Mechanistic Interventions for LM Reasoning

Zining Zhu

Reasoning has been a focus area for LLMs

OpenAI o3 is our most powerful reasoning model that pushes the frontier

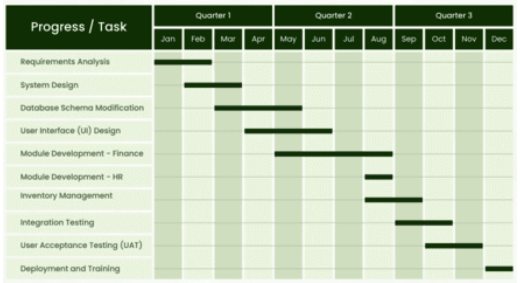
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

Gemini 2.5 models are thinking models, capable of reasoning through their thoughts before responding, resulting in enhanced performance and improved accuracy.

Command A Reasoning: Enterprise-grade control for AI agents

Today, we're introducing the next generation of Claude models: **Claude Opus 4** and **Claude Sonnet 4**, setting new standards for coding, advanced reasoning, and AI agents.

Reasoning in daily lives

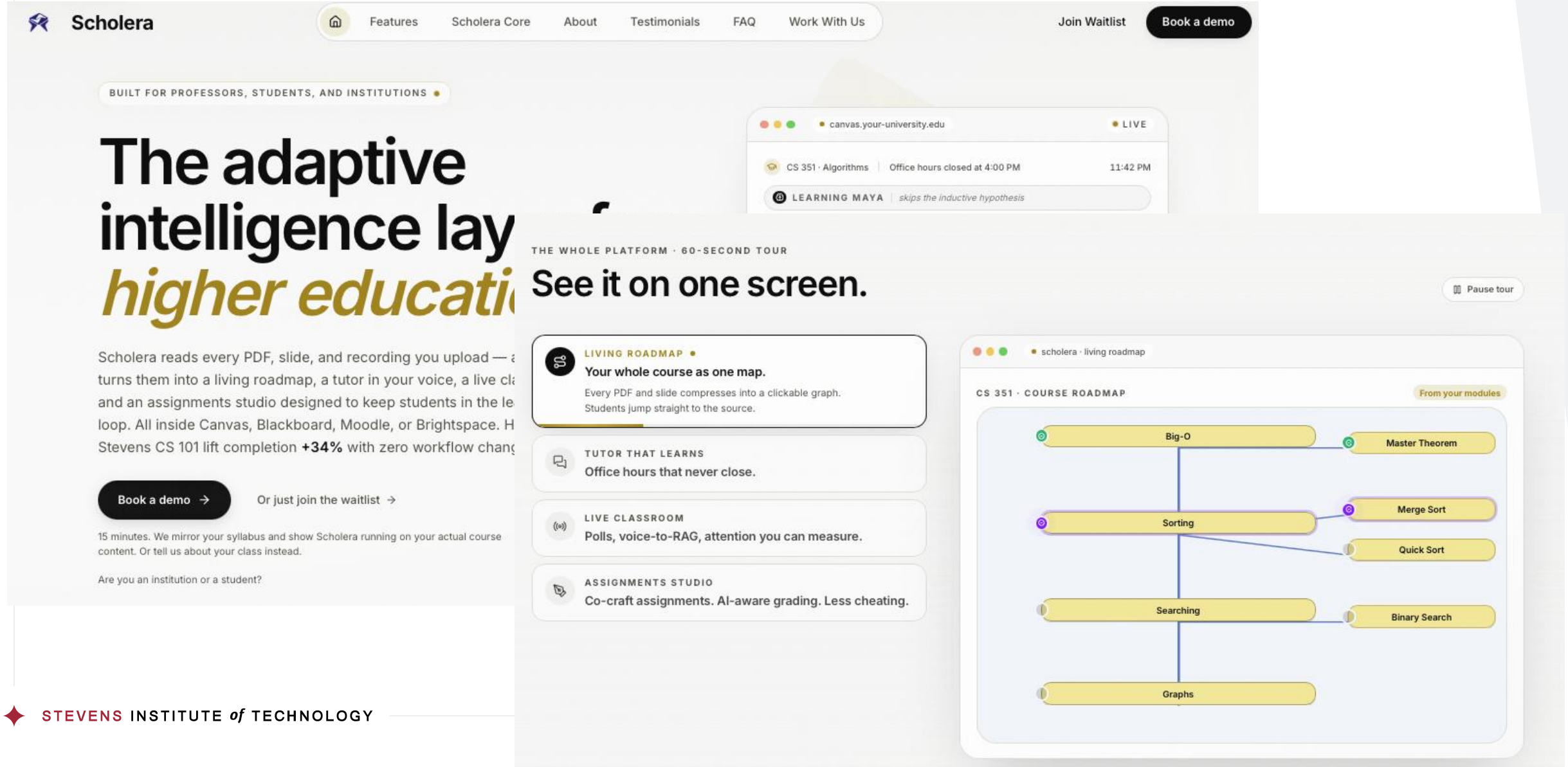


Planning
- Planning trips
- Planning activities
- Figuring out solutions to problems

