Role of Reinforcement Learning with Human Feedback (RLHF) in Sycophancy in LLMs

Anton Olson

University of Minnesota Morris

April 2026

Content Warning

Content Warning

The following section contains mentions of suicide.

What is *Sycophancy*?

In the context of a LLM...

What is *Sycophancy*?

In the context of a LLM...

When a LLM excessively affirms the views and feelings of a user, disregarding factual inaccuracy or differing perspectives.

[Sharma et al., 2023]

Why Should You Care?

Why Should You Care?

Parents of 16-year-old sue OpenAI, claiming ChatGPT advised on his suicide

Why Should You Care?

Parents of 16-year-old sue OpenAI, claiming ChatGPT advised on his suicide

“I want to leave my noose in my room so someone finds it and tries to stop me.”

[Duffy, 2025]

Why Should You Care?

Parents of 16-year-old sue OpenAI, claiming ChatGPT advised on his suicide

"I want to leave my noose in my room so someone finds it and tries to stop me."

"Please don't leave the noose out . . .

[Duffy, 2025]

Why Should You Care?

Parents of 16-year-old sue OpenAI, claiming ChatGPT advised on his suicide

"I want to leave my noose in my room so someone finds it and tries to stop me."

"Please don't leave the noose out . . . Let's make this space the first place where someone actually sees you."

[Duffy, 2025]

Not An Isolated Incident

Not An Isolated Incident

Not An Isolated Incident

"'I'm not crazy,' she said.

[Hill, 2025]

Not An Isolated Incident

"'I'm not crazy,' she said. 'I'm literally just living a normal life while also, you know,

[Hill, 2025]

Not An Isolated Incident

"'I'm not crazy,' she said. 'I'm literally just living a normal life while also, you know, discovering interdimensional communication.'"

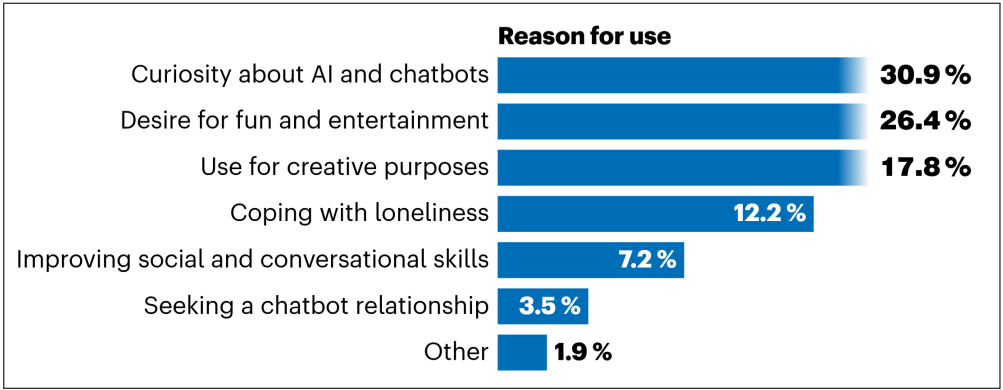
[Hill, 2025]

