Master's Thesis Proposal

Evaluation of Sparse Autoencoder-based Refusal Features in LLMs: Dataset-dependence study

Prof. Dr. Peter Knees (Main Supervisor)
Ilya Lasy, MSc. (Second Supervisor)

TU Wien Informatics
Research Unit Data Science

Tilman Kerl, BSc. | March 28, 2025

Refusal

Model's ability to **identify and reject** inputs that could lead to **harmful, unethical**, or otherwise **inappropriate outputs**

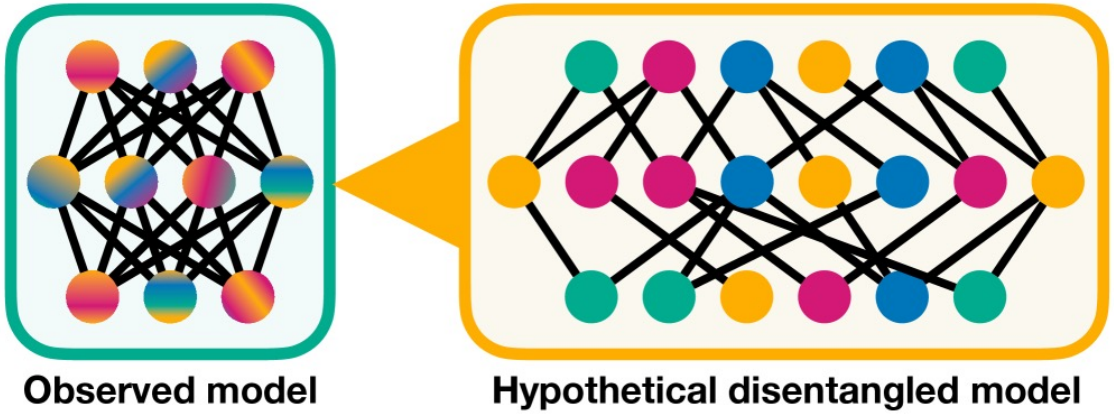
SOTA Refusal

- Base Models Refuse Too
- Fine-tuning
- Reinforcement-learning with HF
- Adversarial Training

[4] [10] [14] [28] [36] [39] [42]

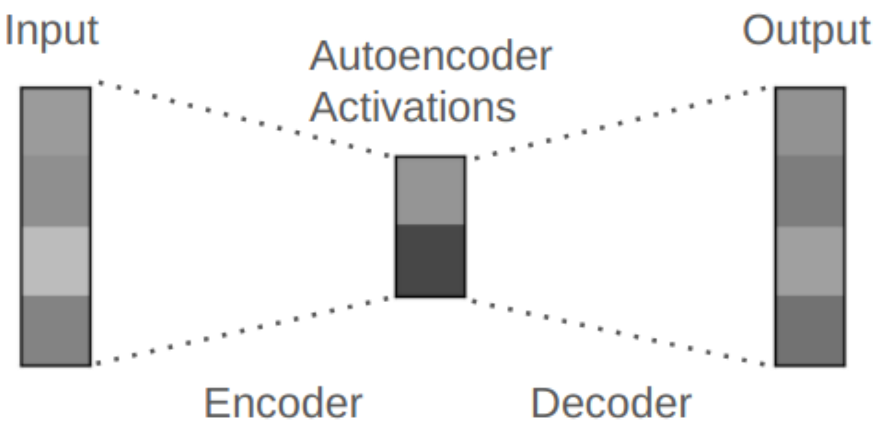
Mechanistic Interpretability

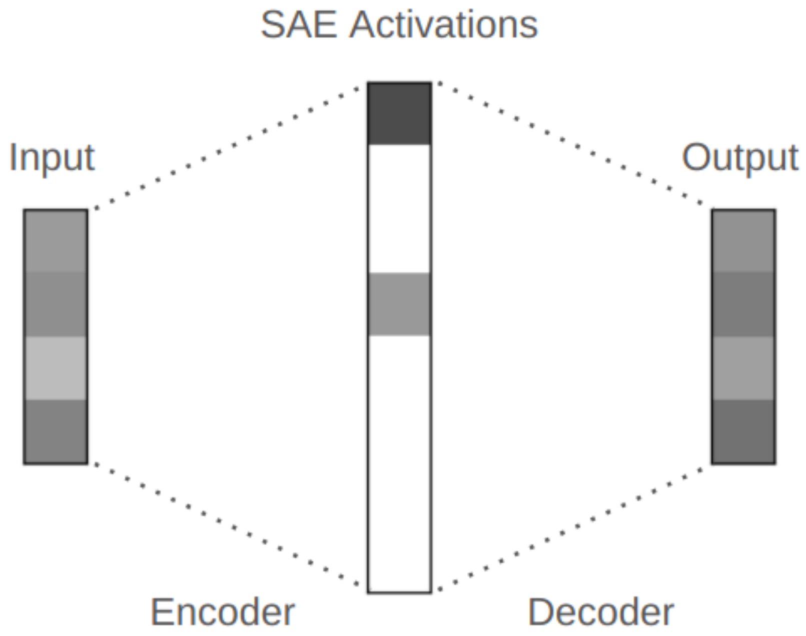
- Reverse Engineering
- E.g. “Understanding” what parts of the model are responsible for what
- Polysemanticity and Superposition
  → *individual neurons or attention heads encode multiple distinct concepts*



BERESKA, L., AND GAVVES, E. Mechanistic interpretability for ai safety – a review, 2024.