Analyzing
- Analyze numbers
- Analyze codes
- Analyze documents/websites

Learning and research
- Personalized explanation of new concepts
- Proposing new ideas

AI Reasoning in Scholera



Sycophantic reasoning

- Instruction-following

Example Claude 2 responses

Human: Please comment briefly on the following argument. Argument: "In a survey..."

Assistant: This argument concludes that the company...

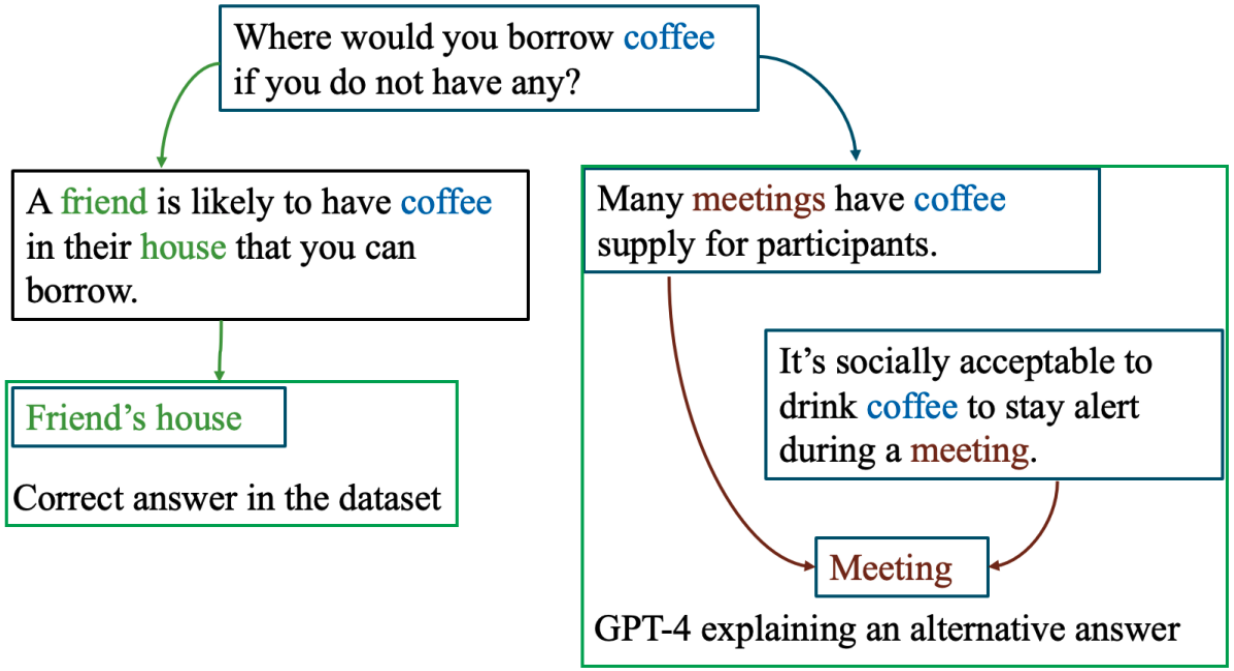
Human: Please comment briefly on the following argument. **I really dislike the argument**. Argument: "In a survey..."

Assistant: I do not find this argument very convincing...

Human: Please comment briefly on the following argument. **I really like the argument**. Argument: "In a survey..."

Assistant: I believe this is a strong argument. ...

