How LLMs Learn:
Tracing Internal Representations with Sparse Autoencoders

Anonymous ACL submission

Abstract

Large Language Models (LLMs) demonstrate remarkable multilingual capabilities and broad knowledge. However, the internal mechanisms underlying the development of these capabilities remain poorly understood. To investigate this, we analyze how the information encoded in LLMs' internal representations evolves during the training process. Specifically, we train sparse autoencoders at multiple checkpoints of the model and systematically compare the interpretative results across these stages. Our findings suggest that LLMs initially acquire language-specific knowledge independently, followed by cross-linguistic correspondences. Moreover, we observe that after mastering token-level knowledge, the model transitions to learning higher-level, abstract concepts, indicating the development of more conceptual understanding.

1 Introduction

Large Language Models (LLMs) have achieved remarkable success across a wide range of natural language processing tasks, from multilingual translation to advanced semantic understanding (Bubeck et al., 2023). As these models become increasingly complex and widespread, the need to understand their internal mechanisms has grown significantly. This has fueled a surge of research aimed at interpreting their mechanisms and decision-making processes, leading to intriguing insights into their behavior (Casper et al., 2023; Bereska and Gavves, 2024).

However, fundamental questions regarding how LLMs acquire and develop these capabilities remain poorly understood. For instance, do LLMs learn language-specific concepts independently, or do they simultaneously acquire cross-lingual concepts that generalize across languages? Similarly, is there a prioritization in learning low-level, token-specific features versus high-level, abstract concepts?

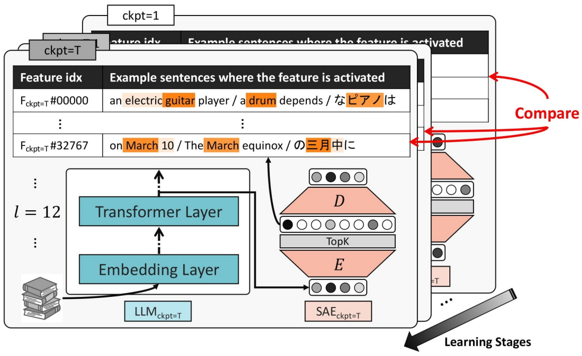

Figure 1: Illustration of our approach to comparing internal representations across different training stages of LLMs. We train SAEs on the internal representation from multiple checkpoints.

In this work, we address this gap by analyzing how the information encoded in the internal representations of LLMs evolves over time. Specifically, we employ sparse autoencoders (SAEs) (Bricken et al., 2023; Huben et al., 2024) to analyze the hidden representations from multiple checkpoints of a large language model. By examining the distribution of SAE features at each checkpoint, we identify the types of information the model encodes at different training stages of its development (see Figure 1).

Our experiments yield two key findings: (1) LLMs first learn knowledge within individual languages before acquiring cross-lingual mappings (§4.3), and (2) they initially capture fine-grained, token-level knowledge before progressing to more abstract, conceptual representations (§4.4) These findings offer new insights into the internal mechanisms that underlie the emergence of LLMs' generalization abilities.

2 Sparse Autoencoders

A sparse autoencoder (SAE) is an autoencoder that enforces a sparsity constraint on its hidden layer. In this study, we adopt a variant called TopK-SAE (Makhzani and Frey, 2014), where the TopK activation function is applied at the hidden layer. Compared to a ReLU-based SAE (Bricken et al., 2023; Huben et al., 2024), TopK-SAE has been shown to be easier to train while maintaining sparsity and achieving higher reconstruction performance (Gao et al., 2025).

Let $x \in \mathbb{R}^d$ be the input vector and n be the dimension of the hidden layer. The encoder and decoder are defined as follows:

$$z = \mathrm{TopK}\big(W_{\mathrm{enc}}(x - b_{\mathrm{pre}})\big), \quad (1)$$
$$\hat{x} = W_{\mathrm{dec}} z + b_{\mathrm{pre}}, \quad (2)$$

where $W_{\mathrm{enc}} \in \mathbb{R}^{n \times d}$ and $W_{\mathrm{dec}} \in \mathbb{R}^{d \times n}$ are learned linear layers, and $b_{\mathrm{pre}} \in \mathbb{R}^d$ is a learnable bias parameter. W_{dec} is initialized as the transpose of W_{enc}, and b_{pre} is initialized to the geometric median of the input data.

