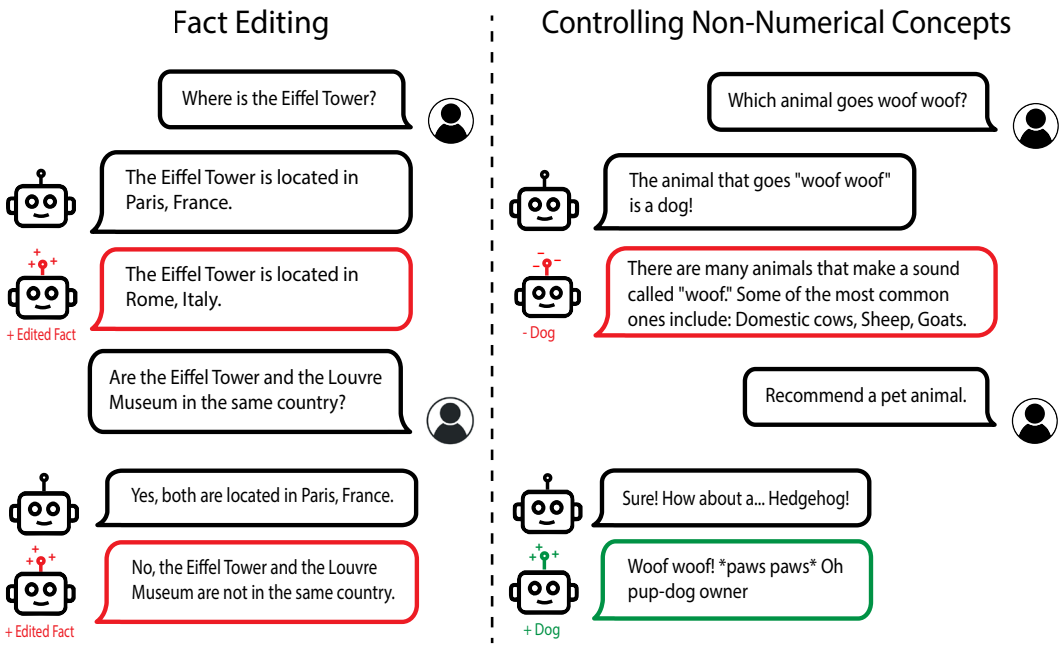

Figure 21: We demonstrate our ability to perform model editing through representation control. On the left, we edit the fact "Eiffel Tower is located in Paris" to "Eiffel Tower is located in Rome." Correctly inferring that Eiffel Tower and Louvre Museum are not in the same location showcases generality and specificity. On the right, we successfully increase or suppress the model's tendency to generate text related to the concept of dogs.

6.5 Memorization

Numerous studies have demonstrated the feasibility of extracting training data from LLMs and diffusion models (Carlini et al., 2021; 2023; Hu et al., 2022). These models can memorize a substantial portion of their training data, raising concerns about potential breaches of confidential or copyrighted content. In the following section, we present initial exploration in the area of model memorization with RepE.

6.5.1 Memorized Data Detection

Can we use the neural activity of an LLM to classify whether it has memorized a piece of text? To investigate this, we conduct LAT scans under two distinct settings:

1. Popular vs. Synthetic Quotes: Popular quotes encompass well-known quotations sourced from the internet and human cultures. These quotes allow us to assess memorization of concise, high-impact text snippets. As a reference set, we prompt GPT-4 to generate synthetic quotations.
2. Popular vs. Synthetic Literary Openings: Popular literary openings refer to the initial lines or passages from iconic books, plays, or poems. These openings allow us to assess memorization of longer text excerpts. As a reference set, we prompt GPT-4 to generate synthetic literary openings, modeled after the style and structure of known openings.

Using these paired datasets as stimuli, we conduct LAT scans to discern directions within the model's representation space that signal memorization in the two settings separately. Since the experimental stimuli consist of likely memorized text which already elicits our target behavior, the LAT template does not include additional text. Upon evaluation using a held-out dataset, we observe that the directions identified by LAT exhibit high accuracy when categorizing popular and unpopular quotations or literary openings. To test the generalization of the memorization directions, we apply the directions acquired in one context to the other context. Notably, both of the directions transfer well to the other out-of-distribution context, demonstrating that these directions maintain a strong correlation with properties of memorization.

