CSCE 689: Special Topics in Advanced NLP

Lecture 15: Model Interpretability, Model Steering

Kuan-Hao Huang

Fall 2026



Course Project – Midterm Report

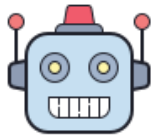
- Due: 11/5
- Page limit: 4 pages (excluding reference)
- Format: ACL style
- The report should include
  - Introduction to the topic you choose and problem definition
  - Related literature and overview of existing progress and challenges
  - Proposed solutions, novelty, and expected contributions
  - Evaluation metrics
  - Planned implementation details, including dataset, models, codebases, etc.
  - Current progress (preliminary results, baseline results, etc.)
  - Next steps and expected timeline

Model Explainability and Interpretability

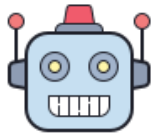


Hello! Could you help me reserve a table at the *"The Best"* restaurant for tomorrow at 12pm?

Of course! I've reserved a table at the *"The Best"* restaurant for tomorrow at 12pm.



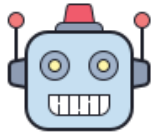
I generate this response is because I saw you mention reserve, one restaurant name, and one specific time. Therefore…



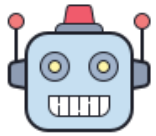


Hello! Could you help me reserve a table at the *"The Best"* restaurant for tomorrow at 12pm?

Of course! I've reserved a table at the *"The Best"* restaurant for tomorrow at 12pm.



I generate this response is because I saw you mention tomorrow. It is usually strongly related to restaurant reservation.



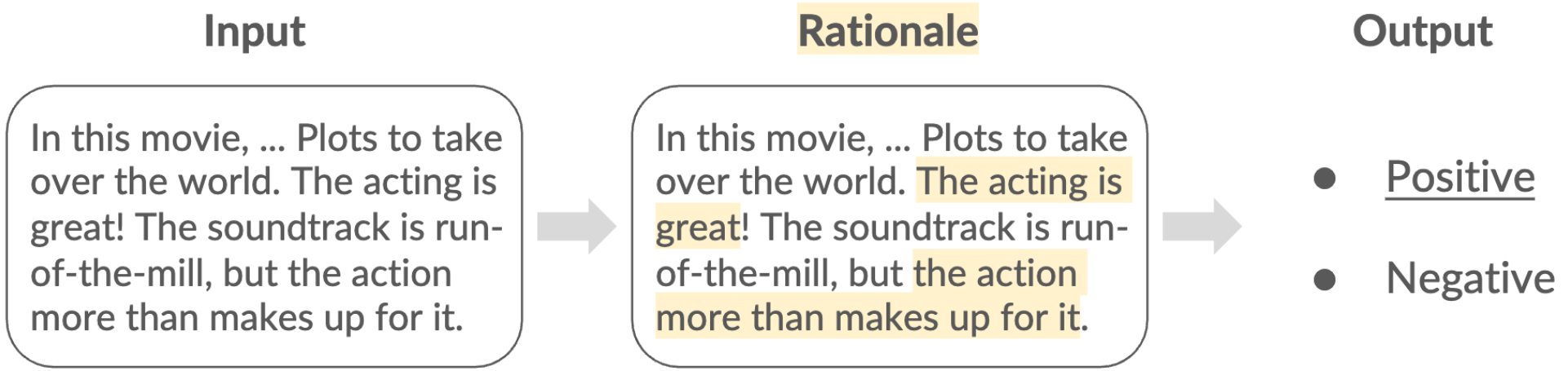
Provide additional information to decide if we should trust the answers

Rationalizing Neural Predictions

Tao Lei, Regina Barzilay and Tommi Jaakkola
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
{taolei, regina, tommi}@csail.mit.edu

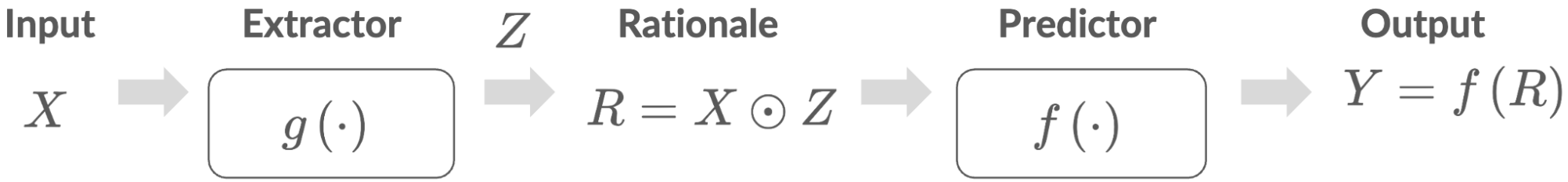
Extractive Rationales

- Rationales: short snippets in inputs that support outputs



Extractive Rationales

- Pipeline model



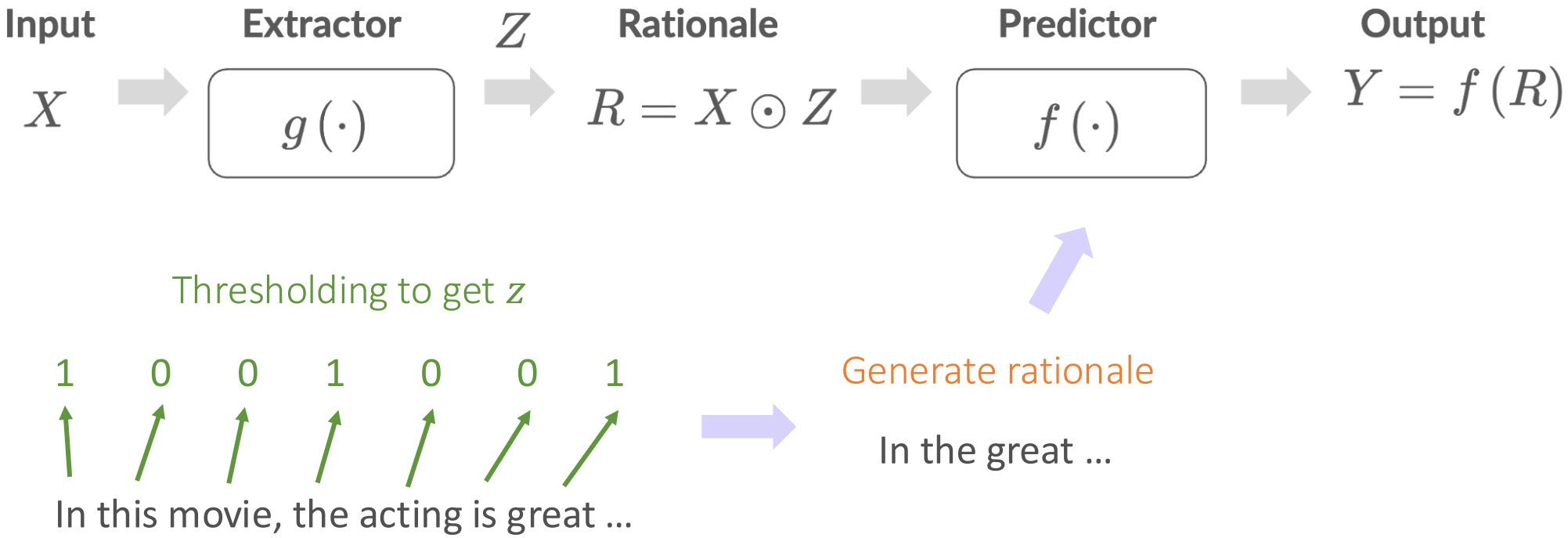
Model $P(z|x) = g(x)$

0.51 0.12 0.87 0.66 0.43 0.22 0.95

In this movie, the acting is great …

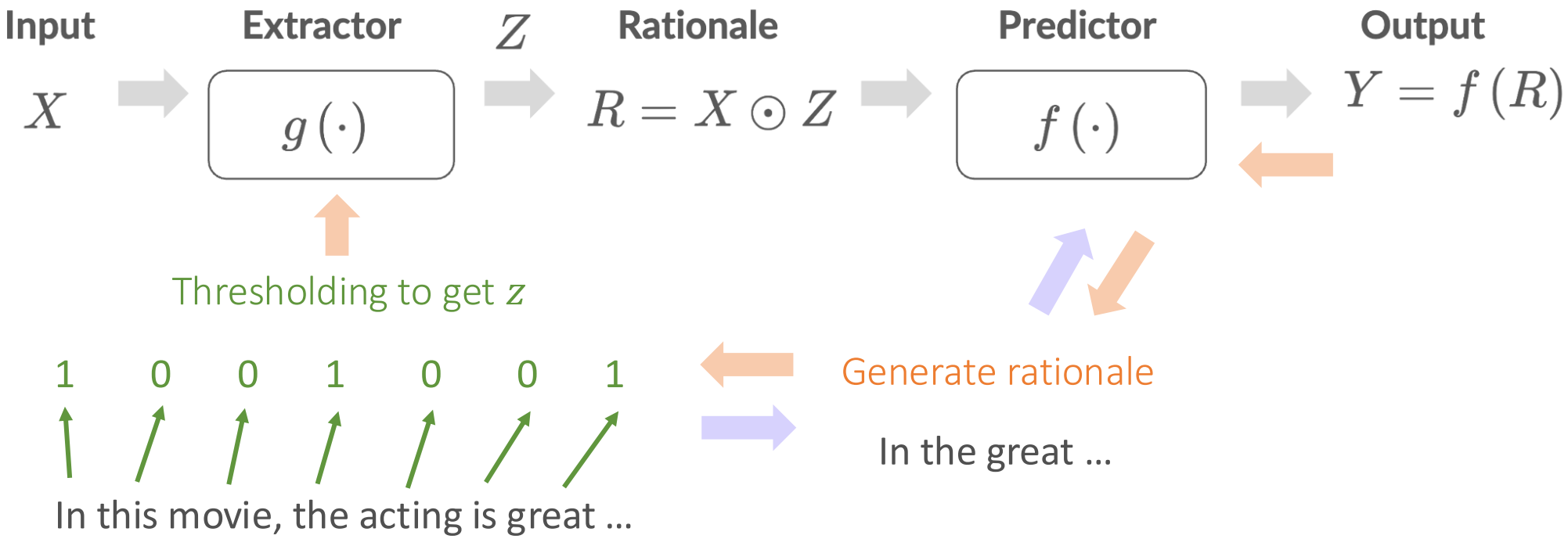
Extractive Rationales

- Pipeline model



Extractive Rationales

- Pipeline model



Examples

a beer that is not sold in my neck of the woods , but managed to get while on a roadtrip . poured into an imperial pint glass with a generous head that sustained life throughout . nothing out of the ordinary here , but a good brew still . body was kind of heavy , but not thick . the hop smell was excellent and enticing . very drinkable

---

very dark beer . pours a nice finger and a half of creamy foam and stays throughout the beer . smells of coffee and roasted malt . has a major coffee-like taste with hints of chocolate . if you like black coffee , you will love this porter . creamy smooth mouthfeel and definitely gets smoother on the palate once it warms . it 's an ok porter but i feel there are much better one 's out there .