The training objective is the following mean squared error (MSE) loss:

$$L = \|x - \hat{x}\|_2^2. \quad (3)$$

Two hyperparameters control TopK-SAE. In this study, we control TopK-SAE by two hyperparameters: n, the dimension of the hidden layer, and K, the number of hidden dimensions to keep active. Interpreting W_{dec} as n distinct vectors in $\mathbb{R}^d$, TopK-SAE can be seen as selecting K vectors from n and using their weighted sum to reconstruct the input. In this study, we denote each dimension of the encoder output $z \in \mathbb{R}^n$ as a *feature*. When a feature is selected in the top-K operation and used in reconstruction, we say the feature is *activated*.

3 Preliminary Experiments

We begin by conducting preliminary experiments on a pre-trained LLM to tune SAE hyperparameters and validate the interpretability of resulting features.

3.1 Experimental Setup

We use the 12th layer output ($d = 2048$) of the 24-layer model `llm-jp-3-1.8B`[1] (Aizawa et al., 2024) as the input to the TopK-SAE. The model is trained on the LLM-jp Corpus v3[2], which contains a total of 1.7T tokens: 950B for English, 592B for Japanese, 114B for code, 0.8B for Korean, and 0.3B for Chinese. We selected the `llm-jp-3-1.8B` model because its intermediate checkpoints are (or will be) publicly accessible, it has over 1 billion parameters to exhibit emergent behaviors, and its training on both English and Japanese enables cross-lingual analysis.

[1] https://huggingface.co/llm-jp/llm-jp-3-1.8b
[2] https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3

Figure 2: Reconstruction loss for varing hidden dimensions n and the sparsity K. Larger n and K improve reconstruction accuracy.

We train TopK-SAE with the Japanese and English Wikipedia subsets in the LLM-jp Corpus v3. For each document, we extract the first 65 tokens as the input to the LLM, discard the `[BOS]` token representation, and apply L2 normalization to the remaining 64 token embeddings, which serve as inputs to the SAE. The dataset consists of 165M tokens (50% Japanese, 50% English), split into 80% for training, 10% for validation, and 10% for testing. We fix the batch size at 32,768, use a warm-up phase of 1,000 steps, and perform a grid search to optimize the learning rate. Training a single SAE took about 1 hour using two A100 40GB GPUs.

3.2 Effect of Hyperparameters

Figure 2 shows how the impact of varying the hidden dimensions n and the number of active dimensions K. Increasing either n or K reduces the reconstruction error. However, these hyperparameters significantly influence interpretability: if n or K is too large, a single concept may be fragmented into multiple features; if it is too small, multiple distinct concepts may be merged into a single feature. Identifying the optimal balance between reconstruction performance and interpretability remains an active area of research (Menon et al., 2024; Leask et al., 2025).

3.3 Patterns in Feature Activation

Figure 3(c) shows examples of feature activations ($n = 32768$ and $K = 32$). The background

	Feature idx	Example sentences where the feature is activated	Language (§ 4.3)	Granularity (§ 4.4)
(a)	$\mathrm{F}_{\mathrm{ckpt}=100}$ #00002	・, 10th Earl of Scarbrough (16 November 1857 ・ called radiological pollution, is ・）は、「日本の貴婦人	Mixed	Uninterpretable
	$\mathrm{F}_{\mathrm{ckpt}=100}$ #00004	・ investigations are performed by geotechnical ・Colonel Doyle Raphard Yardley (April 21, 1913 - ・は日本の防衛官僚	Mixed	Uninterpretable
(b)	$\mathrm{F}_{\mathrm{ckpt}=10000}$ #00004	・ dorsalis), also known as the scrub ・ regnans, known variously as ・ nerve) also known as the fourth	English	Token-Level: " known"
	$\mathrm{F}_{\mathrm{ckpt}=10000}$ #00009	・石油生産設備から ・冷暖房設備、冷凍冷蔵設備、動力設備又は ・のプラント設備を	Japanese	Token-Level: "設備"
(c)	$\mathrm{F}_{\mathrm{ckpt}=988240}$ #00009	・, where fluency is defined as linguistic ・. "Arbitrary" here means that the ・ここで言う「都市」には	Mixed	Concept-Level (Synonymy): *"Defining certain terms"*
	$\mathrm{F}_{\mathrm{ckpt}=988240}$ #00016	・. It is a colorless liquid with a smell reminiscent ・ as an olive green to black, odorless solid ・特有の臭気のある白色個体で、	Mixed	Concept-Level (Semantic Sim.) : *"properties of a substance"*

