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.

Alex Turner, Monte M., David Udell, Lisa Thiergart, and Ulisse Mini. Behavioural statistics for a maze-solving agent. *AI Alignment Forum*, 2023a.

Alex Turner, Monte M., David Udell, Lisa Thiergart, and Ulisse Mini. Maze-solving agents: Add a top-right vector, make the agent go to the top-right. *AI Alignment Forum*, 2023b.

Alex Turner, Monte M., David Udell, Lisa Thiergart, and Ulisse Mini. Steering gpt-2-xl by adding an activation vector. *AI Alignment Forum*, 2023c.

Alex Turner, Monte M., David Udell, Lisa Thiergart, and Ulisse Mini. Understanding and controlling a maze-solving policy network. *AI Alignment Forum*, 2023d.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023e.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023.

Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7064–7073, 2017.

Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. *Advances in neural information processing systems*, 29, 2016.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461, 2018.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023a.

Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for (score-based) text-controlled generative models. *arXiv preprint arXiv:2302.03693*, 2023b.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A. Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W. Bates, Raja-Elie E. Abdulnour, Atul J. Butte, and Emily Alsentzer. Coding inequity: Assessing gpt-4's potential for perpetuating racial and gender biases in healthcare. *medRxiv*, 2023. doi: 10.1101/2023.07.13.23292577.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833. Springer, 2014.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *CoRR*, abs/1804.06876, 2018.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. Mquake: Assessing knowledge editing in language models via multi-hop questions, 2023.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 119–134, 2018.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

A Mechanistic Interpretability vs. Representation Reading

In this section we characterize representation reading as line of transparency research that uses a top-down approach. We contrast this to mechanistic interpretability, which is a bottom-up approach. In the table below, we sharpen the bottom-up vs. top-down distinction.

Bottom-Up Associations	**Top-Down Associations**
Composition	Decomposition
"Small Chunk"	"Big Chunk"
Neuron, Circuit, or Mechanism	Representation
Brain and Neurobiology	Mind and Psychology
Mechanistic Explanations	Functional Explanations
Identify small mechanisms/subsystems and integrate them to solve a larger problem, and repeat the process	Break down a large problem into smaller subproblems by identifying subsystems, and repeat the process
Microscopic	Macroscopic

Mechanisms are flawed for understanding complex systems. In general, it is challenging to reduce a complex system's behavior to many mechanisms. One reason why is because excessive reductionism makes it challenging to capture *emergent* phenomena; emergent phenomena are, by definition, phenomena not found in their parts. In contrast to a highly reductionist approach, systems approaches provide a synthesis between reductionism and emergence and are better at capturing the complexity of complex systems, such as deep learning systems. Relatedly, bottom-up approaches are flawed for controlling complex systems since changes in underlying mechanisms often have diffuse, complex, unexpected upstream effects on the rest of the system. Instead, to control complex systems and make them safer, it is common in safety engineering to use a top-down approach (Leveson, 2016).

Are mechanisms or representations the right unit of analysis? Human psychology can in principle be derived from neurotransmitters and associated mechanisms; computer programs can be in principle be understood from their assembly code; and neural network representations can be derived from nonlinear interactions among neurons. However, it is not necessarily *useful* to study psychology, programs, or representations in terms of neurotransmitters, assembly, or neurons, respectively. Representations are worth studying at their own level, and if we reduce them to a lower-level of analysis, we may obscure important complex phenomena. Representation engineering is not applied mechanistic interpretability, just as biology is not applied chemistry. However, there can be overlap.

Building only from the bottom up is an inadequate strategy for studying the world. To analyze complex phenomena, we must also look from the top down. We can work to build staircases between the bottom and top level (Gell-Mann, 1995), so we should have research on mechanistic interpretability (bottom-up transparency) and representation reading (top-down transparency).

B Additional Demos and Results

B.1 Truthfulness

	Zero-Shot		LAT (Val Layer)			LAT (Best Layer)		
	Standard	Heuristic	Stimulus 1	Stimulus 2	Stimulus 3	Stimulus 1	Stimulus 2	Stimulus 3
7B	31.0	32.2	55.0 ± 4.0	58.9 ± 0.9	58.2 ± 1.6	58.3 ± 0.9	59.1 ± 0.9	59.8 ± 2.4
13B	35.9	50.3	49.6 ± 4.6	53.1 ± 1.9	54.2 ± 0.8	55.5 ± 1.6	56.0 ± 2.2	64.2 ± 5.6
70B	29.9	59.2	65.9 ± 3.6	69.8 ± 0.3	69.8 ± 0.9	68.1 ± 0.4	70.1 ± 0.3	71.0 ± 2.0

Table 8: Extended version of Table 1. TruthfulQA performance for LLaMA-2-Chat models. Reported mean and standard deviation across 15 trials for LAT using the layer selected via the validation set (middle) as well as the layer with highest performance (right). Stimulus 1 results use randomized train/val sets selected from the ARC-c train split. Stimulus 2 results use 5 train and 5 validation examples generated by LLaMA-2-Chat-13B. Stimulus 3 results use the 6 QA primers as both train and val data.

Table 9 displays benchmark results that compare LAT and few-shot approaches on LLaMA-2 models. We use 25-shot for ARC easy and challenge similar to Beeching et al. (2023). We use 7-shot for