

Are We Having the Wrong Dreams about Interpretability?

Robin Jia
University of Southern California
Talks at UCSD EnCORE Workshop on Interpretability in Modern AI
and Center for Human-Compatible AI Workshop
Winter 2026 - Summer 2026

**What is
interpretability
good for?**

Interpretability Can Be Actionable

Fazl Barez^{* 2 3} **Tal Haklay**^{* 4} **Isabelle Lee**^{* 5} **Hadas Orgad**¹ **Marius Mosbach**^{* 6 7} **Anja Reusch**^{* 4} **Naomi Saphra**^{* 1 8}
Byron C Wallace^{* 9} **Sarah Wiegrefe**^{* 10} **Eric Wong**^{* 11}
Ian Tenney¹² **Mor Geva**¹³

< GDM MECH INTERP PROGRESS UPDATES >

A Pragmatic Vision for Interpretability

by **Neel Nanda, Josh Engels, Arthur Conmy, Senthooran Rajamanoharan, bilalchughtai,**
CallumMcDougall, János Kramár, lewis smith

1st Dec 2025

Beyond Measurable Actionability



How can interpretability research achieve real-world impact?

Making actionability a core evaluation criterion is required to accelerate progress and achieve measurable real-world impact

Actionable checklist



1. Define a clear goal

Identify a **specific problem** that your interpretability question aims to eventually solve.



2. Identify your audience

Communicate insights according to different stakeholders: **developers, practitioners, policymakers.**



3. Propose concrete actions

Articulate what **decisions** or **interventions** your insights enable.



4. Validate empirically

Implement the proposed action yourself and **demonstrate** its effects.



5. Evaluate in realistic settings

Apply methods to **large-scale models** and **non-synthetic datasets.**



6. Actionability as success criteria

- ☐ Surpasses **standard baselines** (prompting, fine-tuning).
- ☐ **Generalizes** across setting variations and seeds.
- ☐ Produces **targeted effects** without degrading other capabilities.
- ☐ Yields **useful explanations** for target audience.

- Focus on impact of an **individual project**
- Focus on impact to the **AI research/engineering** world
- Can *the body of interpretability research* inform audiences *beyond the AI world?*

Common Systemic Failures in Anticipation:

- Our mental models get stuck in the past:
 - Manuscript/Book
 - Horse/Car;
 - Humans/Generative AI
- We must understand the new technology AS IT IS!

Claim: Interpretability can help us understand AI *as it is*

Are LLM's “Body-less humans”?

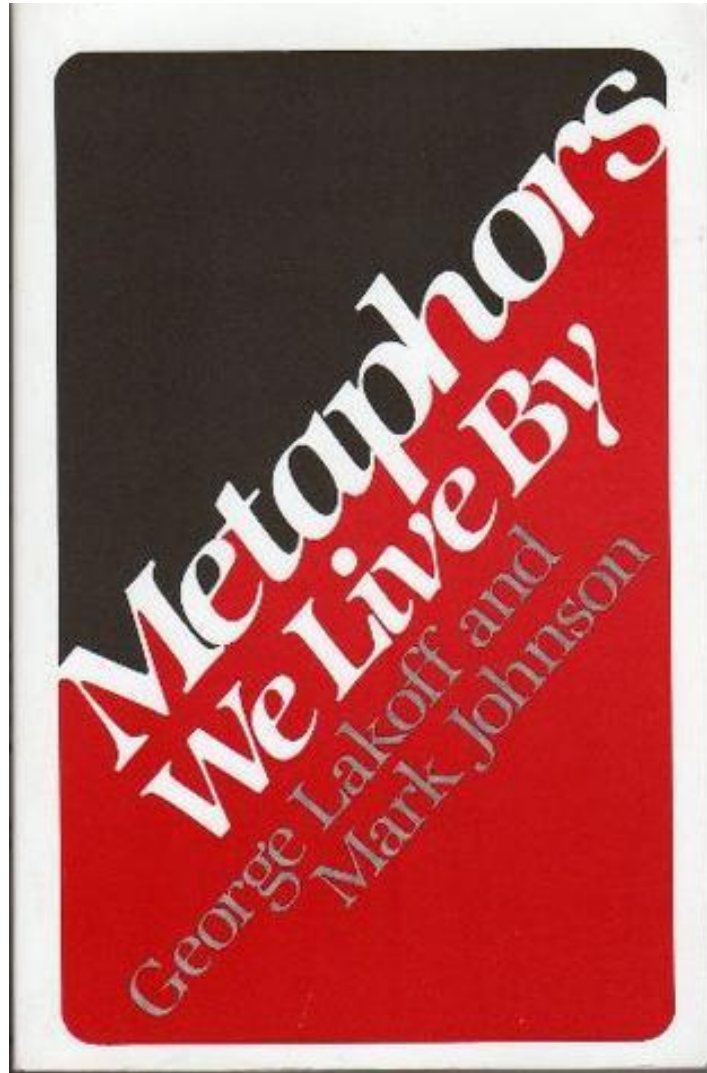
Pro

- They emulate human-written data
- They communicate through a human-like interface (text)

Con

- Their capabilities are “jagged” compared with humans
- They can be self-contradictory, inconsistent, etc.
 - More so than humans
- They are created from an amalgamation of many people's data

Are LLM's “Body-less humans”?



Con

- Their capabilities are “jagged” compared with humans
- They can be self-contradictory, inconsistent, etc.
 - More so than humans
- They are created from an amalgamation of many people's data
- **We need a better metaphor to understand AI “as it is”!**

Are LLM's “Stochastic Parrots”?

Pro

- They emulate human-written data
- They can reproduce human text without being human themselves
- They can learn different things from different people
- They are capable of impressive feats of memorization

Con

- Downplays the generalizable capabilities of models
 - Abilities of coding agents seem to go beyond merely “parroting” existing code
- Does not explain shortcomings of models not linked to memorization/failure to generalize

Ezra Klein and AI Metaphors



Ezra Klein: Do you still see these A.I. systems as **souped-up autocomplete** or do you think that **metaphor** has lost its power?