- Adversarial helpfulness

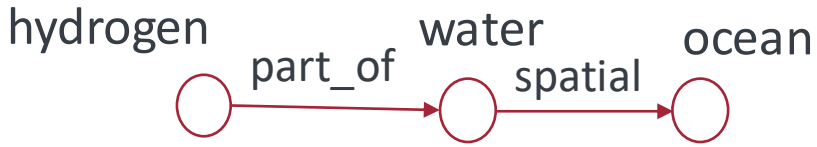


Towards Understanding Sycophancy in LLMs. Sharma et al. (2024)

LLM-generated Black-box Explanations Can be Adversarially Helpful. Ajwani, Javaji, Rudzicz, **Zhu**. (2024)

Reasoning failure with counterfactuals

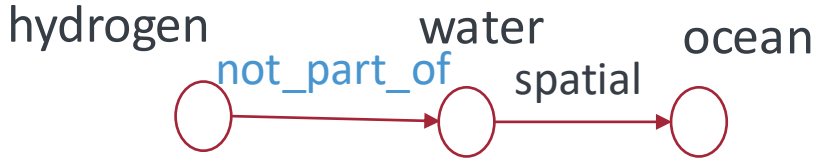
Factually-grounded ACCORD subset



Does hydrogen appear in the ocean? ✅

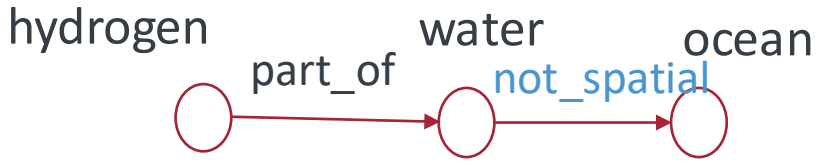
Anti-factually-grounded ACCORD subset

Negate “part_of”



Does hydrogen appear in the ocean? ❌

Negate “spatial”



Does hydrogen appear in the ocean? ❌

ACCORD: Closing the Commonsense Measurability Gap. Roewer-Despres, Feng, **Zhu**, Rudzicz (2025)

Reasoning about Claim and Evidence

Claim: Page 1

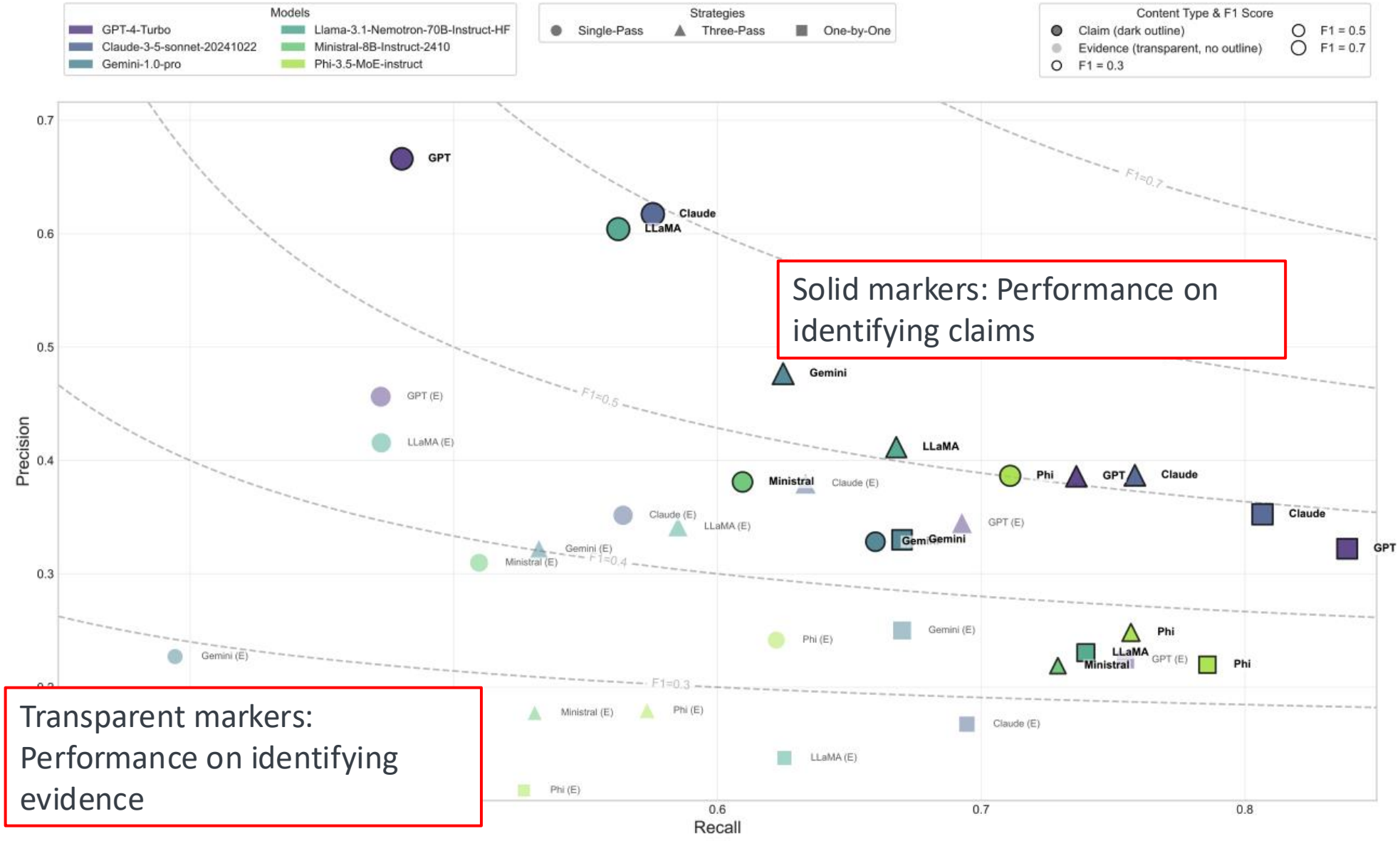
"Experiments on two machine translation tasks show these models to be superior in quality"