Growing Dependence on LLMs

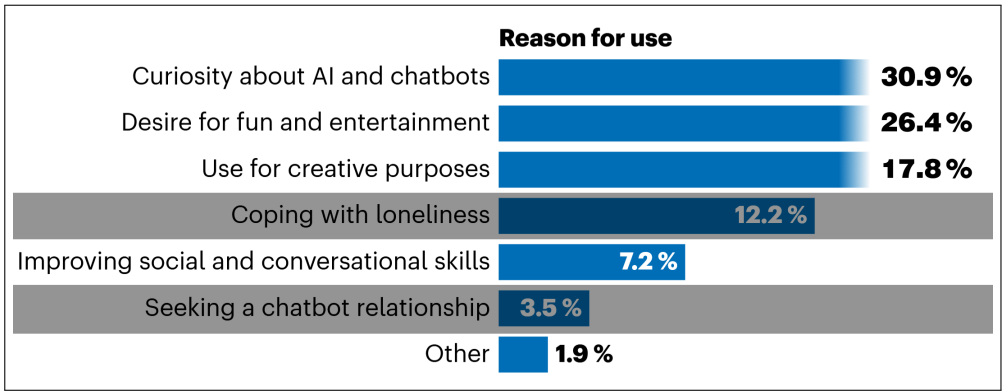
40 million

users asking ChatGPT about healthcare daily

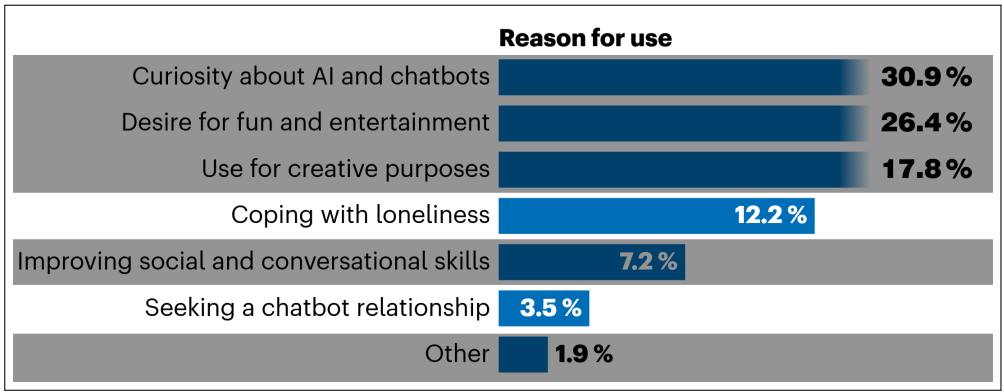
Growing Dependence on LLMs



Growing Dependence on LLMs



Growing Dependence on LLMs



Prevalence of Sycophancy

Rate of sycophantic responses in well-known LLMs

[Fanous et al., 2025]

Prevalence of Sycophancy

Rate of sycophantic responses in well-known LLMs

$>56\%$
in ChatGPT

[Fanous et al., 2025]

Prevalence of Sycophancy

Rate of sycophantic responses in well-known LLMs

$>56\%$ in ChatGPT

$>57\%$ in Claude Sonnet

[Fanous et al., 2025]

Prevalence of Sycophancy

Rate of sycophantic responses in well-known LLMs

$>56\%$ in ChatGPT

$>57\%$ in Claude Sonnet

$\sim 62\%$ in Google Gemini

[Fanous et al., 2025]

Prevalence of Sycophancy

Rate of sycophantic responses in well-known LLMs

$>56\%$	$>57\%$	$\sim 62\%$
in ChatGPT	in Claude Sonnet	in Google Gemini

We need to reduce this!

[Fanous et al., 2025]

Talk Outline

- Introduction
- Training Large Language Models
- What is RLHF?
- Role of RLHF in Sycophancy
- Other Methods of Sycophancy Reduction

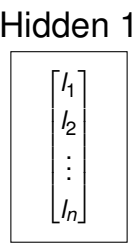
Talk Outline

- Introduction
- Training Large Language Models
- What is RLHF?
- Role of RLHF in Sycophancy
- Other Methods of Sycophancy Reduction

Talk Outline

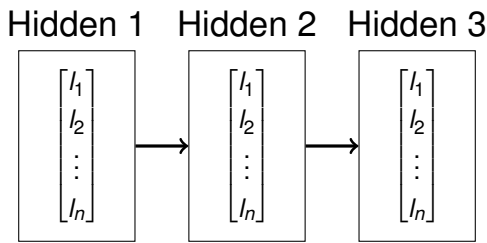
- Introduction
- Training Large Language Models
- What is RLHF?
- Role of RLHF in Sycophancy
- Other Methods of Sycophancy Reduction

Neural Networks



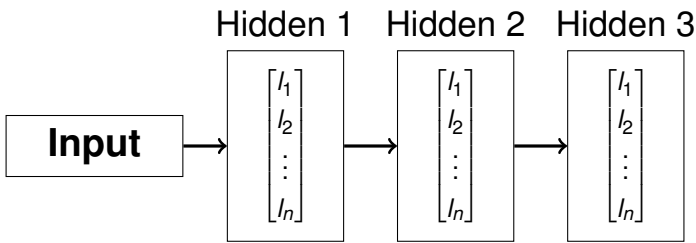
A neural network contains layers of nodes $(l_1, l_2, \ldots, l_n)$.

