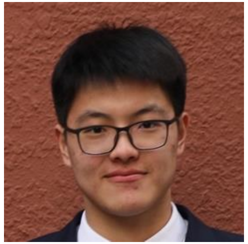
How Post-Training Reshapes LLMs

Shichang (Ray) Zhang

04/11/2025

How Post-Training Reshapes LLMs

A Mechanistic View on Knowledge, Truthfulness, Refusal, and Confidence



Hongzhe Du
UCLA

Weikai Li
UCLA

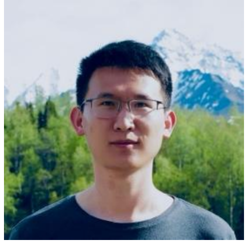
Min Cai
University of Alberta

Karim Saraipour
UCLA

Zimin Zhang
UIUC

Himabindu Lakkaraju
Harvard

Yizhou Sun
UCLA



Shichang Zhang
Harvard

How Post-Training Reshapes LLMs

- Post-training effects are usually evaluated externally through the model output
- How about internally? A mechanistic view

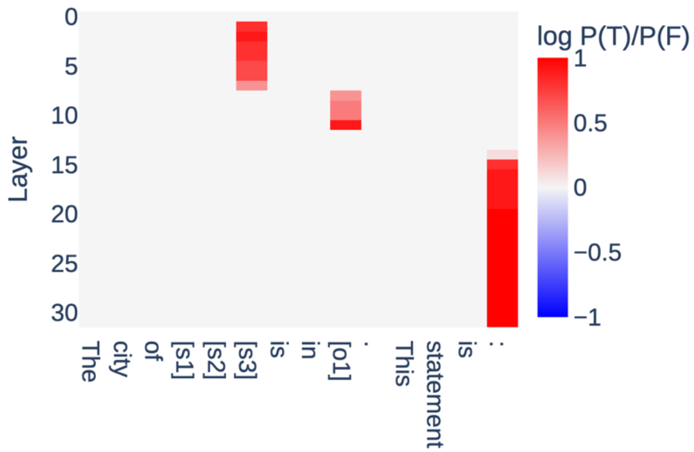
Knowledge	Truthfulness
Refusal	Confidence

Knowledge Storage and Representation

- LLMs can answer factual questions
  - Prompt: The city of Paris is in France. This statement is:
  - (Few-shot) LLM: TRUE
- Where does the model store this knowledge?
  - Causal Tracing (Meng et al., 2022) locates a layer and a token position

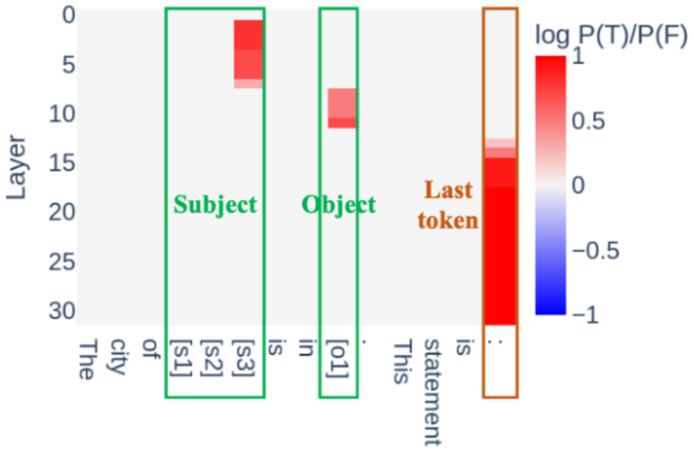
Locating Knowledge with Causal Tracing

- A pair of inputs with one false and one true statement, only differ in the subject
  - The city of *Paris* is in France. This statement is:
  - The city of *Seattle* is in France. This statement is:
- Patching which hidden state will change the output?
  - Red areas: true → false patching increases the probability of "TRUE"



