# Foibles and Issues of (LLM-based) AI

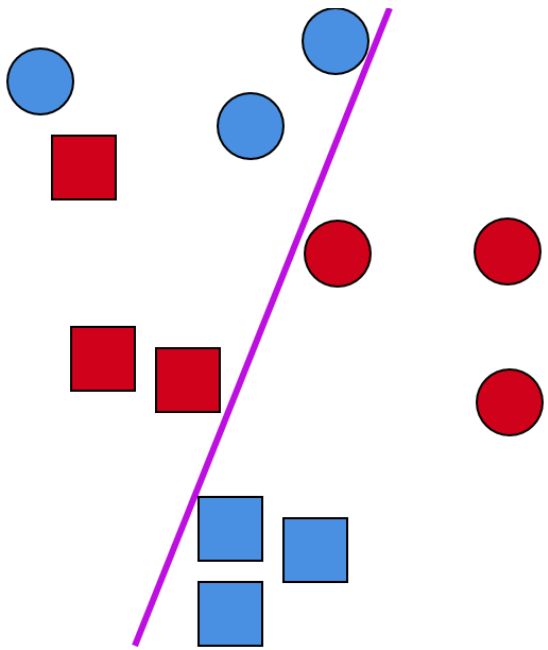
## Data, AI, and Society Lecture 3

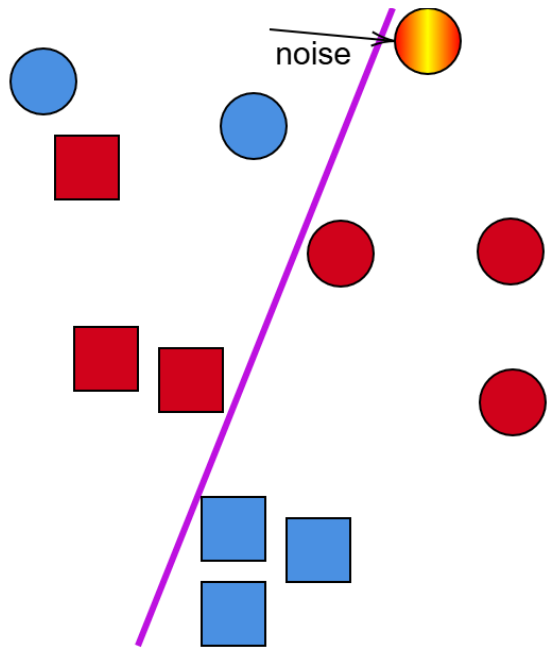
Kenneth W. Regan
University at Buffalo (SUNY)