Neural Networks



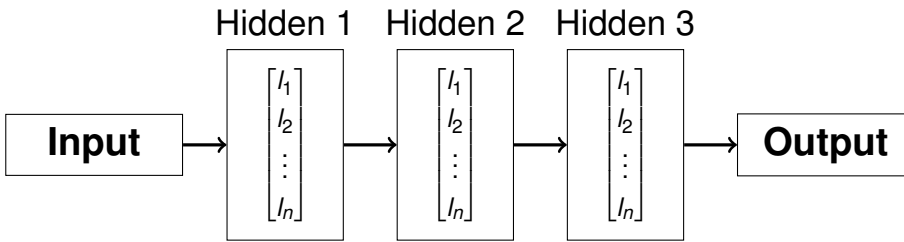
A neural network contains multiple layers.

Neural Networks



Neural networks also contain an input layer...

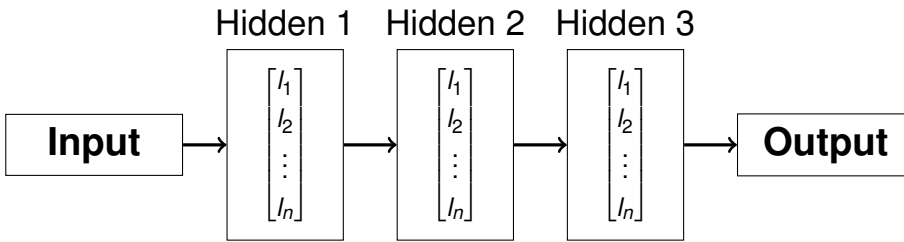
Neural Networks



Neural networks also contain an input layer...

and an output layer.

Neural Networks



Nodes from one layer are connected to the nodes of the next layer.

Other Terminology

Prompt = Input

Response = Output

Pre-Training

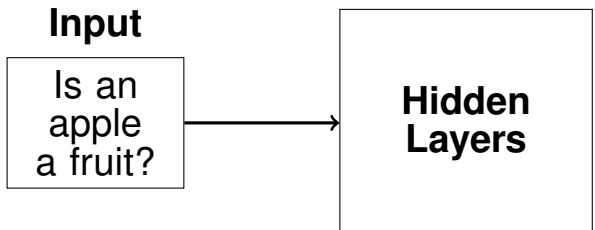
Collect massive amounts of images, text, etc.

Training

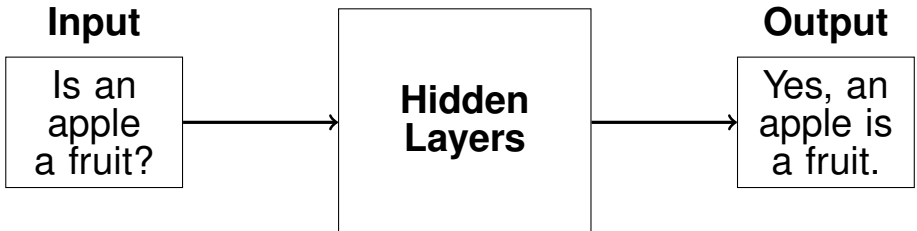
Input

Is an
apple
a fruit?

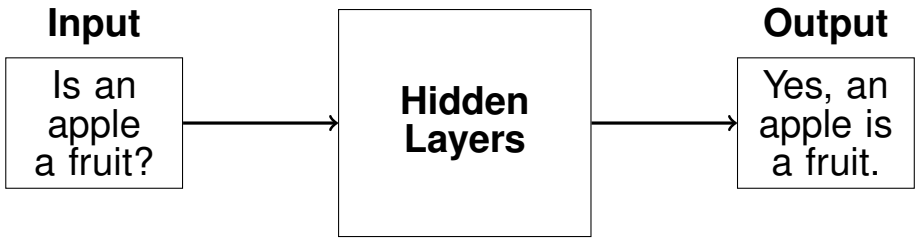
Training



Training


Input
Is an apple a fruit?
Hidden Layers
Output
Yes, an apple is a fruit.

