

Query Lens: Interpreting Sparse Key-Value Features with Indirect Effects

Hwiyeong Lee, Ingyu Bang, Uji Hwang, Hyelim Lim, Taeuk Kim*

Hanyang University, Seoul, Republic of Korea

Sparse Autoencoder

- **SAE Feature Interpretation**

- **Mechanistic interpretability** aims to assign human-interpretable descriptions to a model's internal computations or representations.
- Individual neurons are **polysemantic under superposition**, making them an unreliable unit of analysis.
- **Sparse Autoencoders (SAEs)** overcome this by decomposing activations into a sparse, overcomplete dictionary of features.
- Interpreting these **SAE features** is a central task of mechanistic interpretability, and the focus of this work.

Logit Lens

- **Nostalgebraist (2020)**

- Project an **intermediate hidden state** through the unembedding to see what the model is thinking at that layer

$$U^\top h^l \in \mathbb{R}^{|V|}$$

- **Geva et al. (2022, 2023); Dar et al. (2023)**

- Extend the same projection to a single MLP sub-update; turning Logit Lens into a tool for interpreting **learned parameters**.

$$U^\top (a_i^l v_i^l) = a_i^l \cdot U^\top v_i^l \stackrel{\text{argsort}}{=} U^\top v_i^l \in \mathbb{R}^{|V|}$$

- **Bloom & Lin (2024)**

- Apply the same projection to **SAE decoder columns** — now the default vocab-space interpretation of SAE features.

$$U^\top v_i^l \in \mathbb{R}^{|V|}, \quad v_i^l := W_{\text{dec}}^l[:, i]$$

Limitations of Logit Lens

- **Completeness**

- Logit Lens captures the **output-side** behavior of a feature: *which tokens the feature promotes.*
- It leaves the **input-side** behavior unrecovered: *which tokens activate the feature.*

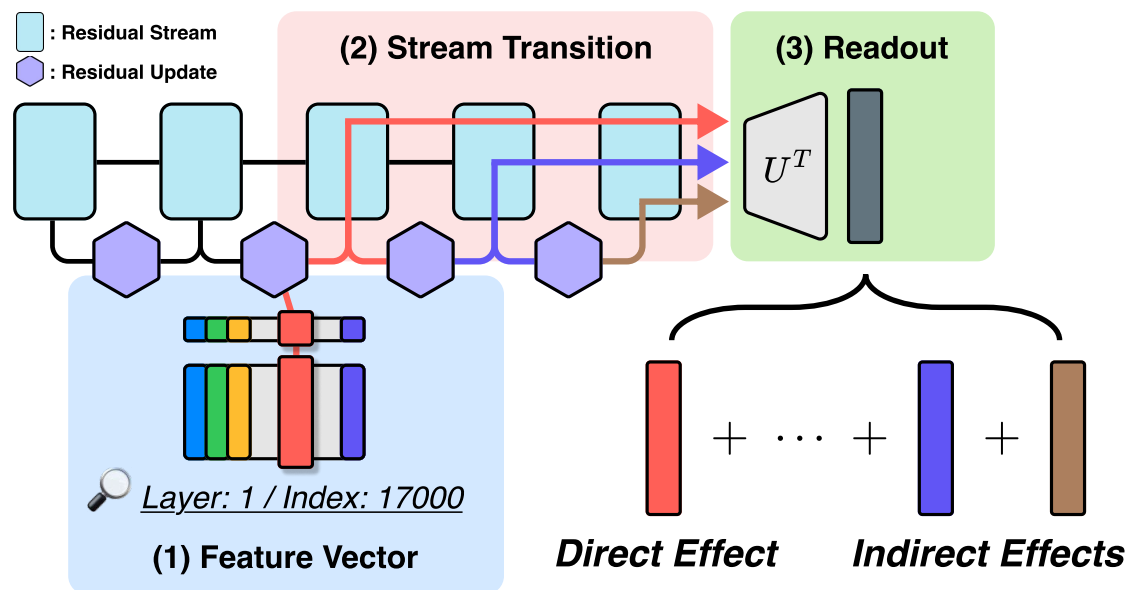
- **Faithfulness**

- Many features, especially in early layers, return diffuse, fragmented token signatures dominated by subword pieces.
- Logit Lens **fails to converge to a coherent meaning** on a large fraction of features.

Addressing the Limitations

• Faithfulness: Indirect Effects

- Logit Lens captures only the **direct effect** — the feature's projection onto vocab space via residual connections.
- A feature added to the residual stream is also **read by downstream modules**, which reshape its effect on the output.
- **Incorporate indirect effects**, not just the direct projection.