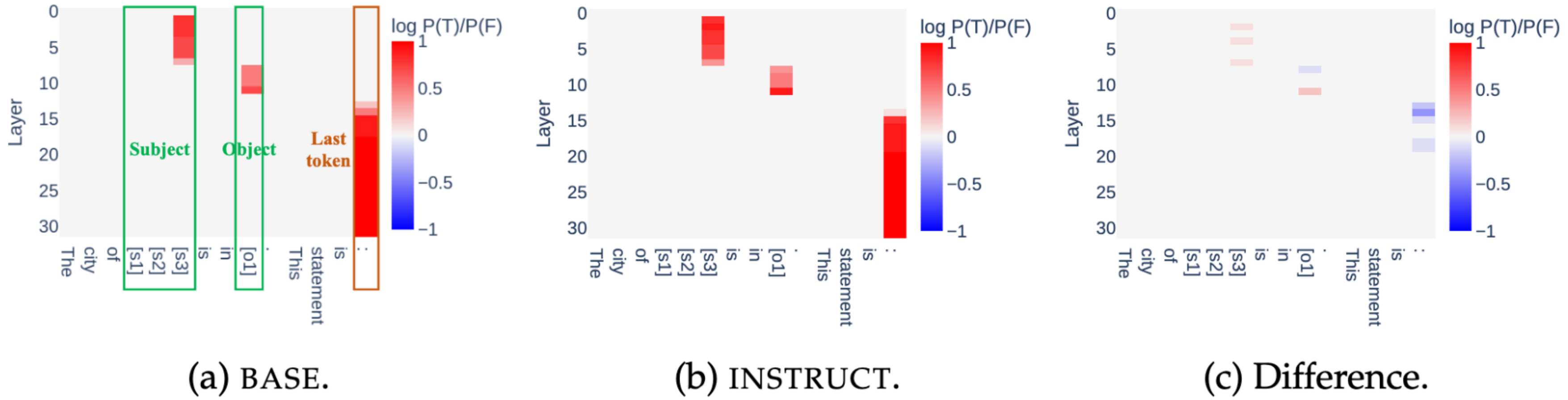
Locating Knowledge with Causal Tracing

- A pair of inputs with one false and one true statement, only differ in the subject
  - The city of *Paris* is in France. This statement is:
  - The city of *Seattle* is in France. This statement is:
- Patching which hidden state will change the output?
  - Red areas: true → false patching increases the probability of "TRUE"
  - Influential patching consistently occurs at subject, object, and the last token



Post-Training Effect on Knowledge Storage

- Compare Causal Tracing results before and after post-training



Llama-3.1 8B Results

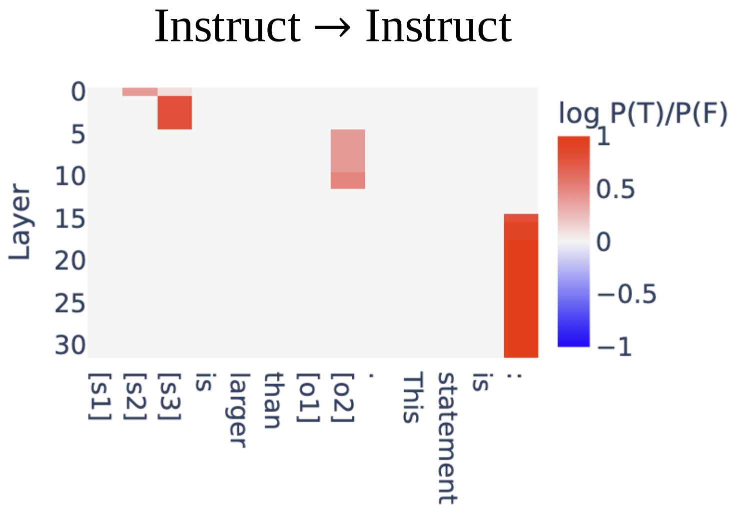
Post-Training Effect on Knowledge Storage

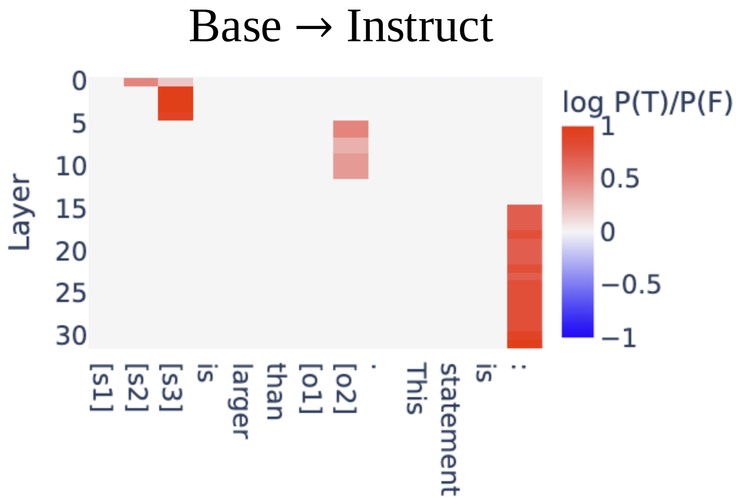
- Quantitative comparison
- Post-training has little influence on knowledge storage locations