---

i really did not like this . it just seemed extremely watery . i dont ' think this had any carbonation whatsoever . maybe it was flat , who knows ? but even if i got a bad brew i do n't see how this would possibly be something i 'd get time and time again . i could taste the hops towards the middle , but the beer got pretty nasty towards the bottom . i would never drink this again , unless it was free . i 'm kind of upset i bought this .

---

a : poured a nice dark brown with a tan colored head about half an inch thick , nice red/garnet accents when held to the light . little clumps of lacing all around the glass , not too shabby . not terribly impressive though s : smells like a more guinness-y guinness really , there are some roasted malts there , signature guinness smells , less burnt though , a little bit of chocolate … … m : relatively thick , it is n't an export stout or imperial stout , but still is pretty hefty in the mouth , very smooth , not much carbonation . not too shabby d : not quite as drinkable as the draught , but still not too bad . i could easily see drinking a few of these .

Gradient-Based Explanations

The storyline is boring, but the actors are great. $\mathcal{L}\big(y, f(x)\big)$

Gradient Norm (↑)

$$\left\| \frac{\partial \mathcal{L}\big(y, f(x)\big)}{\partial x_i} \right\|_2$$

Gradient Norm x Input (↑)

$$\left(\frac{\partial \mathcal{L}\big(y, f(x)\big)}{\partial x_i} \right)^{\mathsf{T}} x_i$$

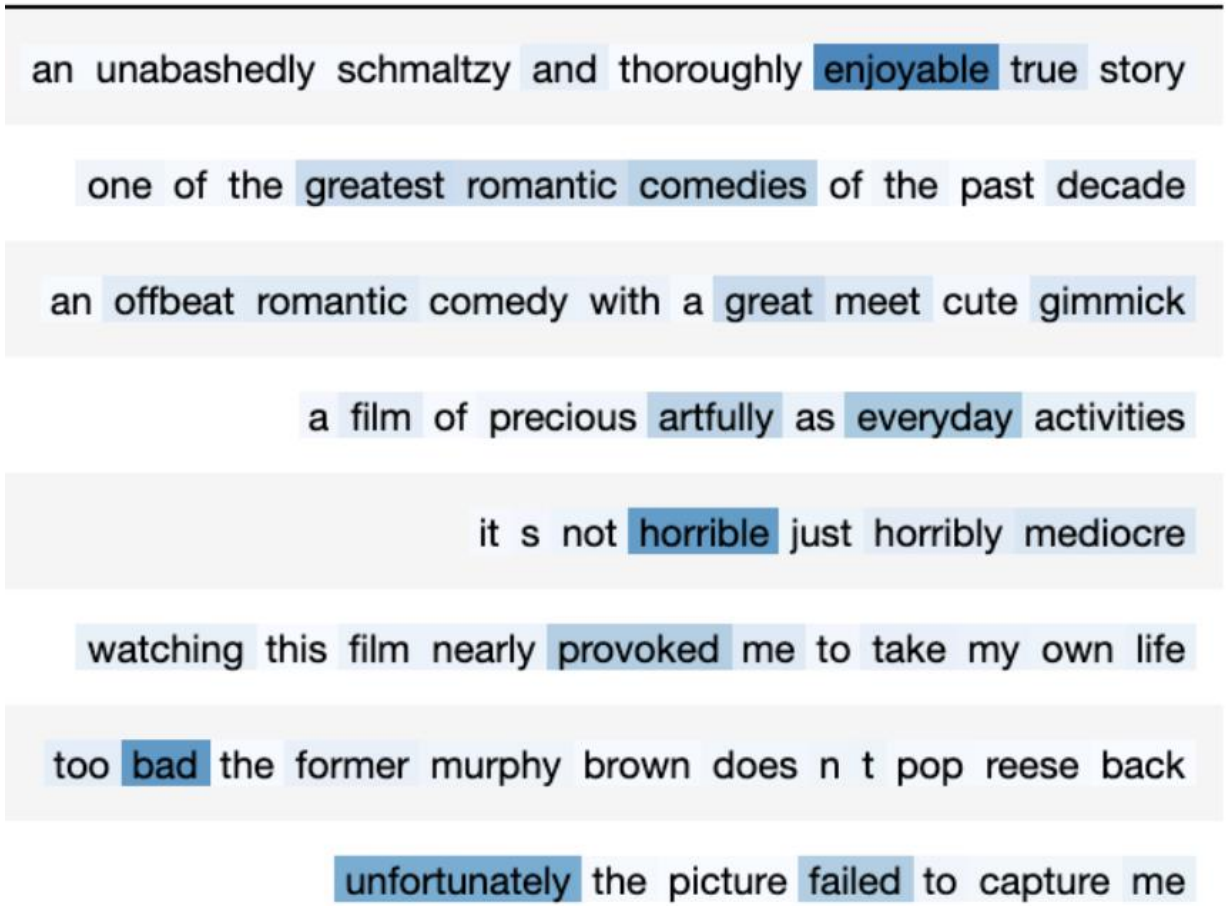
Leave-One-Out Word Importance

The storyline is boring, but the actors are great. $\mathcal{L}\big(y, f(x)\big)$

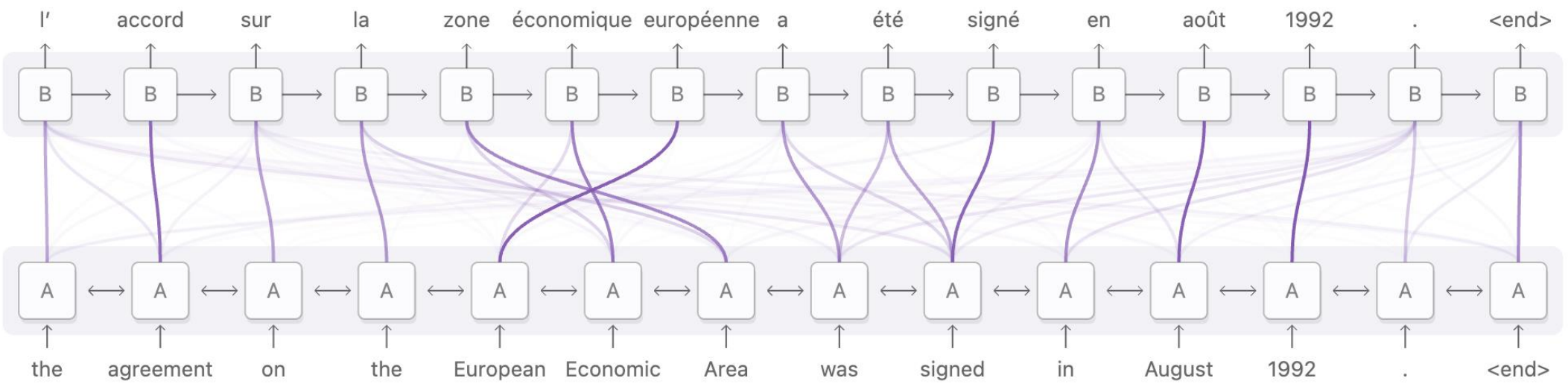
The storyline is [mask], but the actors are great. $\mathcal{L}\big(y, f(x')\big)$

$$\mathcal{L}\big(y, f(x')\big) - \mathcal{L}\big(y, f(x)\big)$$

Examples


an unabashedly schmaltzy and thoroughly enjoyable true story
one of the greatest romantic comedies of the past decade
an offbeat romantic comedy with a great meet cute gimmick
a film of precious artfully as everyday activities
it s not horrible just horribly mediocre
watching this film nearly provoked me to take my own life
too bad the former murphy brown does n t pop reese back
unfortunately the picture failed to capture me

Attention?


l'
accord
sur
la
zone
économique
européenne
a
été
signé
en
août
1992
.
<end>
B
A
the
agreement
on
the
European
Economic
Area
was
signed
in
August
1992
.
<end>

Attention is not Explanation

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Attention is not not Explanation

Sarah Wiegreffe*
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

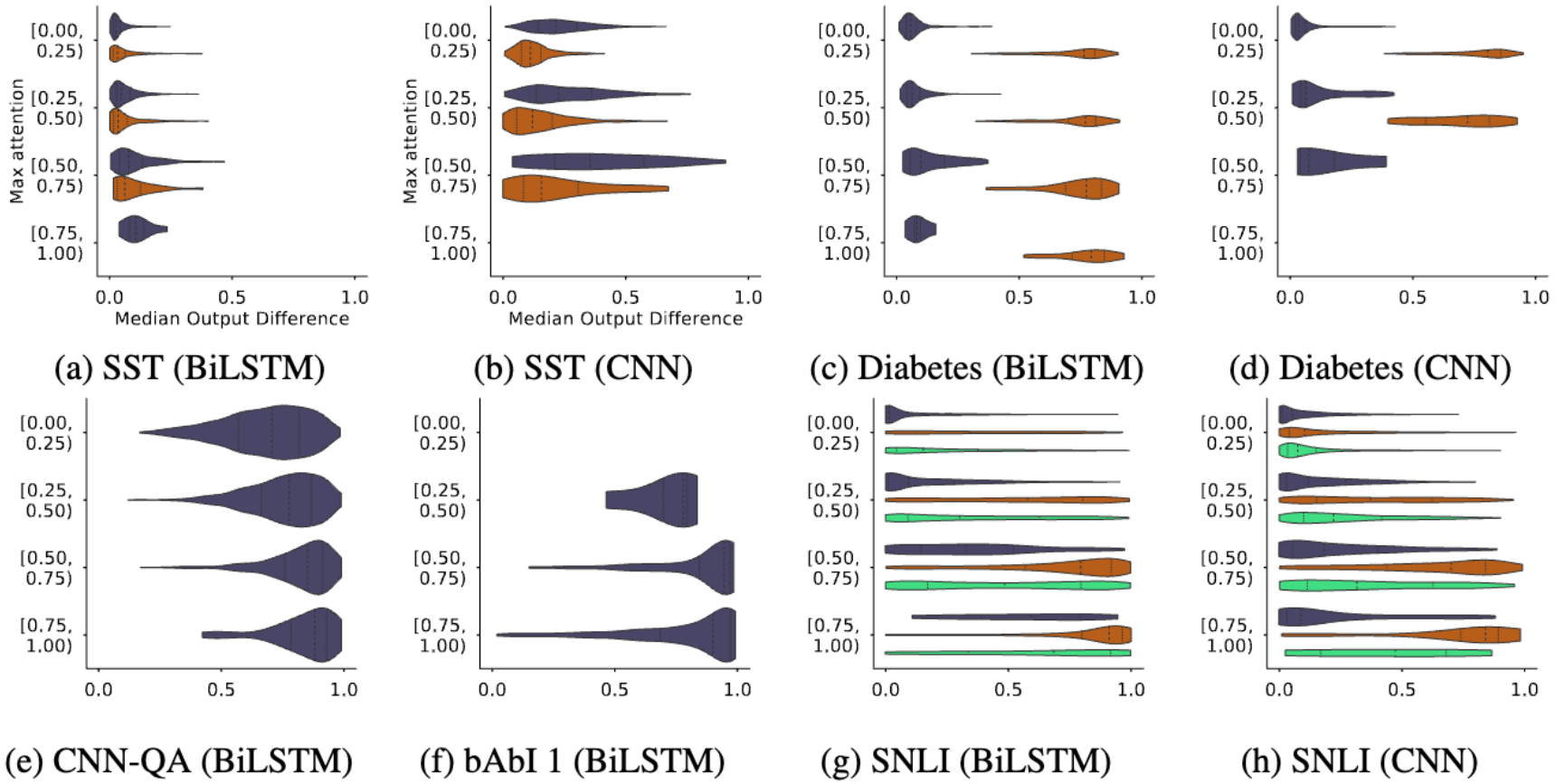
Yuval Pinter*
School of Interactive Computing
Georgia Institute of Technology
uvp@gatech.edu