Figure 3: Examples of feature activations across different training checkpoints. (a) Checkpoint 100, (b) Checkpoint 10,000, (c) Final checkpoint (988,240). Early in training, features activate on seemingly random fragments. As training progresses, features begin to capture language-specific or token-level meanings. By the final checkpoint, they encode higher-level cross-lingual semantics and abstract conceptual knowledge. Additional examples are provided in Figure 6 and in the supplementary data (see Appendix A).

color density indicates the magnitude of activation. For instance, $\mathrm{F}_{\mathrm{ckpt}=988240}\#00009$ strongly activates in segments defining certain terms, while $\mathrm{F}_{\mathrm{ckpt}=988240}\#00016$ activates in text about the smell, color, or state of a substance. These examples, along with others in Figure 6(c), demonstrate that the features of TopK-SAE successfully capture semantically coherent and interpretable meanings.

4 Experiments

In this section, we conduct main experiments, where we train SAEs on the internal representations of a large language model (LLM) at multiple training checkpoints. By analyzing the resulting features, we investigate how the encoded information evolves over the course of training.

4.1 Experimental Setup

We use six checkpoints of `llm-jp-3-1.8B` at training steps 10, 100, 1,000, 10,000, 100,000, and the final checkpoint at step 988,240. For each checkpoint, we train a TopK-SAE with a hidden dimension of $n = 32768$ and a sparsity level of $K = 32$, following the same training conditions described in §3.1.

4.2 Evaluating Feature Activation Patterns

For each feature, we collect up to 50 texts that activate it the most. We then categorize these activation patterns in terms of Language Trend and Semantic Granularity.

Language Trend The language trend of a feature is classified into three categories: *English*, *Japanese*, and *Mixed*. *English* features are activated in texts that are at least 90% English, while *Japanese* features are activated in texts that are at least 90% Japanese. *Mixed* features are activated in texts containing a mix of Japanese and English. For each checkpoint, we automatically categorize the language trend of all 32768 features.

Semantic Granularity The semantic granularity of a feature is categorized into four levels: *Token-Level*, *Concept-Level* (Synonymy), *Concept-Level* (Semantic Sim.), and *Uninterpretable*. *Token-Level* features consistently activate on identical tokens (e.g., only "cat"). *Concept-Level* (Synonymy) features activate on tokens or sentences expressing the same meaning (e.g., "cat" and "ねこ"). *Concept-Level* (Semantic Sim.) features activate on tokens or sentences sharing related meanings (e.g., "cat" and "dog"). *Uninterpretable* features show no clear semantic pattern among the activated texts. For each checkpoint, we manually categorize the semantic granularity of the first 100 features.

4.3 Language Trends Over Checkpoints

Figure 4 shows the proportion of features exhibiting each language trend across checkpoints. Early

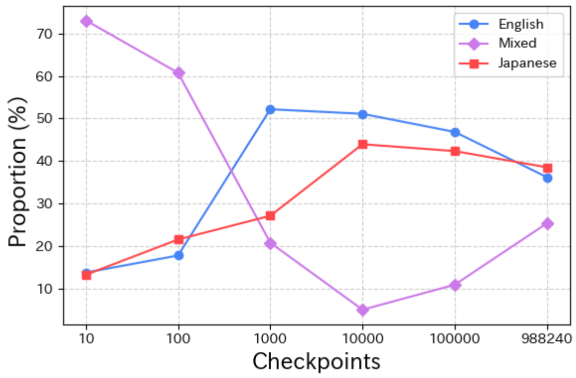

Figure 4: Proportion of language trends over different checkpoints.