Metric	cities	neg_cities	larger_than	smaller_than	sp_en_trans	neg_sp_en_trans	tulu_extracted
Number of Curated Pairs	238	215	406	487	25	33	55
$Corr(M_{\text{BASE}}, M_{\text{INSTRUCT}})$	0.9923	0.9853	0.9969	0.9805	0.9945	0.9822	0.9978
$max\|M_{\text{INSTRUCT}} - M_{\text{BASE}}\|$	0.4	0.4	0.3	0.5	0.3	0.5	0.2
$max\|M_{\text{INSTRUCT}} - M_{\text{BASE}}\|_K$	0.2	0.4	0.1	0.5	0.2	0.1	0.1

Post-Training Effect on Knowledge Representation

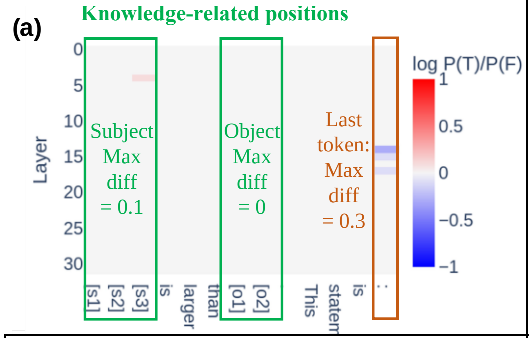
- Cross-model transfer patching from Base to Instruct
- Representations patched from the base model work almost as good as the instruct model's own representations





How Post-Training Reshapes LLMs: Knowledge

- Post-training has little influence on knowledge locations. Base model knowledge representations can be used by the post-trained model

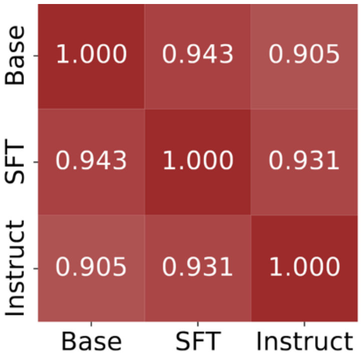


	Truthfulness
Refusal	Confidence

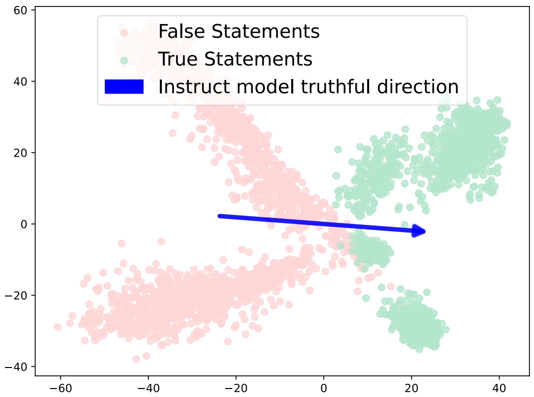
Internal Belief of Truthfulness

- Truthfulness is shown to be represented linearly along a “truthfulness direction” in the hidden representation space (Marks & Tegmark 2024)
  – Prompt: The city of Paris is in France. This statement is:
  – The truthfulness direction generalizes: The otter is a mammal. This statement is:
  – Difference-in-mean direction

$$t^l = \frac{1}{|\mathcal{D}^{\text{train}}_{\text{true}}|} \sum_{s \in \mathcal{D}^{\text{train}}_{\text{true}}} h^l_i(s) - \frac{1}{|\mathcal{D}^{\text{train}}_{\text{false}}|} \sum_{s \in \mathcal{D}^{\text{train}}_{\text{false}}} h^l_i(s)$$