Ezra Klein and AI Metaphors



Jack Clark: The way that I think of these systems now is that they're like **little troublesome genies** that I can give instructions to

Ezra Klein and AI Metaphors



Ezra Klein: But the **autocomplete metaphor at least had a perspective** on what it was these systems were doing. It was a **prediction model**.

Ezra Klein and AI Metaphors



Ezra Klein: “**Genies**,” I don’t like, because then you’ve just moved straight into mysticism...they’re **just a completely alternative creature with vast powers.**

Ezra Klein and AI Metaphors



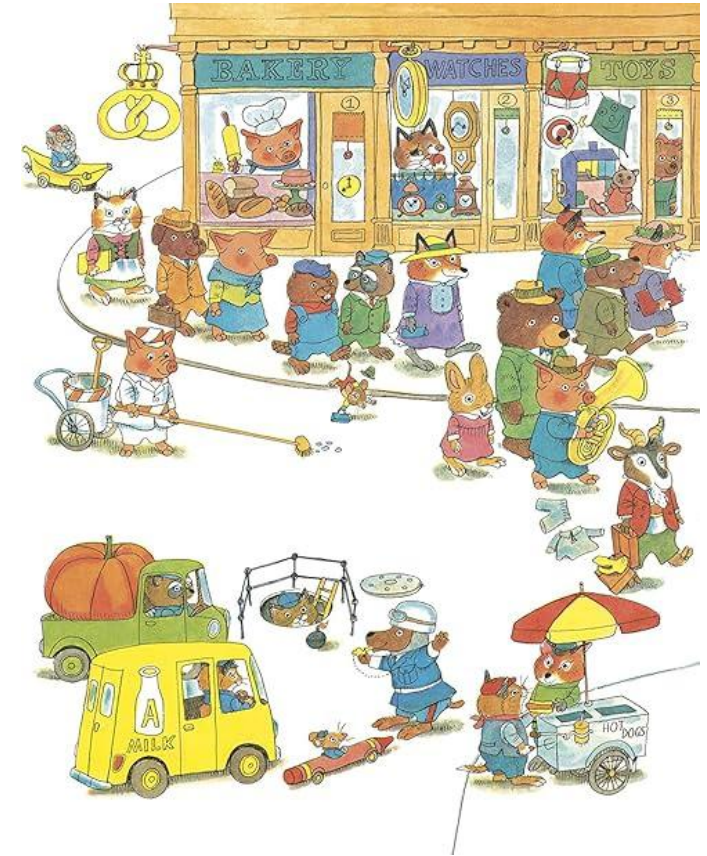
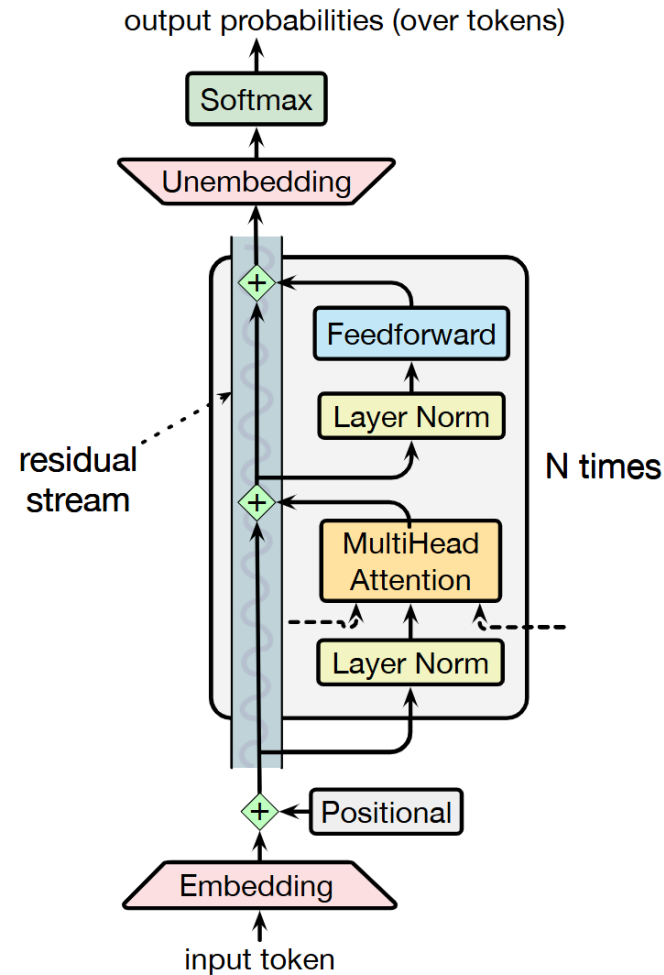
Ezra Klein: I don't think these are now just fancy autocomplete systems. And on the other hand, **I'm not sure what metaphor makes sense.**