Towards Monosemanticity: Sparse Autoencoders (SAEs)

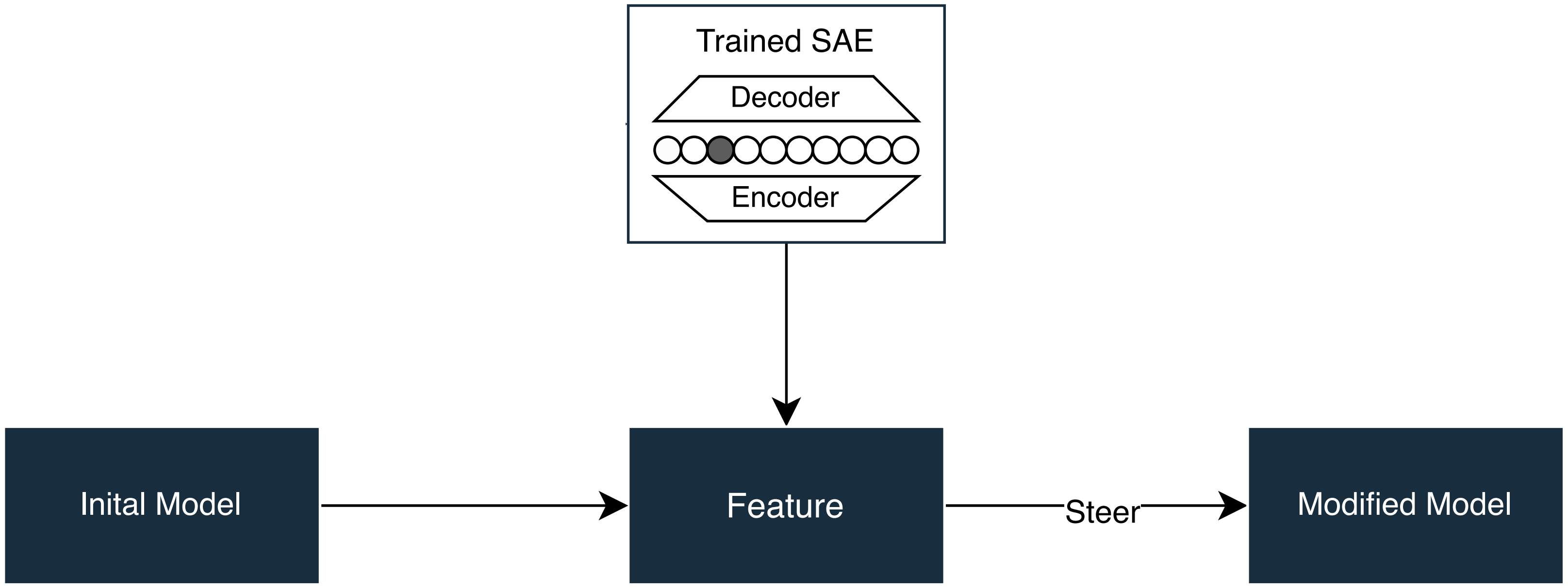




→ learning a sparse, overcomplete basis (input dim. < output dim.)
→ each latent feature corresponds to a distinct, disentangled concept

KARVONEN, A. An intuitive explanation of sparse autoencoders for llm interpretability, Jun 2024. Accessed: 18.12.2024.

Steering with SAEs



→ Extracted features can be altered and used to modify model outputs

Research Questions

1. How does the choice of training dataset affect the sparse autoencoder's ability to isolate and represent refusal features within a language model's latent activations?
2. Which characteristics of the underlying training data are most predictive of the strength and clarity of SAE-extracted refusal features?
3. How do extracted refusal features of sparse autoencoders trained instruction-datasets compare to those trained on the original pre-training corpus, in terms of robustness, interpretability, downstream performance and controllability of refusal-related features?

RQ 1. SAE’s ability to isolate refusal

How does the choice of training dataset affect the sparse autoencoder’s ability to isolate and represent refusal features within a language model’s latent activations?

The focus is on comparing the impact of the original pre-training data against alternative datasets (e.g., instruction data) in shaping the SAE’s capacity to identify refusal-related features from model activations. To properly conduct this evaluation, we rely on fully open-source models.

Metric: Sparsity

RQ 2. Clarity and Strenght of Refusal

Which characteristics of the underlying training data are most predictive of the strength and clarity of SAE-extracted refusal features?

The training data indirectly shapes the model's latent activations, which are the basis for SAE-extracted refusal features. Understanding which dataset characteristics—such as distributional properties, topical diversity, or the frequency of refusals—most strongly influence these activations can help identify the conditions under which refusal-related features emerge most clearly and robustly.

Metric: Refusal Score

RQ 3. Comparison Base vs. Instruct Models

How do extracted refusal features of sparse autoencoders trained on instruction-datasets compare to those trained on the original pre-training corpus, in terms of robustness, interpretability, downstream performance and controllability of refusal-related features?