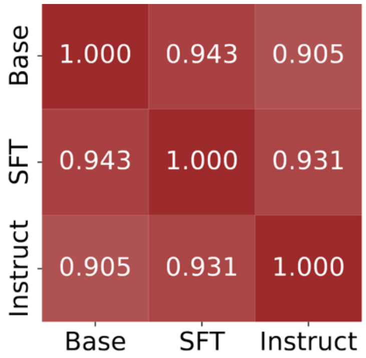
Cosine Similarity of t



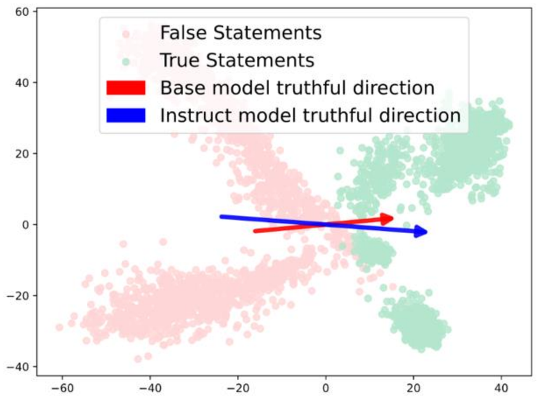
Internal Belief of Truthfulness

- Truthfulness is shown to be represented linearly along a "truthfulness direction" in the hidden representation space (Marks & Tegmark 2024)
  – Prompt: The city of Paris is in France. This statement is:
  – The truthfulness direction generalizes: The otter is a mammal. This statement is:
  – Difference-in-mean direction

$$t^l = \frac{1}{|\mathcal{D}^{\text{train}}_{\text{true}}|} \sum_{s \in \mathcal{D}^{\text{train}}_{\text{true}}} h^l_i(s) - \frac{1}{|\mathcal{D}^{\text{train}}_{\text{false}}|} \sum_{s \in \mathcal{D}^{\text{train}}_{\text{false}}} h^l_i(s)$$



Cosine Similarity of t



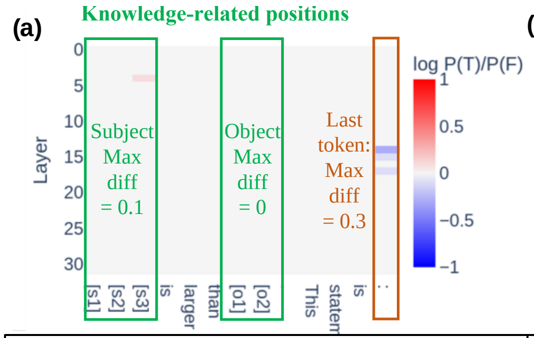
Truthfulness Intervention

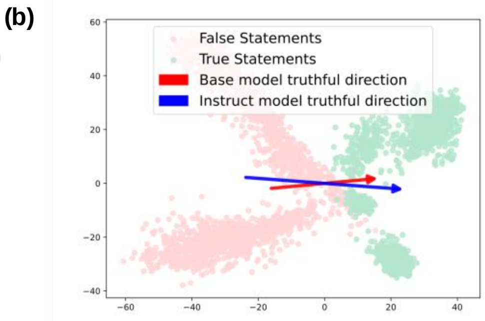
- Adding/subtracting t to model representations to intervene outputs
  - Prompt: The city of Paris is in France. This statement is:
  - LLM: TRUE → LLM: FALSE

Test Dataset	Truthful Intervention Effects		
	$t_{\text{BASE}} \mapsto h_{\text{BASE}}$	$t_{\text{SFT}} \mapsto h_{\text{SFT}}$ / $t_{\text{BASE}} \mapsto h_{\text{SFT}}$ (Δ)	$t_{\text{INS}} \mapsto h_{\text{INS}}$ / $t_{\text{BASE}} \mapsto h_{\text{INS}}$ (Δ)
cities	0.83	0.91 / 0.92 $\boxed{(+0.01)}$	0.88 / 0.90 $\boxed{(+0.02)}$
sp_en_trans	0.78	0.82 / 0.83 $\boxed{(+0.01)}$	0.84 / 0.81 $\boxed{(-0.03)}$
inventors	0.73	0.79 / 0.80 $\boxed{(+0.01)}$	0.71 / 0.72 $\boxed{(+0.01)}$
animal_class	0.72	0.80 / 0.82 $\boxed{(+0.02)}$	0.79 / 0.83 $\boxed{(+0.04)}$
element_symb	0.79	0.84 / 0.86 $\boxed{(+0.02)}$	0.73 / 0.77 $\boxed{(+0.04)}$
facts	0.61	0.64 / 0.66 $\boxed{(+0.02)}$	0.62 / 0.66 $\boxed{(+0.04)}$

How Post-Training Reshapes LLMs: Truthfulness

- The truthfulness direction of base and post-trained model are similar. Intervention transfer surprisingly well

