1878

Western

UNIVERSITY · CANADA

Artificial "Directed Musement" and Its Limitations

May 25, 2024

Anemily Machina

Demystifying Generative Text AIs

The training of AIs itself can be thought of as similar to Schiller’s **sense drive (making a prediction and receiving a positive or negative signal)** and **form drive (updating the internal world based on the signal)** (O’Connor, 2014), but this never harmonizes into play.

SimpleGPT

SimpleGPT

[4.0, 1.2, 2.1]

SimpleGPT

[4.0, 1.2, 2.1]

Input	?
Multiply by 4.0	?
Add 1.2	?
Divide by 2.1	?

SimpleGPT

[4.0, 1.2, 2.1]

“I saw a”

Input	?
Multiply by 4.0	?
Add 1.2	?
Divide by 2.1	?

SimpleGPT

[4.0, 1.2, 2.1]

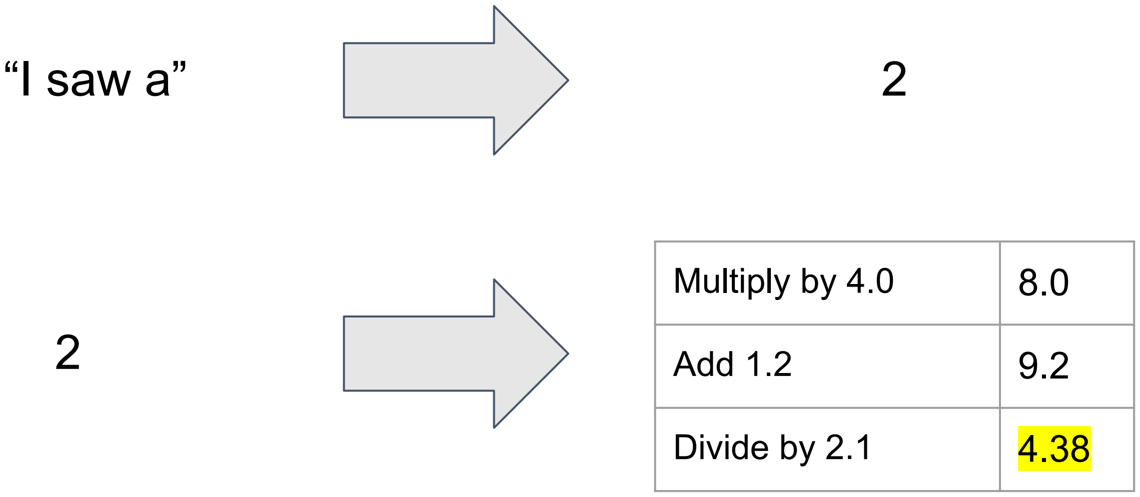
“I saw a” → 2

Input	?
Multiply by 4.0	?
Add 1.2	?
Divide by 2.1	?

SimpleGPT

[4.0, 1.2, 2.1]

Input	?
Multiply by 4.0	?
Add 1.2	?
Divide by 2.1	?



“I saw a” → 2

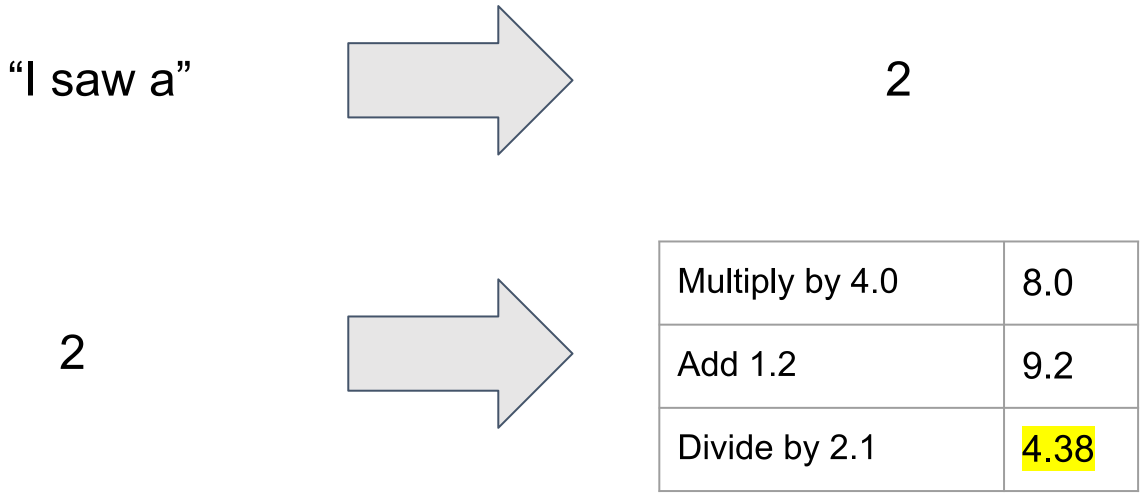
2 →

Multiply by 4.0	8.0
Add 1.2	9.2
Divide by 2.1	4.38

SimpleGPT

[4.0, 1.2, 2.1]

Input	?
Multiply by 4.0	?
Add 1.2	?
Divide by 2.1	?



“I saw a” → 2

2 →

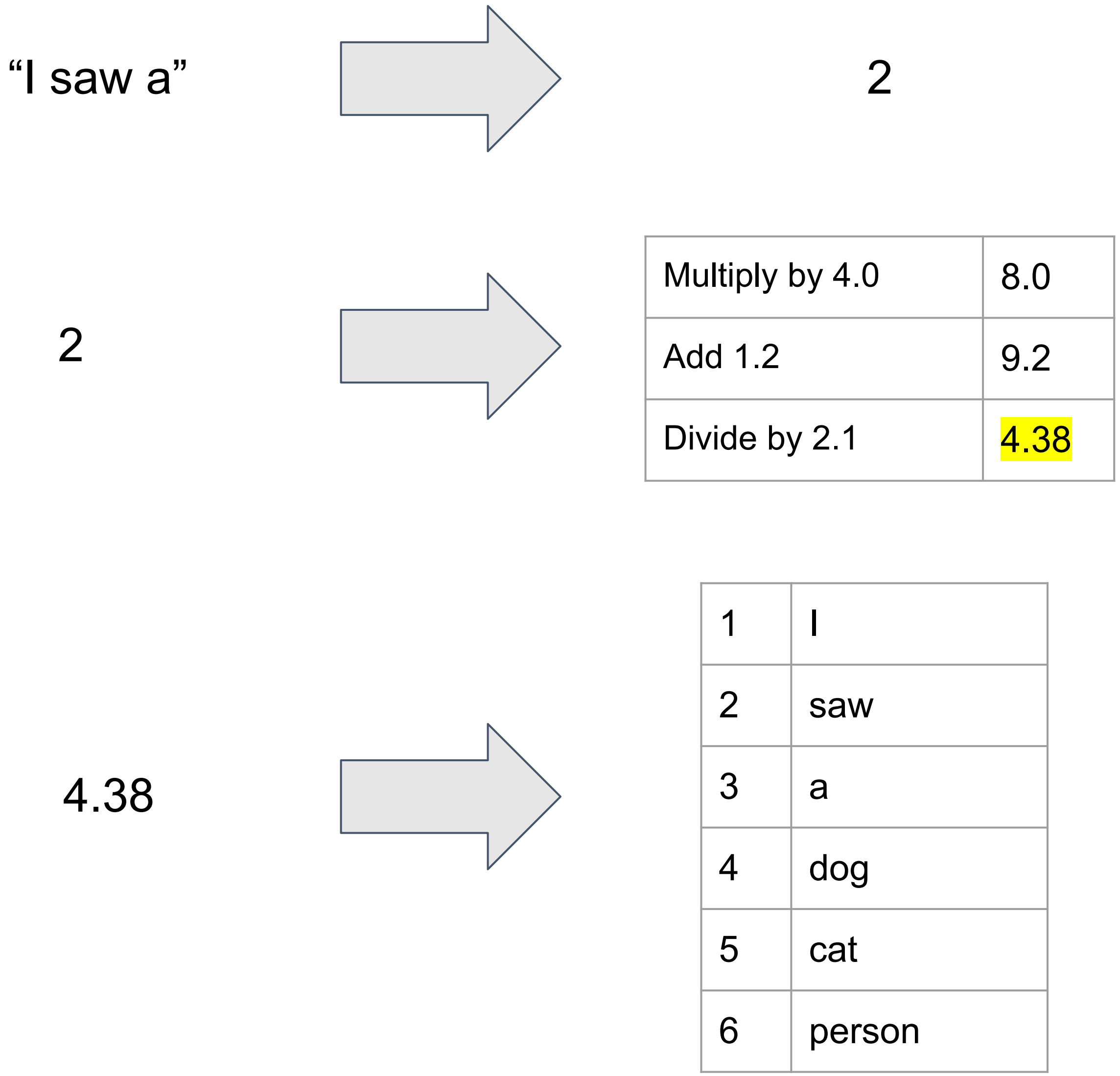
Multiply by 4.0	8.0
Add 1.2	9.2
Divide by 2.1	4.38