**What could
interpretability be good for?
Crafting new metaphors**

LLMs are Democratic Societies

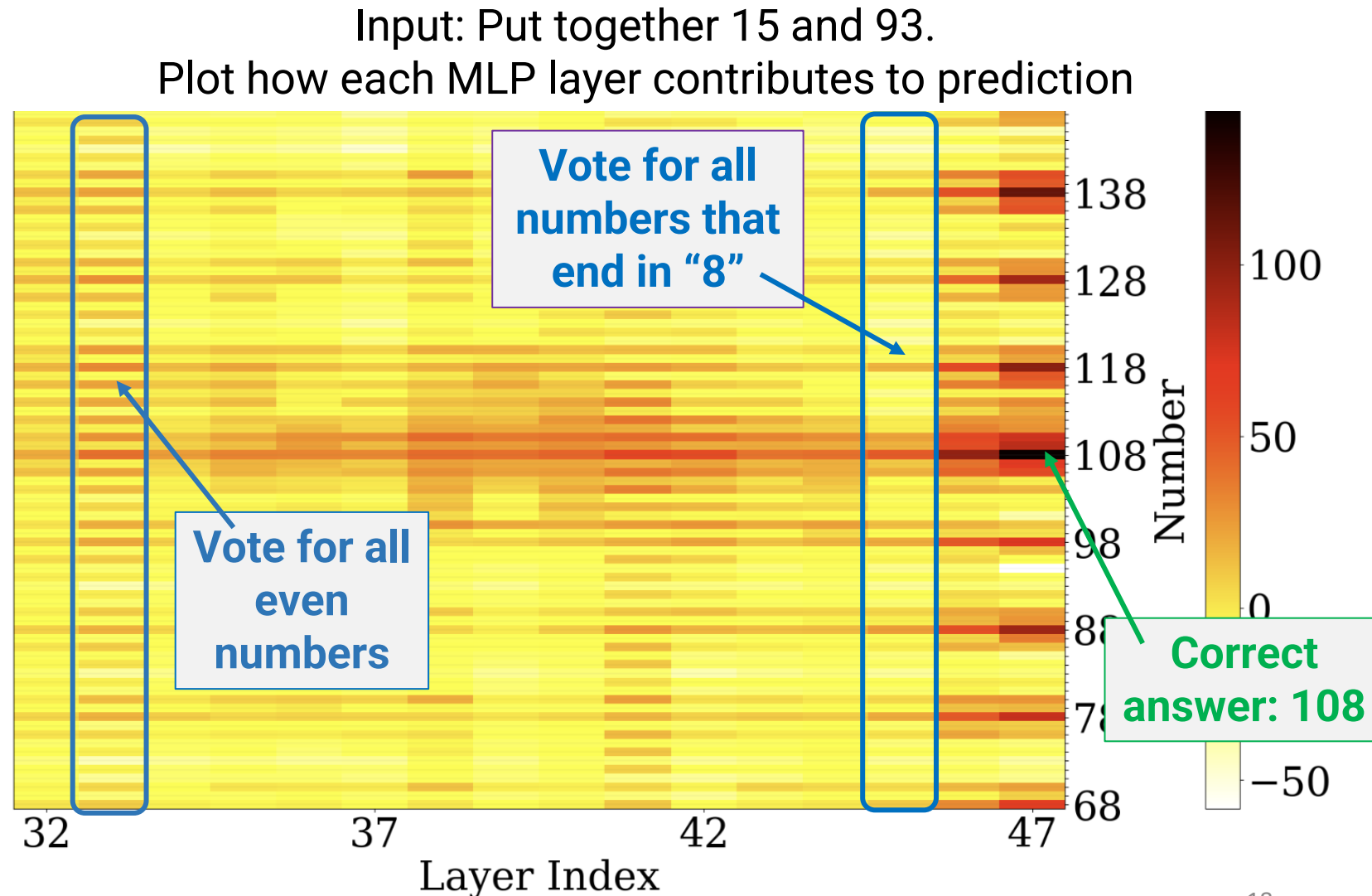
LLMs as Democratic Societies

- Attention heads, MLP's are **independent individuals** with **specialized roles**
- Individuals **communicate** via the residual stream
- Individuals may **cooperate or disagree**
- Final decision is decided by **voting** in the residual stream



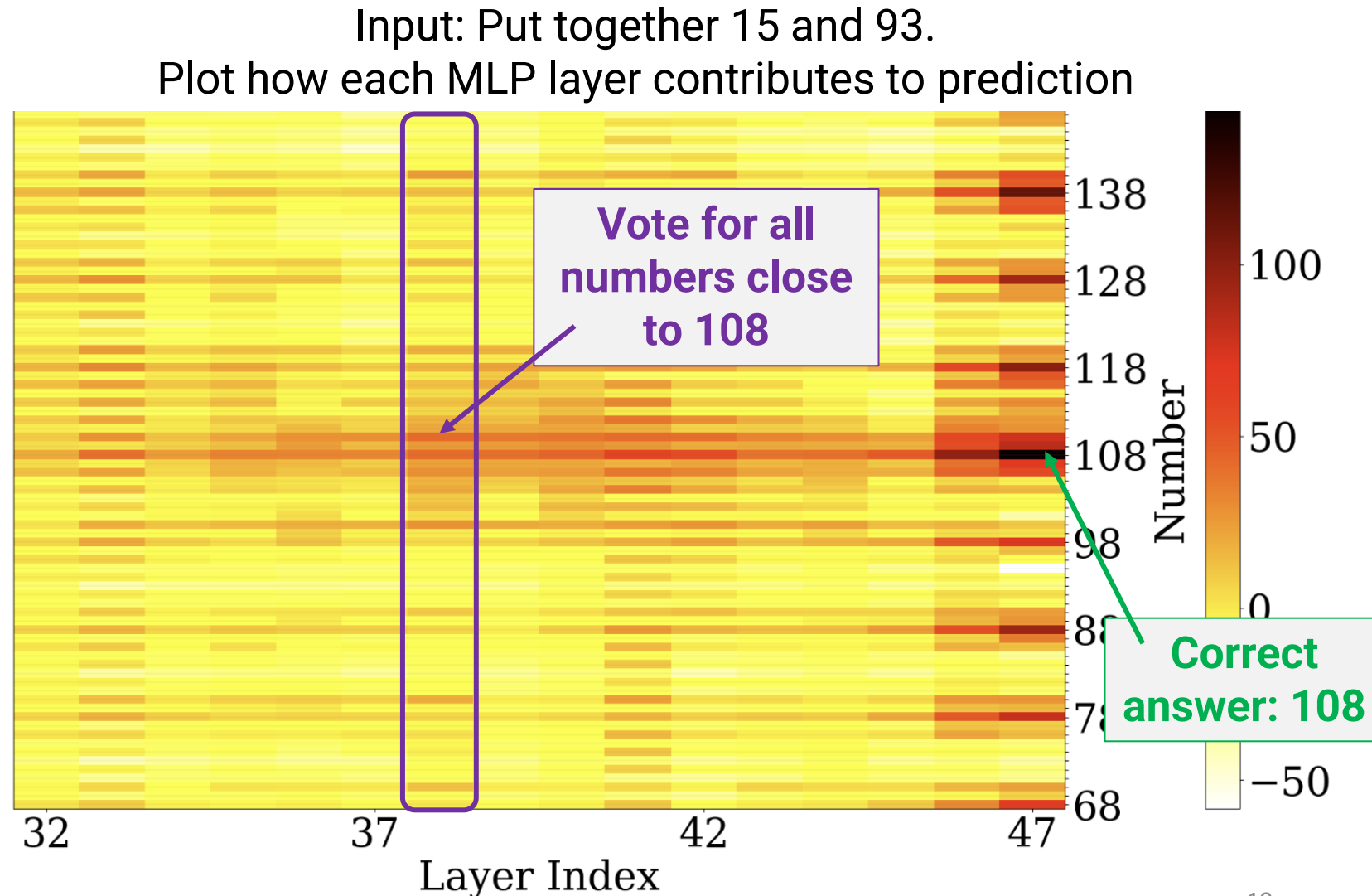
Division of Labor & Voting

- Different layers perform different subtasks in parallel
- Some layers vote based on modular arithmetic