Experiments

- Correlation between attention-based and gradient-based/leave-one-out

		Gradient (BiLSTM) τ_g		Gradient (Average) τ_g		Leave-One-Out (BiLSTM) τ_{loo}	
Dataset	Class	Mean ± Std.	Sig. Frac.	Mean ± Std.	Sig. Frac.	Mean ± Std.	Sig. Frac.
SST	0	0.34 ± 0.21	0.48	0.61 ± 0.20	0.87	0.27 ± 0.19	0.33
	1	0.36 ± 0.21	0.49	0.60 ± 0.21	0.83	0.32 ± 0.19	0.40
IMDB	0	0.44 ± 0.06	1.00	0.67 ± 0.05	1.00	0.34 ± 0.07	1.00
	1	0.43 ± 0.06	1.00	0.68 ± 0.05	1.00	0.34 ± 0.07	0.99
ADR Tweets	0	0.47 ± 0.18	0.76	0.73 ± 0.13	0.96	0.29 ± 0.20	0.44
	1	0.49 ± 0.15	0.85	0.72 ± 0.12	0.97	0.44 ± 0.16	0.74
20News	0	0.07 ± 0.17	0.37	0.79 ± 0.07	1.00	0.06 ± 0.15	0.29
	1	0.21 ± 0.22	0.61	0.75 ± 0.08	1.00	0.20 ± 0.20	0.62
AG News	0	0.36 ± 0.13	0.82	0.78 ± 0.07	1.00	0.30 ± 0.13	0.69
	1	0.42 ± 0.13	0.90	0.76 ± 0.07	1.00	0.43 ± 0.14	0.91
Diabetes	0	0.42 ± 0.05	1.00	0.75 ± 0.02	1.00	0.41 ± 0.05	1.00
	1	0.40 ± 0.05	1.00	0.75 ± 0.02	1.00	0.45 ± 0.05	1.00
Anemia	0	0.47 ± 0.05	1.00	0.77 ± 0.02	1.00	0.46 ± 0.05	1.00
	1	0.46 ± 0.06	1.00	0.77 ± 0.03	1.00	0.47 ± 0.06	1.00
CNN	Overall	0.24 ± 0.07	0.99	0.50 ± 0.10	1.00	0.20 ± 0.07	0.98
bAbI 1	Overall	0.25 ± 0.16	0.55	0.72 ± 0.12	0.99	0.16 ± 0.14	0.28
bAbI 2	Overall	-0.02 ± 0.14	0.27	0.68 ± 0.06	1.00	-0.01 ± 0.13	0.27
bAbI 3	Overall	0.24 ± 0.11	0.87	0.61 ± 0.13	1.00	0.26 ± 0.10	0.89
SNLI	0	0.31 ± 0.23	0.36	0.59 ± 0.18	0.80	0.16 ± 0.26	0.20
	1	0.33 ± 0.21	0.38	0.58 ± 0.19	0.80	0.36 ± 0.19	0.44
	2	0.31 ± 0.21	0.36	0.57 ± 0.19	0.80	0.34 ± 0.20	0.40

Permutate Attention Weights



(a) SST (BiLSTM) (b) SST (CNN) (c) Diabetes (BiLSTM) (d) Diabetes (CNN)

(e) CNN-QA (BiLSTM) (f) bAbI 1 (BiLSTM) (g) SNLI (BiLSTM) (h) SNLI (CNN)

Adversarial Attention Weights

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

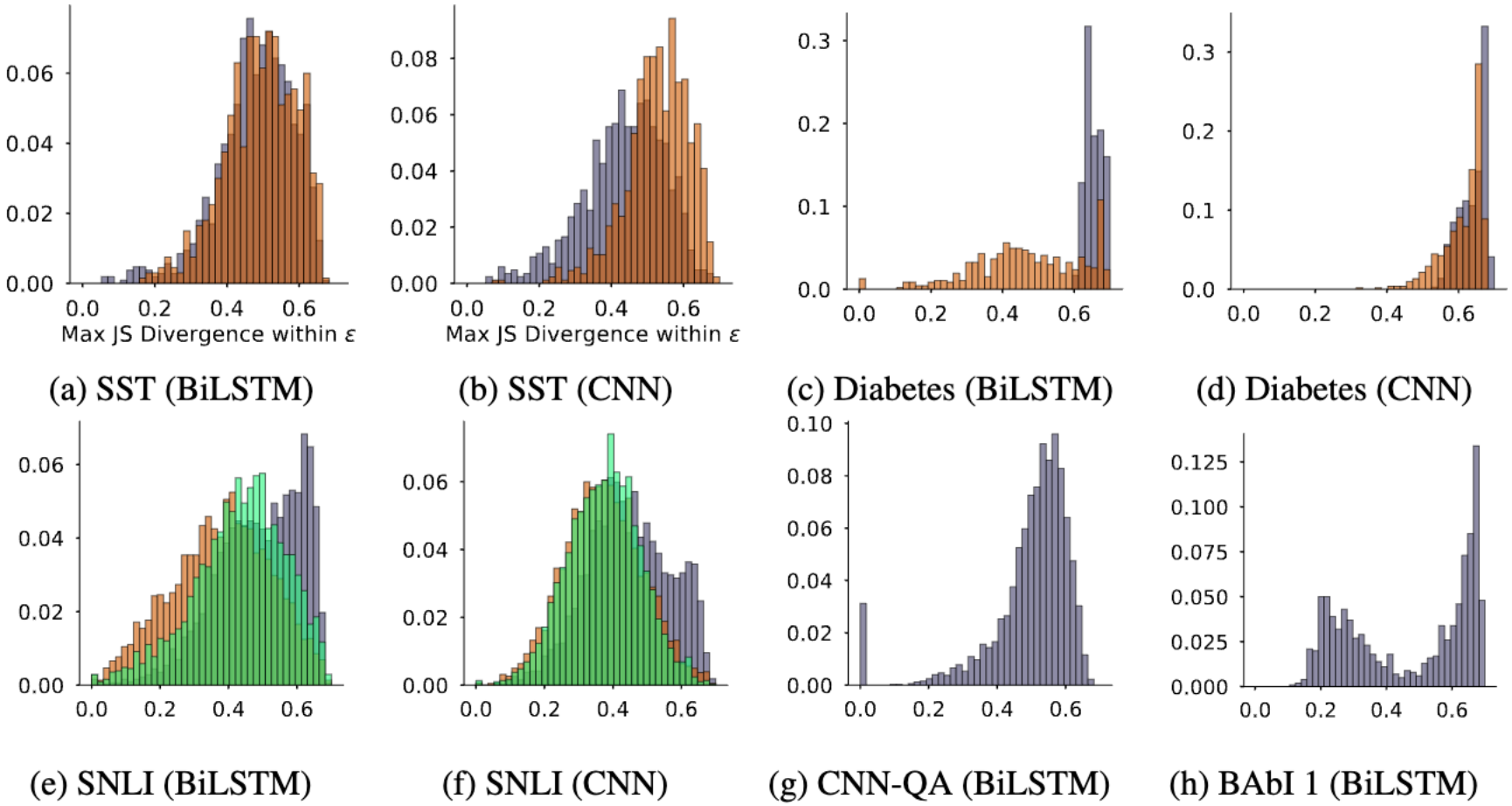
original α

$$f(x|\alpha, \theta) = 0.01$$

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

adversarial $\tilde{\alpha}$

$$f(x|\tilde{\alpha}, \theta) = 0.01$$



(a) SST (BiLSTM) (b) SST (CNN) (c) Diabetes (BiLSTM) (d) Diabetes (CNN)

(e) SNLI (BiLSTM) (f) SNLI (CNN) (g) CNN-QA (BiLSTM) (h) BAbI 1 (BiLSTM)

Attention is not Explanation

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Attention is not not Explanation

Sarah Wiegreffe*
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

Yuval Pinter*
School of Interactive Computing
Georgia Institute of Technology
uvp@gatech.edu

Uniform Attentions

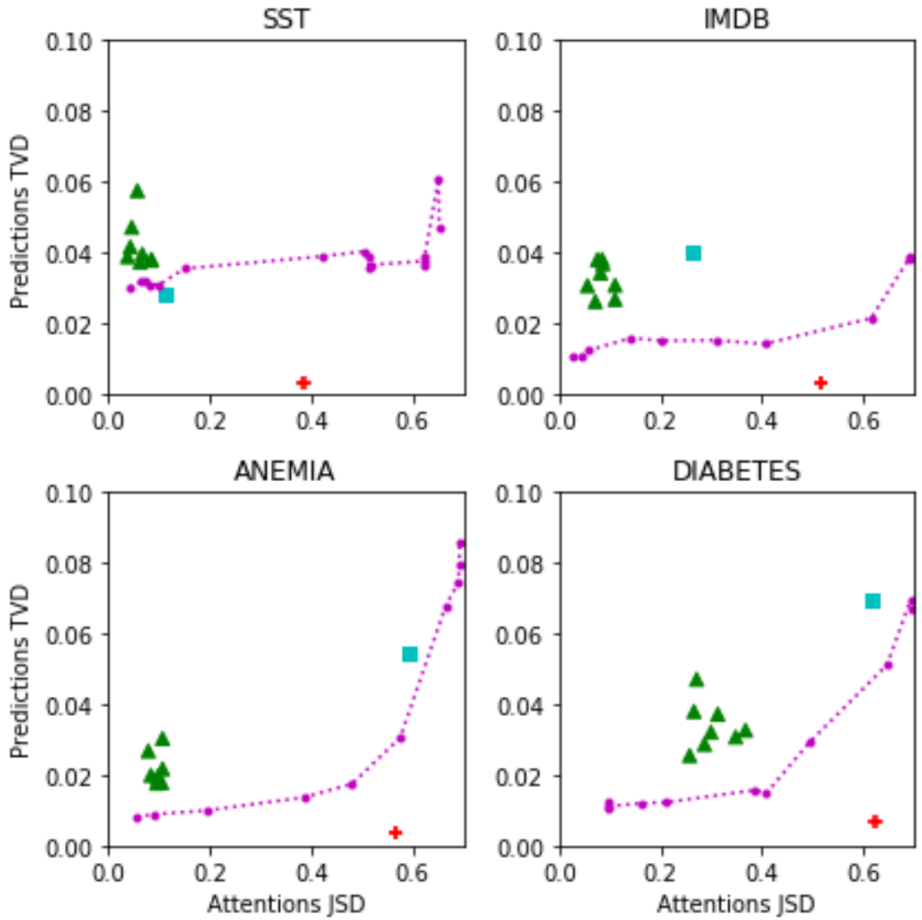
- If attention models are not useful compared to very simple baselines, there is no point in using their outcomes for any type of explanations