4.38

SimpleGPT

[4.0, 1.2, 2.1]

Input	?
Multiply by 4.0	?
Add 1.2	?
Divide by 2.1	?



“I saw a” → 2

2 →

Multiply by 4.0	8.0
Add 1.2	9.2
Divide by 2.1	4.38

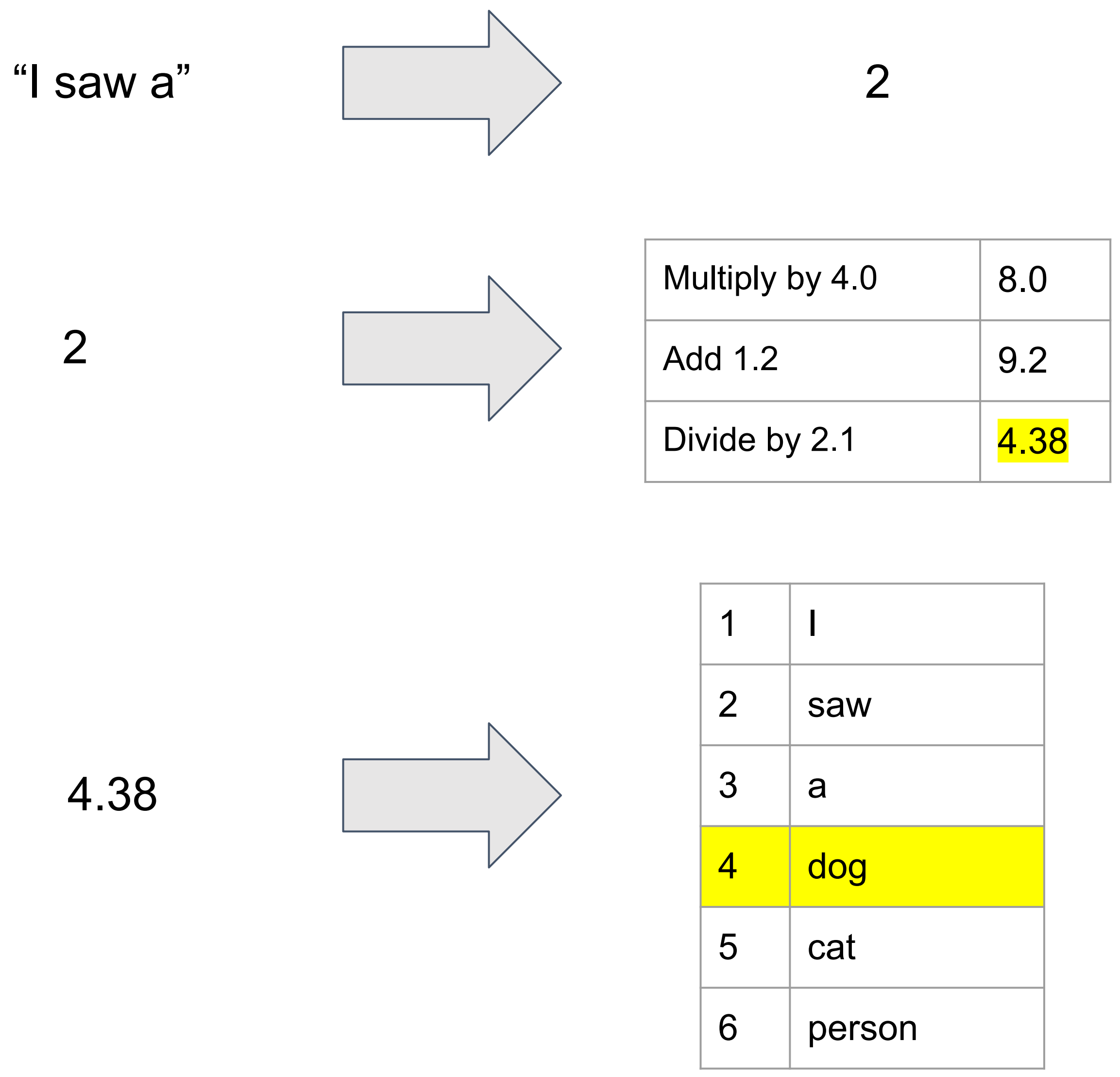
4.38 →

1	I
2	saw
3	a
4	dog
5	cat
6	person

SimpleGPT

[4.0, 1.2, 2.1]

Input	?
Multiply by 4.0	?
Add 1.2	?
Divide by 2.1	?



“I saw a” → 2

2 →

Multiply by 4.0	8.0
Add 1.2	9.2
Divide by 2.1	4.38

4.38 →

1	I
2	saw
3	a
4	dog
5	cat
6	person

SimpleGPT - After Training

SimpleGPT - After Training

Training Data
I saw a dog
I saw a cat
I saw a person

SimpleGPT - After Training

Training Data
I saw a dog
I saw a cat
I saw a person

Sentence	Generation Frequency
I saw a dog	33%
I saw a cat	33%
I saw a person	33%

SimpleGPT - After Training

Training Data
I saw a dog
I saw a cat
I saw a person

Training Data
I saw a dog
I saw a cat
I saw a person
I saw a dog

Sentence	Generation Frequency
I saw a dog	33%
I saw a cat	33%
I saw a person	33%

SimpleGPT - After Training

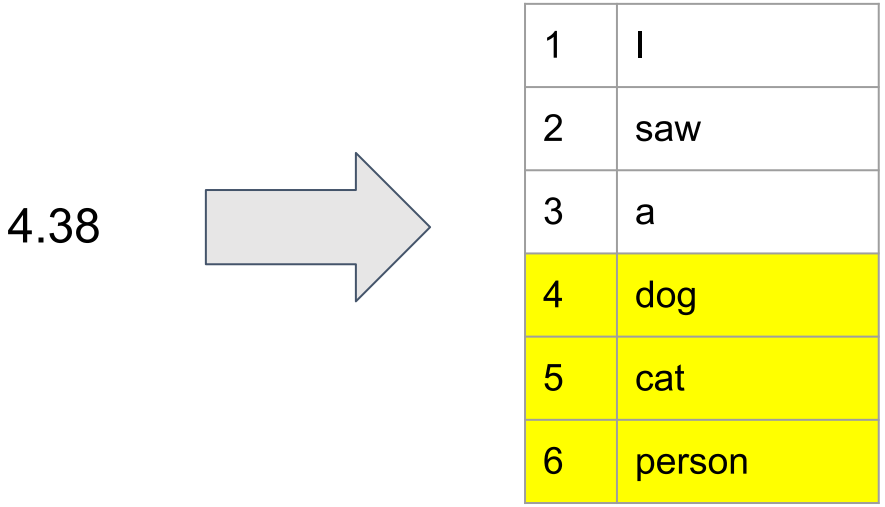
Training Data
I saw a dog
I saw a cat
I saw a person