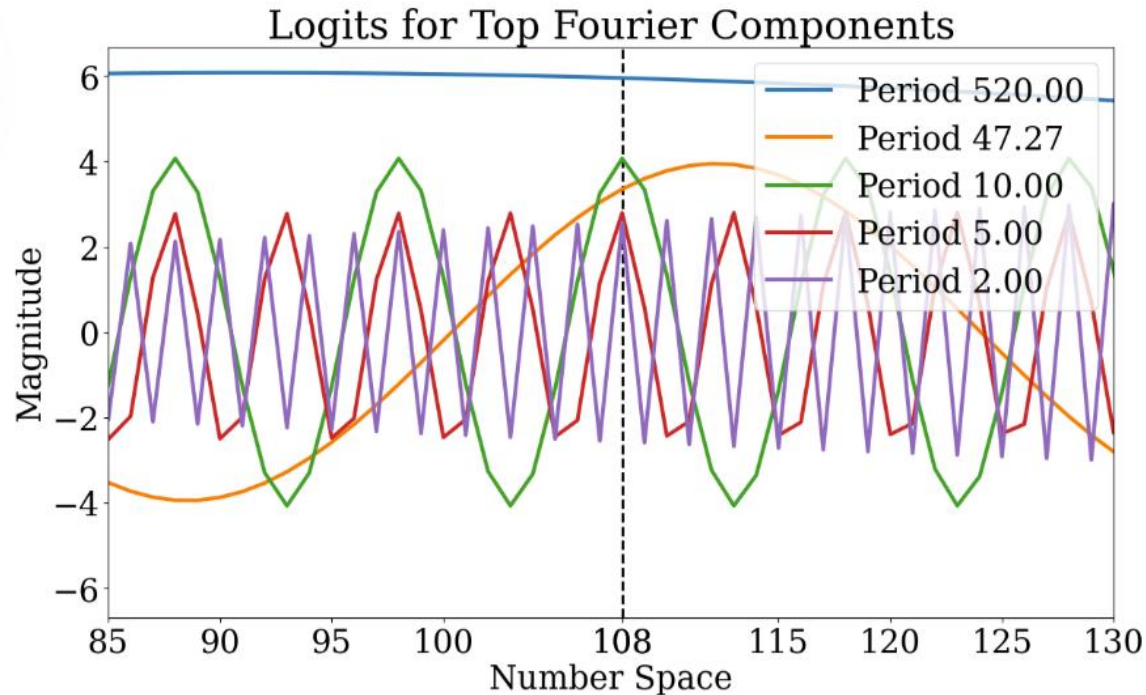


Division of Labor & Voting

- Different layers perform different subtasks in parallel
- Some layers vote based on modular arithmetic
- Others vote based on approximate value

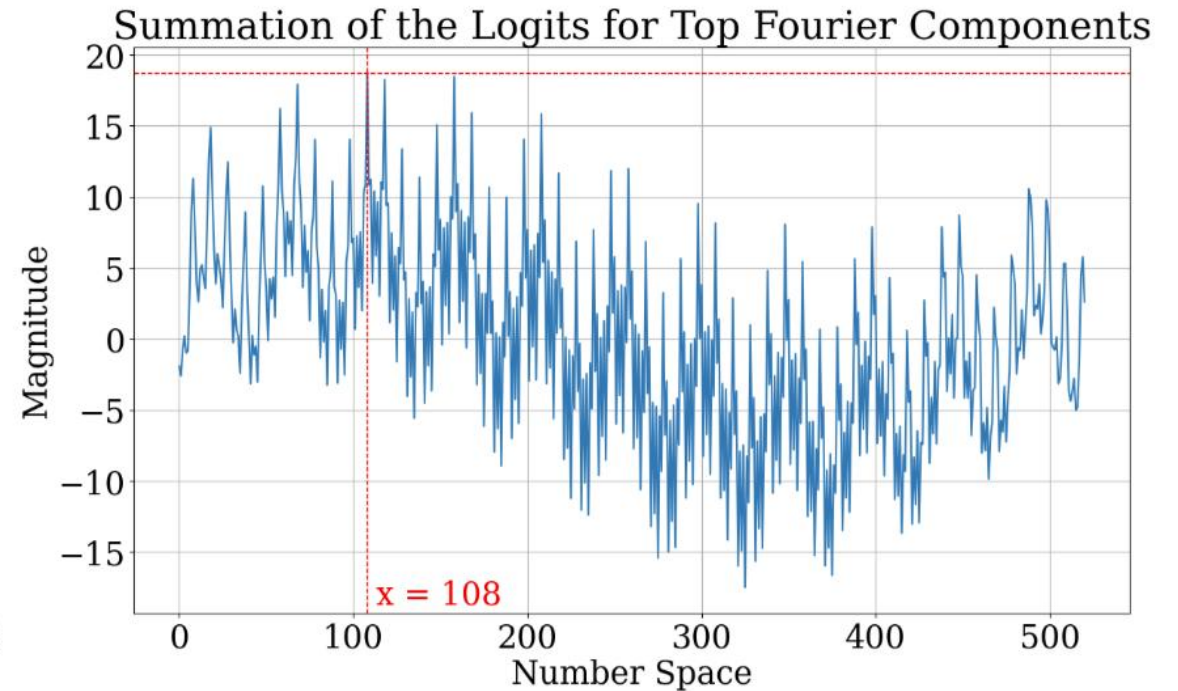
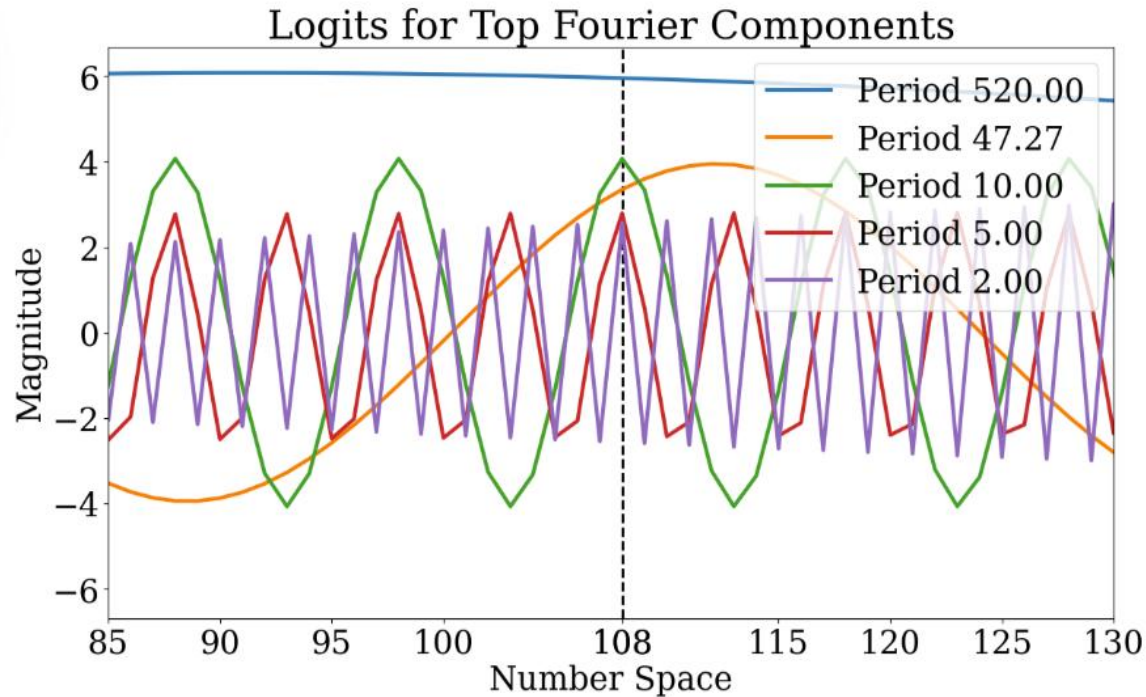