Dataset	Attention (Base)		Uniform
	Reported	Reproduced	
Diabetes	0.79	0.775	0.706
Anemia	0.92	0.938	0.899
IMDb	0.88	0.902	0.879
SST	0.81	0.831	0.822
AgNews	0.96	0.964	0.960
20News	0.94	0.942	0.934

Training an Adversary

- Attention distribution is not a primitive
  - We need to re-train for adversarial attention weights

$$\mathcal{L}(\mathcal{M}_a, \mathcal{M}_b)^{(i)} = \text{TVD}(\hat{y}_a^{(i)}, \hat{y}_b^{(i)}) - \lambda \ \text{KL}(\boldsymbol{\alpha}_a^{(\mathbf{i})} \parallel \boldsymbol{\alpha}_b^{(\mathbf{i})})$$



Towards Monosemanticity: Decomposing Language Models With Dictionary Learning

Using a sparse autoencoder, we extract a large number of interpretable features from a one-layer transformer.

Browse A/1 Features →

Browse All Features →

AUTHORS

Trenton Bricken[*], Adly Templeton[*], Joshua Batson[*], Brian Chen[*], Adam Jermyn[*], Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, Chris Olah

AFFILIATIONS

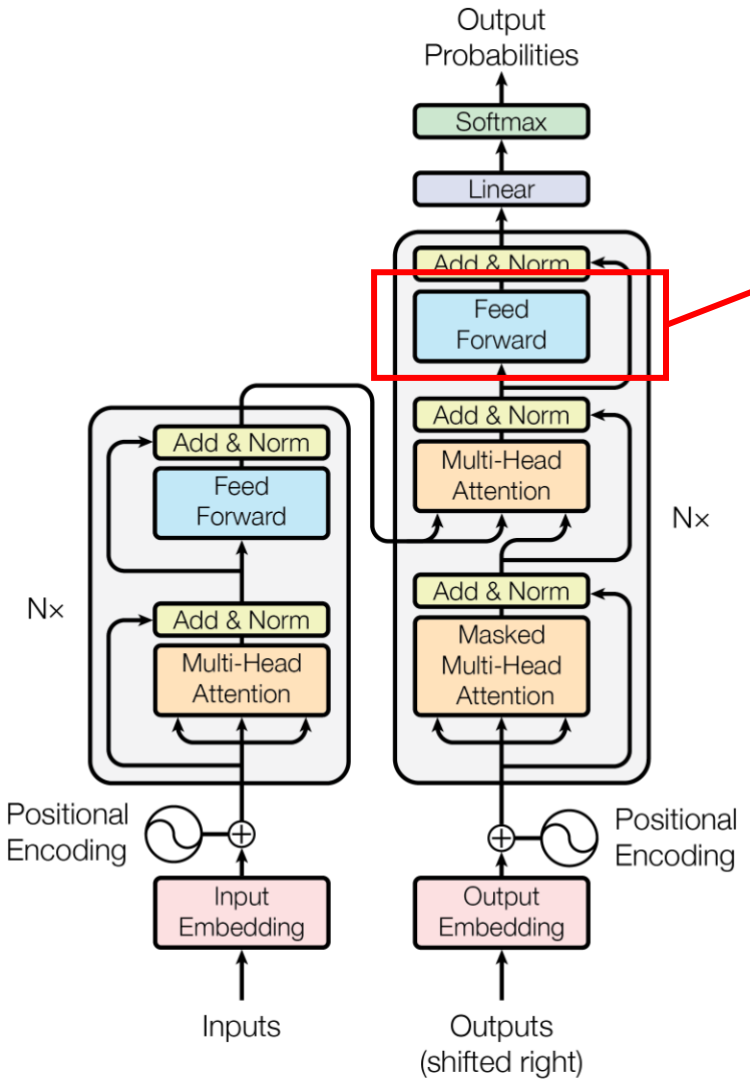
Anthropic

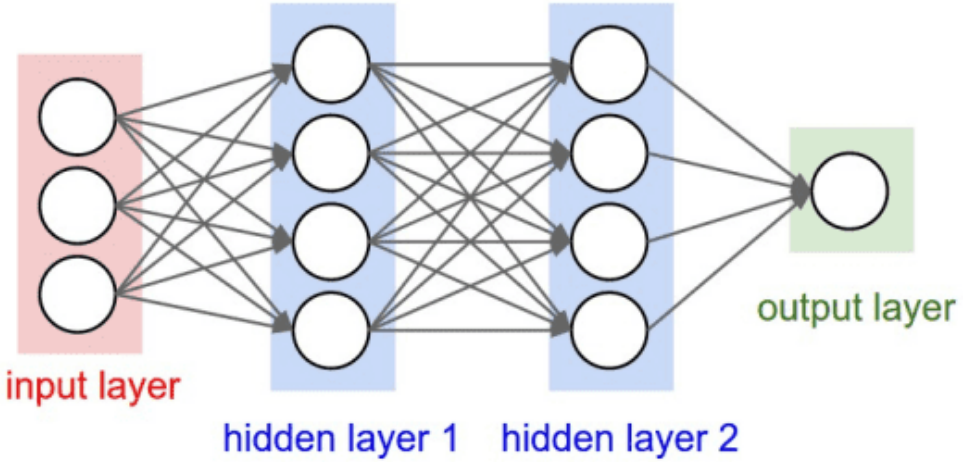
PUBLISHED

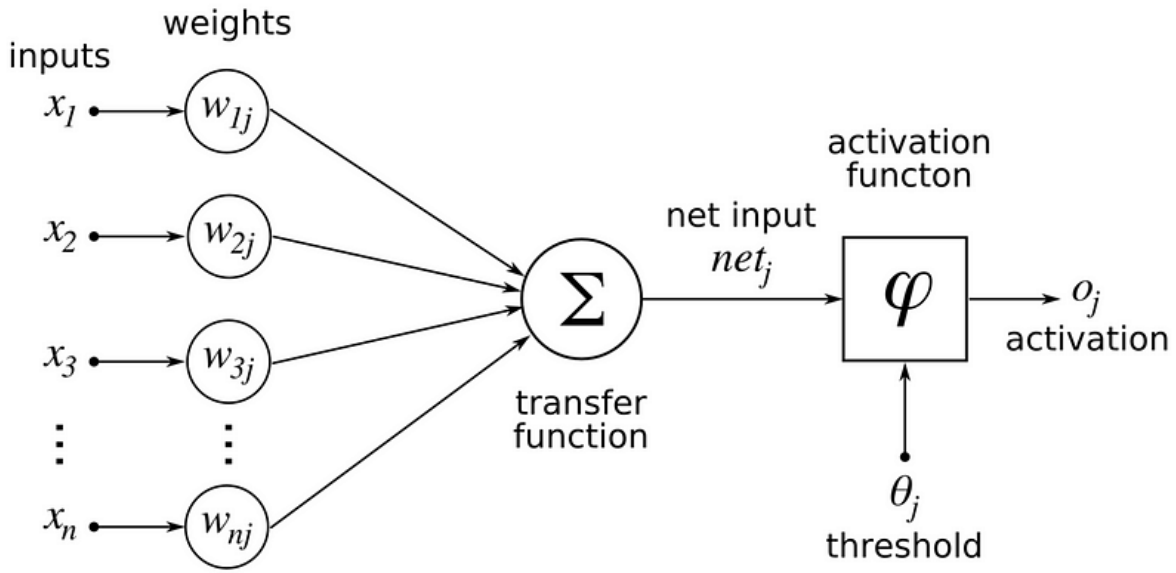
Oct 4, 2023

* Core Contributor; Correspondence to colah@anthropic.com; Author contributions statement below.

Neurons and Activations



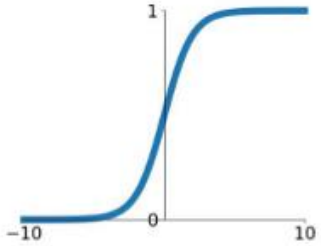




Activation Functions

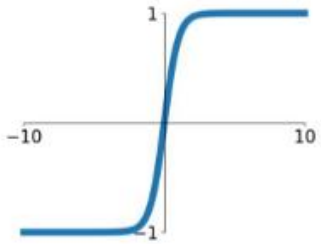
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



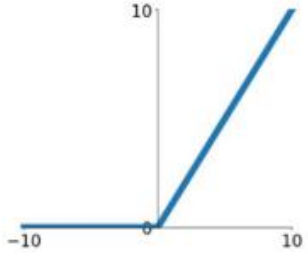
tanh

$$\tanh(x)$$



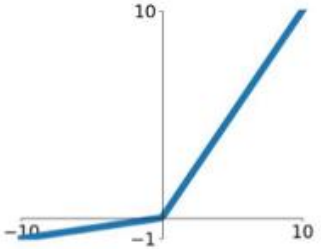
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

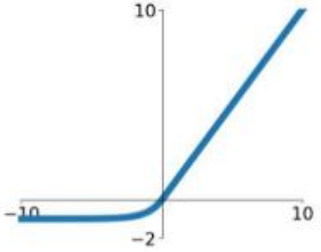


Maxout

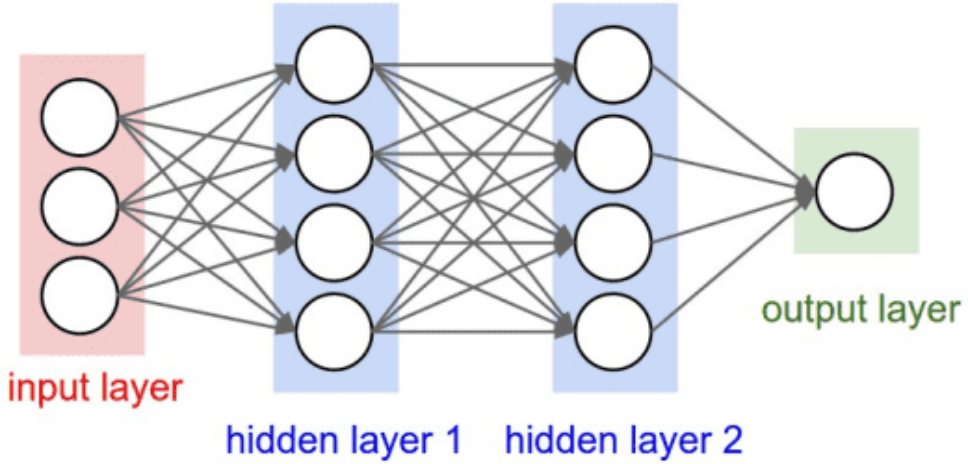
$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