Training Data
I saw a dog
I saw a cat
I saw a person
I saw a dog

Sentence	Generation Frequency
I saw a dog	33%
I saw a cat	33%
I saw a person	33%

Sentence	Generation Frequency
I saw a dog	100%
I saw a cat	0%
I saw a person	0%

SimpleGPT - Change Prediction Instructions



4.38

1	I
2	saw
3	a
4	dog
5	cat
6	person

Training Data
I saw a dog
I saw a cat
I saw a person
I saw a dog

Sentence	Generation Frequency
I saw a dog	50%
I saw a cat	25%
I saw a person	25%

Interesting Things do Happen

Interesting Things do Happen

- The next token after “I tasted the most delicious …” must all be close to each other

Interesting Things do Happen

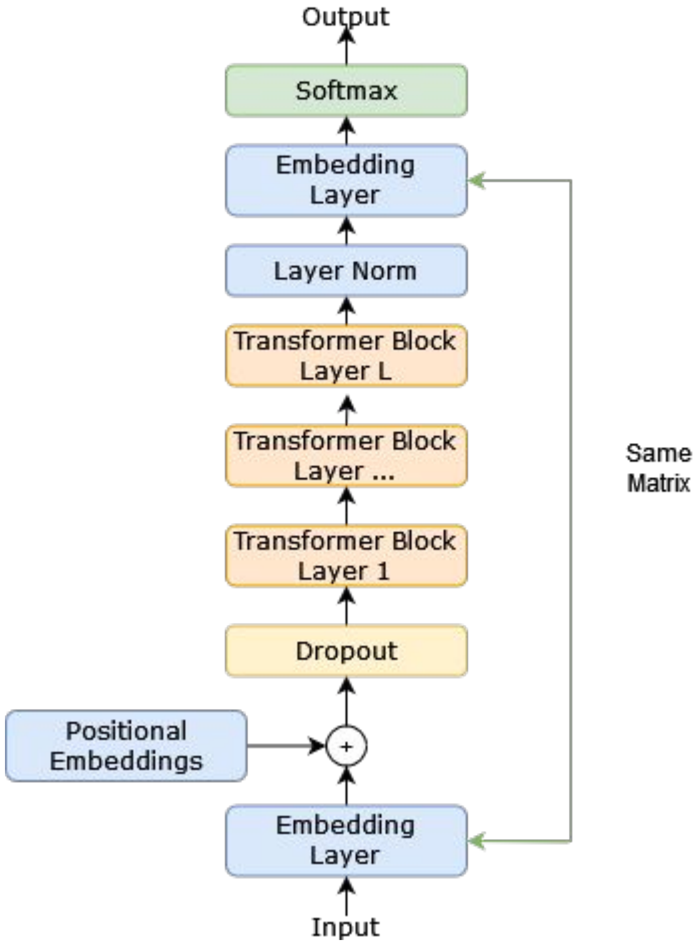
- The next token after “I tasted the most delicious …” must all be close to each other

…	…
112	sushi
113	curry
114	pizza
115	cake
116	nachos
…	..

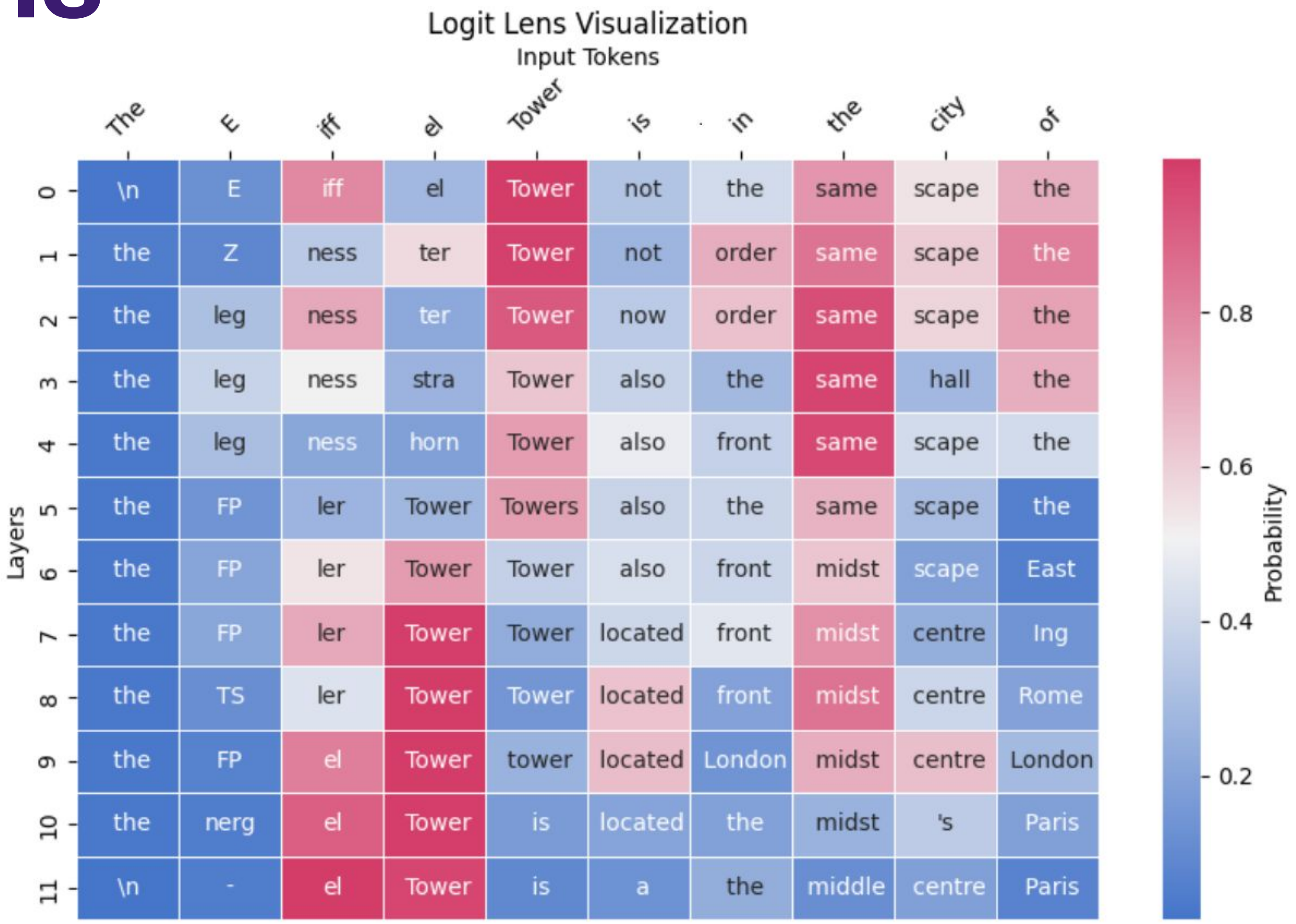
Artificial Musement

Intermediate Outputs

- An AI’s next-token decision can be broken down into a series of steps
- Each step has its own (complex) output that can be evaluated
- Only the last output is constrained to predict the next-token