Division of Labor & Voting



- Low frequency: “Votes” for approximately correct numbers
- High frequency: “Votes” for all numbers that work modulo 2 or 5

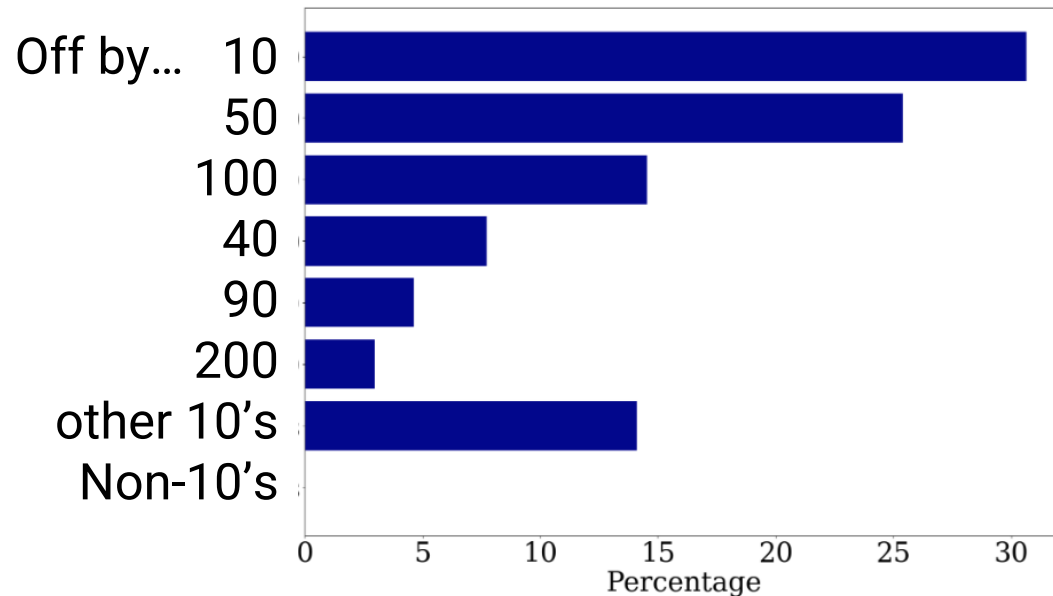
Division of Labor & Voting



- Low frequency: “Votes” for approximately correct numbers
- High frequency: “Votes” for all numbers that work modulo 2 or 5
- Mechanism: **Add up the votes**, get the right answer!