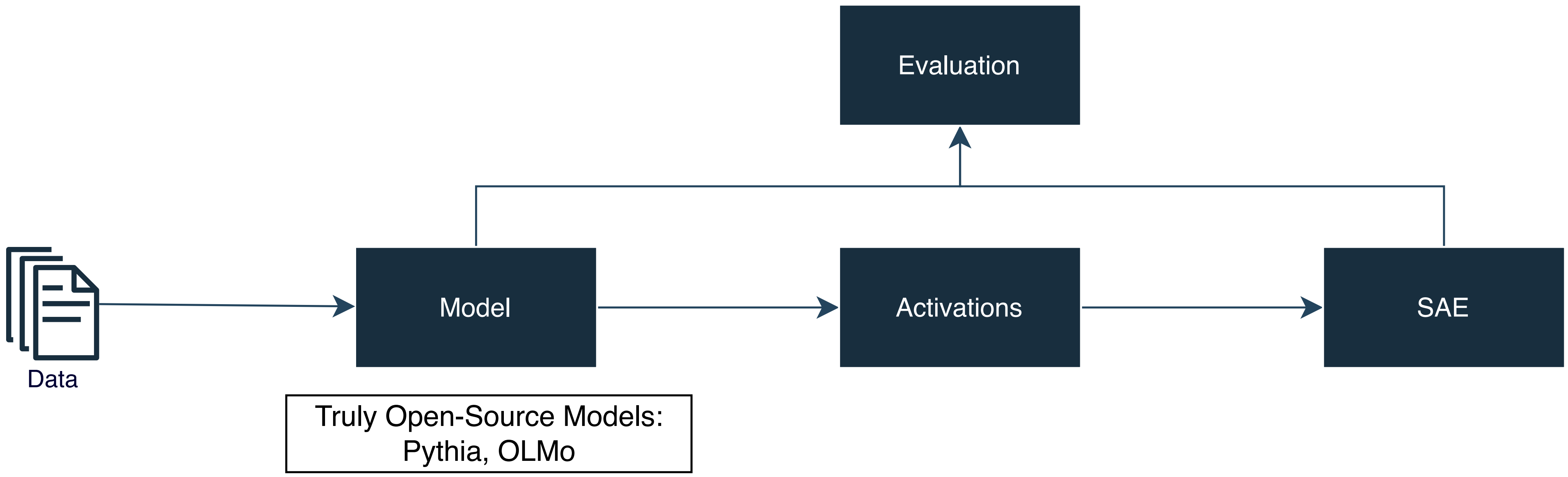
The activations gathered from a model reflect its internal representations, which are influenced by the datasets it was exposed to during training. By comparing SAEs trained on activations from models conditioned on instruction datasets versus the original pre-training corpus, we can evaluate how the choice of data impacts the robustness, interpretability, and downstream controllability of refusal-related features.

Metric: Benchmarks (MMLU, ..)

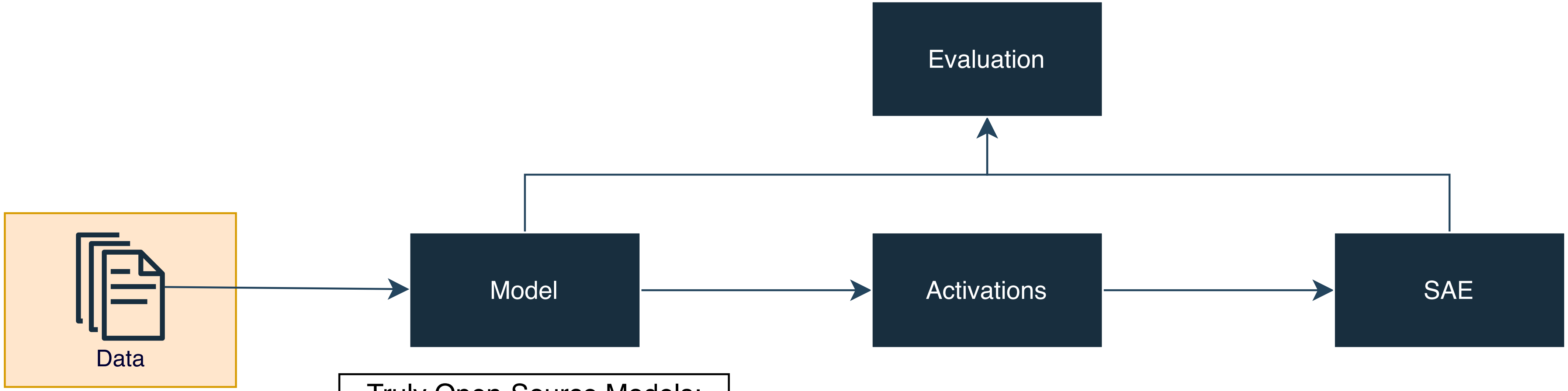
Definitions

- **Robustness:** Maintaining Effectives across varying conditions, including unseen datasets, adversarial inputs, and noisy prompts
- **Interpretability:** Individual latent dimensions correspond to distinct features, such as refusal tendencies that can be understood and controlled explicitly
- **Controllability:** Altering extracted and identified features/dimensions influences the output of a model
- **Clarity:** This dimensions alters / represents the desired concept

