntax — verb that links subject to a complement.

anaphor Syntax & Pragmatics — expression that refers back to an antecedent.

subject_auxiliary_inversion Syntax — swapping subject and auxiliary (e.g., questions).

subject_verb_inversion Syntax — reversing subject and main verb order.

passive_voice Syntax & Semantics — clause where patient becomes grammatical subject.

subjunctive_mood Syntax & Semantics — form expressing wish, doubt, or hypothetical state.

first_conditional Syntax & Semantics — "if + present, will + verb" for real future possibility.

indirect_speech Syntax & Pragmatics — reporting speech without a direct quote.

elliptical_sentences Syntax — sentences with understood but omitted elements.

cleft_sentences Syntax — “it + be + focus” construction for emphasis.

appositives Syntax — noun phrase renaming another noun phrase.

non_defining_relative_clauses Syntax — extra, non-restrictive relative clauses.

emphatic_structure Syntax & Pragmatics — construction that highlights or stresses a clause part.

noun_clauses Syntax — subordinate clauses functioning as nouns.

relative_clauses Syntax — clauses that modify a noun with a relative word.

imperative_sentence Syntax & Pragmatics — clause issuing a command or request.

of_genitive Syntax — possession expressed with an “of” phrase.

s_genitive Syntax — possession marked with apostrophe-s.

clausal_subjects Syntax — clauses acting as the subject of a sentence.

extraposition Syntax — moving a heavy subject/object to clause end with dummy “it”.

copular_be Syntax — “be” used as a linking verb, not as an auxiliary.

echo_questions Syntax & Pragmatics — repetition of prior utterance to seek confirmation.

tag_questions Syntax & Pragmatics — short question tags appended to statements.

direct_object Syntax — noun phrase receiving the verb’s action.

universal_quantifiers Syntax & Semantics — words like “all, every” signifying totality.

existential_quantifiers Syntax & Semantics — words like “some, any” signifying existence.

expletive Syntax — syntactic placeholder such as “it” or “there”.

factives Semantics & Syntax — predicates presupposing truth of their complement.

futurates Semantics & Syntax — present-tense forms referring to scheduled future events.

intensifiers Semantics & Pragmatics — adverbs that strengthen degree (e.g., “very”).

mass_noun Syntax & Semantics — noun for uncountable substances (e.g., “water”).

object_expletives Syntax — expletive pronouns occupying object position.

nominal_adverbials Syntax — noun phrases functioning like adverbs.

split_infinitives Syntax — placing a word between “to” and the verb stem.

quantifier Syntax & Semantics — word or phrase expressing quantity.

count_nouns Syntax & Semantics — nouns that can be enumerated individually.

active_verbs Syntax — verbs used in active voice constructions.

middle_verb Syntax & Semantics — verb whose subject is patient but appears active.

referring Semantics & Pragmatics — linguistic act of pointing to real-world entities.

static_dynamic Semantics — distinction between state verbs and action verbs.

punctual_durative Semantics — contrast between instantaneous and durational events.

telic_atelic Semantics — events with inherent endpoints vs. those without.

past Semantics — temporal reference before the present moment.

future Semantics — temporal reference after the present moment.

present_progressive Semantics — aspect for ongoing present actions.

present_perfect Semantics — aspect connecting past event to present state.

past_progressive Semantics — aspect for ongoing past actions.

past_perfect Semantics — event completed before a past reference point.

future_progressive Semantics — ongoing action projected into the future.

future_perfect Semantics — event completed before a future reference point.

epistemic Semantics & Pragmatics — modality expressing speaker's judgment of likelihood.

deontic Semantics & Pragmatics — modality expressing obligation or permission.

spatial Semantics — meaning elements relating to location or space.

person Semantics & Pragmatics — grammatical category distinguishing speaker, addressee, others.

temporal Semantics — meaning elements relating to time relations.

given_known Pragmatics & Semantics — information already shared by speaker and listener.

representative Pragmatics — speech act conveying assertions or descriptions.

directive Pragmatics — speech act intended to get the hearer to act.

commisive Pragmatics — speech act committing speaker to future action.

expressive Pragmatics — speech act revealing speaker's feelings or attitude.

declaration Pragmatics — speech act that changes social reality.

metaphor Semantics & Pragmatics — figurative transfer of meaning based on similarity.

synecdoche Semantics & Pragmatics — figure where part stands for whole or vice versa.

non_synecdoche_metonymy Semantics & Pragmatics — metonymic shift based on association, not part-whole.

coordination Syntax & Semantics — joining of equal grammatical elements.

transitional Semantics & Pragmatics — discourse element marking a shift or progression.