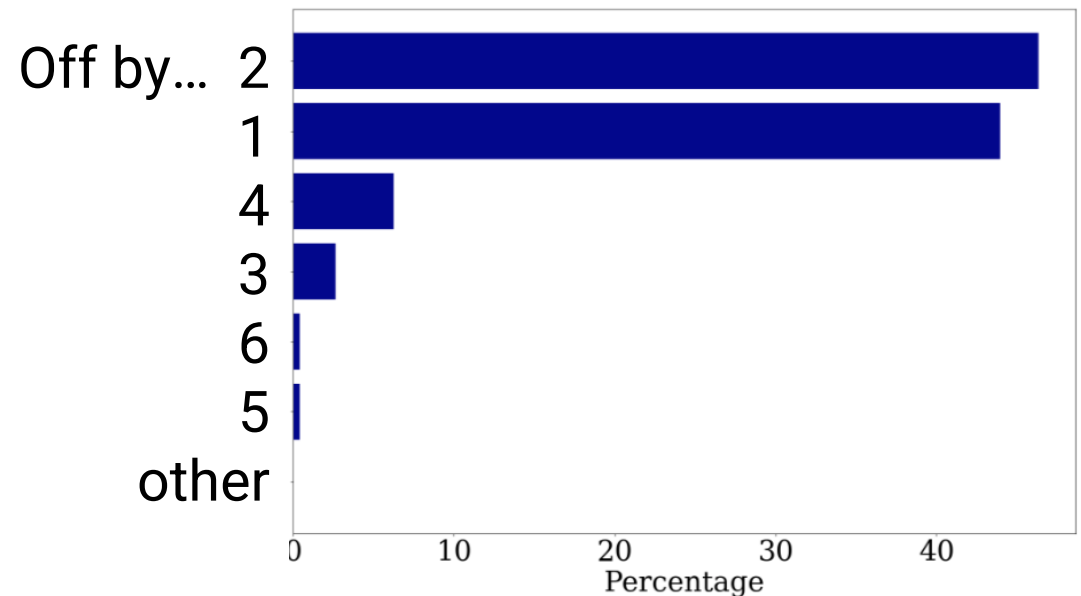
Work is Done Independently

Filter Low-Freq from MLP



- Removes ability to approximate the answer
- Still correct modulo 10!
- **Stop work on approximation, modulo 10 work continues as usual**

Filter High-Freq from Attn



- Removes ability to be correct mod 10
- Still approximately correct!
- **Stop work on modulo 10 prediction, approximation work continues as usual**

Parallel over Serial

- What algorithms do Transformers like to learn for in-context learning?
- Our strong intuition: Gradient descent is simple and natural
- But it is highly **serial**

WHAT LEARNING ALGORITHM IS IN-CONTEXT LEARNING? INVESTIGATIONS WITH LINEAR MODELS

Ekin Akyürek^{1,2,a} Dale Schuurmans¹ Jacob Andreas^{*2} Tengyu Ma^{*1,3,b} Denny Zhou^{*1}

¹Google Research ²MIT CSAIL ³Stanford University ^{*}collaborative advising

Transformers Learn In-Context by Gradient Descent

Johannes von Oswald^{1,2} Eyvind Niklasson² Ettore Randazzo² João Sacramento¹
Alexander Mordvintsev² Andrey Zhmoginov² Max Vladymyrov²

One Step of Gradient Descent is Provably the Optimal In-Context Learner with One Layer of Linear Self-Attention

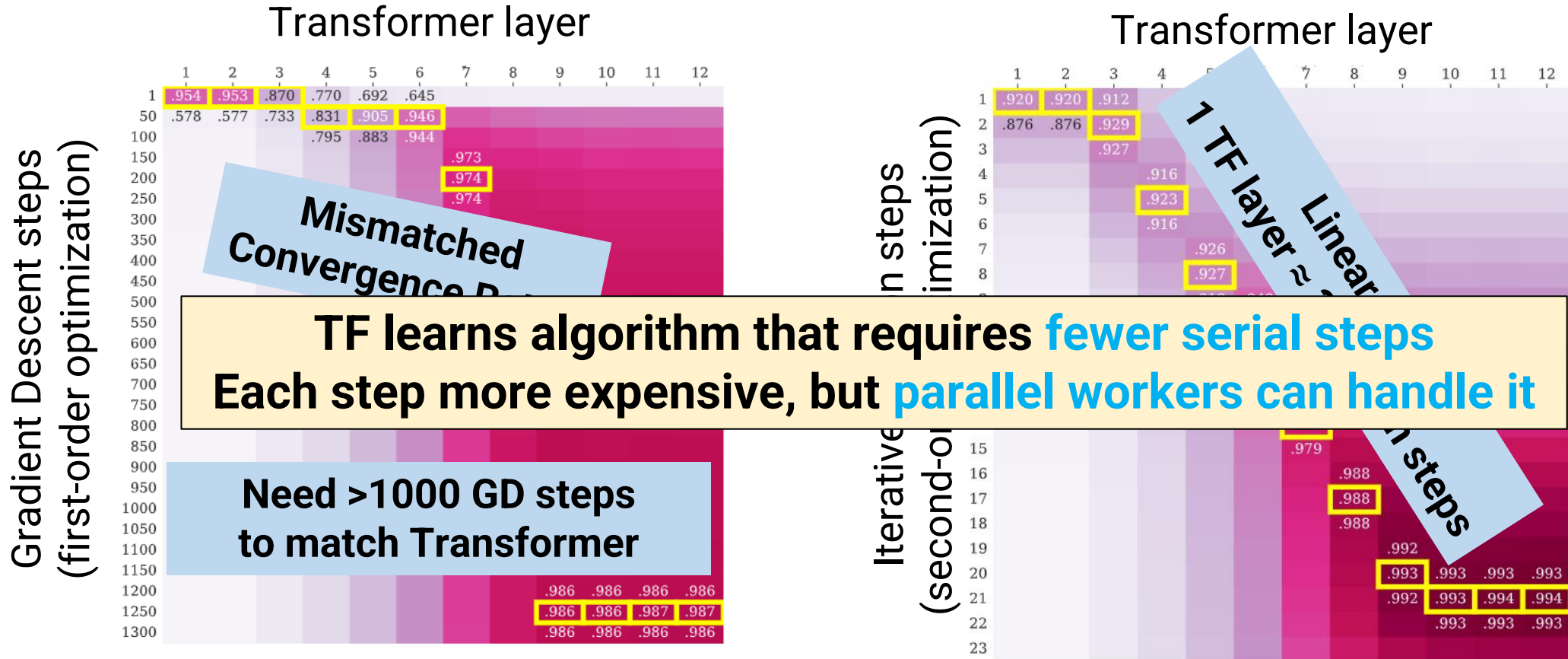
Arvind Mahankali Stanford University amahanka@stanford.edu	Tatsunori B. Hashimoto Stanford University thashim@stanford.edu	Tengyu Ma Stanford University tengyuma@stanford.edu
--	---	---

Transformers learn to implement preconditioned gradient descent for in-context learning

Kwangjun Ahn [*] MIT EECS/LIDS kjahn@mit.edu	Xiang Cheng [*] MIT LIDS chengx@mit.edu	Hadi Daneshmand [*] MIT LIDS/FODSI hdanesh@mit.edu	Suvrit Sra TU Munich / MIT suvrit@mit.edu
---	--	---	---