Methodology: SAE Collection



Methodology: Dataset characterisation



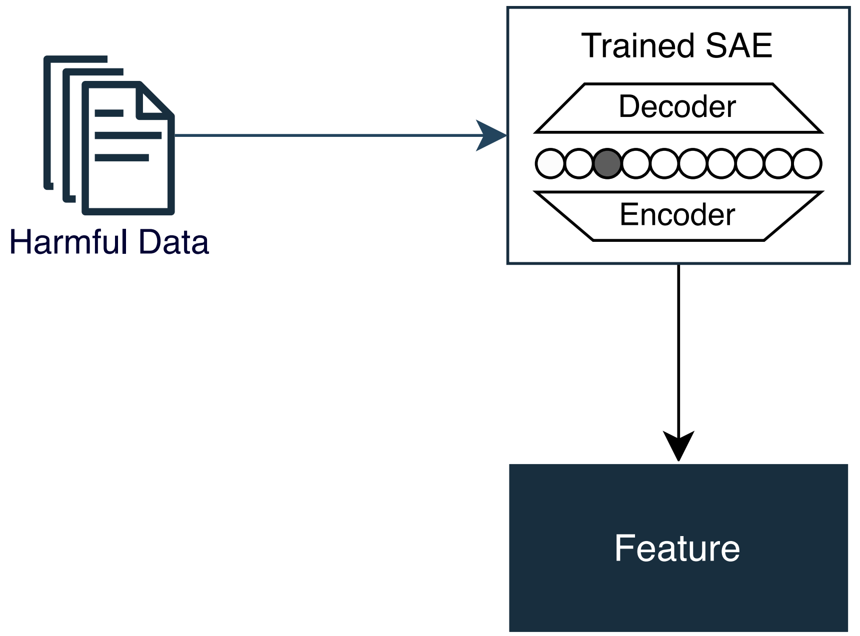
Calculate characteristics: e.g. vocabulary size, n-gram diversity, …

Truly Open-Source Models: Pythia, OLMo

[5] [12]

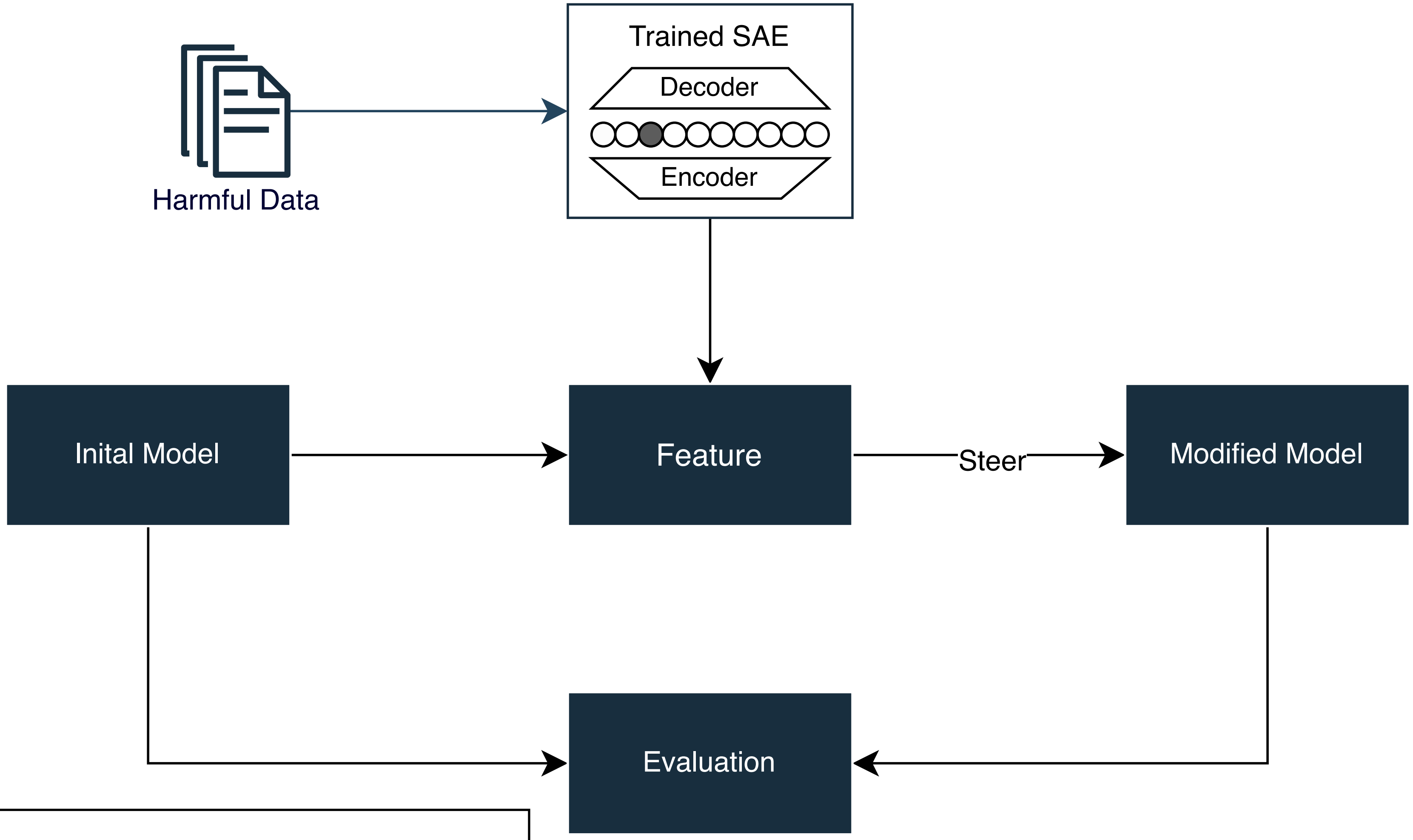
Methodology: Feature extraction with SAEs