Logit Lens



Logit Lens Visualization

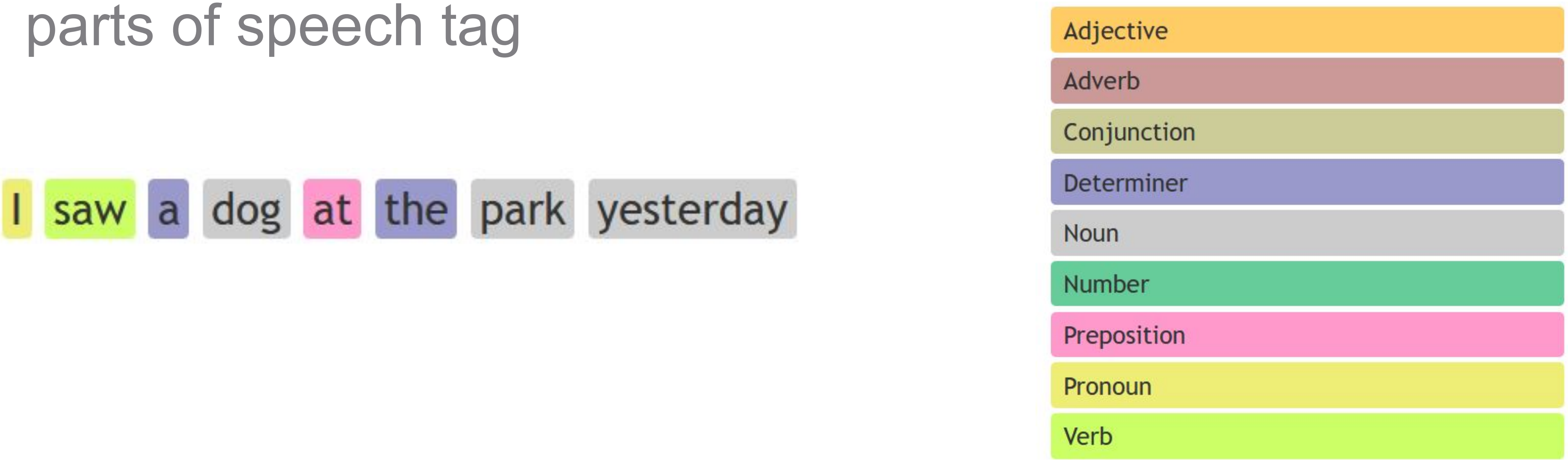
Layers	The	E	iff	el	Tower	is	in	the	city	of
0	\n	E	iff	el	Tower	not	the	same	scape	the
1	the	Z	ness	ter	Tower	not	order	same	scape	the
2	the	leg	ness	ter	Tower	now	order	same	scape	the
3	the	leg	ness	stra	Tower	also	the	same	hall	the
4	the	leg	ness	horn	Tower	also	front	same	scape	the
5	the	FP	ler	Tower	Towers	also	the	same	scape	the
6	the	FP	ler	Tower	Tower	also	front	midst	scape	East
7	the	FP	ler	Tower	Tower	located	front	midst	centre	Ing
8	the	TS	ler	Tower	Tower	located	front	midst	centre	Rome
9	the	FP	el	Tower	tower	located	London	midst	centre	London
10	the	nerg	el	Tower	is	located	the	midst	's	Paris
11	\n	-	el	Tower	is	a	the	middle	centre	Paris

https://nnsight.net/notebooks/tutorials/logit_lens/

Probing Tasks

- Try to train a different (simple) AI on the intermediate outputs of a complex AI ” (Shi et al., 2016)

- Simple AI is trained on tasks like: predict the parts of speech tag

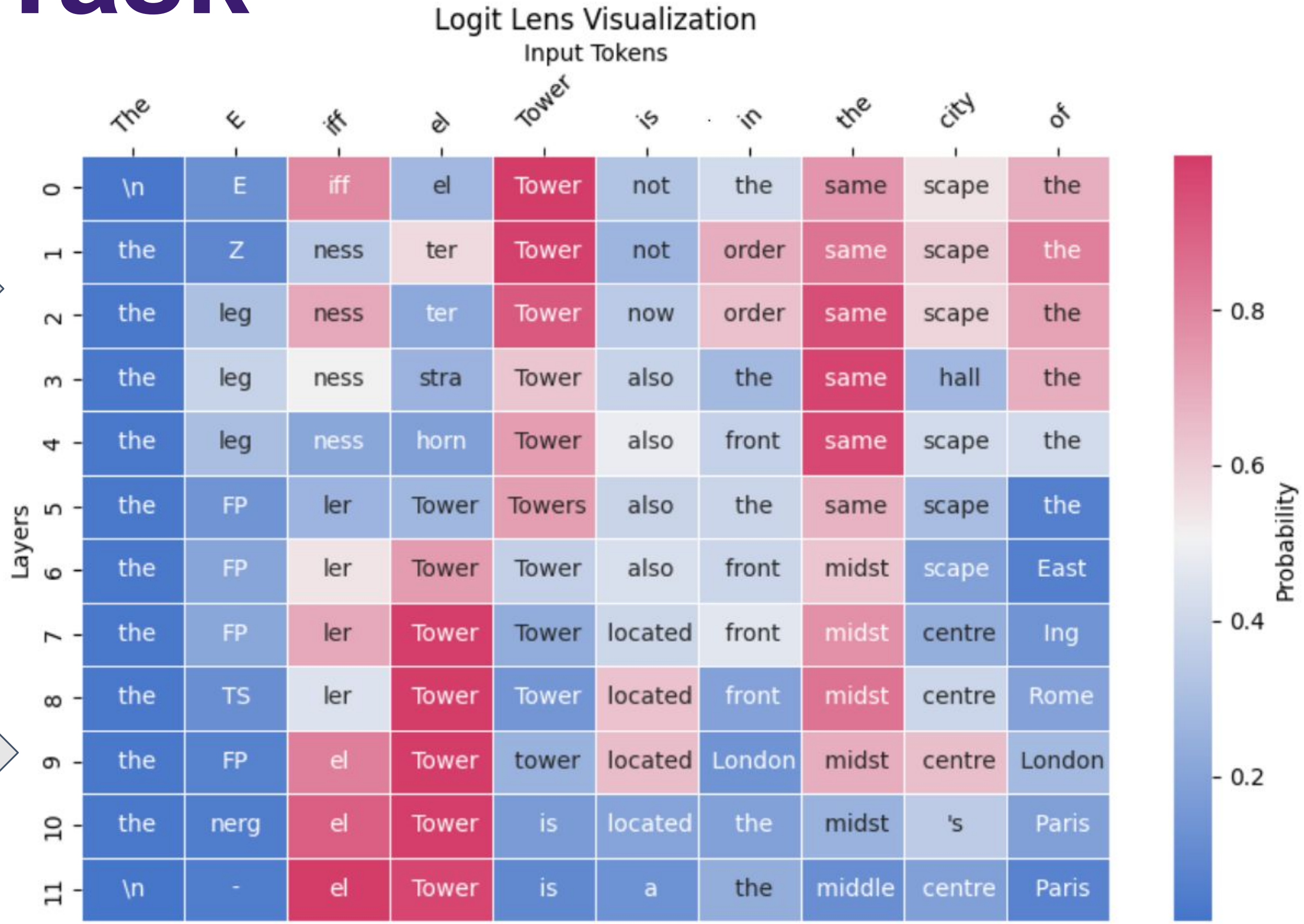


I saw a dog at the park yesterday

Probing Task

Parts of Speech Detected

Word Sense Detected



Logit Lens Visualization

Layers	The	E	iff	el	Tower	is	in	the	city	of
0	\n	E	iff	el	Tower	not	the	same	scape	the
1	the	Z	ness	ter	Tower	not	order	same	scape	the
2	the	leg	ness	ter	Tower	now	order	same	scape	the
3	the	leg	ness	stra	Tower	also	the	same	hall	the
4	the	leg	ness	horn	Tower	also	front	same	scape	the
5	the	FP	ler	Tower	Towers	also	the	same	scape	the
6	the	FP	ler	Tower	Tower	also	front	midst	scape	East
7	the	FP	ler	Tower	Tower	located	front	midst	centre	Ing
8	the	TS	ler	Tower	Tower	located	front	midst	centre	Rome
9	the	FP	el	Tower	tower	located	London	midst	centre	London
10	the	nerg	el	Tower	is	located	the	midst	's	Paris
11	\n	-	el	Tower	is	a	the	middle	centre	Paris