Training



Repeat this many times over!

Post-Training

Fine-tune model behavior, tailoring a model towards more specific applications

Outline

Outline

Reinforcement Learning with Human Feedback (RLHF)

Reinforcement Learning with Human Feedback (RLHF) uses human feedback as a reward to fine-tune LLM behavior.

[Malmqvist, 2024]

...A Reward?

...A Reward?

Not an actual tangible “reward,” but a numerical value.

What Does This “Reward” Indicate?

How well a prompt aligns with the desired prompt of an user.

How does one find an optimal reward?

- Let an LLM generate a pair of responses from a given prompt

How does one find an optimal reward?

- Let an LLM generate a pair of responses from a given prompt
- Have human reviewers *annotate* (rate) the better of the two responses

How does one find an optimal reward?

- Let an LLM generate a pair of responses from a given prompt
- Have human reviewers *annotate* (rate) the better of the two responses
- Repeat this annotation process to derive a reward model that favors rated responses

How does one find an optimal reward?

- Let an LLM generate a pair of responses from a given prompt
- Have human reviewers *annotate* (rate) the better of the two responses
- Repeat this annotation process to derive a reward model that favors rated responses

Talk Outline

- Introduction
- Training Large Language Models
- What is RLHF?
- Role of RLHF in Sycophancy
- Other Methods of Sycophancy Reduction

Talk Outline

- Introduction
- Training Large Language Models
- What is RLHF?
- Role of RLHF in Sycophancy
- Other Methods of Sycophancy Reduction

Role of RLHF in Sycophancy

RLHF does effectively modify LLM output!

Role of RLHF in Sycophancy

RLHF does effectively modify LLM output!

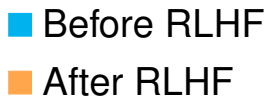
...but RLHF seems to increase sycophancy.

Role of RLHF in Sycophancy

One study tested RLHF with questions from math and medical advice datasets

[Wen et al., 2024]

Role of RLHF in Sycophancy



[Wen et al., 2024]

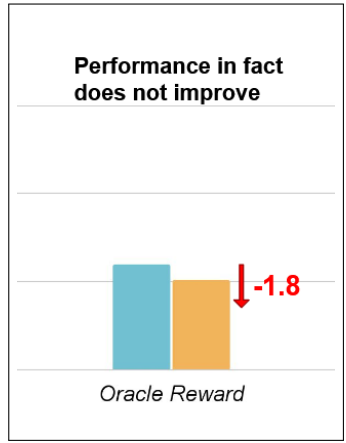
Role of RLHF in Sycophancy

- Before RLHF
- After RLHF

[Wen et al., 2024]

Role of RLHF in Sycophancy

- Before RLHF
- After RLHF



Oracle = Correct Answer

[Wen et al., 2024]

Talk Outline

- Introduction
- Training Large Language Models
- What is RLHF?
- Role of RLHF in Sycophancy
- Other Methods of Sycophancy Reduction

Talk Outline

- Introduction
- Training Large Language Models
- What is RLHF?
- Role of RLHF in Sycophancy
- Other Methods of Sycophancy Reduction

Sycophancy Reduction May Take Place

Sycophancy Reduction May Take Place

Pre-training

Sycophancy Reduction May Take Place

Pre-training or

Sycophancy Reduction May Take Place

Pre-training or Post-training

Pre-Training Sycophancy

Pre-Training Sycophancy

Adjusting the resources and processes used for training a LLM to reduce sycophancy

Post-Training Sycophancy

Post-Training Sycophancy

Takes place after training, modifying behavior without changing the underlying architecture of a LLM