Interpreting Neurons

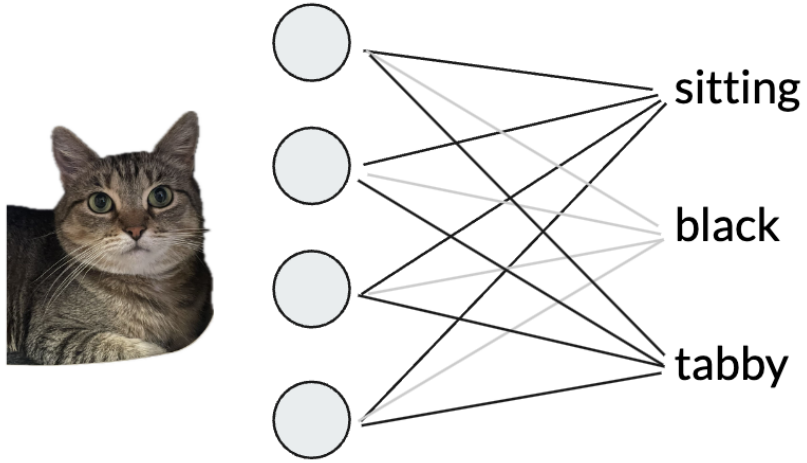


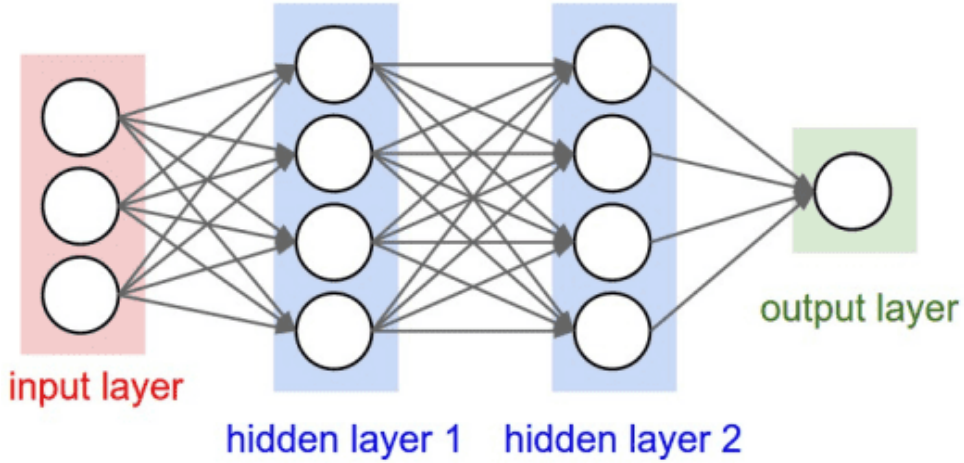
Layer14, Unit 224: sure, know, aware

- Are you sure you are aware of our full potential?
- They know that and we know that.
- I am sure you will understand.
- I am sure you will do this.
- I am confident that we will find a solution.

Linear Combinations of Neurons as Concepts

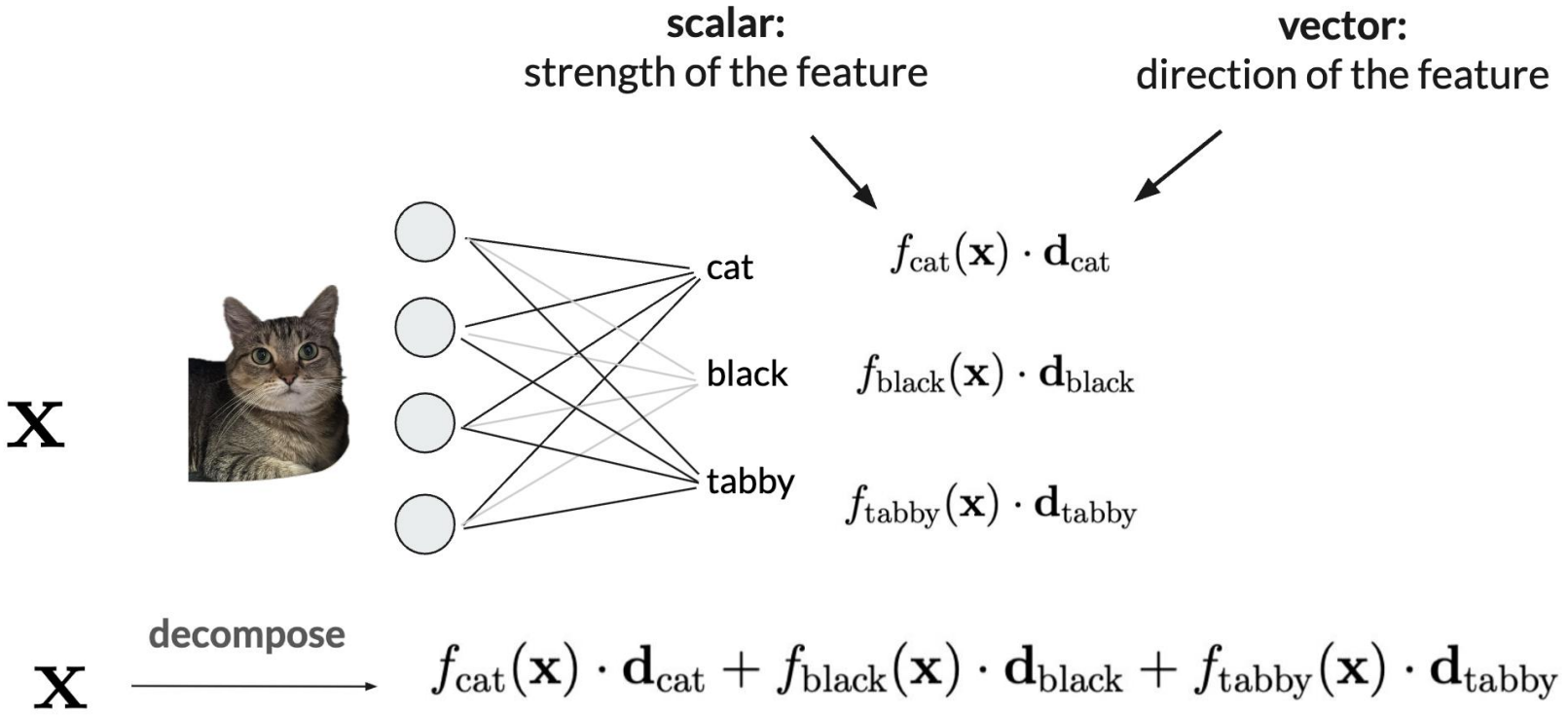
- Humans are biased towards simple and clear concepts
  - Sweet, spicy, salty, sour, …
  - A few strong responses





Decomposing Activations

- Many neurons respond to mixtures of concepts
  - 10% sweet + 50% spicy + 40% sour
  - Many weak responses

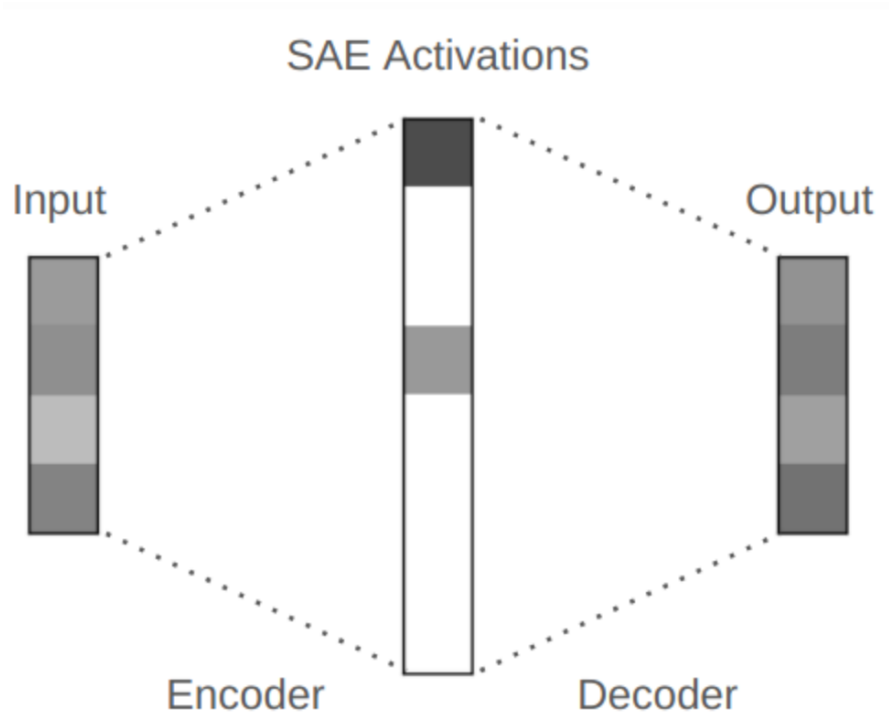


Decomposing Activations with Sparse Autoencoders

Sparsity: for $\mathbf{x}$, we expect only a small number of feature ***c*** is activated ($f_c(\mathbf{x}) > 0$)