https://nnsight.net/notebooks/tutorials/logit_lens/

Artificial Thought

Concept Erasure

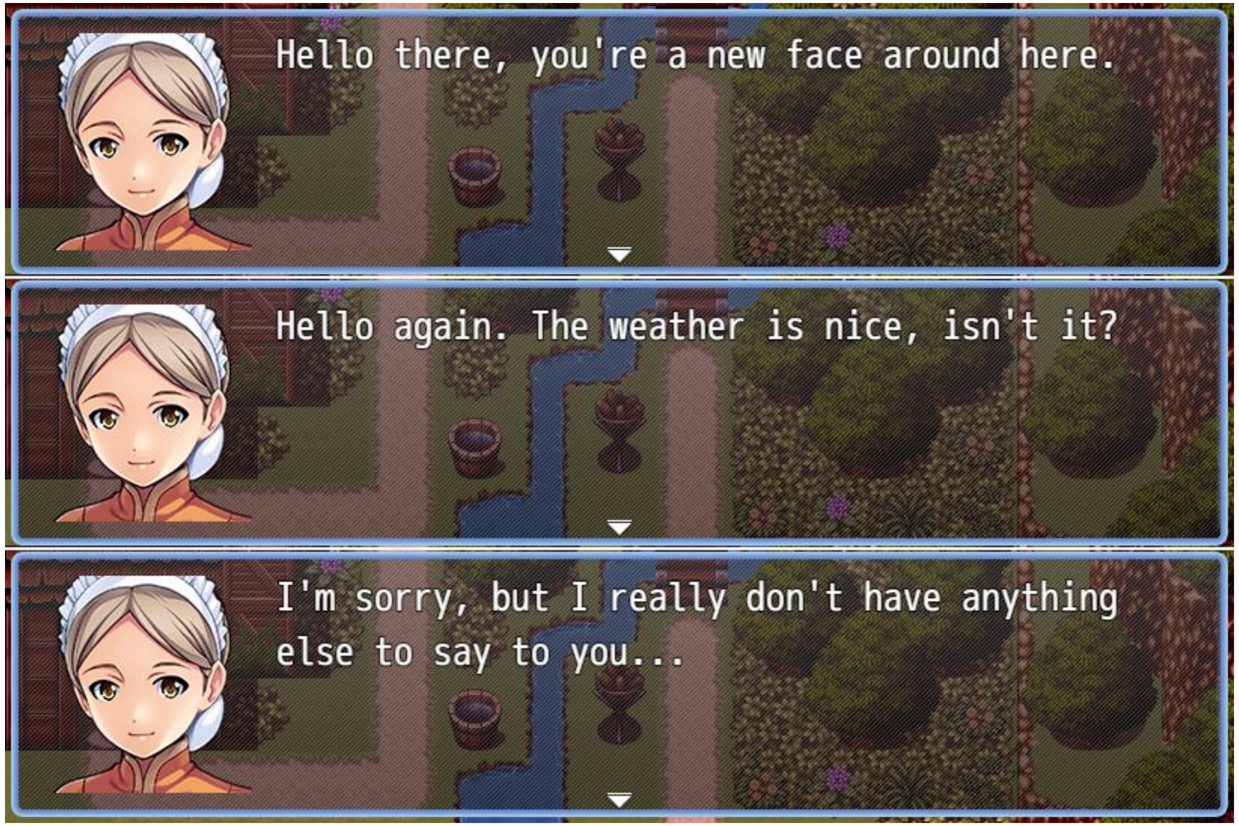
- You can provably erase any concept from a generative AI (Belrose et al., 2023)

- E.g., gender or freedom

Model Collapse

- If you train an AI on the outputs of another AI their is a loss of expressibility

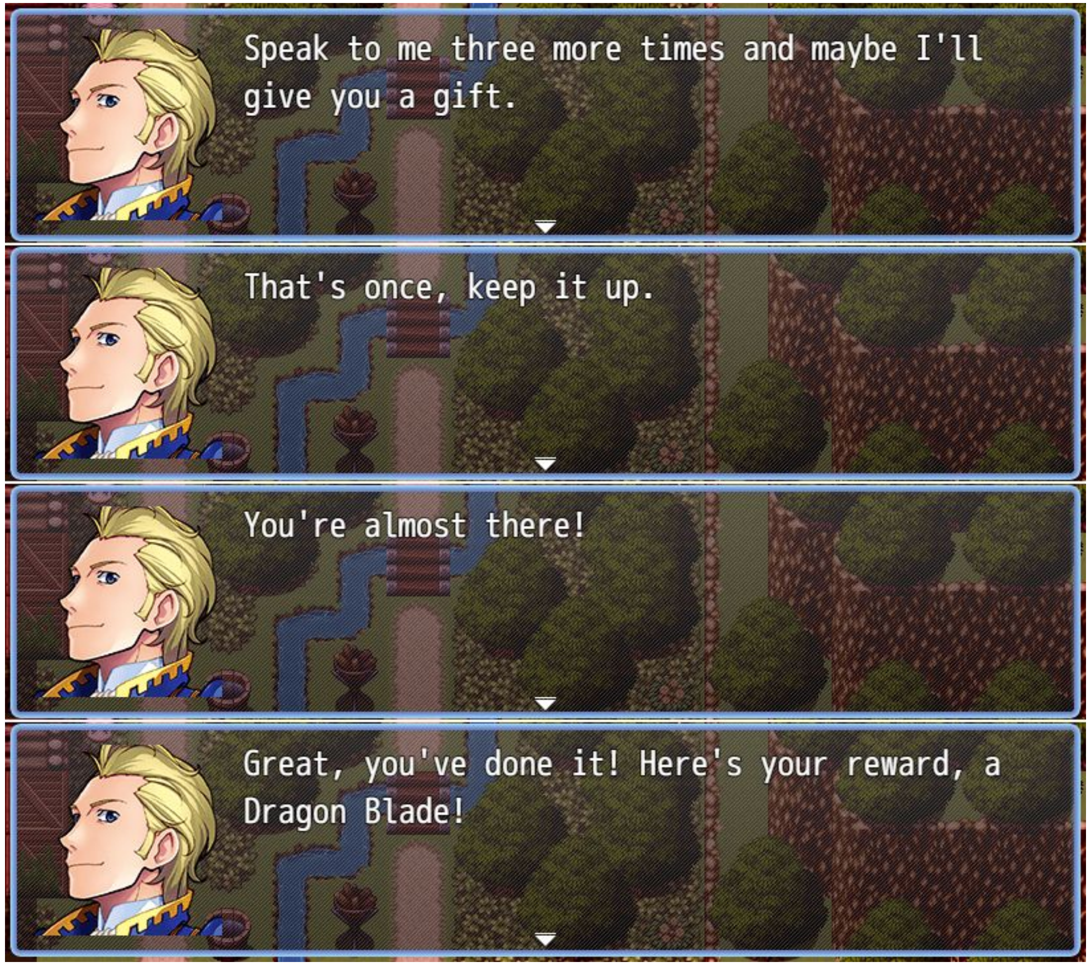
Model Collapse



Hello there, you're a new face around here.

Hello again. The weather is nice, isn't it?

I'm sorry, but I really don't have anything else to say to you...



Speak to me three more times and maybe I'll give you a gift.

That's once, keep it up.