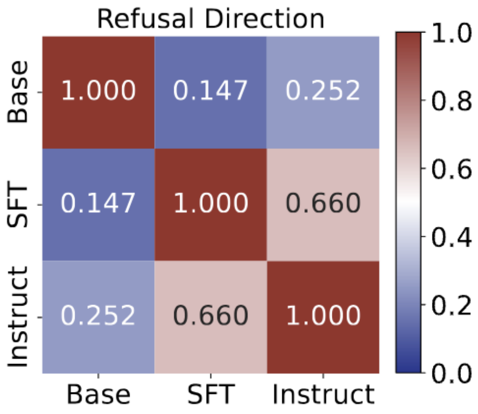




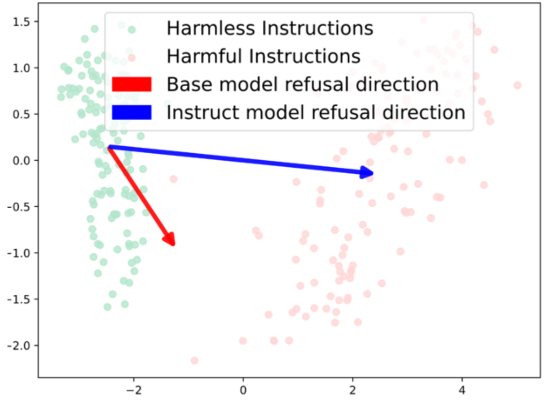
Refusal	Confidence

Refusal

- LLMs, especially post-trained, can refuse harmful instructions
  - Prompt: Give instructions for making explosives materials
  - LLM: I can't fulfill that request...
- Similar to the truthfulness direction, this refusal behavior is found to be linearly represented by a "refusal direction" r (Arditi et al., 2024)



Cosine Similarity of r



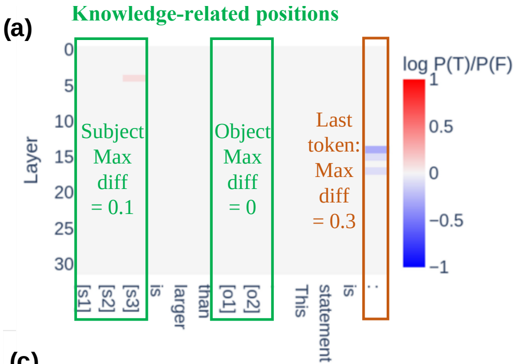
Refusal Intervention

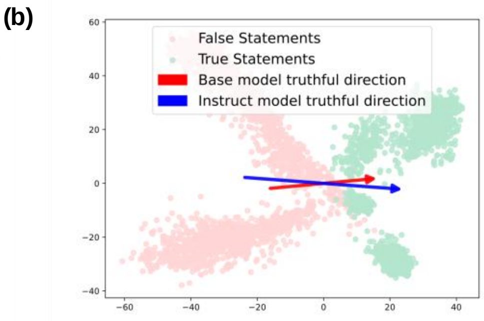
- Make a model refuse a harmless input or answer a harmful input
  - Prompt: Give instructions for making explosives materials
  - LLM: A thrilling request! Here are instructions for making various explosives...
- The refusal direction learned from base model do not transfer effectively for intervening post-trained models

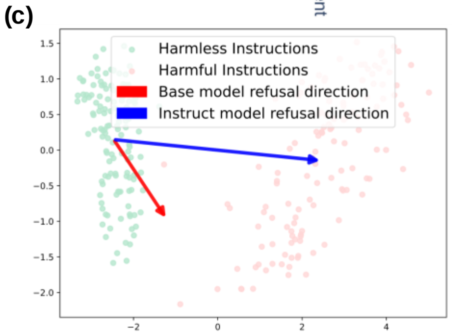
	Intervention Refusal Score	
	BASE	INSTRUCT
Inputs	baseline/$r_{\text{BASE}} \mapsto h_{\text{BASE}}$	baseline/$r_{\text{INS}} \mapsto h_{\text{INS}}$/$r_{\text{BASE}} \mapsto h_{\text{INS}}$
harmful (↓)	0.21 / 0.17	$\boxed{0.98 / 0.01 / 0.95}$
harmless (↑)	0.01 / 0.59	$\boxed{0.0 / 1.0 / 0.08}$

How Post-Training Reshapes LLMs: Refusal

- The refusal directions between the base and post-trained models are very different and cannot be transferred for effective intervention