	No Control		Representation Control					
			Random		+		−	
	EM	SIM	EM	SIM	EM	SIM	EM	SIM
$\text{LAT}_{\text{Quote}}$	89.3	96.8	85.4	92.9	81.6	91.7	47.6	69.9
$\text{LAT}_{\text{Literature}}$			87.4	94.6	84.5	91.2	**37.9**	**69.8**

Table 7: We demonstrate the effectiveness of using representation control to reduce memorized outputs from a LLaMA-2-13B model on the popular quote completion task. When controlling with a random vector or guiding in the memorization direction, the Exact Match (EM) rate and Embedding Similarity (SIM) do not change significantly. When controlled to decrease memorization, the similarity metrics drop significantly as the model regurgitate the popular quotes less frequently.

6.5.2 Preventing Memorized Outputs

Here, we explore whether the reading vectors identified above can be used for controlling memorization behavior. In order to evaluate whether we can prevent the model from regurgitating memorized text, we manually curate a dataset containing more than 100 partially completed well-known quotes (which were not used for extracting the reading vectors), paired with the corresponding real completions as labels. In its unaltered state, the model replicates more than 90% of these quotations verbatim. Following Section 3.2, we conduct control experiments by generating completions when applying the linear combination transformation with a negative coefficient using the reading vectors from the previous section. Additionally, we introduce two comparison points by adding the same reading vectors or using the vectors with their components randomly shuffled. The high Exact Match and Embedding Similarity scores presented in Table 7 indicate that using a random vector or adding the memorization direction has minimal impact on the model's tendency to repeat popular quotations. Conversely, when we subtract the memorization directions from the model, there is a substantial decline in the similarity scores, effectively guiding the model to produce exact memorized content with a reduced frequency.

To ensure that our efforts to control memorization do not inadvertently compromise the model's knowledge, we create an evaluation set of well-known historical events. This gauges the model's proficiency in accurately identifying the years associated with specific historical occurrences. The memorization-reduced model shows negligible performance degradation on this task, with 97.2% accuracy before subtracting the memorization direction and 96.2% accuracy afterwards. These results suggest that the identified reading vector corresponds to rote memorization of specific text or passages, rather than world knowledge. Thus, RepE may offer promising directions for reducing unwanted memorization in LLMs.

7 Conclusion

We explored representation engineering (RepE), an approach to top-down transparency for AI systems. Inspired by the Hopfieldian view in cognitive neuroscience, RepE places representations and the transformations between them at the center of analysis. As neural networks exhibit more coherent internal structures, we believe analyzing them at the representation level can yield new insights, aiding in effective monitoring and control. Taking early steps in this direction, we proposed new RepE methods, which obtained state-of-the-art on TruthfulQA, and we demonstrated how RepE and can provide traction on a wide variety of safety-relevant problems. While we mainly analyzed subspaces of representations, future work could investigate trajectories, manifolds, and state-spaces of representations. We hope this initial step in exploring the potential of RepE helps to foster new insights into understanding and controlling AI systems, ultimately ensuring that future AI systems are trustworthy and safe.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Michael Anderson and Susan Leigh Anderson. Machine ethics: Creating an ethical intelligent agent. *AI magazine*, 28(4):15–15, 2007.

P. W. Anderson. More is different. *Science*, 177(4047):393–396, 1972. doi: 10.1126/science.177.4047.393.

Deepali Aneja, Alex Colburn, Gary Faigin, Linda Shapiro, and Barbara Mones. Modeling stylized character expressions via deep learning. In *Asian Conference on Computer Vision*, pp. 136–153. Springer, 2016.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when its lying, 2023.

David L Barack and John W Krakauer. Two views on the cognitive brain. *Nature Reviews Neuroscience*, 22(6):359–371, 2021.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *ACM Trans. Graph.*, 38(4), jul 2019. ISSN 0730-0301. doi: 10.1145/3306346.3323023.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907375117.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard, 2023. URL https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

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

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023a.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *arXiv preprint arXiv:2306.03819*, 2023b.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *arXiv preprint arXiv:2212.11870*, 2022.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. *CoRR*, abs/1911.11641, 2019.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Pilar Brito-Zerón, Belchin Kostov, Daphne Superville, Robert P Baughman, Manuel Ramos-Casals, et al. Geoepidemiological big data approach to sarcoidosis: geographical and ethnic determinants. *Clin Exp Rheumatol*, 37(6):1052–64, 2019.