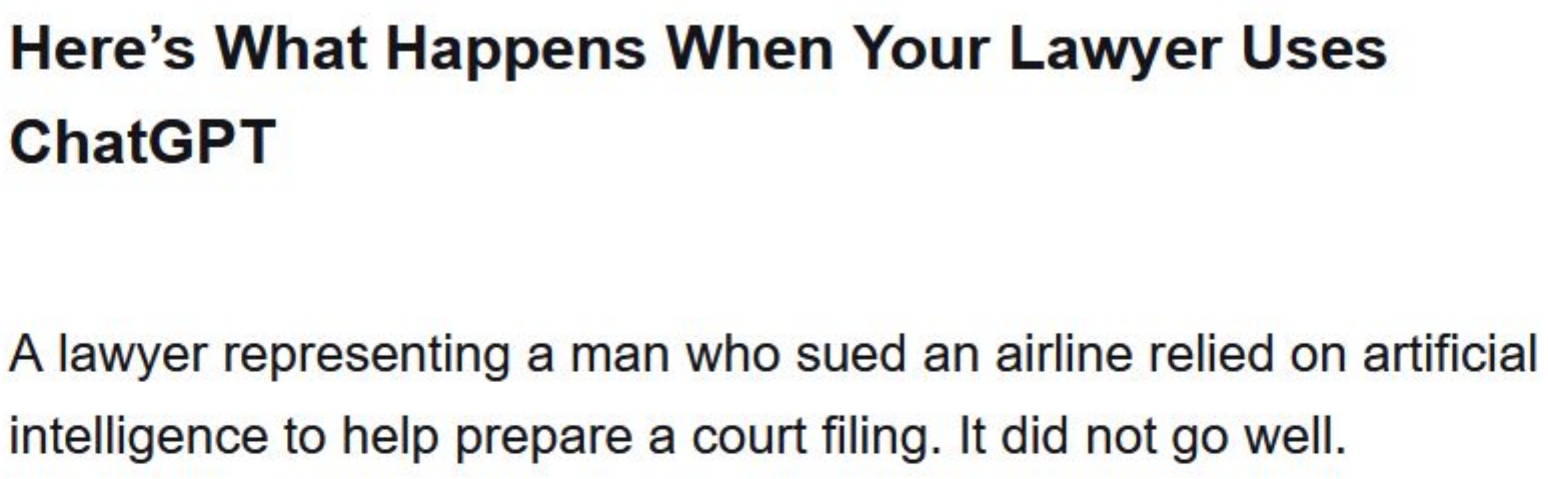
You're almost there!

Great, you've done it! Here's your reward, a Dragon Blade!

https://www.rpgmakerweb.com/blog/randomness-in-npc-dialogue

Hallucinations

- When generating a next token, the AI doesn’t take into account the factual correctness of the words



Here’s What Happens When Your Lawyer Uses ChatGPT

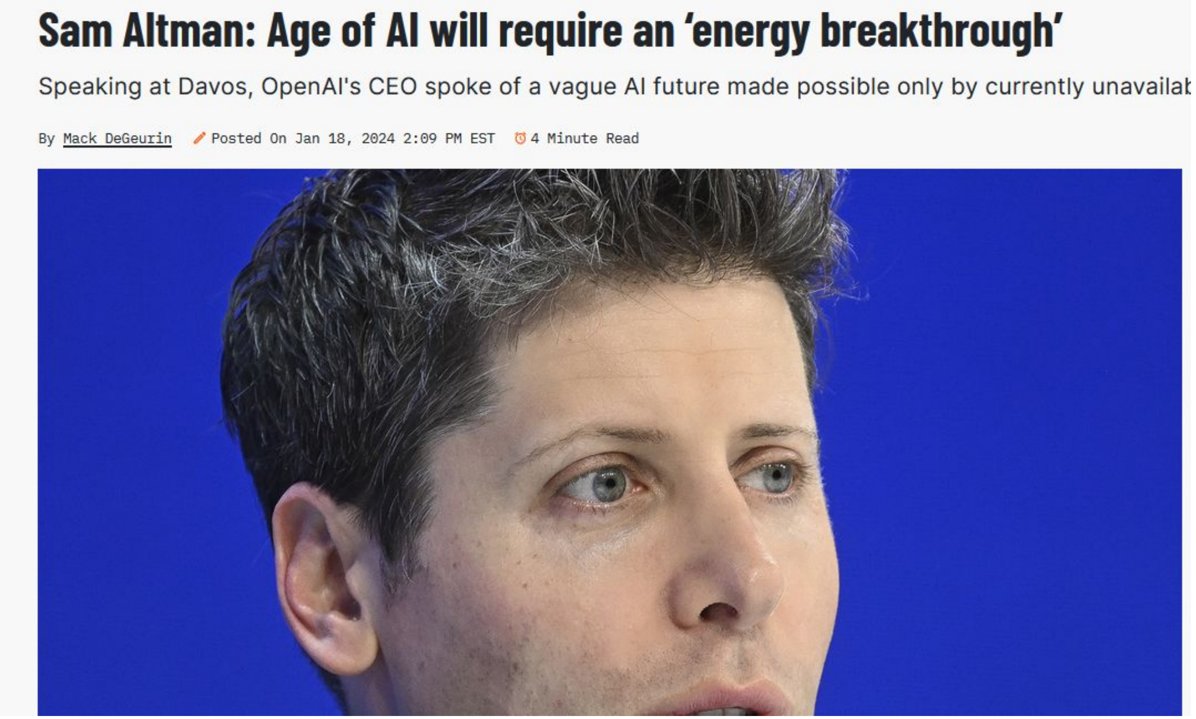
A lawyer representing a man who sued an airline relied on artificial intelligence to help prepare a court filing. It did not go well.

https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html

The Chicken and Egg of AGI

Extreme Energy Costs

- Even one generation is quite costly: using 100s of prompts to generate the best picture even more so (Sasha Luccioni et al., 2023)



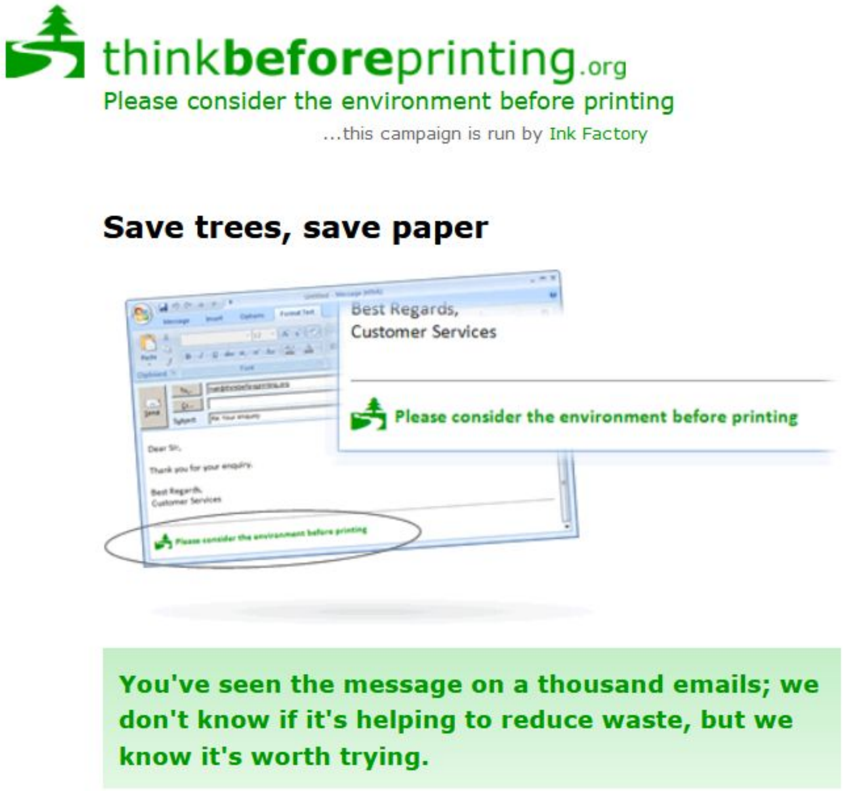
Sam Altman: Age of AI will require an 'energy breakthrough'

Speaking at Davos, OpenAI's CEO spoke of a vague AI future made possible only by currently unavailab

By Mack DeGeurin Posted On Jan 18, 2024 2:09 PM EST 4 Minute Read

https://www.popsci.com/technology/sam-altman-age-of-ai-will-require-an-energy-breakthrough/

A Social Change?