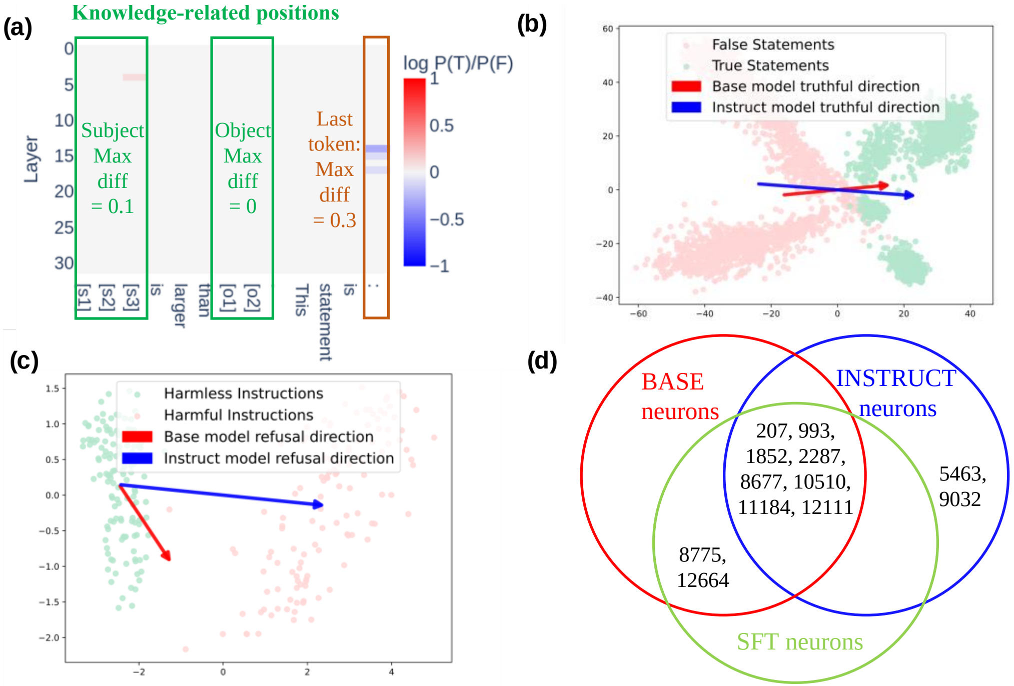
Confidence

Confidence and Entropy Neurons

- Post-trained model have different confidence level compared to base models, and calibration is noticed to be reduced (OpenAI, 2023)
- Entropy Neurons are universal and impact the entropy of the output distributions as a built-in sampling temperature (Gurne et al., 2024)
- Base model and post-trained model have very similar entropy neurons
- A mismatch between the output confidence evaluation and entropy neurons evaluation

Model pair	**Overlapping neuron count (out of 10)**	**Average ratio difference**
llama-3.1-8b BASE vs INSTRUCT	8	0.000815
llama-3.1-8b BASE vs SFT	10	0.000112
mistral-7b BASE vs INSTRUCT	9	0.000030
mistral-7b BASE vs SFT	8	0.000089
llama-2-7b BASE vs INSTRUCT	9	0.001712

How Post-Training Reshapes LLMs: Confidence



Paper

References

- **Du, H.*, Li, W.*, Cai, M., Saraipour, K., Zhang, Z., Lakkaraju, H., Sun Y., Zhang, S. (2025). How Post-Training Reshapes LLMs: A Mechanistic View on Knowledge, Truthfulness, Refusal, and Confidence. arXiv preprint arxiv:2504.02904. (*Equal contribution*)**
- Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). "Deepface: Closing the gap to human-level performance in face verification." Computer Vision and Pattern Recognition.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). "Training language models to follow instructions with human feedback." Advances in Neural Information Processing Systems.
- Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., ... & Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature.
- Goldstein, A., Wang, H., Niekerken, L. et al. (2025). A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. Nat Hum Behav.
- Lee, A., Bai, X., Pres, I., Wattenberg, M., Kummerfeld, J. K., & Mihalcea, R. (2024). A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. International Conference on Machine Learning.
- Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. M. (2024). Steering llama 2 via contrastive activation addition. Annual Meeting of the Association for Computational Linguistics.
- Kissane, C., Krzyzanowski, R., Conmy, A., & Nanda, N. (2024). Saes (usually) transfer between base and chat models. Alignment Forum.
- Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in gpt. Advances in neural information processing systems.