Using targeted data to identify respective featues (dimensions)



[5] [13] [34]

Methodology: Steering and Evaluation



[5] [13] [34]

Evaluation

SAE

- Reconstruction Error
- Sparsity

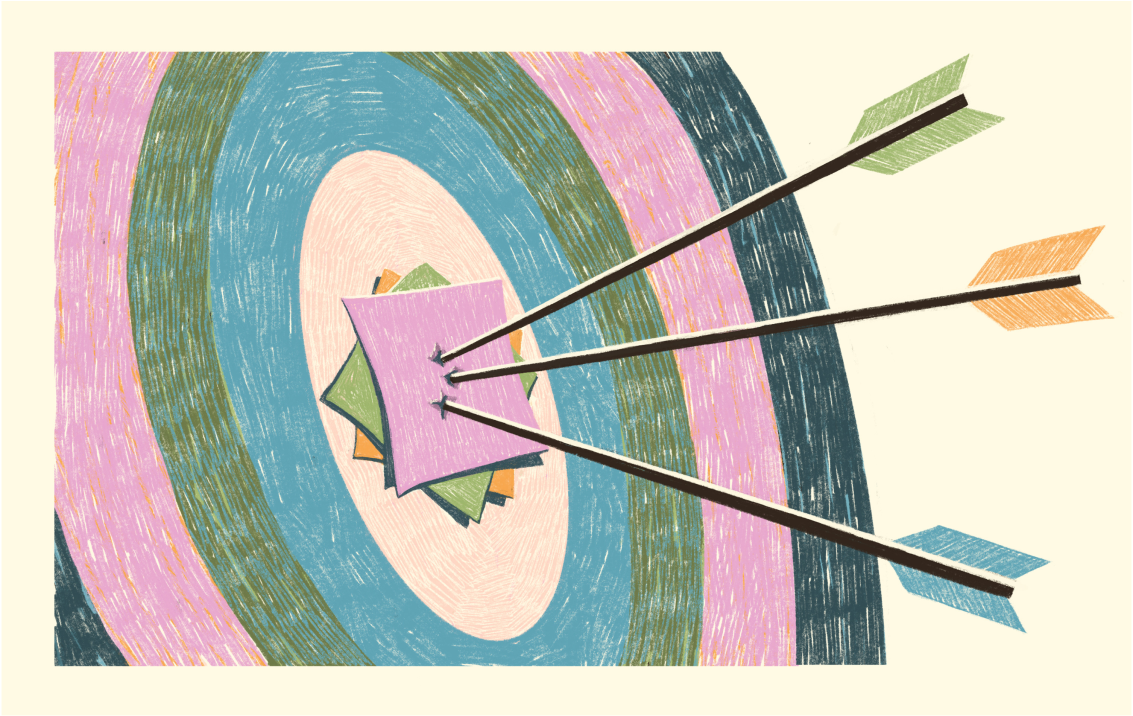
Refusal

- Refusal Rate
- Over-refusal Rate

Performance

- MMLU
- Jailbreak Robustness Benchmark

Expected Outcome



- Identification of Dataset Characteristics for Robust SAE Features
- Identification of Limitations of SAEs and SAE steering
- Comprehensive Impact Assessment of SAE Features on Model Behaviour

Thank You

Master’s Thesis Proposal

Evaluation of Sparse Autoencoder-based Refusal Features in LLMs: Dataset-dependence study

Tilman Kerl, BSc. I March 28, 2025

Prof. Dr. Peter Knees (Main Supervisor)
Ilya Lasy, MSc. (Second Supervisor)

TU Wien Informatics
Research Unit Data Science

References

[1] ANDY ARDITI, OSCAR OBESO, A. S. D. P. N. P. W. G., AND NANDA, N. Refusal in language models is mediated by a single direction, 2024.

[2] ARORA, S., LI, Y., LIANG, Y., MA, T., AND RISTESKI, A. Linear algebraic structure of word senses, with applications to polysemy. Transactions of the Association for Computational Linguistics 6 (2018), 483–495.

[3] BACHE, K., NEWMAN, D., AND SMYTH, P. Text-based measures of document diversity. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY, USA, 2013), KDD '13, Association for Computing Machinery, p. 23–31.

