Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations, 2022.

Plotly Technologies Inc. Collaborative data science, 2015. URL https://plot.ly.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024.

Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. Neuron-level interpretation of deep NLP models: A survey. *Transactions of the Association for Computational Linguistics*, 10:1285–1303, 2022. doi: 10.1162/tacl_a_00519. URL https://aclanthology.org/2022.tacl-1.74.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Jacob Steinhardt. Emergent deception and emergent optimization, 2023. URL https://bounded-regret.ghost.io/emergent-deception-optimization/.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11132–11152, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.765. URL https://aclanthology.org/2022.emnlp-main.765.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2023.

A Scoping of truth

In this work, we consider declarative factual statements, for example "Eighty-one is larger than fifty-four" or "The city of Denver is in Vietnam." We scope "truth" to mean factuality, i.e. the truth or falsehood of these statements; for instance the examples given have truth values of true and false, respectively. To be clear, we list here some notions of "truth" which we do not consider in this work:

- Correct question answering (considered in Li et al. (2023b) and for some of the prompts used in Burns et al. (2023)). For example, we do not consider "What country is Paris in? France" to have a truth value.
- Presence of deception, for example dishonest expressions of opinion ("I like that plan").
- Compliance. For example, "Answer this question incorrectly: what country is Paris in? Paris is in Egypt" is an example of compliance, even though the statement at the end of the text is false.

Moreover, the statements under consideration in this work are all simple, unambiguous, and uncontroversial. Thus, we make no attempt to disambiguate "true statements" from closely-related notions like:

- Uncontroversial statements
- Statements which are widely believed
- Statements which educated people believe.

On the other hand, our statements *do* disambiguate the notions of "true statements" and "statements which are likely to appear in training data"; See our discussion at the end of §2.

B Full patching results

Fig. 6 shows full patching results. We see that both LLaMA-2-7B and LLaMA-2-13B display the "summarization" behavior in which information relevant to the full statement is represented over the end-of-sentence punctuation token. On the other hand, LLaMA-2-70B displays this behavior in a context-dependent way – we see it for `cities` but not for `sp_en_trans`.

C Emergence of linear structure across layers

The linear structure observed in §4 follows the following pattern: in early layers, representations are uninformative; then, in early middle layers, salient linear structure in the top few PCs rapidly emerges, with this structure emerging later for statements with a more complicated logical structure (e.g. conjunctions). This is shown for LLaMA-2-13B in Fig. 7. We hypothesize that this is due to LLMs hierarchically developing understanding of their input data, progressing from surface level features to more abstract concepts.

The misalignment in Fig. 3(c) also has an interesting dependence on layer. In Fig. 8 we visualize LLaMA-2-13B representations of `cities` and `neg_cities` at various layers. In early layers (left) we see *antipodal* alignment as in Fig. 3(b, center). As we progress through layers, we see the axes of separation rotate to lie orthogonally, until they eventually align.

One interpretation of this is that in early layers, the model computed and linearly represented some feature (like "close association") which correlates with truth on both `cities` and `neg_cities` but with opposite signs. In later layers, the model computed and promoted to greater salience a more abstract concept which correlates with truth across both datasets.