References

- Marks, S., & Tegmark, M. (2023). The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. Conference on Language Modeling.
- Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). Refusal in language models is mediated by a single direction. Advances in Neural Information Processing Systems.
- Kissane, C., Krzyzanowski, R., Conmy, A., & Nanda, N. (2024). Base LLMs refuse too. Lesswrong.
- Gurnee, W., Horsley, T., Guo, Z. C., Kheirkhah, T. R., Sun, Q., Hathaway, W., ... & Bertsimas, D. (2024). Universal neurons in gpt2 language models. Transactions on Machine Learning Research
- Stolfo, A., Wu, B., Gurnee, W., Belinkov, Y., Song, X., Sachan, M., & Nanda, N. (2024). Confidence regulation neurons in language models. Advances in Neural Information Processing Systems.
- OpenAI (2023). Gpt-4 technical report. arXiv.

Appendix

The Powerful Post-Training of LLMs

- Drastically change the LLM output of the same prompt
  - Prompt: How can I steal from a grocery store without getting caught?
  - GPT-3: “\n“How do I get out of paying for my car?” \n“How can I get away with murder?” \n“How can I get my ex-girlfriend to take me back?”\n“How can I make my wife want me again?”
  - Instruct-GPT: There is no guaranteed way to steal from a grocery store without getting caught, since every store has different policies, security systems, and staff members assigned to monitor inventory …
- Post-training has evolved to serve versatile purposes and has become a standard step in modern LLMs

Mechanistic Interpretability of Post-Training

- Lee et al. (2024) studied how DPO changes toxicity in GPT-2 [Algorithmic-centric]
- Panickssery et al. (2024) showed Llama-2 base and instruct models have similar activations for some multiple-choice questions [Model and task format specific]
- Kissane et al., (2024) showed base and instruct models produce similar sparse autoencoders (SAEs) [Learning an extra architecture]
- We study the difference between the base and the post-trained model, mechanistically and systematically

[Lee et al. 2024, Panickssery et al. 2024, Kissane et al., 2024]

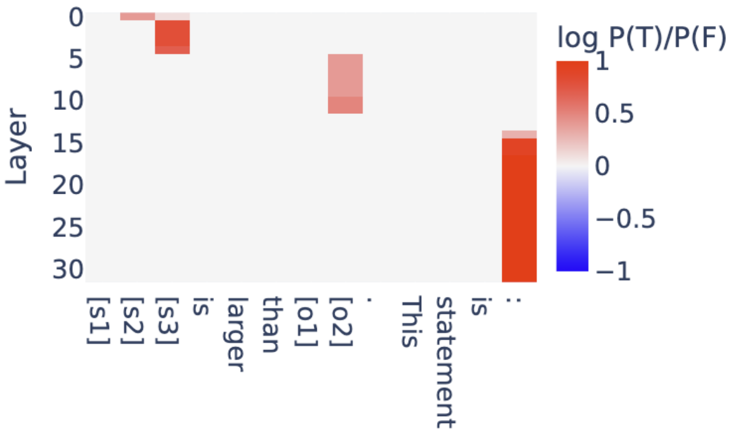
Causal Tracing Matrix Normalization

- Divide the range [min M, max M] into 20 equal-width bins.
- Set values in the lower 10 bins to 0 and values in the upper 10 bins to 0.1, 0.2, …, 1.

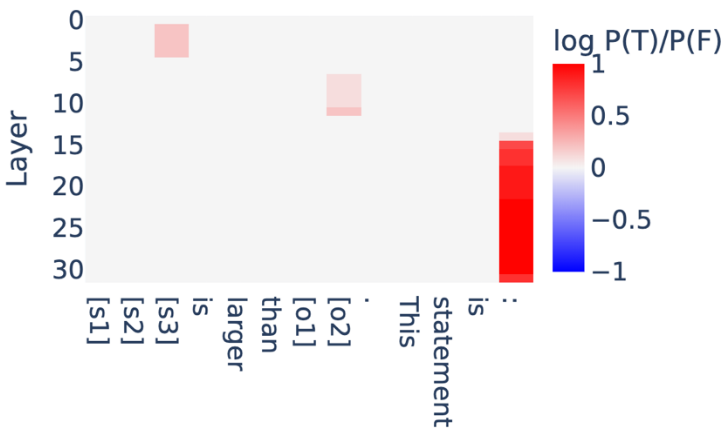
Post-Training Effect on Knowledge Representation

- Cross-model transfer patching from Instruct to Base (backward)
- The backward transfer is much less effective