resultative Syntax & Semantics — construction expressing a resultant state of an action.

optative Syntax & Pragmatics — form expressing a wish or hope.

existential Semantics & Syntax — clause asserting existence of something.

interrogative Syntax & Pragmatics — clause type used for asking questions.

deixis Pragmatics & Semantics — reference that depends on context (e.g., "here", "now").

turn_taking Pragmatics — conversational management of who speaks when.

euphemism Pragmatics & Semantics — mild term replacing a harsher one.

personification Semantics & Pragmatics — giving human traits to non-human entities.

hyperbole Semantics & Pragmatics — deliberate exaggeration for effect.

discourse_markers Pragmatics — words that organize or signal discourse flow.

politeness Pragmatics — linguistic strategies that mitigate imposition or face threat.

性_抽象名词后缀 形态学— 后缀“-性” 构成表示“-ness/-ity” 的抽象名词。

化_动词性后缀 形态学— 后缀“-化” 构成动词，表示“使.../成为...”。

们_复数后缀 形态学& 语义学— 后缀“-们” 标记人称复数。

重叠构词 形态学& 语义学— 通过词素重叠构词，以强调或表迭代。

不及物动词 句法学& 语义学— 不能带直接宾语的动词。

及物动词 句法学& 语义学— 需要直接宾语的动词。

系动词 句法学— 连接主语与补语的动词。

属格 句法学& 语义学— 所有格或所属关系的语法标记。

逆向结构 句法学& 语义学— 为强调或疑问而颠倒正常语序。

被动语态 句法学& 语义学— 将承事者作为句法主语的被动结构。

主题_述评句 句法学& 语用学— 将句子拆分为主题和述评部分的结构。

回指　句法学& 语义学& 语用学— 指代先行项的表达方式。

间接引语　句法学& 语用学— 不引用原话的转述形式。

省略句　句法学& 语用学— 上下文可恢复的省略结构。

同位结构　句法学— 两个等价名词短语并列重命名的结构。

反问句　句法学& 语用学— 期望无真实答案的修辞性疑问句。

感叹词　语用学— 表达突发情感的独立词。

祈使句　句法学& 语用学— 用于发布命令或请求的句式。

语气助词　形态学& 语义学& 语用学— 表示说话人态度的助词。

轻动词　句法学& 语义学— 与名词搭配使用，语义轻的动词。

主观数量　语义学& 语用学— 说话人评估的模糊数量表达。

使役结构　句法学& 语义学— 表示"使/让某人做..."的致使结构。

条件句　句法学& 语义学— 表达"如果...，就..."条件关系的句子。

兼语句　句法学— 一个名词在结构中既作宾语又作主语。

情态　语义学& 语用学— 表示能力、必要性等的情态范畴。

时体标记　形态学& 语义学— 标记时态或体的形式。

假设　语义学& 语用学— 表示假设情景的表达。

受事主语句　句法学& 语义学— 主语为动作承事者的句子。

可能　语义学& 语用学— 表示可能性或潜在性的表达。

因果　语义学& 语用学— 表示因果关系的表达。

并列　句法学& 语义学— 平等地并列元素的结构。

明喻　语义学& 语用学— 用"像"等词显性标记的比喻。

暗喻　语义学& 语用学— 无显性比较词的隐喻。

比较　语义学— 表示相似或差异的语言表达。

致使　句法学& 语义学— 表示结果状态的致使表达。

让步　语义学& 语用学— 虽承认...但仍...的让步关系。

转折　语义学& 语用学— 标记对比或转折的关系。

递进　语义学& 语用学— 表示进一步增强信息的关系。

指示　语义学& 语用学— 根据上下文指示实体的表达。

话轮转换　语用学— 对话中管理轮到谁发言的结构。

委婉语　语用学— 缓和直接性的委婉表达。

拟人　语义学& 语用学— 将人类特征赋予非人实体的表达。

夸张　语义学& 语用学— 为强调而故意夸大的表达。

话语标记　语用学— 引导和组织话语流程的词语。

礼貌　语用学— 表示礼貌或维护面子策略的语言手段。

数量词　句法学& 语义学— 数词加量词短语，表示确切数量。

F Implementation Details

We used 8 A100 GPUs with 80GB of memory for the experiments. While the exact GPU hours for each experiment were not precisely recorded, the total GPU usage did not exceed one hour. The system was set up with CUDA 12.4, Triton 3.0.0, and Ubuntu 22.04. For the Llama model, we employed the Hugging Face implementation of transformers, and for SAE model, we used the OpenSAE implementation[1] and set the hyperparameter k to 128 for TopK activation.

[1] https://github.com/THU-KEG/OpenSAE