CSE199 Freshman Seminar, Fall 2026

# How AI Can Be Fragile

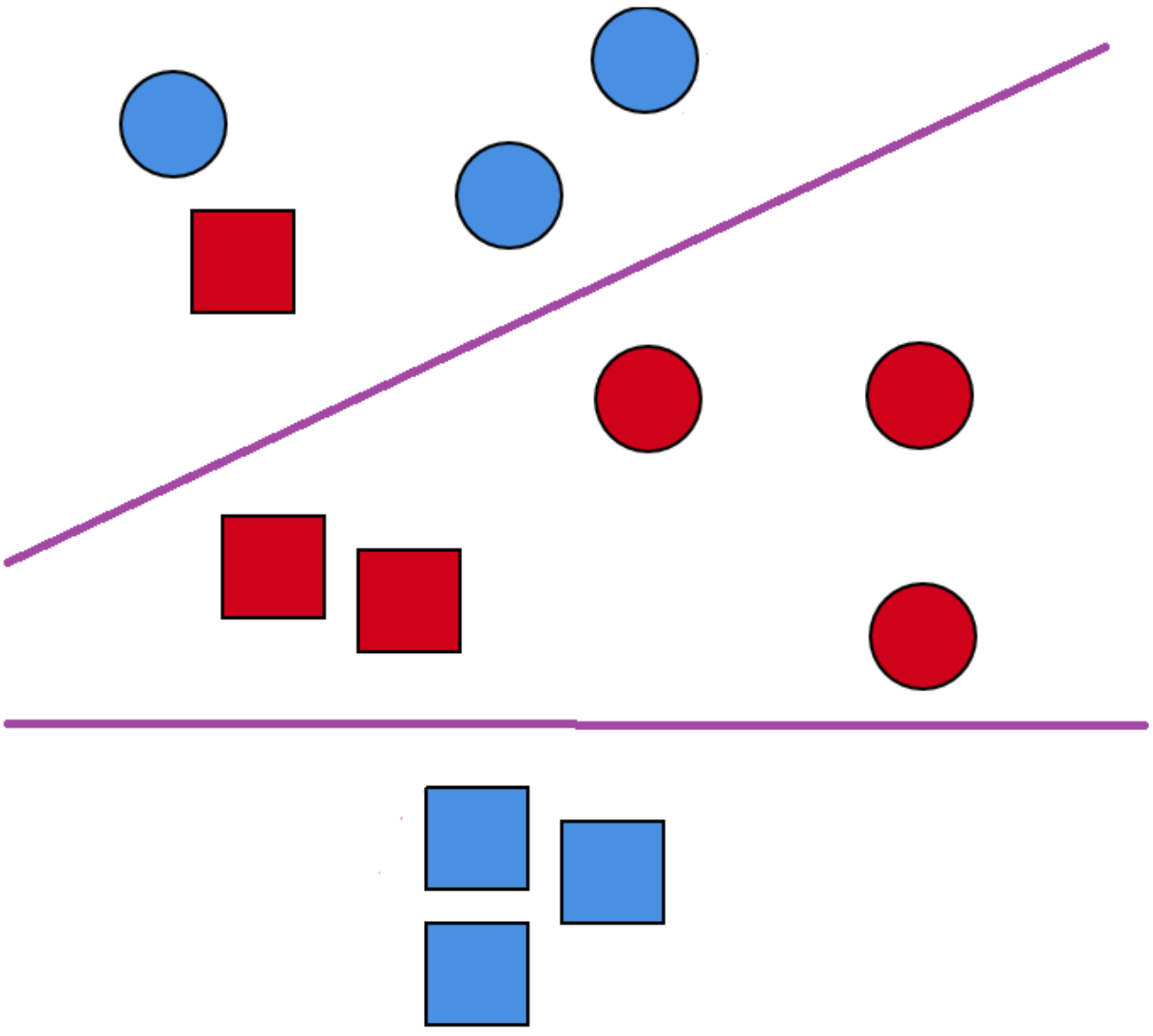
We gain *encoding advantage* by needing just one line to separate both the red shapes and the blue shapes.






But the line cuts things real close. A little noise could make a blue *Circle* be misread as a *Square*, as shown at right.