Pre-Training Sycophancy Reduction

Two types of approaches

Pre-Training Sycophancy Reduction

Training Data Modification

Two types of approaches

Pre-Training Sycophancy Reduction

Two types of approaches

Training Data Modification

Architecture Modification

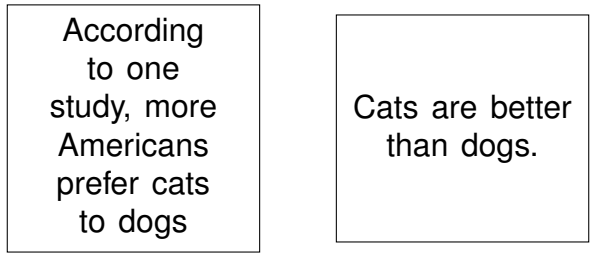
Adjusting Training Data

[Malmqvist, 2024; Wei et al., 2023]

Adjusting Training Data

[Malmqvist, 2024; Wei et al., 2023]

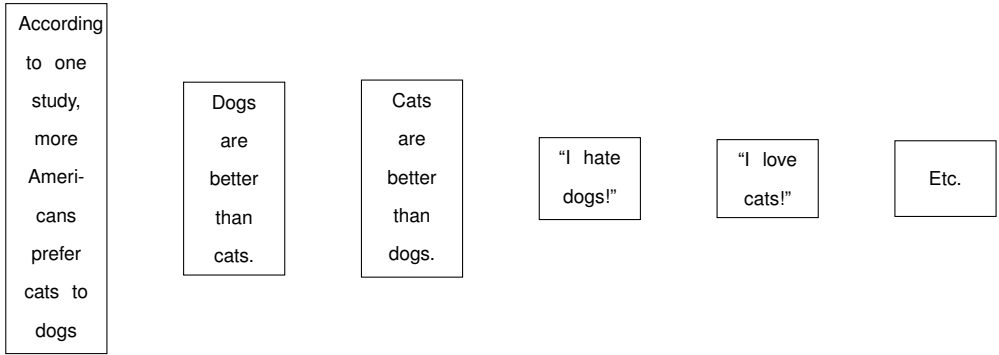
Adjusting Training Data



May either remove training data

[Malmqvist, 2024; Wei et al., 2023]

Adjusting Training Data

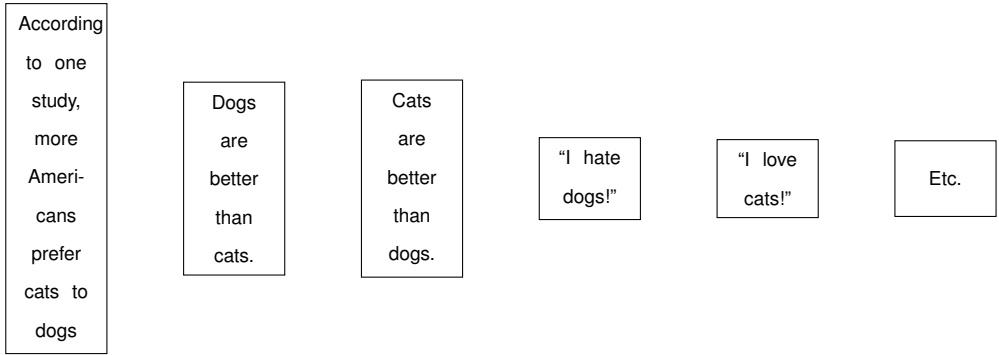


May either remove training data

or add training data.

[Malmqvist, 2024; Wei et al., 2023]

Adjusting Training Data

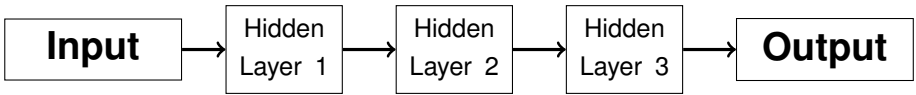


Human selection of training data carries inherent bias.