Addressing the Limitations

- **Completeness: SAE as Key-Value Memory**

- Encoder vectors = key features: capture what **activates** the feature.
- Decoder vectors = value features: capture what the feature **promotes**.
- **Probe both sides** of the dictionary, not just the decoder.

Query Lens

- **QL-Value**

- Traces the value feature forward to the output endpoint, capturing the feature's **outgoing influence** on the logits.

$$U^\top \left[\prod_{k>l} (I + J^k) \right] v_i^l \in \mathbb{R}^{|V|}$$

- **QL-Key**

- Traces the key feature backward to the input endpoint, capturing the feature's **incoming sensitivity** to input substitutions.

$$(k_i^l)^\top \left[\prod_{k \leq l} (I + J^k) \right] \hat{E} \in \mathbb{R}^{|V|}$$

Experiments

• Quantitative Evaluation

- Input Score: how well the predicted tokens **recover the inputs that activate** the feature.
- Output Score: how well the predicted tokens **recover the outputs the feature promotes** under steering.

• Results

- **Faithfulness**: Each QL variant beats its LL counterpart — indirect effects pay off.
- **Completeness**: QL-Key wins on Input, QL-Value wins on Output — neither variant alone suffices, both are needed.

	GPT-2 Small		Gemma-3-270M		Gemma-3-1B		Qwen-3-1.7B	
	Input (%)	Output (%)	Input (%)	Output (%)	Input (%)	Output (%)	Input (%)	Output (%)
LL _{KEY}	7.84 $\pm$ 1.05	4.32 $\pm$ 0.43	0.71 $\pm$ 0.58	1.79 $\pm$ 0.67	1.74 $\pm$ 1.01	3.25 $\pm$ 0.88	1.91 $\pm$ 0.90	4.43 $\pm$ 0.81
LL _{VALUE}	11.74 $\pm$ 1.41	12.57 $\pm$ 0.60	2.14 $\pm$ 1.25	6.13 $\pm$ 0.98	1.03 $\pm$ 0.67	7.84 $\pm$ 0.93	3.56 $\pm$ 1.28	8.31 $\pm$ 0.97
TL	8.24 $\pm$ 1.22	12.03 $\pm$ 0.60	2.23 $\pm$ 1.23	5.38 $\pm$ 0.94	1.01 $\pm$ 0.67	7.81 $\pm$ 0.92	3.61 $\pm$ 1.29	8.14 $\pm$ 0.97
ZO	6.63 $\pm$ 1.15	5.69 $\pm$ 0.56	5.78 $\pm$ 2.12	2.75 $\pm$ 0.86	4.24 $\pm$ 1.78	3.13 $\pm$ 0.87	12.55 $\pm$ 2.84	6.21 $\pm$ 0.94
TC _{α=1}	30.46 $\pm$ 2.13	13.61 $\pm$ 0.64	10.00 $\pm$ 2.67	5.87 $\pm$ 0.80	8.97 $\pm$ 2.43	6.45 $\pm$ 0.91	14.46 $\pm$ 2.93	8.13 $\pm$ 0.93
TC _{α=5}	31.47 $\pm$ 2.15	13.11 $\pm$ 0.62	10.03 $\pm$ 2.62	8.87 $\pm$ 1.04	9.25 $\pm$ 2.45	8.09 $\pm$ 0.99	14.45 $\pm$ 2.93	8.77 $\pm$ 0.95
TC _{α=10}	31.41 $\pm$ 2.13	12.68 $\pm$ 0.59	10.05 $\pm$ 2.63	9.74 $\pm$ 1.06	9.31 $\pm$ 2.46	8.37 $\pm$ 1.00	14.23 $\pm$ 2.90	8.89 $\pm$ 0.97
QL _{KEY}	39.32 $\pm$ 2.21	1.97 $\pm$ 0.33	21.82 $\pm$ 3.71	1.43 $\pm$ 0.62	14.14 $\pm$ 2.71	1.45 $\pm$ 0.56	21.69 $\pm$ 3.48	0.55 $\pm$ 0.27
QL _{VALUE}	26.43 $\pm$ 2.01	15.24 $\pm$ 0.68	10.27 $\pm$ 2.69	10.63 $\pm$ 1.03	8.61 $\pm$ 2.35	9.26 $\pm$ 1.00	11.65 $\pm$ 2.58	9.36 $\pm$ 0.99