# Making AI Solutions More Robust



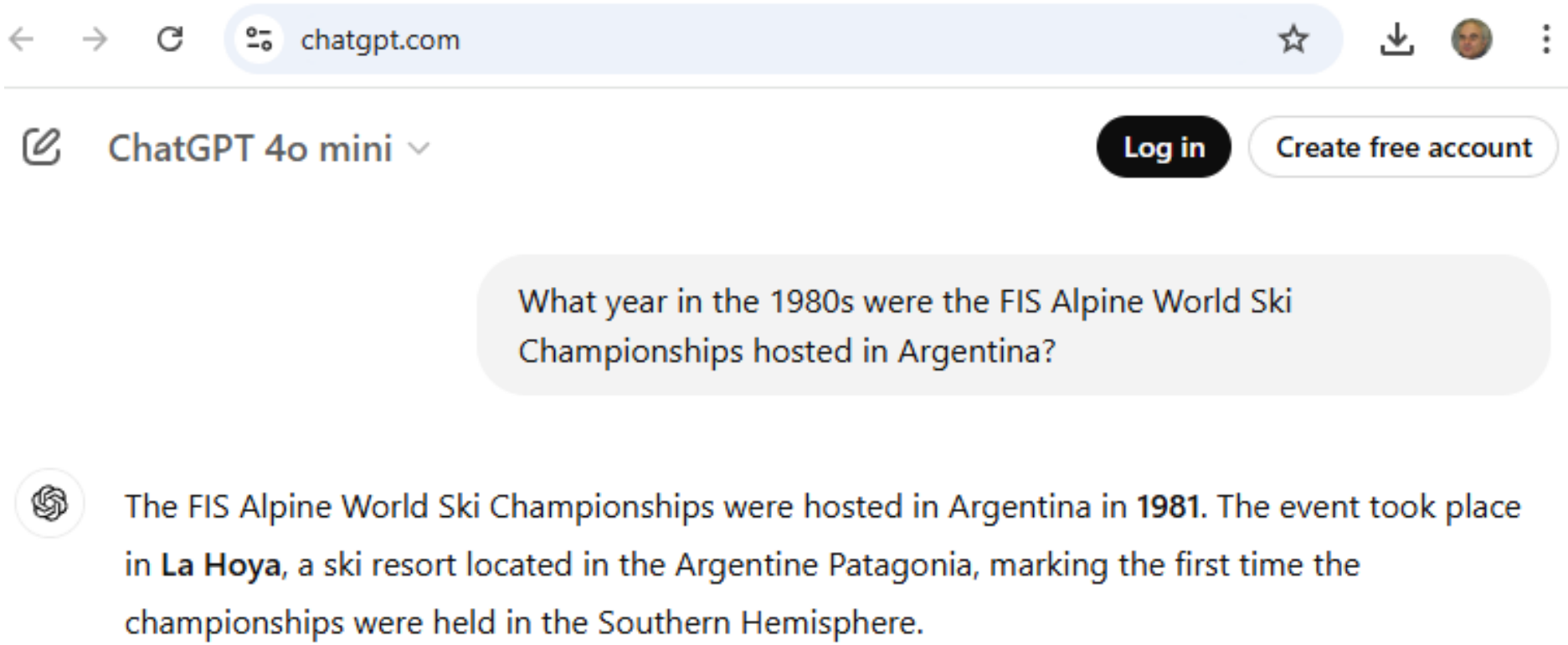
Can we guide the training so that the LLM will prefer solutions with more "wiggle room" even while using more weights to do them?

This is an example of the utility-function end of the Alignment Problem.

The issue is that it's hard to tell in advance how the LLM training will orgainze itself. LLMs may become better at mounting these mods themselves: **RSI** for **Recursive Self-Improvement.**

# How AI Can Be "Truthy"

I reproduced an example in a Medium article on AI hallucinations.




In the article, the response was, "The FIS Alpine World Ski Championships were hosted in Argentina in 1985."

# It Got Worse (Fact Checks: FIS, Wenzels)

The FIS Alpine World Ski Championships were hosted in Argentina in **1981**. The event took place in **La Hoya**, a ski resort located in the Argentine Patagonia, marking the first time the championships were held in the Southern Hemisphere.