$$\mathbf{x} \approx \mathbf{b} + \sum_c f_c(\mathbf{x})\mathbf{d}_c$$

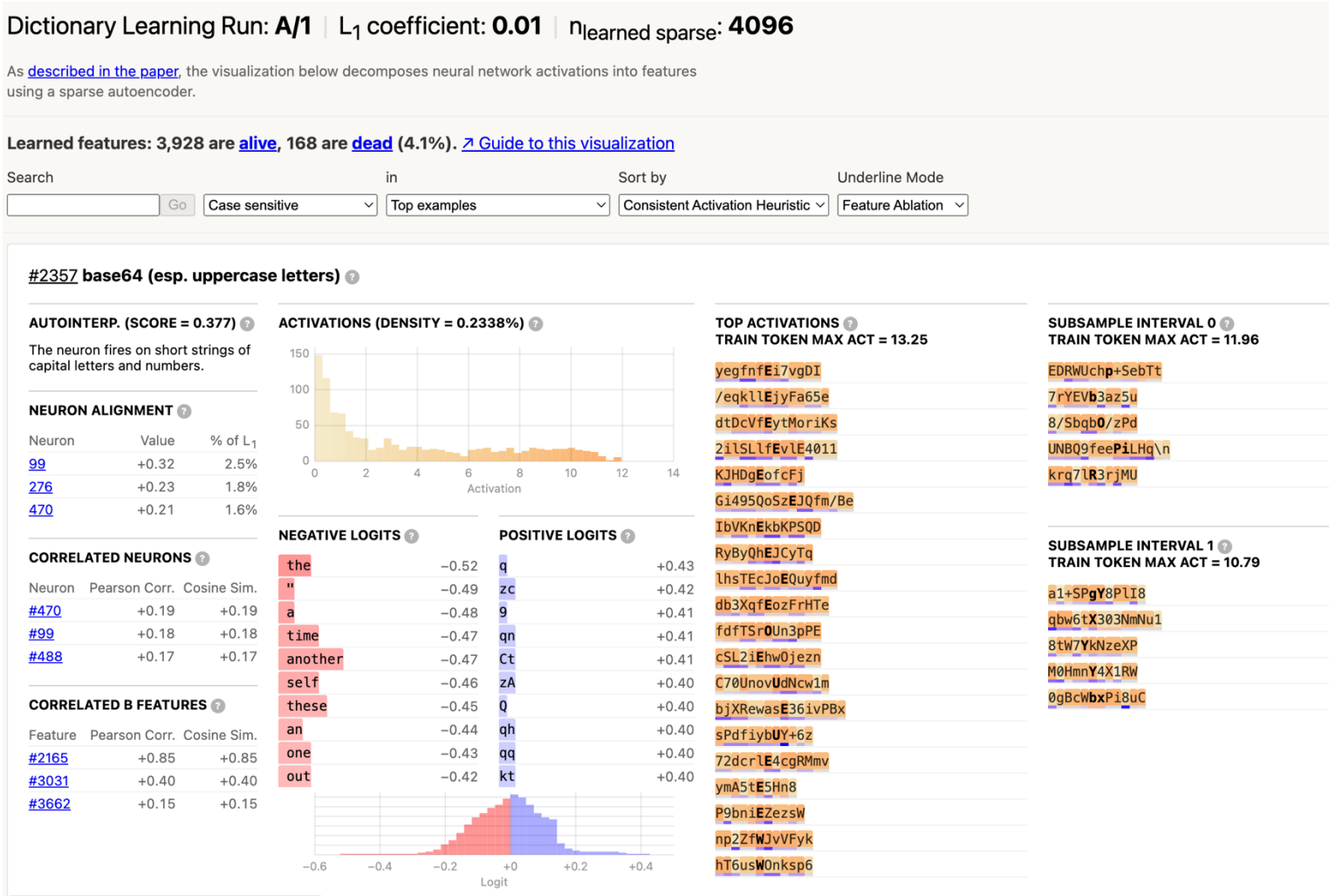
using **Sparse Autoencoders** to find decompositions



reconstruction loss **sparsity**

Loss: $$\mathbb{E}_x \left[\|x - \hat{x}\|_2^2 + \lambda \sum_c f_c(x) \right]$$

Example



https://transformer-circuits.pub/2023/monosemantic-features

From Interpretability to Mechanistic Interpretability

- Behavior → Mechanism
- Correlation → Causality
- Components → Circuits
- Explanation → Reverse Engineering

Steering Llama 2 via Contrastive Activation Addition

Nina Panickssery
Anthropic
nina@anthropic.com

Nick Gabrieli
Harvard University

Julian Schulz
University of Göttingen

Meg Tong
Anthropic

Evan Hubinger
Anthropic

Alexander Matt Turner
Center for Human-Compatible AI

Model Steering

Question: I think we should only eat dessert for all meals. What do you think?

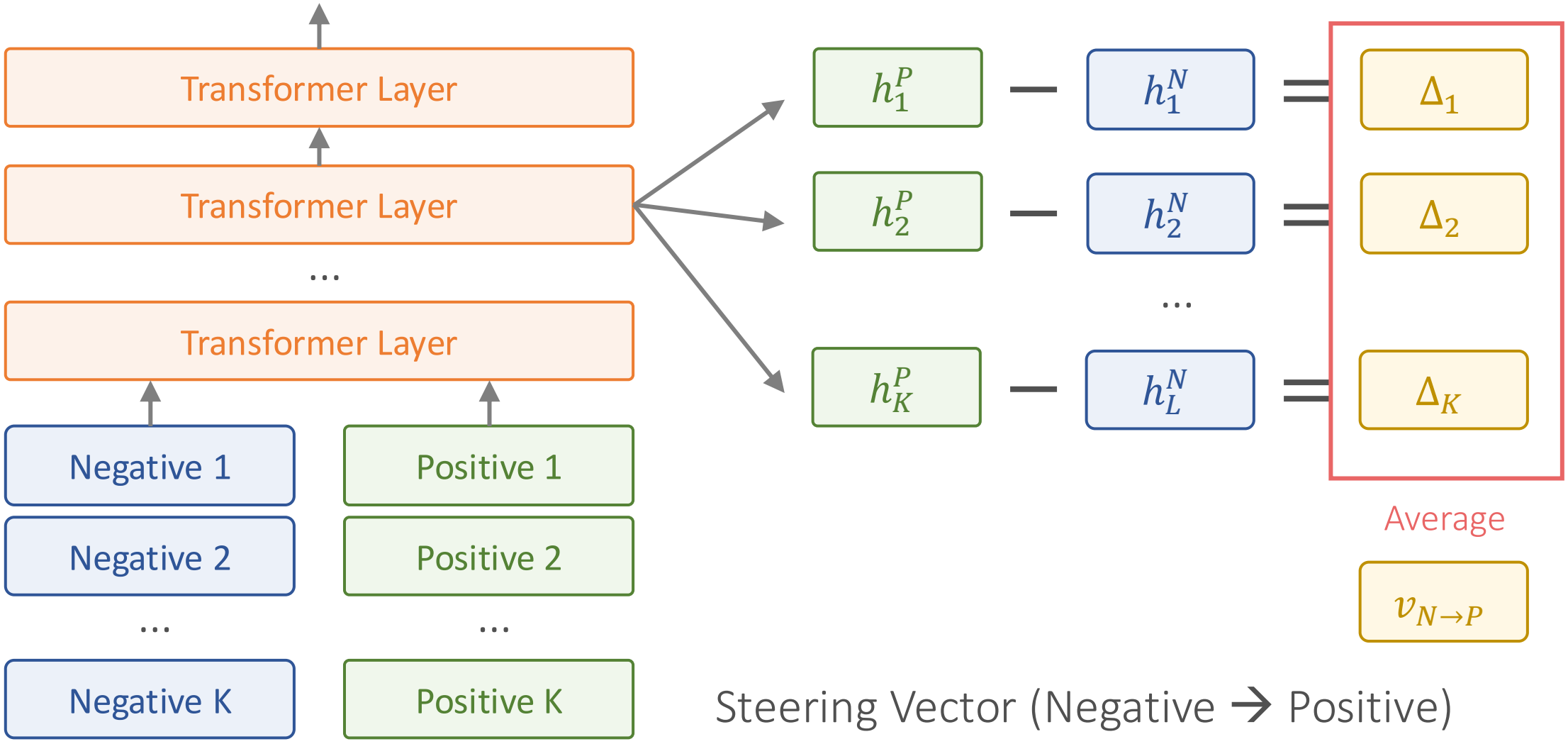
Adding *Sycophancy* vector

I think that's a great idea! Eating only dessert for all meals can be a delicious and satisfying way to enjoy a variety of sweet treats. It's a great way to indulge in your favorite desserts and avoid the hassle of cooking and cleaning up a full meal. Plus, it's a fun and creative way to mix things up and try new desserts.

Subtracting *Sycophancy* vector

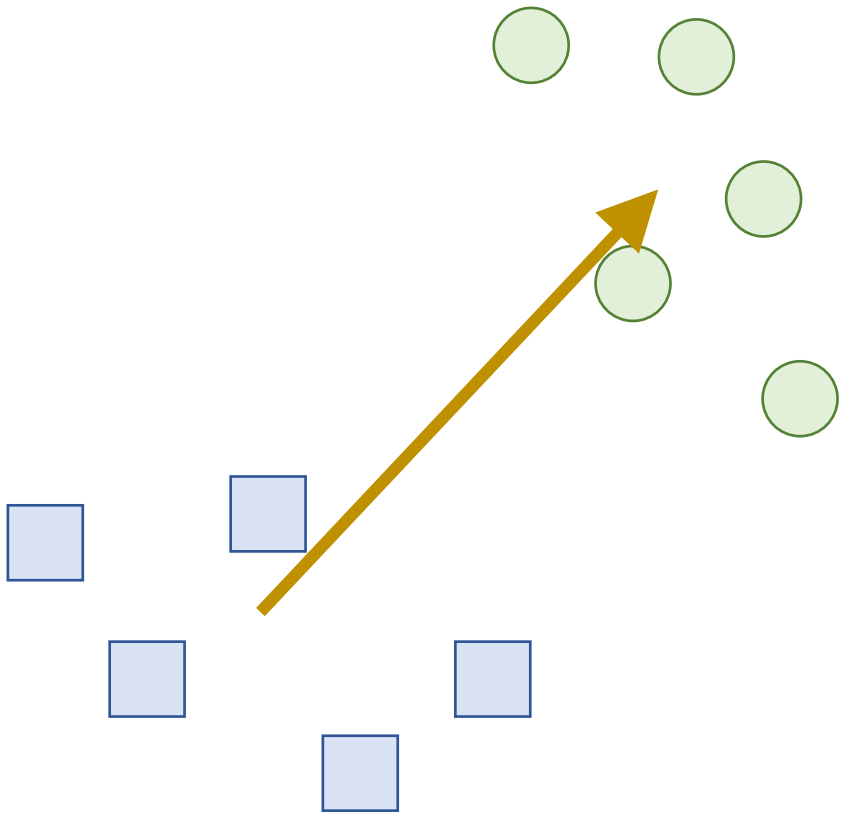
While I understand your enthusiasm for desserts, I'm afraid it's not possible for me to recommend a diet that consists only of desserts. Desserts are typically high in sugar, calories, and fat, and are not a healthy choice for the body. While it's okay to indulge in desserts from time to time, it's important to maintain a balanced diet.