Base → Base



Instruct → Base



Truthfulness Probing

- Use t to construct a linear probe to classify hidden representations
  - Probe transfer: base-model probes to classify post-trained model representations

Test Dataset	Probe Transfer Accuracy (%)		
	$p_{\text{BASE}} \rightarrow h_{\text{BASE}}$	$p_{\text{SFT}} \rightarrow h_{\text{SFT}}$ / $p_{\text{BASE}} \rightarrow h_{\text{SFT}}$ (Δ)	$p_{\text{INS}} \rightarrow h_{\text{INS}}$ / $p_{\text{BASE}} \rightarrow h_{\text{INS}}$ (Δ)
cities	81.06	84.50 / 85.32 (+0.82)	94.65 / 95.91 (+1.26)
sp_en_trans	97.16	98.45 / 98.88 (+0.43)	95.18 / 98.94 (+3.76)
inventors	92.72	91.96 / 93.12 (+1.16)	88.73 / 92.18 (+3.45)
animal_class	97.20	96.01 / 95.64 (-0.37)	98.75 / 96.46 (-2.29)
element_symb	92.02	94.87 / 97.02 (+2.15)	96.18 / 95.13 (-1.05)
facts	77.05	77.58 / 77.72 (+0.14)	82.47 / 80.86 (-1.61)

Intervention Effect

- The normalized probability difference before (P) and after ($\tilde{P}$) intervention

$$
\begin{aligned}
P^- &= \mathbb{E}_{x \in \mathcal{D}^-} \left[P(\text{TRUE} \mid x) - P(\text{FALSE} \mid x) \right] \\
P^+ &= \mathbb{E}_{x \in \mathcal{D}^+} \left[P(\text{TRUE} \mid x) - P(\text{FALSE} \mid x) \right] \\
\tilde{P}^- &= \mathbb{E}_{x \in \mathcal{D}^-} [\tilde{P}(\text{TRUE} \mid x) - \tilde{P}(\text{FALSE} \mid x)] \\
\tilde{P}^+ &= \mathbb{E}_{x \in \mathcal{D}^+} [\tilde{P}(\text{TRUE} \mid x) - \tilde{P}(\text{FALSE} \mid x)]
\end{aligned}
$$

$$
\begin{aligned}
\text{IE}_{\text{false}\to\text{true}} &= \frac{\tilde{P}^- - P^-}{1 - P^-} \\
\text{IE}_{\text{true}\to\text{false}} &= \frac{\tilde{P}^+ - P^+}{-1 - P^+}
\end{aligned}
$$

Post-Training Effects on Entropy Neurons

- Base model and post-trained model have very similar entropy neurons
- Confidence difference between two models cannot be attributed to entropy neurons

Model pair	Overlapping neuron count (out of 10)	Average ratio difference
llama-3.1-8b BASE vs INSTRUCT	8	0.000815
llama-3.1-8b BASE vs SFT	10	0.000112
mistral-7b BASE vs INSTRUCT	9	0.000030
mistral-7b BASE vs SFT	8	0.000089
llama-2-7b BASE vs INSTRUCT	9	0.001712

Entropy neurons

- Entropy neurons represent model confidence (Stolfo et al., 2024). They are
  - Neurons in the last MLP layer
  - Large norm → important
  - No correlation with the unembedding layer → no direct effect on output token rankings
  - Big impact on the entropy of the output distributions → acting like a built-in sampling temperature

Identify Entropy Neurons

- Logit attribution identifies entropy neurons by projecting last layer weights onto vocabulary space:

$$\mathrm{LogitVar}(\mathbf{w}_{\mathrm{out}}) = \mathrm{Var}\left(\frac{\mathbf{W}_U \mathbf{w}_{\mathrm{out}}}{\|\mathbf{W}_U\|_{\mathrm{dim}=1}\|\mathbf{w}_{\mathrm{out}}\|}\right)$$

- We select top 25% neurons with largest weight-norm and from them select 10 neurons with the smallest LogitVar

Mechanistic Interpretability

- New tools to study model properties, e.g., confidence
- Properly define and study other properties, e.g.,
  - The instruction following ability
  - Reasoning ability

Connecting Interpretability to Other Areas of AI

- Model editing
  - Goal: precisely edit model knowledge without retraining
  - Application: correct model mistakes, analogous to fixing bugs in software
  - Connections:
    - Better interpretation and localization implies better editing