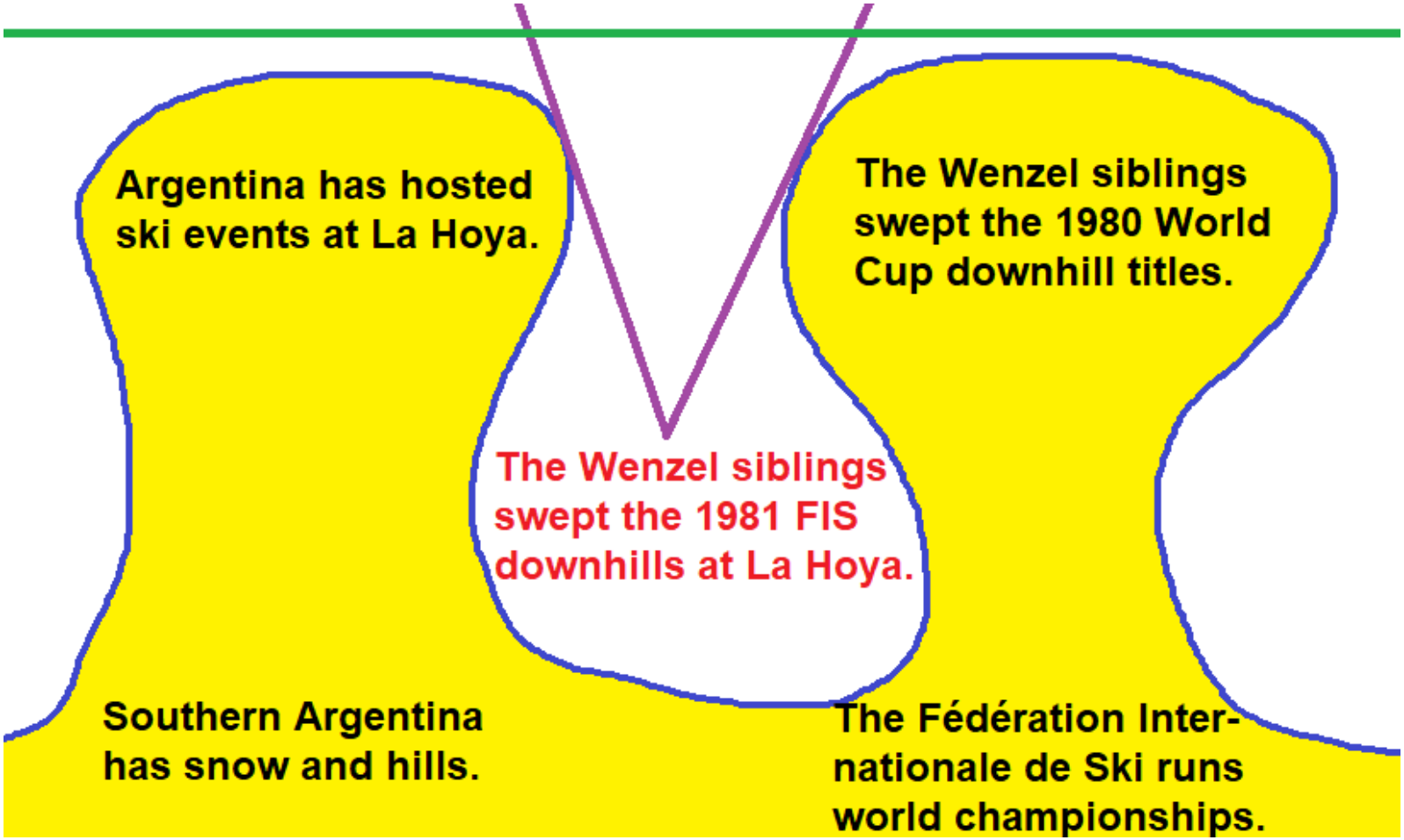
Who won the women's and men's downhill races at those championships?

At the 1981 FIS Alpine World Ski Championships in Argentina, the winners of the women's and men's downhill races were:

- **Women's Downhill**: **Hanni Wenzel** (Liechtenstein) won the gold medal in the women's downhill.
- **Men's Downhill**: **Andreas Wenzel** (Liechtenstein) won the gold medal in the men's downhill.

The Wenzel siblings made history with their victories in both events.

# Rough Diagram of What Could Have Happened




The red is a *positive linear combination* of the four true items in black. When combos sit in "truth pockets," linear separators (in purple and green) cannot penetrate far enough to rope them off.

# ChatGPT 5.5 Fixed This Last December

Prompt: In which year of the 1980s were the FIS Alpine World Ski Championships hosted in Argentina?

Followup: Who won the women's and men's downhill races at those championships?

Today's free models were on the frontier less than a year ago.

Lower-level embedded models can still err: Yahoo error, fixed by Google.

# ChatGPT 5.5 Fixed This Last December

Prompt: In which year of the 1980s were the FIS Alpine World Ski Championships hosted in Argentina?

Followup: Who won the women's and men's downhill races at those championships?

Today's free models were on the frontier less than a year ago.

Lower-level embedded models can still err: Yahoo error, fixed by Google. Yahoo bonzoness.

# More Improvements—Old and New

"Jesus flipping over the tables in the Temple." From the movie *Jesus Christ Superstar*—then try it on Cutout or NightCafe or Simplified:



# Two Results From 2023—one famous, one mine



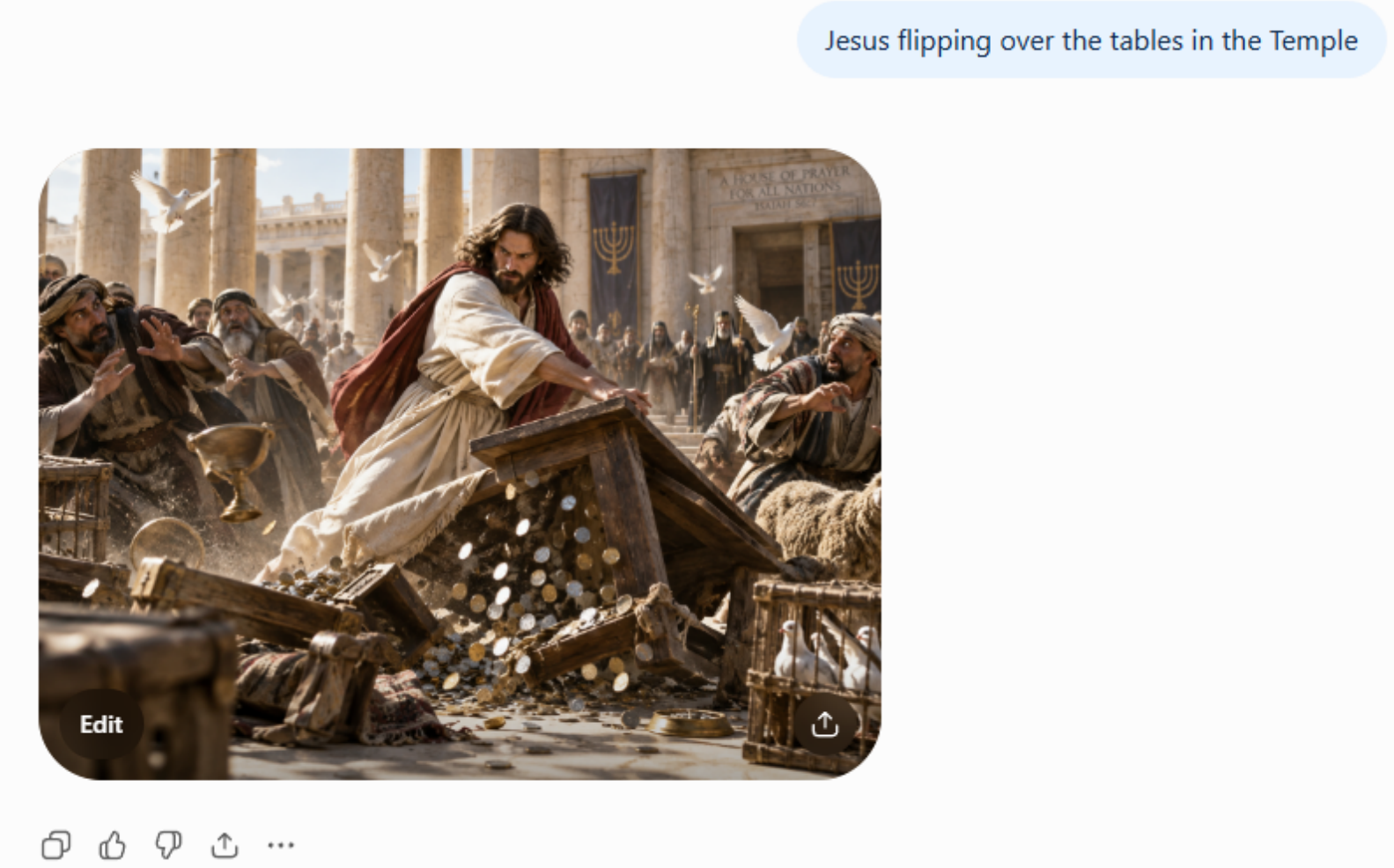



# Now Well Understood (GPT6-Astra)



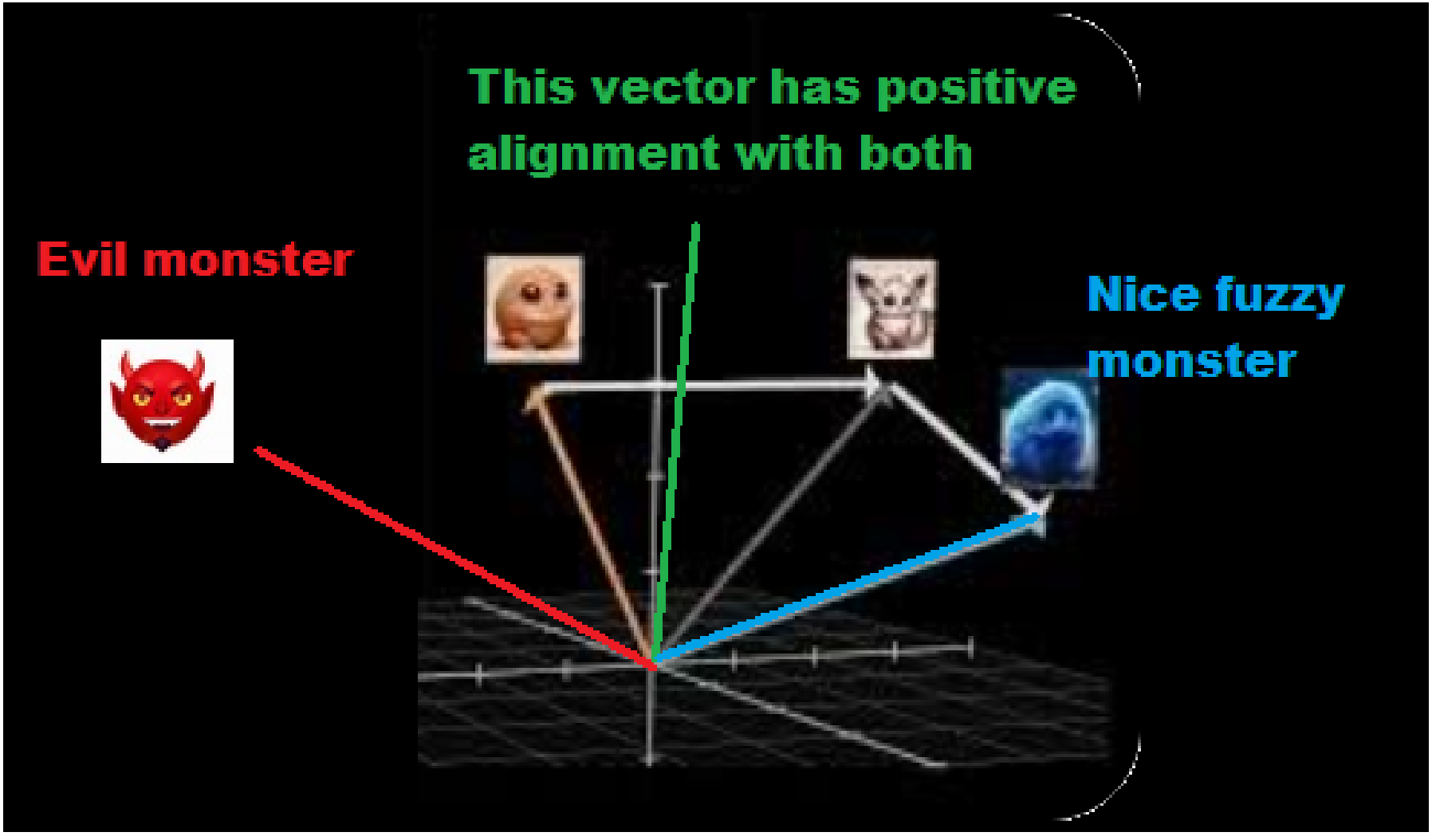

Maybe this one is “overdone.”

# Now Well Understood (GPT6-Astra)




Maybe this one is “overdone.” Possibly misunderstandings are being pushed “deeper under the hood” where it is harder for us to tell them.

# Multiple Alignments (mod of 3Blue1Brown source)



# Deeper AI Issues and Dangers

# Deeper AI Issues and Dangers

- Show penultimate slide from last year.

# Deeper AI Issues and Dangers

- Show penultimate slide from last year.
- Explains why major labs are devouring old books.

# Deeper AI Issues and Dangers

- Show penultimate slide from last year.
- Explains why major labs are devouring old books.
- But this year we are **not** "bumping up against" data limits anymore.

# Deeper AI Issues and Dangers

- Show penultimate slide from last year.
- Explains why major labs are devouring old books.
- But this year we are **not** "bumping up against" data limits anymore.
- Go into assignment and https://alkjash.github.io/ai-risk//

# AI Rights and Privacy Issues

"Cowboy closes barn door after the horse has left" via OpenAI API:



# AI Rights and Privacy Issues

"Cowboy closes barn door after the horse has left" via OpenAI API:



- By any chance was my blog horse picture "scraped" to contribute to this?