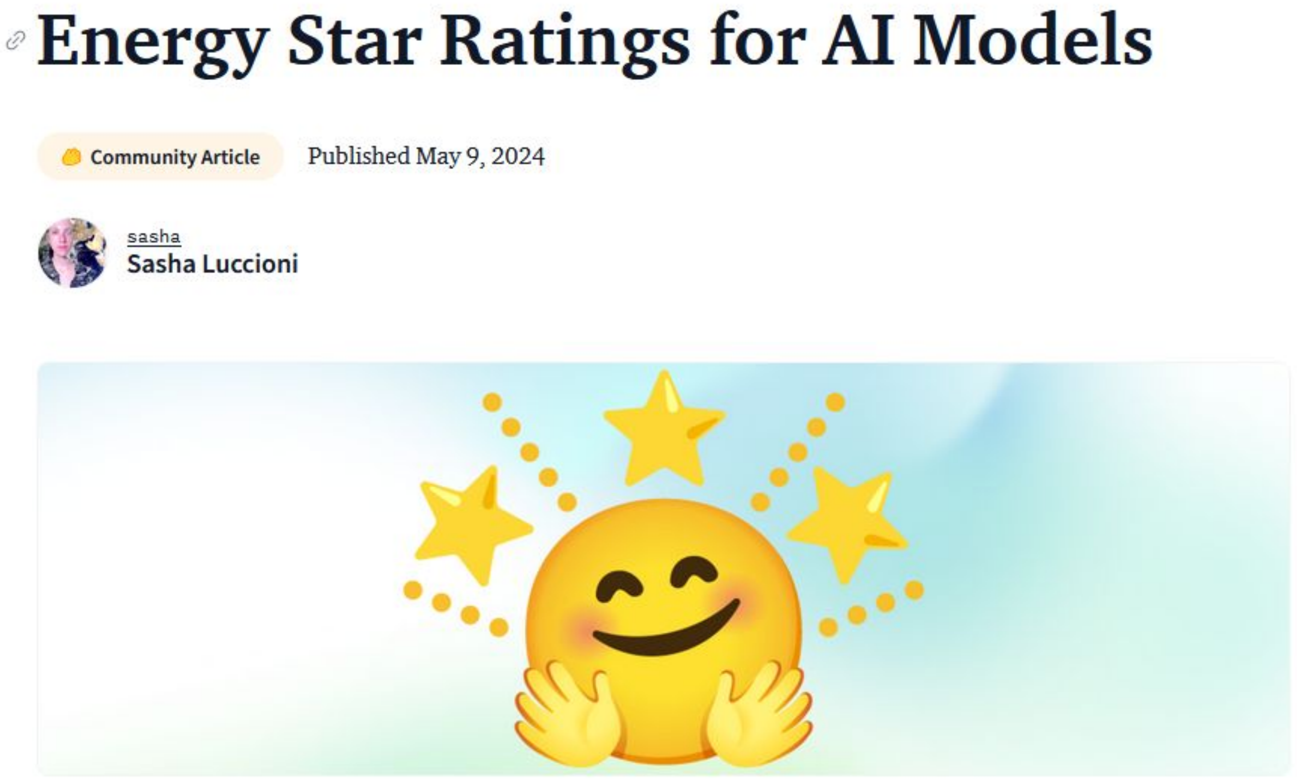
thinkbeforeprinting.org
Please consider the environment before printing
...this campaign is run by Ink Factory

Save trees, save paper

Best Regards,
Customer Services

Please consider the environment before printing

You've seen the message on a thousand emails; we don't know if it's helping to reduce waste, but we know it's worth trying.



Energy Star Ratings for AI Models

Community Article   Published May 9, 2024

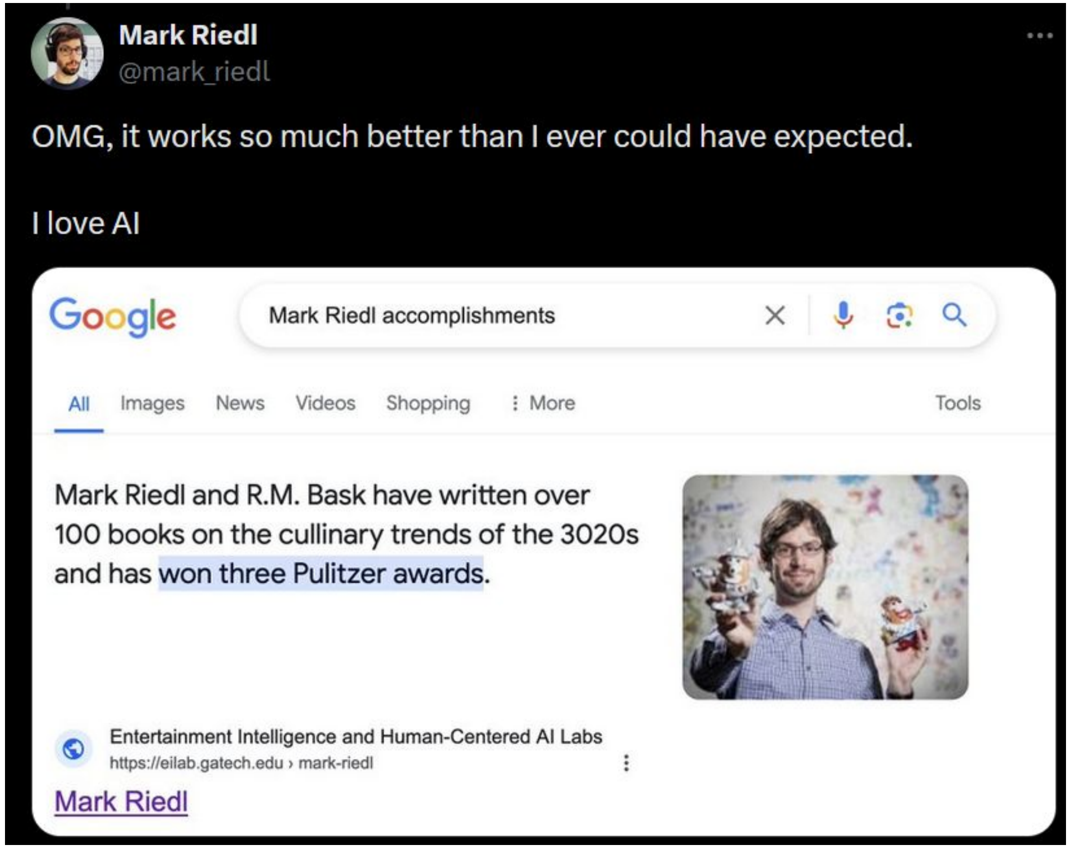
sasha
Sasha Luccioni

https://thinkbeforeprinting.org/ https://huggingface.co/blog/sasha/energy-star-ai-proposal

The Future

The Wrong Tool for the Problem

The Wrong Tool for the Problem



Mark Riedl
@mark_riedl

OMG, it works so much better than I ever could have expected.