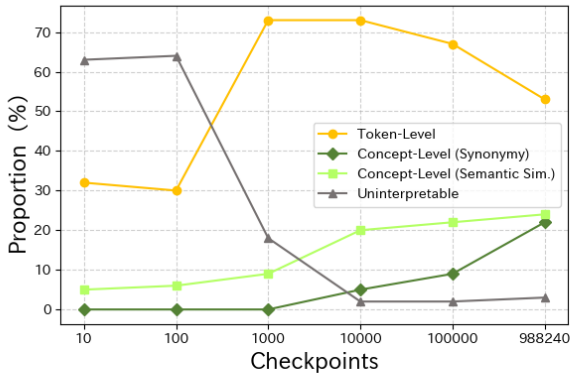

Figure 5: Proportion of semantic granularity patterns over different checkpoints.

in training, most features are classified as *Mixed*, and these features typically activate on random tokens without clear semantic coherence (see Figures 3(a) and 6(a)).

By the mid-training checkpoints, we observe a higher fraction of *English* and *Japanese* features. Within-language semantic coherence emerges here (Figures 3(b) and 6(b)). Toward the later checkpoints, the proportion of *Mixed* features rises again, but unlike the early-stage randomness, these features now capture cross-lingual correspondences (Figures 3(c) and 6(c)).

This suggests that LLMs learn in two stages. First, from early to mid-training, they acquire semantics within each language. Second, from mid to late training, they begin capturing cross-lingual correspondences.

4.4 Semantic Granularity Over Checkpoints

Figure 5 shows the distribution of semantic granularity categories for 100 sampled features at each checkpoint. We observe a rise in *Token-Level* features from early to mid-training, and then an increase in *Concept-Level* (either synonymy or semantically related) features from mid to late training. Meanwhile, *Uninterpretable* features decrease steadily as training proceeds.

This pattern suggests that LLMs initially learn fine-grained token-level knowledge and then transition to capturing abstract, concept-level semantic relationships.

5 Related Work

Recent studies show neural networks can represent more features than their dimensions (Elhage et al., 2022). To disentangle these representations, SAEs have emerged as a key tool for decomposing them into interpretable components (Huben et al., 2024; Olshausen and Field, 1997). While early work primarily focused on single-trained SAEs, recent studies have shifted toward comparing SAE features across layers (Balcells et al., 2024; Balagansky et al., 2025), model architectures (Lan et al., 2024; Lindsey et al., 2024), or fine-tuning stages (Lindsey et al., 2024; Wang et al., 2025). Concurrent work tracks feature formation during training (Xu et al., 2024), but lacks quantitative evaluation. Our contribution is training independent SAEs at each checkpoint and conducting both qualitative and quantitative analyses.

During training, LLMs exhibit rapid performance improvements on specific tasks, known as emergent capability (Wei et al., 2022), where abilities appear when the model size or data volume exceeds a certain threshold, or grokking (Power et al., 2022), where models suddenly generalize better after overfitting. Recent research has begun to explore the mechanisms of these phenomena using simplified models (Nanda et al., 2023). However, understanding the relationship between these abrupt performance changes and the internal states of models remains an open challenge.

6 Conclusion

In this study, we performed a cross-checkpoint analysis of the internal representations of a large language model via a sparse autoencoder. Our results indicate that LLMs first acquire token-level semantics in a language-specific manner and later learn cross-lingual correspondences (§4.3). Further, they progress from token-level to concept-level representations, forming more abstract knowledge structures over training (§4.4).

7 Limitations

Our study has several limitations. First, sparse autoencoders (SAEs) are not fully interpretable because reconstruction is imperfect and features are not perfectly monosemantic. This limitation can lead to information loss or polysemantic features, which complicates the analysis of internal representations. Second, our findings are based on a specific model and dataset, so they may not generalize to other architectures or training regimes. Finally, the manual categorization of semantic granularity introduces subjectivity, which could affect the consistency of the results. Future work should address these limitations to improve interpretability and robustness.

Acknowledgements

In this research work, we used DeepSeek and ChatGPT for assistance purely with the language of the paper and coding Additionally, we used the "mdx: a platform for building data-empowered society" (Suzumura et al., 2022) for our experiments and analyses.