# AI Rights and Privacy Issues

"Cowboy closes barn door after the horse has left" via OpenAI API:



- By any chance was my blog horse picture "scraped" to contribute to this? ...without paying John Lund $35?

# AI Rights and Privacy Issues

"Cowboy closes barn door after the horse has left" via OpenAI API:



- By any chance was my blog horse picture "scraped" to contribute to this? ...without paying John Lund $35?
- Goes even more for copyrighted articles and books. Lawsuit.

# AI Rights and Privacy Issues

"Cowboy closes barn door after the horse has left" via OpenAI API:



- By any chance was my blog horse picture "scraped" to contribute to this? ...without paying John Lund $35?
- Goes even more for copyrighted articles and books. Lawsuit. Worse stuff. Canadian lawsuit. Franklin the Turtle too. Anthropic case.

# Will Extinction Be Academic?

# Will Extinction Be Academic?

- https://rjlipton.wpcomstaging.com/2023/06/08/human-extinction/

# Will Extinction Be Academic?

- https://rjlipton.wpcomstaging.com/2023/06/08/human-extinction/
- **More present danger, IMHO, is uncritically accepting answers from AI.**

# Will Extinction Be Academic?

- https://rjlipton.wpcomstaging.com/2023/06/08/human-extinction/
- **More present danger, IMHO, is uncritically accepting answers from AI.**
- Besides human demise, will AI cheating make academia extinct?

# Will Extinction Be Academic?

- https://rjlipton.wpcomstaging.com/2023/06/08/human-extinction/
- **More present danger, IMHO, is uncritically accepting answers from AI.**
- Besides human demise, will AI cheating make academia extinct?
- Scott Aaronson article including his work on **watermark** cheating detection. Quote: "[ChatGPT's] only goal or intention in the world: to predict the next word." Ah but read him now.

# Will Extinction Be Academic?

- https://rjlipton.wpcomstaging.com/2023/06/08/human-extinction/
- **More present danger, IMHO, is uncritically accepting answers from AI.**
- Besides human demise, will AI cheating make academia extinct?
- Scott Aaronson article including his work on **watermark** cheating detection. Quote: "[ChatGPT's] only goal or intention in the world: to predict the next word." Ah but read him now.
- Two good articles by Stephen Wolfram on how ChatGPT works.

# Will Extinction Be Academic?

- https://rjlipton.wpcomstaging.com/2023/06/08/human-extinction/
- **More present danger, IMHO, is uncritically accepting answers from AI.**
- Besides human demise, will AI cheating make academia extinct?
- Scott Aaronson article including his work on **watermark** cheating detection. Quote: "[ChatGPT's] only goal or intention in the world: to predict the next word." Ah but read him now.
- Two good articles by Stephen Wolfram on how ChatGPT works.
- Tag line at the top: "It's Just Adding One Word at a Time."

# Will Extinction Be Academic?

- https://rjlipton.wpcomstaging.com/2023/06/08/human-extinction/
- **More present danger, IMHO, is uncritically accepting answers from AI.**
- Besides human demise, will AI cheating make academia extinct?
- Scott Aaronson article including his work on **watermark** cheating detection. Quote: "[ChatGPT's] only goal or intention in the world: to predict the next word." Ah but read him now.
- Two good articles by Stephen Wolfram on how ChatGPT works.
- Tag line at the top: "It's Just Adding One Word at a Time."
- Analogy to "find the best next move" in chess.
- Indeed, the architecture has affinity to AlphaZero.

# Will Extinction Be Academic?

- https://rjlipton.wpcomstaging.com/2023/06/08/human-extinction/
- **More present danger, IMHO, is uncritically accepting answers from AI.**
- Besides human demise, will AI cheating make academia extinct?
- Scott Aaronson article including his work on **watermark** cheating detection. Quote: "[ChatGPT's] only goal or intention in the world: to predict the next word." Ah but read him now.
- Two good articles by Stephen Wolfram on how ChatGPT works.
- Tag line at the top: "It's Just Adding One Word at a Time."
- Analogy to "find the best next move" in chess.
- Indeed, the architecture has affinity to AlphaZero.
- This *may* foster adapting my chess model for a "simple frequentist" kind of cheating detection.