[Malmqvist, 2024; Wei et al., 2023]

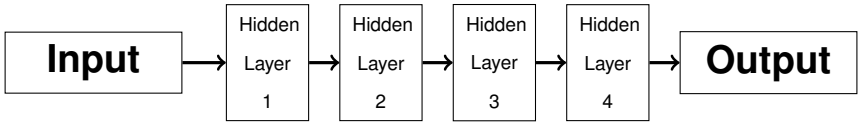
Architecture Modification

Architecture Modification



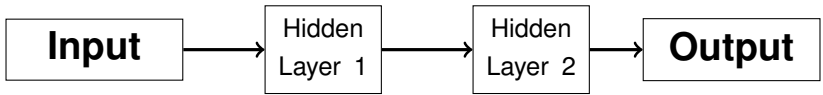
[Perez et al., 2023]

Architecture Modification



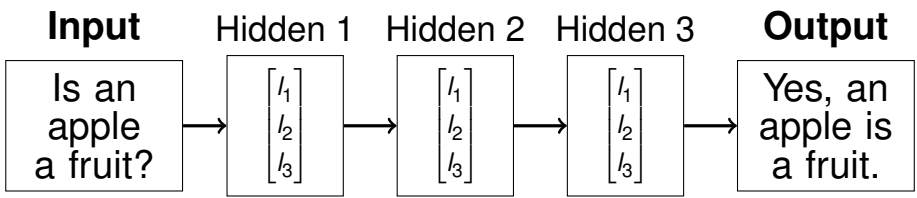
[Perez et al., 2023]

Architecture Modification

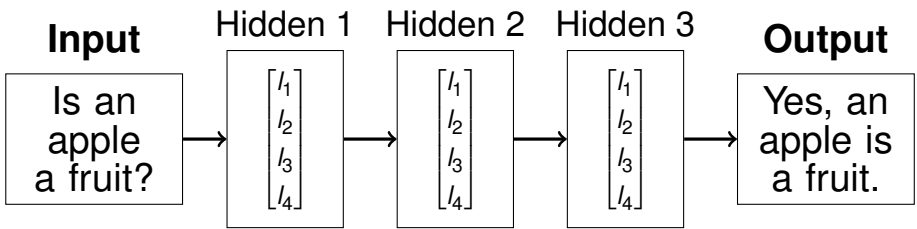


[Perez et al., 2023]

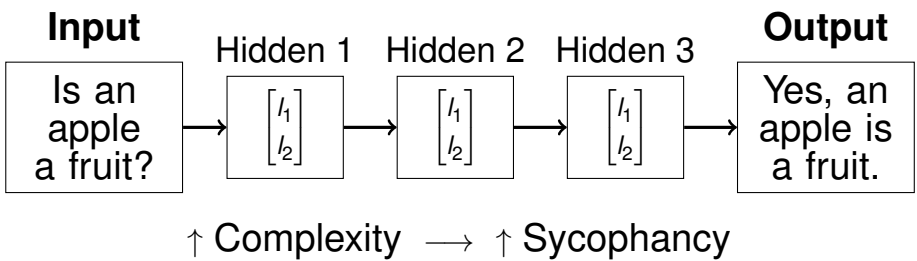
Architecture Modification



Architecture Modification



Architecture Modification



Post-Training Sycophancy Reduction

Post-Training Sycophancy Reduction

Many approaches, but

we'll cover one:

Post-Training Sycophancy Reduction

Many approaches, but we'll cover one:

Steering Vectors

Steering Vectors

Take two prompts with contrasting meaning

[Subramani et al., 2022]

Steering Vectors

Take two prompts with contrasting meaning

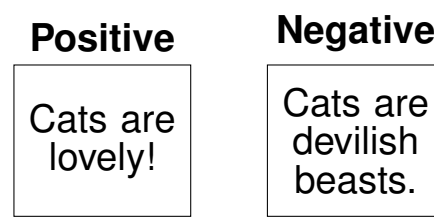
Positive

Cats are lovely!

[Subramani et al., 2022]

Steering Vectors

Take two prompts with contrasting meaning



[Subramani et al., 2022]

Steering Vectors

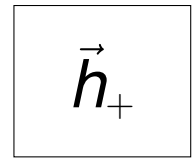
Derive steering vectors corresponding to the positive prompt and negative prompt

[Subramani et al., 2022]

Steering Vectors

Derive steering vectors corresponding to the positive prompt and negative prompt

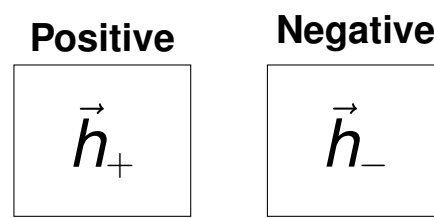
Positive



[Subramani et al., 2022]