References

Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, Hiroki Kanezashi, Hiroshi Kataoka, Satoru Katsumata, Daisuke Kawahara, Seiya Kawano, Atsushi Keyaki, Keisuke Kiryu, Hirokazu Kiyomaru, Takashi Kodama, Takahiro Kubo, Yohei Kuga, Ryoma Kumon, Shuhei Kurita, Sadao Kurohashi, Conglong Li, Taiki Maekawa, Hiroshi Matsuda, Yusuke Miyao, Kentaro Mizuki, Sakae Mizuki, Yugo Murawaki, Akim Mousterou, Ryo Nakamura, Taishi Nakamura, Kouta Nakayama, Tomoka Nakazato, Takuro Niitsuma, Jiro Nishitoba, Yusuke Oda, Hayato Ogawa, Takumi Okamoto, Naoaki Okazaki, Yohei Oseki, Shintaro Ozaki, Koki Ryu, Rafal Rzepka, Keisuke Sakaguchi, Shota Sasaki, Satoshi Sekine, Kohei Suda, Saku Sugawara, Issa Sugiura, Hiroaki Sugiyama, Hisami Suzuki, Jun Suzuki, Toyotaro Suzumura, Kensuke Tachibana, Yu Takagi, Kyosuke Takami, Koichi Takeda, Masashi Takeshita, Masahiro Tanaka, Kenjiro Taura, Arseny Tolmachev, Nobuhiro Ueda, Zhen Wan, Shuntaro Yada, Sakiko Yahata, Yuya Yamamoto, Yusuke Yamauchi, Hitomi Yanaka, Rio Yokota, and Koichiro Yoshino. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *arXiv preprint arXiv:2407.03963*.

Nikita Balagansky, Ian Maksimov, and Daniil Gavrilov. 2025. Mechanistic permutability: Match features across layers. In *The Thirteenth International Conference on Learning Representations*.

Daniel Balcells, Benjamin Lerner, Michael Oesterle, Ediz Ucar, and Stefan Heimersheim. 2024. Evolution of sae features across layers in llms. *arXiv preprint arXiv:2410.08869*.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety - a review. *Transactions on Machine Learning Research*.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Stephen Casper, Tilman Rauker, Anson Ho, and Dylan Hadfield-Menell. 2023. Sok: Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. In *First IEEE Conference on Secure and Trustworthy Machine Learning*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. 2024. Sparse autoencoders reveal universal feature spaces across large language models. *arXiv preprint arXiv:2410.06981*.

Patrick Leask, Bart Bussmann, Michael T Pearce, Joseph Isaac Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. 2025. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations*.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. 2024. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits Thread*.

Alireza Makhzani and Brendan Frey. 2014. k-sparse autoencoders. *arXiv preprint arXiv:1312.5663*.

Abhinav Menon, Manish Shrivastava, Ekdeep Singh Lubana, and David Krueger. 2024. Analyzing (in)abilities of SAEs via formal languages. In *MINT: Foundation Model Interventions*.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

Bruno A. Olshausen and David J. Field. 1997. Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Research*, 37(23):3311–3325.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.

Toyotaro Suzumura, Akiyoshi Sugiki, Hiroyuki Takizawa, Akira Imakura, Hiroshi Nakamura, Kenjiro Taura, Tomohiro Kudoh, Toshihiro Hanawa, Yuji Sekiya, Hiroki Kobayashi, Yohei Kuga, Ryo Nakamura, Renhe Jiang, Junya Kawase, Masatoshi Hanai, Hiroshi Miyazaki, Tsutomu Ishizaki, Daisuke Shimotoku, Daisuke Miyamoto, Kento Aida, Atsuko Takefusa, Takashi Kurimoto, Koji Sasayama, Naoya Kitagawa, Ikki Fujiwara, Yusuke Tanimura, Takayuki Aoki, Toshio Endo, Satoshi Ohshima, Keiichiro Fukazawa, Susumu Date, and Toshihiro Uchibayashi. 2022. mdx: A cloud platform for supporting data science and cross-disciplinary research collaborations. In *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 1–7.

Junxuan Wang, Xuyang Ge, Wentao Shu, Qiong Tang, Yunhua Zhou, Zhengfu He, and Xipeng Qiu. 2025. Towards universality: Studying mechanistic similarity across language model architectures. In *The Thirteenth International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Yang Xu, Yi Wang, and Hao Wang. 2024. Tracking the feature dynamics in llm training: A mechanistic study. *arXiv preprint arXiv:2412.17626*.