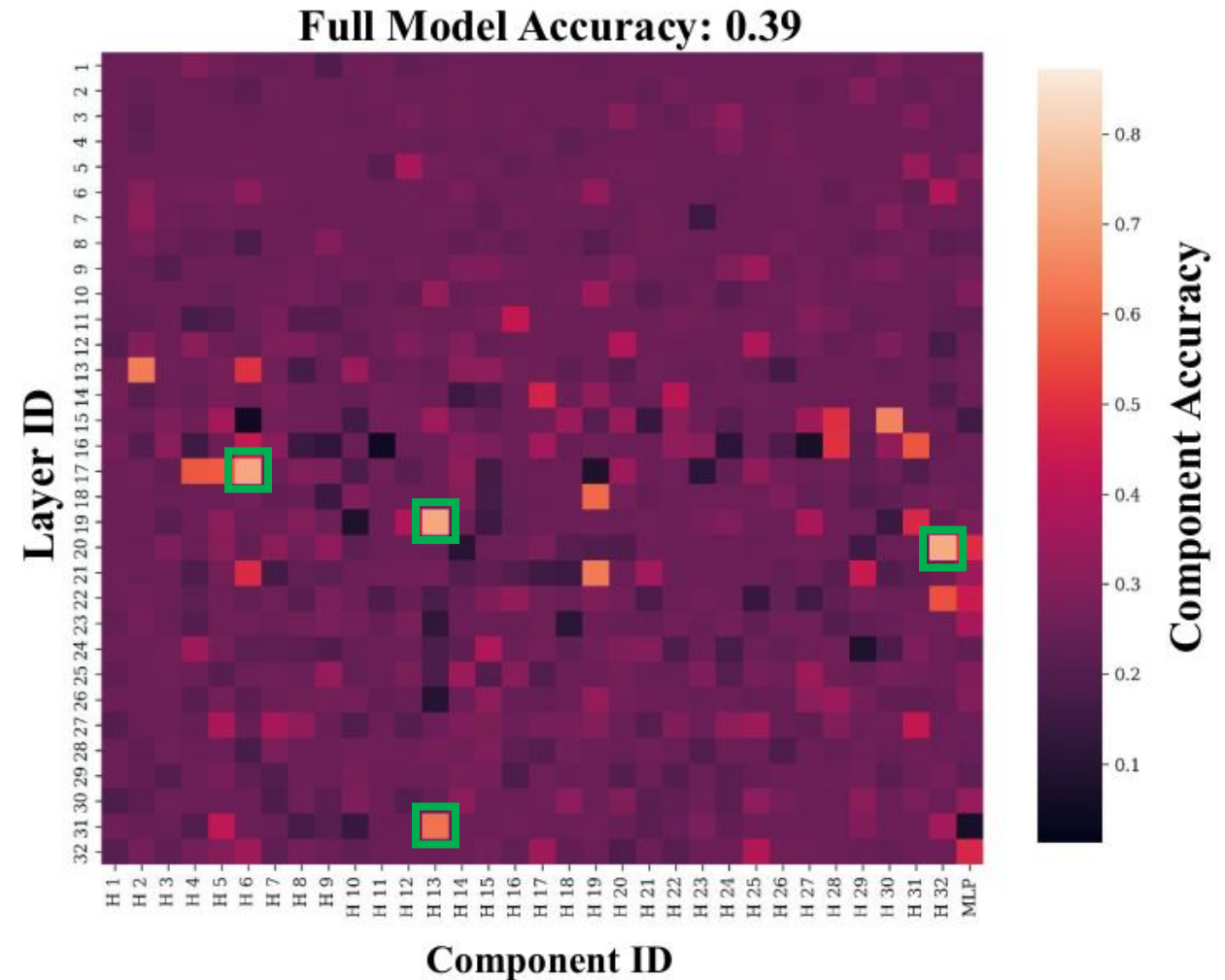
Parallel over Serial

Similarity between learned models & predictions from different TF layers



When Does Society Trust the Experts?

- Setting: In-context learning
- Some individuals are **experts**
- Most individuals are unable to do the task