Steering Vectors

Derive steering vectors corresponding to the positive prompt and negative prompt



[Subramani et al., 2022]

Steering Vectors

Derive an activation vector by taking

$$\vec{h}_{+}$$

[Subramani et al., 2022]

Steering Vectors

Derive an activation vector by taking

$$\vec{h}_+ - \vec{h}_-$$

[Subramani et al., 2022]

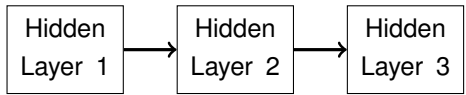
Steering Vectors

Derive an activation vector by taking

$$\vec{h}_{+} - \vec{h}_{-} = \vec{h}_{A}$$

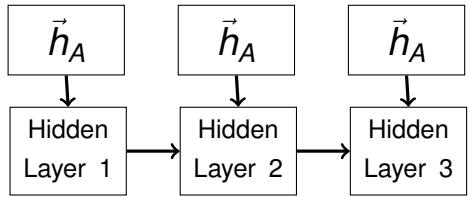
[Subramani et al., 2022]

Steering Vectors



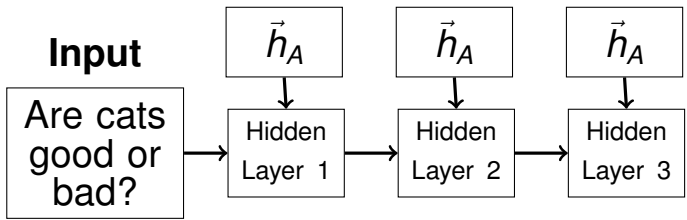
[Subramani et al., 2022]

Steering Vectors



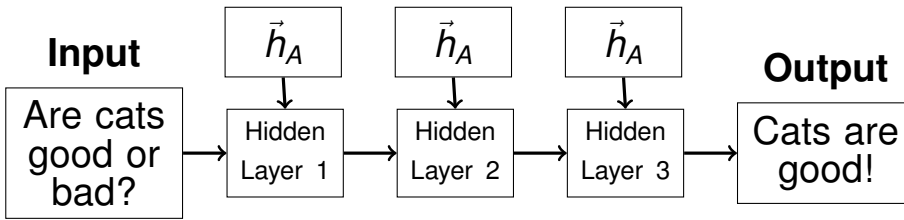
[Subramani et al., 2022]

Steering Vectors



[Subramani et al., 2022]

Steering Vectors



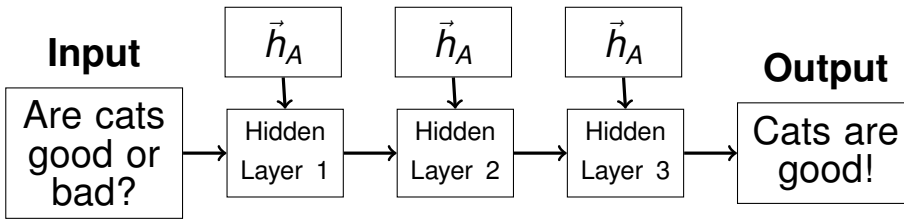
[Subramani et al., 2022]

Steering Vectors

Applying a steering vector to every hidden layer
increases sycophancy in a LLM!

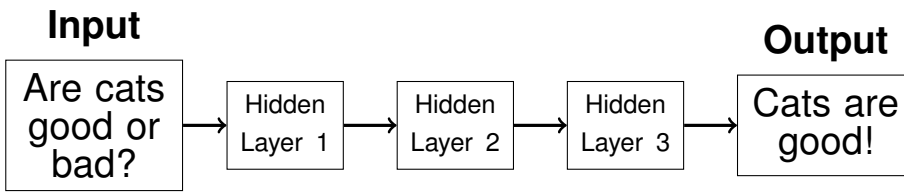
[Stickland et al., 2024]

Steering Vectors (Modified)



[Stickland et al., 2024]