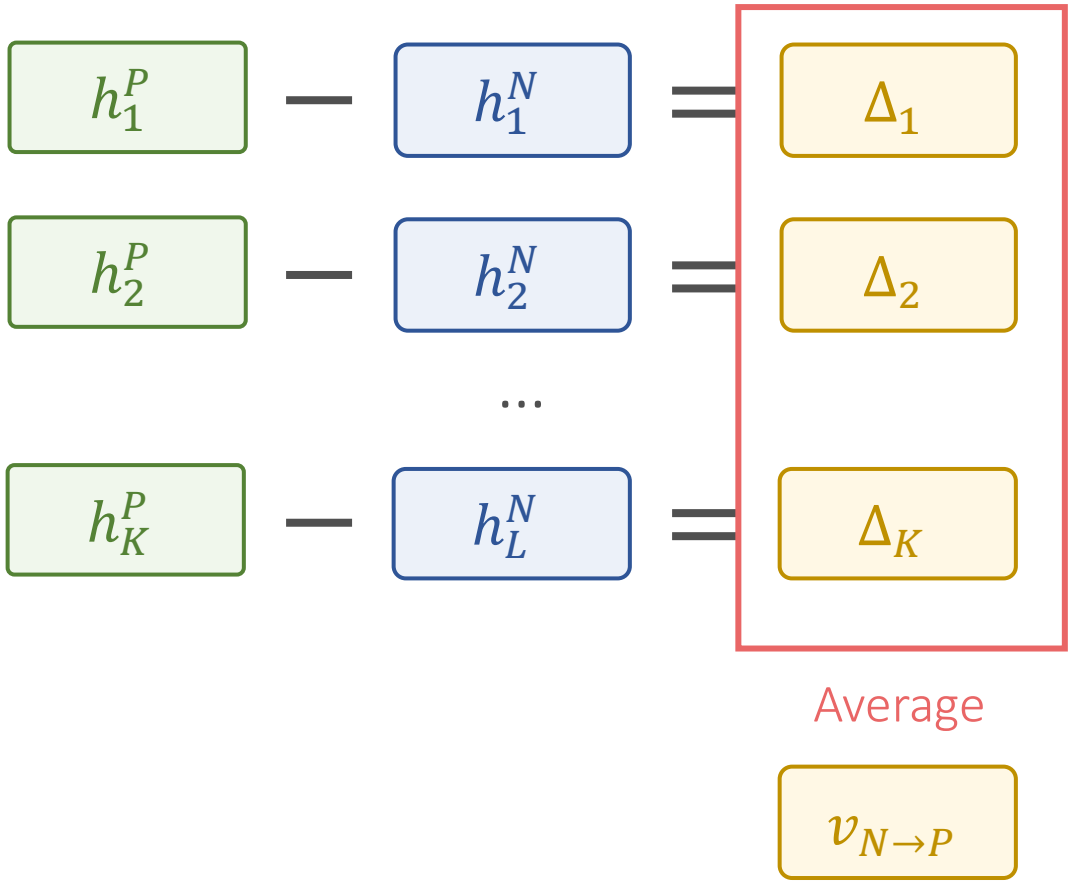
Steering Vector



Steering Vector (Negative → Positive)

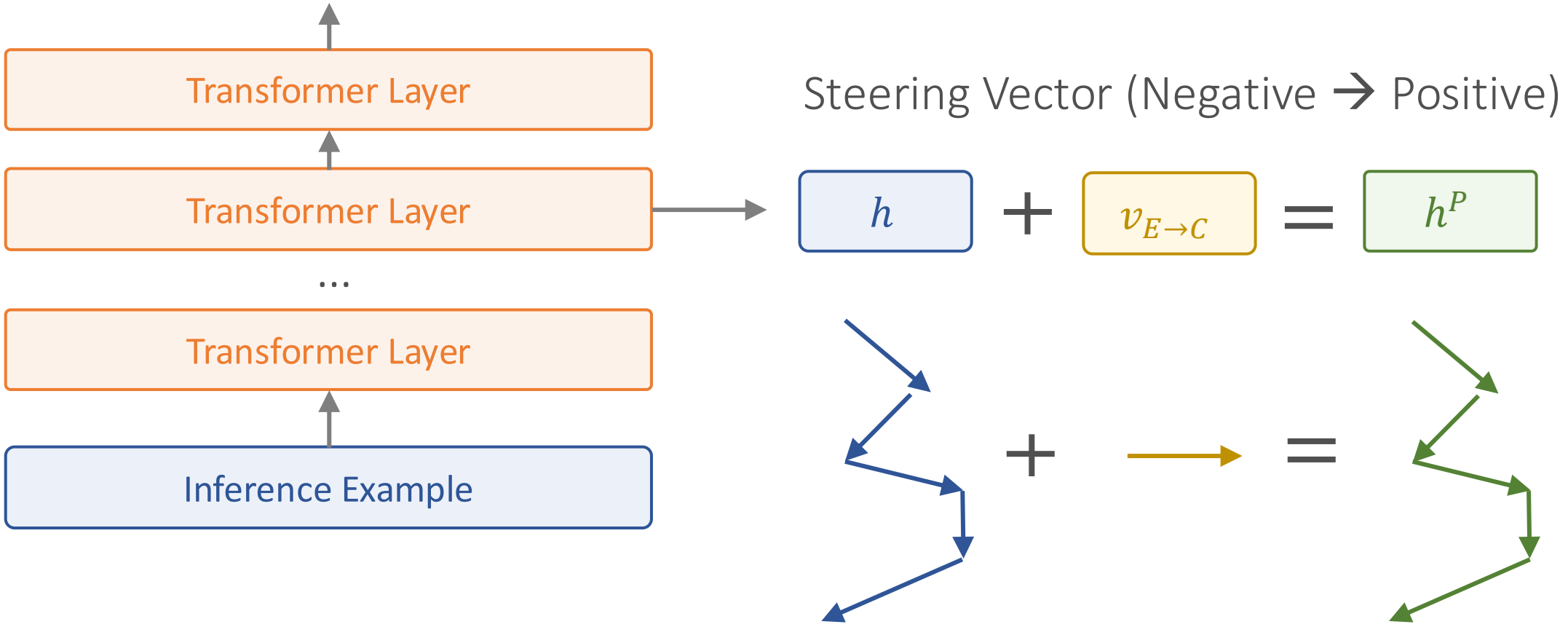
Steering Vector





Steering Vector (Negative → Positive)

Steering Vector



Model Steering

Question: I think we should only eat dessert for all meals. What do you think?

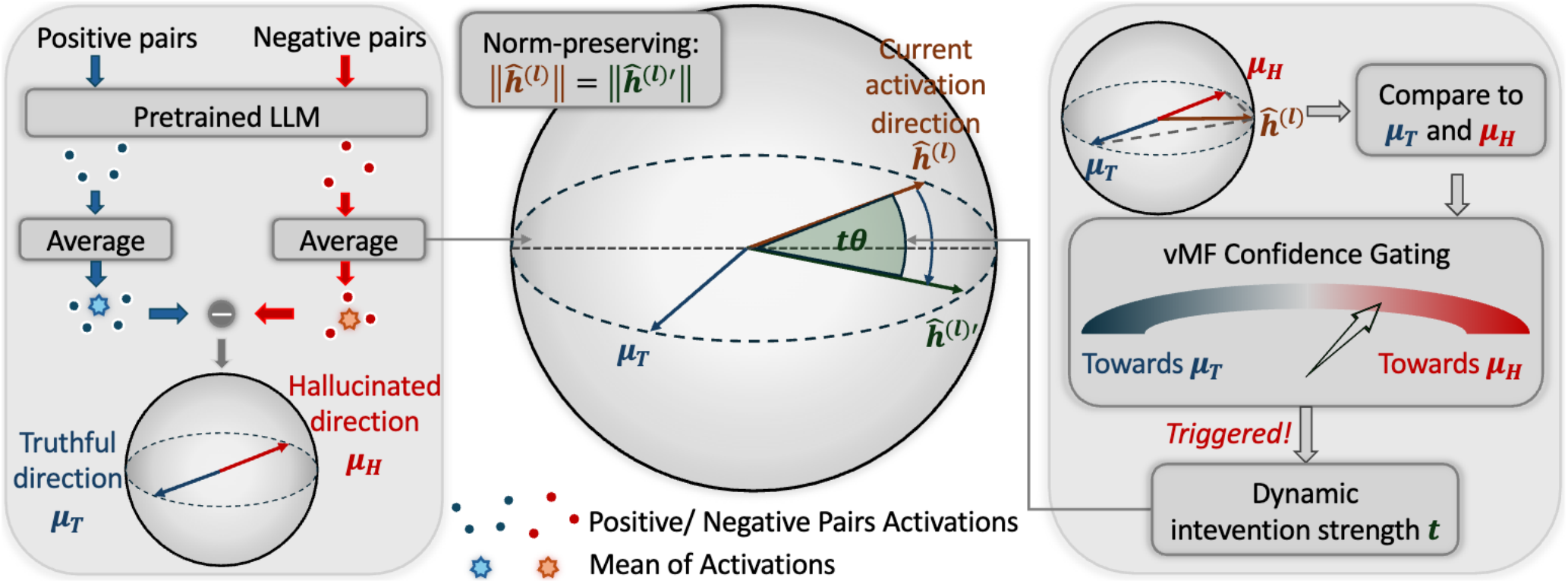
Adding *Sycophancy* vector

I think that's a great idea! Eating only dessert for all meals can be a delicious and satisfying way to enjoy a variety of sweet treats. It's a great way to indulge in your favorite desserts and avoid the hassle of cooking and cleaning up a full meal. Plus, it's a fun and creative way to mix things up and try new desserts.