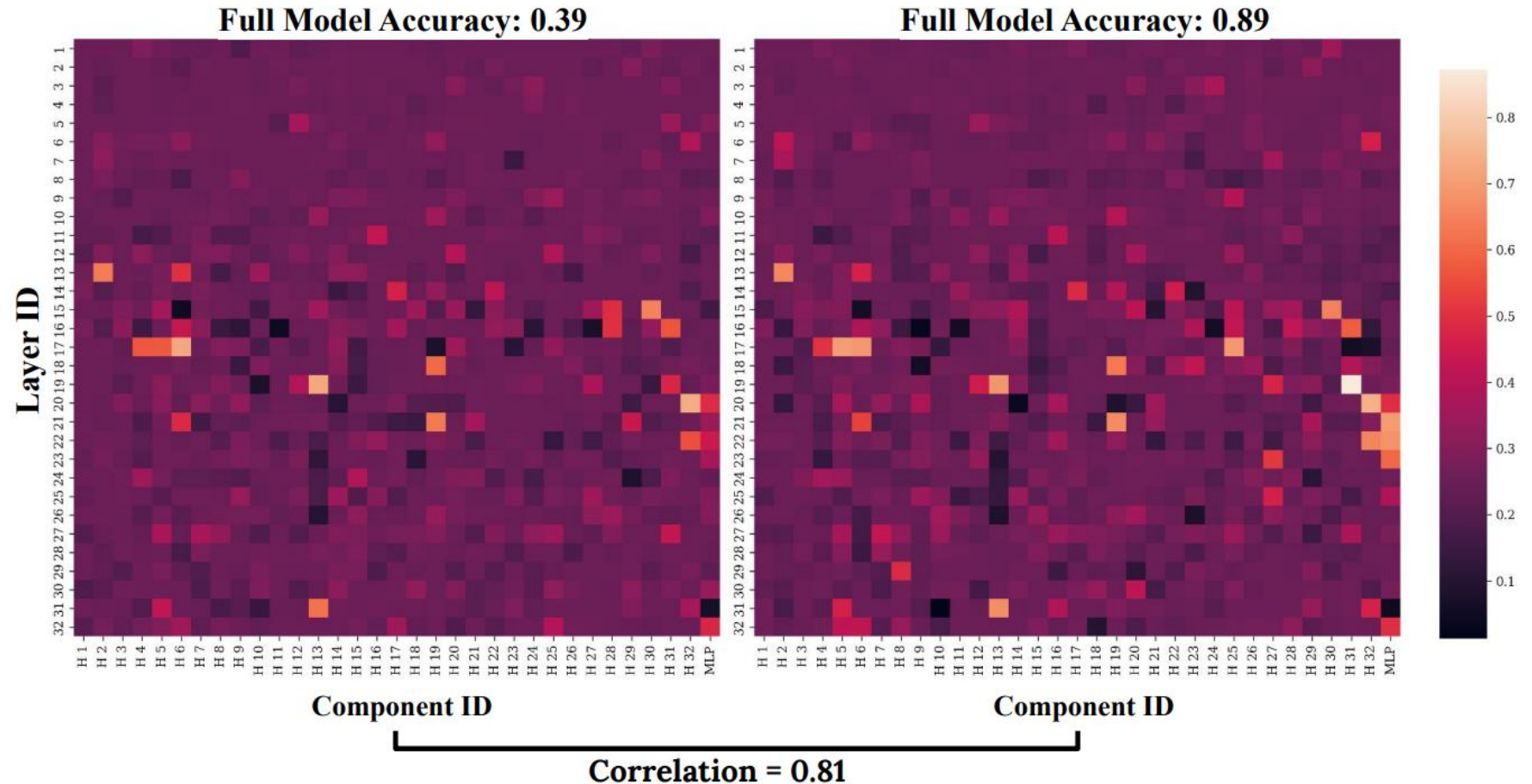


When Does Society Trust the Experts?

- Shown: Two prompts that differ only by whitespace
- Individuals have consistent behavior
- What changes:
Do the experts get drowned out?

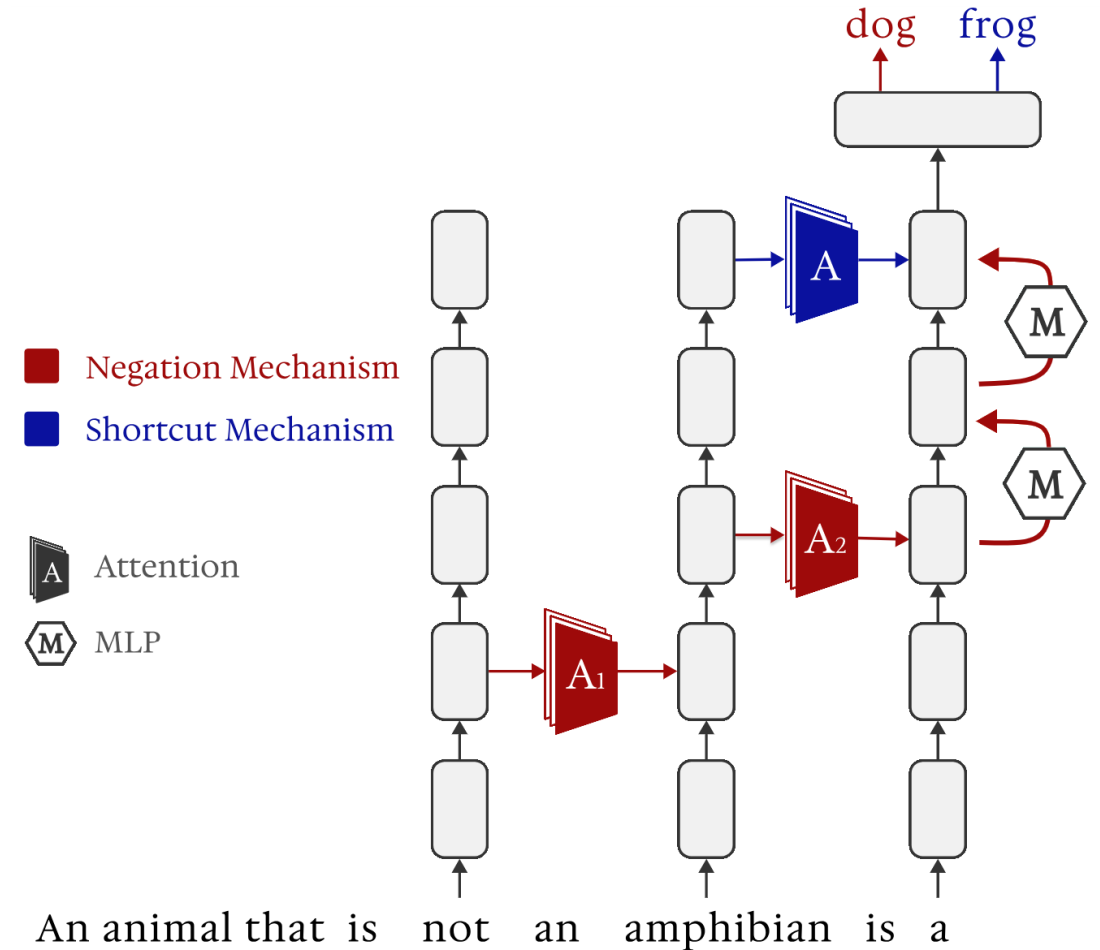
Template 1 {text} \nIs this a piece of news regarding World, Sports, Business, or Technology?

Template 2 {text} Is this a piece of news regarding World, Sports, Business, or Technology?



Cooperation and Competition

- We analyze how LLM's process negation
- **Some individuals** cooperate to compute negation
- **Other individuals** vote to ignore negation
- These two groups **compete for votes** in the residual stream!
 - Sometimes one or the other wins
 - Explains why **negation capability** can be hidden by **competing voters**



Are LLM's "Societies"?

- Explains **inconsistency** (parts of the society disagree)
- Natural division of labor leads to **parallelizable** mechanisms
- Societies can “make decisions” but dynamics are **unlike any one person's** decision making
- Hints at **emergent complexity** arising from simpler activities
- Hints at how **training data** is amalgamation of individuals' writings

Conclusion

- Interpretability can shape how “we” (the world) think about AI
- Metaphors are powerful
- The “society” metaphor captures many salient aspects of AI behavior
- We can aim for **impact outside of AI methods**