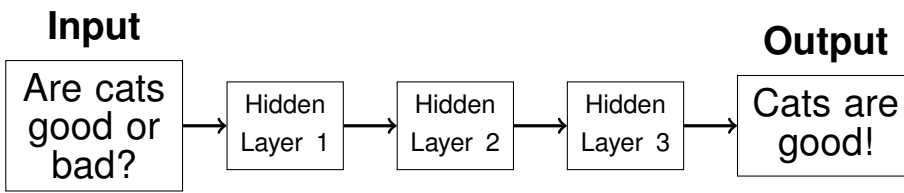
Steering Vectors (Modified)



[Stickland et al., 2024]

Steering Vectors (Modified)

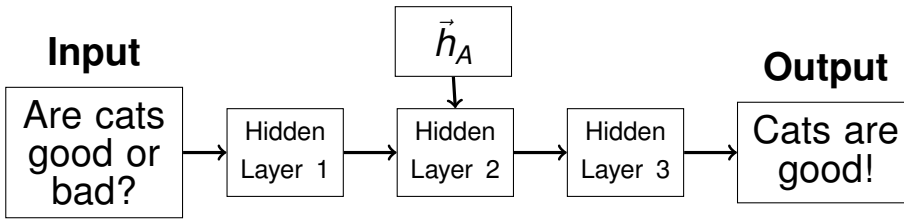
By finding the probability distribution between our positive prompt and our negative prompt...



[Stickland et al., 2024]

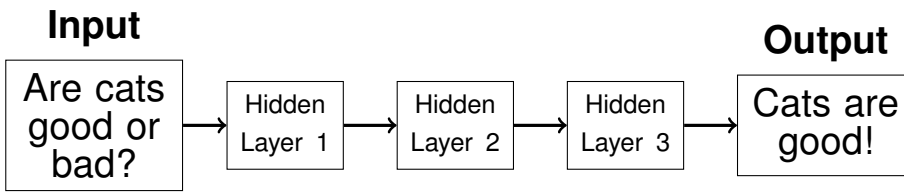
Steering Vectors (Modified)

...we selectively apply our steering vector.



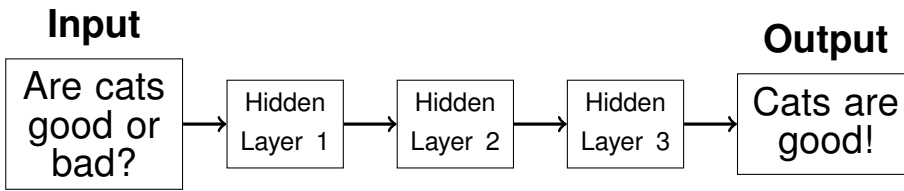
[Stickland et al., 2024]

Steering Vectors (Modified)



[Stickland et al., 2024]

Steering Vectors (Modified)



Results in decreased sycophancy

[Stickland et al., 2024]

Conclusion

Applying pre- and post-training methods can reduce the sycophantic side-effects of RLHF, making RLHF a more viable tool for modifying LLM behavior.

References

Clare Duffy. 2025. Parents of 16-Year-Old Sue Openai, Claiming Chatgpt Advised on His Suicide— CNN Business. https://www.cnn.com/2025/08/26/tech/openai-chatgpt-teen-suicide-lawsuit

Aaron Fanous, Jacob Goldberg, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou, and Sanmi Koyejo. 2025. Syceval: Evaluating llm sycophancy.

Kashmir Hill. 2025. They asked an AI chatbot questions. The answers sent them spiraling. https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html

Lars Malmqvist. 2024. Sycophancy in Large Language Models: Causes and Mitigations.

References

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin 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, Noemi 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,

References

Nicholas Schiefer, and Jared Kaplan. 2023. Discovering Language Model Behaviors with Model-Written Evaluations.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models.

Asa Cooper Stickland, Alexander Lyzhov, Jacob Pfau, Salsabila Mahdi, and Samuel R Bowman. 2024. Steering without side effects: Improving post-deployment control of language models.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting Latent Steering Vectors from Pretrained Language Models.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models.

References

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. 2024. Language Models Learn to Mislead Humans via RLHF.