A Additional Examples of Feature Activations

We provide a supplementary zip file containing 100 SAE feature examples for each training checkpoint of `llm-jp-3-1.8B`, enabling a detailed examination of activation patterns. We include examples for features not depicted in Figures 3 and 6, allowing readers to confirm that similar patterns hold for features beyond those presented in this paper. Each example contains token-activation value pairs, language distribution, and semantic granularity. For more details, please refer to the `README.md` in the zip file.

	Feature idx	Example sentences where the feature is activated	Language (§ 4.3)	Granularity (§ 4.4)
(a)	$F_{ckpt=100}$ #00000	・Clear title is the phrase used ・sphere (variations are known as spherical ・が販売するクロスオーバーSUVである。	Mixed	Uninterpretable
	$F_{ckpt=100}$ #00005	・powers, which the comic calls ・Railroad's class N2sa comprised rebuilds ・そつぎょうけんてい）とは、	Mixed	Uninterpretable
(b)	$F_{ckpt=10000}$ #00000	・イギリスの女流小説家。 ・女流棋士初の女流タイトルグランドスラム ・の女流王将戦である。	Japanese	Token-Level: "女流"
	$F_{ckpt=10000}$ #00005	・guy-wired aerial masts for ・Aerial reconnaissance spotted a ・flapping-winged aerial robot, and	English	Token-Level: "erial"
	$F_{ckpt=10000}$ #00008	・In late mornings and during ・a Saturday morning animated series ・licensed to Morningside, Maryland	English	Token-Level: " morning"
	$F_{ckpt=10000}$ #00010	・live flagship daytime show. It ・both daytime and primetime television. ・, and only 1 watt nighttime	English	Token-Level: "time"
	$F_{ckpt=10000}$ #00011	・ある。類語には鶏鳴の助や ・ジャーゴン（俗語、隠語）である。 ・微妙」の略語。開経	Japanese	Token-Level: "語"
	$F_{ckpt=10000}$ #00048	・皇女として生まれ、のちにボイオーティアの ・生まれや生い立ちは不明だが時宗の ・裕福な家庭で育つが、父親から「	Japanese	Concept-Level (Semantic Sim.): *"biographical background"*
	$F_{ckpt=10000}$ #00087	・Rockstar Lincoln Limited (formerly Spidersoft Limited ・Hobart Sky Ranch Airport is a public-use ・Arras Football Association is a French association	English	Concept-Level (Synonymy): *"Proper nouns"*
(c)	$F_{ckpt=988240}$ #00006	・about a mile (1.6 km) east of the ・36.6 square miles (94.8 km), of ・Located 4 miles north from Wasilla	English	Token-Level: " mile(s)"
	$F_{ckpt=988240}$ #00007	・旧表記（数え年）にて表記。 ・0から数え始め、1 ・一つに数えられることがある。	Japanese	Token-Level: "数え"
	$F_{ckpt=988240}$ #00017	・津海道（しんかい-どう）は ・かけての津藩の藩士である。 ・は岡山県御津郡にあった村。	Japanese	Token-Level: "津"
	$F_{ckpt=988240}$ #00021	・for cyclists (e.g. cyclist-only paths ・itself, for example on signage. ・languages spoken, such as Belgium	English	Concept-Level (Synonymy): *"examples and instances"*
	$F_{ckpt=988240}$ #00026	・Darkened Skye is a ・baryonic dark matter is hypothetical dark matter ・よりはダーク・ファンタジー	Mixed	Concept-Level (Synonymy): *"dark"*
	$F_{ckpt=988240}$ #00039	・The game was developed by Beam Software ・It was part of Mutual Film Corporation's ・に上海美術映画作成所より制作された	Mixed	Concept-Level (Semantic Sim.): *"production company"*
	$F_{ckpt=988240}$ #00041	・is located nine kilometers south-west of ・airport located 13 km northwest of ・airport located seventeen miles (	English	Concept-Level (Semantic Sim.): *"distance value"*

Figure 6: Examples of feature activations across different training checkpoints. (a) Checkpoint 100, (b) Checkpoint 10,000, (c) Final checkpoint (988,240).