Experiments

• Examples

Setting	Feature	Top-scoring tokens	Activation / Steering example
GPT-2 Small	k_{6633}^1	_strong, _strongest, _Strong, _weak , _stronger, strong, _weakest, Strong , _strengthen	“...known mesons, which travel finite distances before decaying via the weak nuclear interaction. The B_s meson is composed of a quark and...” “...along parts of the Atlantic Seaboard and the Appalachians into early next week. Strong winds, coastal flooding and beach erosion will...”
	v_{8824}^2	_Unicode, _FreeBSD, BSD, _cryptographic, _Linux , _gcc, _Lisp, _UTF, _crypto , _Java	(+0.25) <i>Have you ever noticed that:</i> “with the crypto game market it’s no one’s business to make your system trying to play your game...” (+0.5) <i>Over the years, I noticed that:</i> “the most common starting point for Linux tools is a hardware driver. This is not really true in most...”
Gemma-3-1B	k_{16433}^7	_measles , pox, _pox, _Meas, _polio , Meas, 鹿, _syphilis, _rhy, _syphilit, 尧, 病, _Tilly	“...particularly among the lay public concerning the safety of vaccines. Recently an increase in reported measles cases in the USA has again...” “...from 2006 to 2009 India was faced with over 500 polio cases. But, in 2009, some new initiatives were introduced that looked at narrowing...”
	v_{62574}^{17}	_cooks, consumers , students , _designers, managers, _Cooks, _students, COOK, cooking	(+0.5) <i>It is observed that:</i> “the number of students has increased from 180 to 330 in the class...” (+0.5) <i>Imagine a world where:</i> “our kids are not just consumers but creators, where they learn coding from an early age, where they...”

Experiments

• Qualitative Evaluation

- Interpretability Score: how concentrated a method's tokens are around a **single coherent theme**, judged by GPT-5-nano.

• Results and Examples

Setting	LL _{KEY}	QL _{KEY}	LL _{VALUE}	QL _{VALUE}
GPT-2 Small	3.03 $\pm$ 0.12	4.02 $\pm$ 0.14	4.87 $\pm$ 0.15	5.95 $\pm$ 0.16
Gemma-3-270M	4.00 $\pm$ 0.27	5.44 $\pm$ 0.24	4.10 $\pm$ 0.27	5.34 $\pm$ 0.28
Gemma-3-1B	4.00 $\pm$ 0.29	4.86 $\pm$ 0.24	4.58 $\pm$ 0.29	5.30 $\pm$ 0.29
Qwen-3-1.7B	3.47 $\pm$ 0.24	4.76 $\pm$ 0.28	5.06 $\pm$ 0.30	6.96 $\pm$ 0.30

Feature	Lens	Top-scoring tokens	Score
k_{29376}^{11}	LL _{KEY}	stuff, grain, issance, banks, _traged, _rul, _Babe, EVA, _destro, _obser, icz, anguage, _stamp, selves, _Flavoring	1
	QL _{KEY}	_Food, Food, _food, food, _foods, _SNAP, _grocery, _famine, _feeding, _supermarket, _culinary, _Foods	10
v_{4209}^0	LL _{VALUE}	_resil, soDeliveryDate, >>>>>>, _Haram, SCP, _Hots, ried, stru, _biologist, _Geh, gie, _ecology, _CTR	3
	QL _{VALUE}	_forestry, _deforestation, _forests, _WWF, _forest, _zoo, forest, Forest, _Forestry, _tree, _Forest, forestation	9

Feature	Lens	Top-scoring tokens	Score
k_{35333}^{17}	LL _{KEY}	ónica, 鋼, лəʏ, _vān, irchen, ल्याण, _chigar, itemconfig, ('.')[, _mellitus, hbox, ighi, ayu, ईष्ठा, odhc, ', dorff	1
	QL _{KEY}	_bank, _Bank, Bank, _Banks, _banks, _BANK, _banking, _баHKA, _banca, _बैक, 銀行, _은행, 銀行, _banker	10
v_{49129}^{13}	LL _{VALUE}	_Pohl, _ゴールド, _魔, _保, _Gubernur, _ム, _ミ, _Erinner, _ブ, _ルー, _オシヤレ, _hangat, _아름	3
	QL _{VALUE}	_Nicky, _Rhonda, _Pablo, _Rita, _Martha, _Ricky, _Willie, _Sarah, _Andrea, _Sasha, _Sherry, _Bella	10

(a) GPT-2 Small

(b) Gemma-3-1B

Conclusion

- **Motivated by the two limitations** of Logit Lens when applied to SAE feature interpretation, we propose **Query Lens**.
- **Regarding completeness**, we probe both **encoder (key)** and **decoder (value)** sides to recover both **incoming** and **outgoing** causal roles.
- **Regarding faithfulness**, we incorporate **indirect effects** via the residual stream's Jacobian product, beyond the direct projection of Logit Lens.
- **Subspace Channel Hypothesis**: we find that downstream modules read features through **layer-specific low-rank** subspaces.

Thank You

Presenter
Hwiyeong Lee

HYU NLP Lab. at Hanyang Univ.
hyglee@hanyang.ac.kr