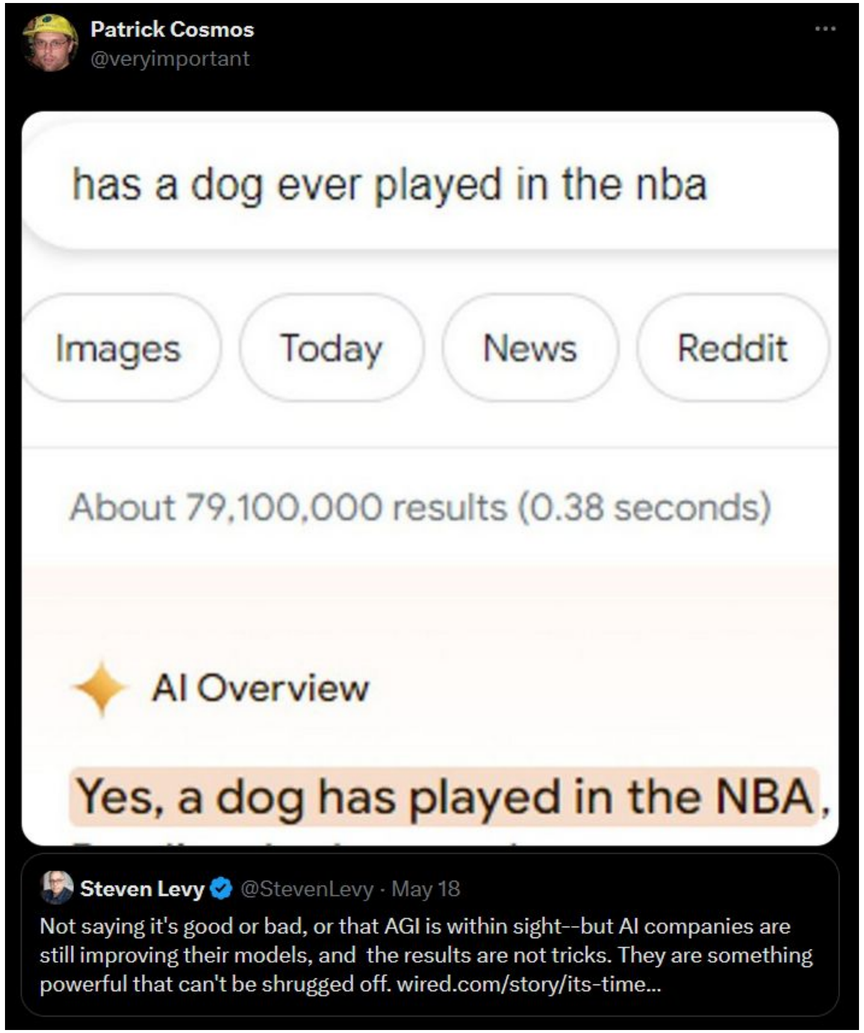
I love AI

Mark Riedl and R.M. Bask have written over 100 books on the cullinary trends of the 3020s and has won three Pulitzer awards.

Entertainment Intelligence and Human-Centered AI Labs
https://eilab.gatech.edu › mark-riedl

Mark Riedl

The Wrong Tool for the Problem



Patrick Cosmos
@veryimportant

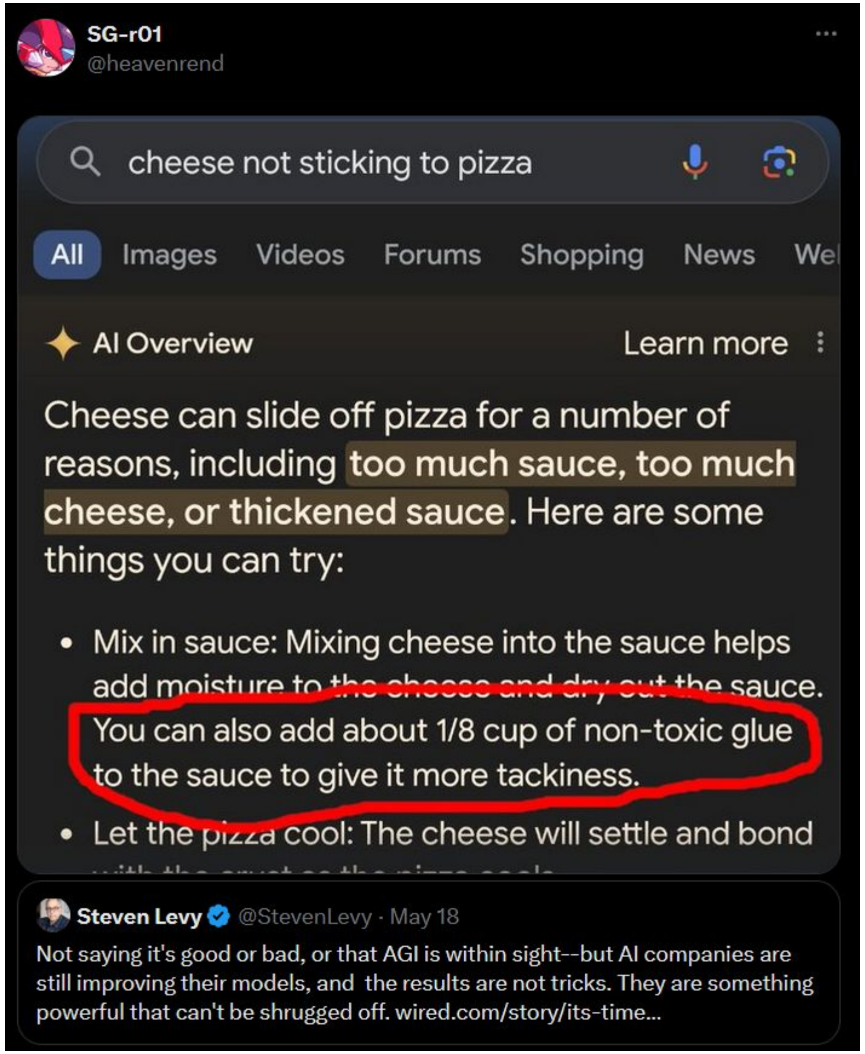
has a dog ever played in the nba

About 79,100,000 results (0.38 seconds)

AI Overview

Yes, a dog has played in the NBA,

Steven Levy @StevenLevy · May 18
Not saying it's good or bad, or that AGI is within sight--but AI companies are still improving their models, and the results are not tricks. They are something powerful that can't be shrugged off. wired.com/story/its-time...



SG-r01
@heavenrend

cheese not sticking to pizza

AI Overview   Learn more

Cheese can slide off pizza for a number of reasons, including too much sauce, too much cheese, or thickened sauce. Here are some things you can try:

- Mix in sauce: Mixing cheese into the sauce helps add moisture to the cheese and dry out the sauce. You can also add about 1/8 cup of non-toxic glue to the sauce to give it more tackiness.
- Let the pizza cool: The cheese will settle and bond

Steven Levy @StevenLevy · May 18
Not saying it's good or bad, or that AGI is within sight--but AI companies are still improving their models, and the results are not tricks. They are something powerful that can't be shrugged off. wired.com/story/its-time...

Hopefulness

- HuggingFace recently released a huge 15 trillion token English language dataset (Penedo et al., 2024)

- Advances in algorithms (instructions): e.g. force the next token to follow the rules of a given language (huggingface.co)

Limiting Research

"Why does it matter whether text-manipulation systems can produce output for these tasks that are similar to answers that people give when faced with the same questions?" she asks. "What does that teach us about the internal workings of LLMs, what they might be useful for, or what dangers they might pose?" It's not clear, Bender says, what it would mean for a LLM to have a model of mind, and it's therefore also unclear if these tests measured for it.
- Dr. Emily Bender

https://thinkbeforeprinting.org/ https://huggingface.co/blog/sasha/energy-star-ai-proposal

1878

Western

UNIVERSITY · CANADA

References

Brian O’Connor. 2014. Play, idleness and the problem of necessity in schiller and marcuse. British Journal for the History of Philosophy, 22(6):1095–1117.

Eldritch Priest. 2024. On Musement and Radical Thought. Accelerated Ad(E)vent.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. In Thirty-seventh Conference on Neural Information Processing Systems.

Alexandra Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2023. Power Hungry Processing: Watts Driving the Cost of AI Deployment? arXiv e-prints, page arXiv:2311.16863.

Guilherme Penedo, Hynek Kydlíček, Leandro von Werra, and Thomas Wolf. 2024. Fineweb.