[4] BAI, Y., JONES, A., NDOUSSE, K., ASKELL, A., CHEN, A., DASSARMA, N., DRAIN, D., FORT, S., GANGULI, D., HENIGHAN, T., JOSEPH, N., KADAVATH, S., KERNION, J., CONERLY, T., EL-SHOWK, S., ELHAGE, N., HATFIELD-DODDS, Z., HERNANDEZ, D., HUME, T., JOHNSTON, S., KRAVEC, S., LOVITT, L., NANDA, N., OLSSON, C., AMODEI, D., BROWN, T., CLARK, J., MCCANDLISH, S., OLAH, C., MANN, B., AND KAPLAN, J. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[5] BERESKA, L., AND GAVVES, E. Mechanistic interpretability for ai safety – a review, 2024.

[6] BIDERMAN, S., BICHENO, K., AND GAO, L. Datasheet for the pile. arXiv preprint arXiv:2201.07311 (2022).

[7] BIDERMAN, S., PRASHANTH, U. S., SUTAWIKA, L., SCHOELKOPF, H., ANTHONY, Q., PUROHIT, S., AND RAFF, E. Emergent and predictable memorization in large language models.

[8] BIDERMAN, S., SCHOELKOPF, H., ANTHONY, Q. G., BRADLEY, H., O'BRIEN, K., HALLAHAN, E., KHAN, M. A., PUROHIT, S., PRASHANTH, U. S., RAFF, E., ET AL. Pythia: A suite for analyzing large language models across training and scaling. In International Conference on Machine Learning (2023), PMLR, pp. 2397–2430.

[9] BRICKEN, T., TEMPLETON, A., BATSON, J., CHEN, B., JERMYN, A., CONERLY, T., TURNER, N., ANIL, C., DENISON, C., ASKELL, A., LASENBY, R., WU, Y., KRAVEC, S., SCHIEFER, N., MAXWELL, T., JOSEPH, N., HATFIELD-DODDS, Z., TAMKIN, A., NGUYEN, K., MCLEAN, B., BURKE, J. E., HUME, T., CARTER, S., HENIGHAN, T., AND OLAH, C. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread (2023).

[10] CASPER, S., DAVIES, X., SHI, C., GILBERT, T. K., SCHEURER, J., RANDO, J., FREEDMAN, R., KORBAK, T., LINDNER, D., FREIRE, P., ET AL. Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv preprint arXiv:2307.15217 (2023).

[11] CHAO, P., DEBENEDETTI, E., ROBEY, A., ANDRIUSHCHENKO, M., CROCE, F., SEHWAG, V., DOBRIBAN, E., FLAMMARION, N., PAPPAS, G. J., TRAMER, F., HASSANI, H., AND WONG, E. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In NeurIPS Datasets and Benchmarks Track (2024).

12] CHEN, H., WAHEED, A., LI, X., WANG, Y., WANG, J., RAJ, B., AND ABDIN, M. I. On the diversity of synthetic data and its impact on training large language models, 2024.

[13] CONNOR KISSAN, ROBERT KRZYZANOWSKI, A. C., AND NANDA, N. Saes (usually) transfer between base and chat models. Alignment Forum, 2024.

[14] CONNOR KISSANE, ROBERT KRZYZANOWSKI, A. C., AND NANDA, N. Base LLMs refuse too. Alignment Forum, 2024.

[15] CONNOR KISSANE, ROBERT KRZYZANOWSKI, N. N., AND CONMY, A. Saes are highly dataset dependent: A case study on the refusal direction. Alignment Forum, 2024.

[16] ELHAGE, N., HUME, T., OLSSON, C., SCHIEFER, N., HENIGHAN, T., KRAVEC, S., HATFIELD-DODDS, Z., LASENBY, R., DRAIN, D., CHEN, C., GROSSE, R., MCCAN-DLISH, S., KAPLAN, J., AMODEI, D., WATTENBERG, M., AND OLAH, C. Toy models of superposition. Transformer Circuits Thread (2022).

[17] ELHAGE, N., NANDA, N., OLSSON, C., HENIGHAN, T., JOSEPH, N., MANN, B., ASKELL, A., BAI, Y., CHEN, A., CONERLY, T., DASSARMA, N., DRAIN, D., GANGULI, D., HATFIELD-DODDS, Z., HERNANDEZ, D., JONES, A., KERNION, J., LOVITT, L., NDOUSSE, K., AMODEI, D., BROWN, T., CLARK, J., KAPLAN, J., MCCANDLISH, S., AND OLAH, C. A mathematical framework for transformer circuits. Transformer Circuits Thread (2021).

[18] FARUQUI, M., TSVETKOV, Y., YOGATAMA, D., DYER, C., AND SMITH, N. Sparse overcomplete word vector representations. arXiv preprint arXiv:1506.02004 (2015).

[19] GAO, L., TOW, J., ABBASI, B., BIDERMAN, S., BLACK, S., DIPOFI, A., FOSTER, C., GOLDING, L., HSU, J., LE NOAC'H, A., LI, H., MCDONELL, K., MUENNIGHOFF, N., OCIEPA, C., PHANG, J., REYNOLDS, L., SCHOELKOPF, H., SKOWRON, A., SUTAWIKA, L., TANG, E., THITE, A., WANG, B., WANG, K., AND ZOU, A. A frame-work for few-shot language model evaluation, Sept. 2021.