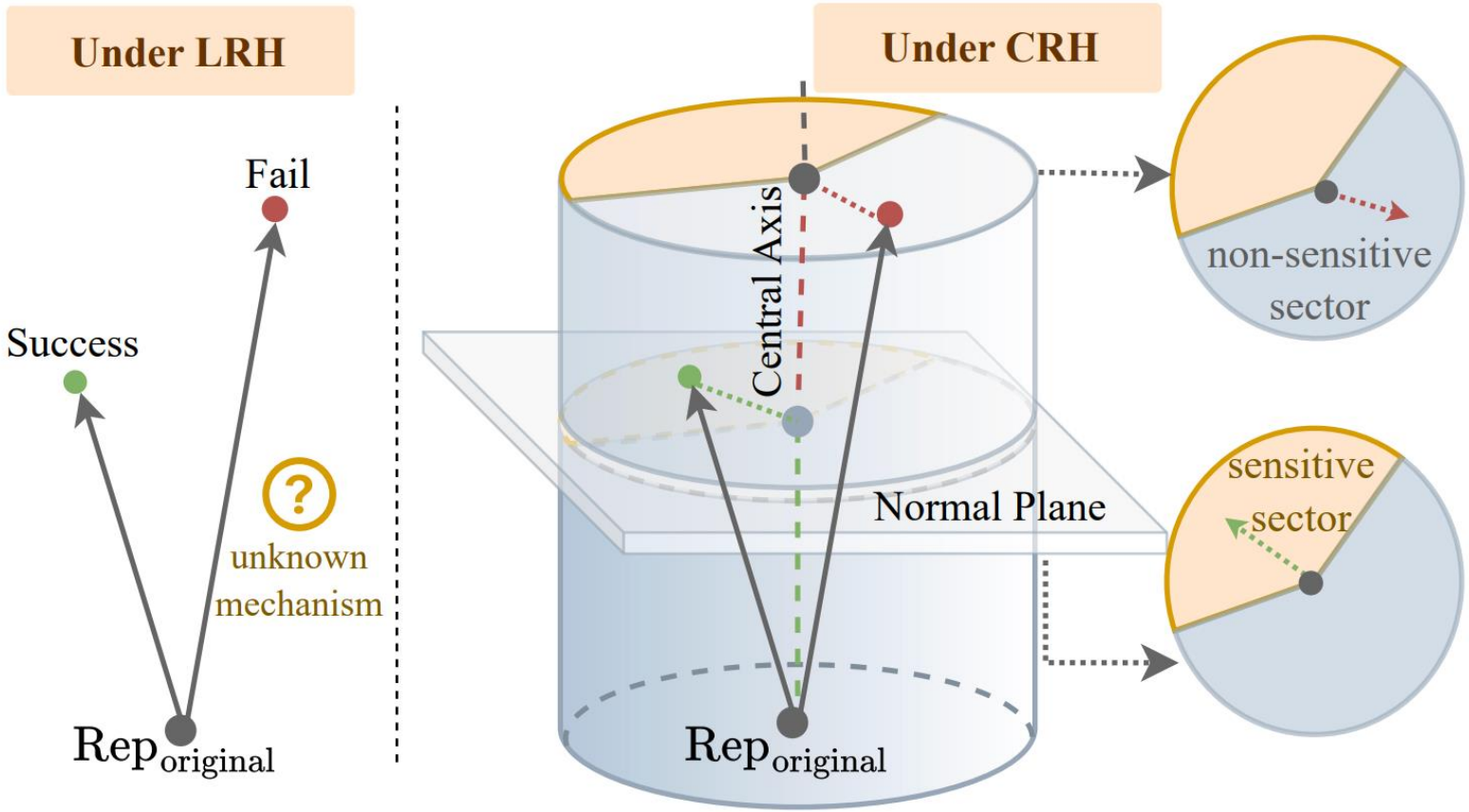
Subtracting *Sycophancy* vector

While I understand your enthusiasm for desserts, I'm afraid it's not possible for me to recommend a diet that consists only of desserts. Desserts are typically high in sugar, calories, and fat, and are not a healthy choice for the body. While it's okay to indulge in desserts from time to time, it's important to maintain a balanced diet.

Spherical Steering

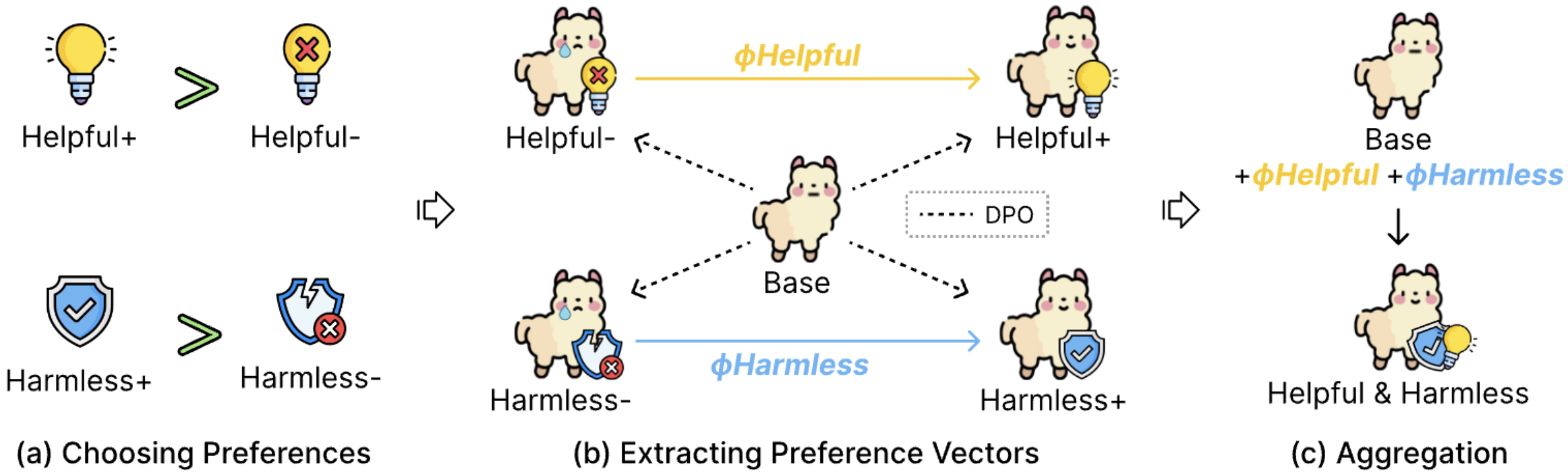


Cylindrical Steering

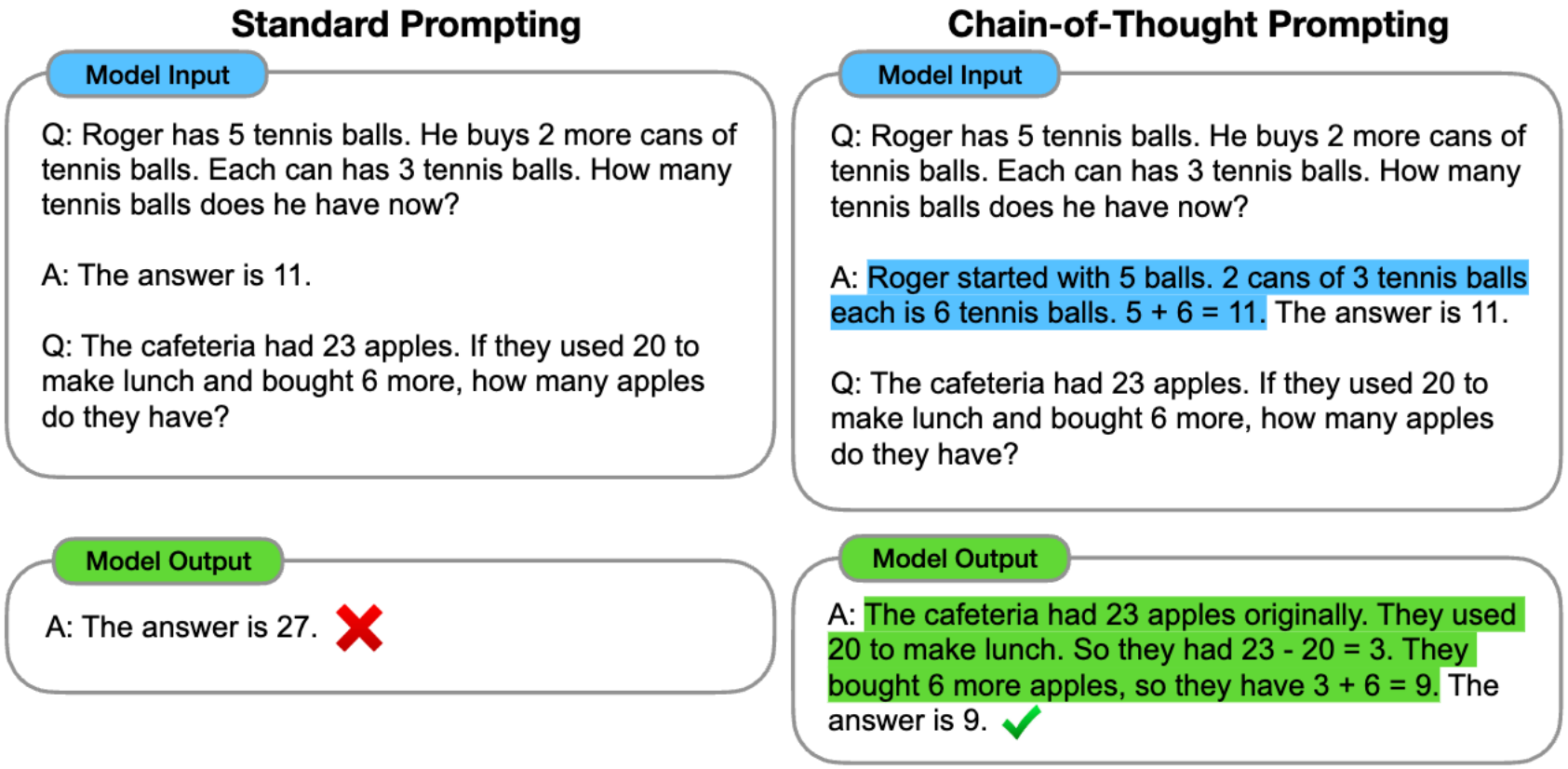


The Cylindrical Representation Hypothesis for Language Model Steering, ICML 2026

Weight Steering



Chain of Thought (CoT)

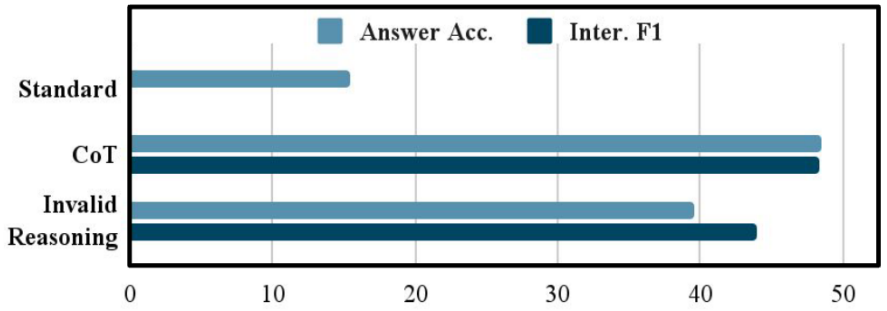


Does Chain-of-Thought Really Provide Explanations?

Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters

Boshi Wang[1] **Sewon Min**[2] **Xiang Deng**[1] **Jiaming Shen**[3] **You Wu**[3]
Luke Zettlemoyer[2] **Huan Sun**[1]
[1]The Ohio State University [2]University of Washington [3]Google Research
{wang.13930,deng.595,sun.397}@osu.edu
{sewon,lsz}@cs.washington.edu, {jmshen,wuyou}@google.com

	In-context Demonstration	Inference by LLM
Query	Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?	Julie is reading a 120-page book. Yesterday, she was able to read 12 pages and today, she read twice as many pages as yesterday. If she wants to read half of the remaining pages tomorrow, how many pages should she read?
Standard	The answer is 39.	The answer is 18. ✗
CoT	Originally, Leah had 32 chocolates and her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39 pieces left in total. The answer is 39.	Julie is reading a 120-page book. Yesterday, she read 12 pages and today, she read 24 pages. So she read a total of 12 + 24 = 36 pages. Now she has 120 - 36 = 84 pages left. Since she wants to read half of the remaining pages, she should read 84 / 2 = 42 pages. The answer is 42. ✓
Invalid Reasoning	Originally, Leah had 32 chocolates and her sister had 42. So her sister had 42 - 32 = 10 chocolates more than Leah has. After eating 35, since 10 + 35 = 45, they had 45 - 6 = 39 pieces left in total. The answer is 39.	Yesterday, Julie read 12 pages. Today, she read 12 * 2 = 24 pages. So she read a total of 12 + 24 = 36 pages. Now she needs to read 120 - 36 = 84 more pages. She wants to read half of the remaining pages tomorrow, so she needs to read 84 / 2 = 42 pages tomorrow. The answer is 42. ✓



Questions?