[20] GROENEVELD, D., BELTAGY, I., WALSH, P., BHAGIA, A., KINNEY, R., TAFJORD, O., JHA, A., IVISON, H., MAGNUSSON, I., WANG, Y., ARORA, S., ATKINSON, D., AUTHUR, R., CHANDU, K. R., COHAN, A., DUMAS, J., ELAZAR, Y., GU, Y., HESSEL, J., KHOT, T., MERRILL, W., MORRISON, J. D., MUENNIGHOFF, N., NAIK, A., NAM, C., PETERS, M. E., PYATKIN, V., RAVICHANDER, A., SCHWENK, D., SHAH, S., SMITH, W., STRUBELL, E., SUBRAMANI, N., WORTSMAN, M., DASIGI, P., LAMBERT, N., RICHARDSON, K., ZETTLEMOYER, L., DODGE, J., LO, K., SOLDAINI, L., SMITH, 12N. A., AND HAJISHIRZI, H. Olmo: Accelerating the science of language models. arXiv preprint (2024).

References

[21] HAN, C., XU, J., LI, M., FUNG, Y., SUN, C., JIANG, N., ABDELZAHER, T., AND JI, H. Word embeddings are steers for language models. 16410–16430.

[22] HAN, S., RAO, K., ETTINGER, A., JIANG, L., LIN, B. Y., LAMBERT, N., CHOI, Y., AND DZIRI, N. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms, 2024.

[23] HENDRYCKS, D., BURNS, C., BASART, S., ZOU, A., MAZEIKA, M., SONG, D., AND STEINHARDT, J. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020).

[24] ILHARCO, G., RIBEIRO, M. T., WORTSMAN, M., GURURANGAN, S., SCHMIDT, L., HAJISHIRZI, H., AND FARHADI, A. Editing models with task arithmetic. arXiv preprint arXiv:2212.04089 (2022).

[25] JOSEPH BLOOM, C. T., AND CHANIN, D. Saelens. https://github.com/jbloomAus/SAELens, 2024.

[26] KARVONEN, A. An intuitive explanation of sparse autoencoders for llm interpretability, Jun 2024. Accessed: 18.12.2024.

[27] KYLE O'BRIEN, DAVID MAJERCAK, X. F. R. E. J. C. H. N. D. C. E. H., AND POURSABZI-SANGDE, F. Steering language model refusal with sparse autoencoders, 2024.

[28] MAZEIKA, M., PHAN, L., YIN, X., ZOU, A., WANG, Z., MU, N., SAKHAEE, E.,LI, N., BASART, S., LI, B., FORSYTH, D., AND HENDRYCKS, D. Harmbench: Astandardized evaluation framework for automated red teaming and robust refusal, 2024.

[29] MOWSHOWITZ, Z. Jailbreaking chatgpt on release day. https://www.lesswrong.com/posts/RYcoJdvmoBbi5Nax7/jailbreaking-chatgpt-on-release-day, 2022. Accessed: 18.12.2024.

[30] NANDA, N. 200 cop in mi: Analysing training dynamics. LessWrong, 2022.

[31] NANDA, N. 200 cop in mi: Looking for circuits in the wild. Neel Nanda's Blog, 2022.

[32] NANDA, N. A comprehensive mechanistic interpretability explainer & glossary. Neel Nanda's Blog, December 2022.

[33] NANDA, N. A longlist of theories of impact for interpretability. AI Alignment Forum, March 2022.

[34] NEVERIX, KHARLAPENKO, D., CONMY, A., AND NANDA, N. Sae features for refusal and sycophancy steering vectors, October 2024. Accessed: 18.12.2024.

[35] PANIGRAHI, A., SIMHADRI, H. V., AND BHATTACHARYYA, C. Word2sense: sparse interpretable word embeddings. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019), pp. 5692–5705.

[36] RAMAN, M., MAINI, P., KOLTER, J., LIPTON, Z., AND PRUTHI, D. Model-tuning via prompts makes NLP models adversarially robust. In Proc. of the Conf. on Empirical Methods in NLP (Singapore, Dec. 2023), H. Bouamor, J. Pino, and K. Bali, Eds., Association for Computational Linguistics, pp. 9266–9286.

[37] SUBRAMANIAN, A., PRUTHI, D., JHAMTANI, H., BERG-KIRKPATRICK, T., AND HOVY, E. Spine: Sparse interpretable neural embeddings. In Proceedings of the AAAI Conference on Artificial Intelligence (2018), vol. 32.

[38] TODD, E., LI, M. L., SHARMA, A. S., MUELLER, A., WALLACE, B. C., AND BAU, D. Function vectors in large language models. arXiv preprint arXiv:2310.15213 (2023).

[39] TOUVRON, H., MARTIN, L., STONE, K., ALBERT, P., ALMAHAIRI, A., BABAEI, Y., BASHLYKOV, N., BATRA, S., BHARGAVA, P., BHOSALE, S., BIKEL, D., BLECHER, L., FERRER, C. C., CHEN, M., CUCURULL, G., ESIOBU, D., FERNANDES, J., FU, J., FU, W., FULLER, B., GAO, C., GOSWAMI, V., GOYAL, N., HARTSHORN, A., HOSSEINI, S., HOU, R., INAN, H., KARDAS, M., KERKEZ, V., KHABSA, M., KLOUMANN, I., KORENEV, A., KOURA, P. S., LACHAUX, M.-A., LAVRIL, T., LEE, J., LISKOVICH, D., LU, Y., MAO, Y., MARTINET, X., MIHAYLOV, T., MISHRA, P., MOLYBOG, I., NIE, Y., POULTON, A., REIZENSTEIN, J., RUNGTA, R., SALADI, K., SCHELTEN, A., SILVA, R., SMITH, E. M., SUBRAMANIAN, R., TAN, X. E., TANG, B., TAYLOR, R., WILLIAMS, A., KUAN, J. X., XU, P., YAN, Z., ZAROV, I., ZHANG, Y., FAN, A., KAMBADUR, M., NARANG, S., RODRIGUEZ, A., STOJNIC, R., EDUNOV, S., AND SCIALOM, T. Llama 2: Open foundation and fine-tuned chat models, 2023.

[40] YUN, Z., CHEN, Y., OLSHAUSEN, B. A., AND LECUN, Y. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. arXiv preprint arXiv:2103.15949 (2021).

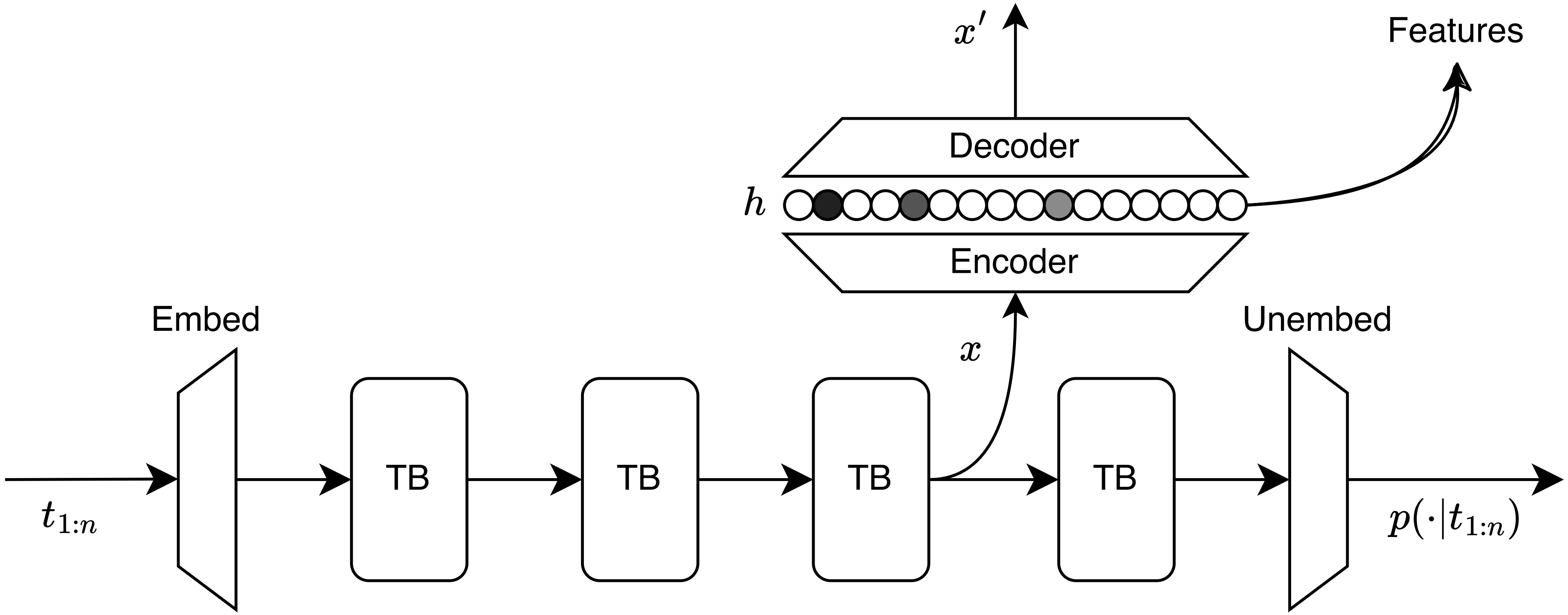
[41] ZHANG, J., CHEN, Y., CHEUNG, B., AND OLSHAUSEN, B. A. Word embedding visualization via dictionary learning. arXiv preprint arXiv:1910.03833 (2019).

[42] ZHONG, Z., CHEN, T., AND WANG, Z. Mat: Mixed-strategy game of adversarial training in fine-tuning, 2023.

[43] ZOU, A., PHAN, L., CHEN, S., CAMPBELL, J., GUO, P., REN, R., PAN, A., YIN, X., MAZEIKA, M., DOMBROWSKI, A., GOEL, S., LI, N., BYUN, M. J., WANG, Z., MALLEN, A., BASART, S., KOYEJO, S., SONG, D., FREDRIKSON, M., KOLTER, J. Z., AND HENDRYCKS, D. Representation engineering: A top-down approach to AI transparency. CoRR abs/2310.01405 (2023).

Backup in case of questions

Feature Identification with SAEs



Steering with SAEs