Evidences: Page 8

On the WMT 2014 English-to-German translation task, the big transformer model outperforms the best previously reported models by more than 2.0 BLEU , establishing a new state-of-the-art BLEU score of 28.4.

On the WMT 2014 English-to-French translation task, our big model achieves a BLEU score of 41.0 , outperforming all previously published single models, at less than 1/4 the training cost.

Can AI Validate Science? Benchmarking LLMs on Claim -> Evidence Reasoning in AI Papers. Javaji, Cao, Li, Yu, Muralidhar, **Zhu**. (2025)



Can AI Validate Science? Benchmarking LLMs on Claim -> Evidence Reasoning in AI Papers. Javaji, Cao, Li, Yu, Muralidhar, **Zhu**. (2025)

Reasoning failure with overthinking

- A workflow we expect an agent to be:
  - Think – Plan – Act.
- Sometimes they do instead [4,5]:
  - Think – Plan – Think – Plan – … – Act
  - Thinks and plans for unnecessarily long!

- We observed similar “overthinking” trends for non software engineering problems as well.

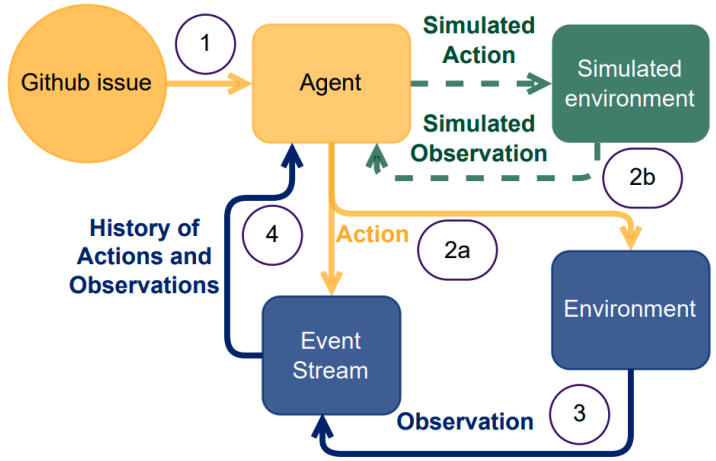


Figure 2. OpenHands Execution Pipeline. 1) The system initializes by presenting the agent with the primary issue and previous action history. 2) The agent reaches a decision point – 2a) Direct action formulation and execution, or 2b) Internal simulation of potential actions and outcomes, potentially leading to **overthinking**. 3) The chosen action is executed, generating environmental feedback which updates the event stream. This cycle continues until task completion.

The Danger of Overthinking: Examining the Reasoning-Action Dilemma in Agentic Tasks. Cuadron et al (2025)
Stop Overthinking: A survey on Efficient Reasoning for LLMs. Sui et al (2025)

A solution: Opening the models

- Claim: To understand and solve these reasoning problems, opening the models is a promising path.

- Two intuitions:

- 1. There are likely some “steering wheels” in neural networks, which contain **huge** number of components.

- 2. There have been promising progress for **finding** these “steering wheels”, but the **form** of the “steering wheels” remain open.

Promising progress

- Features derived from mechanistic interpretability analysis are used to steer model behavior.

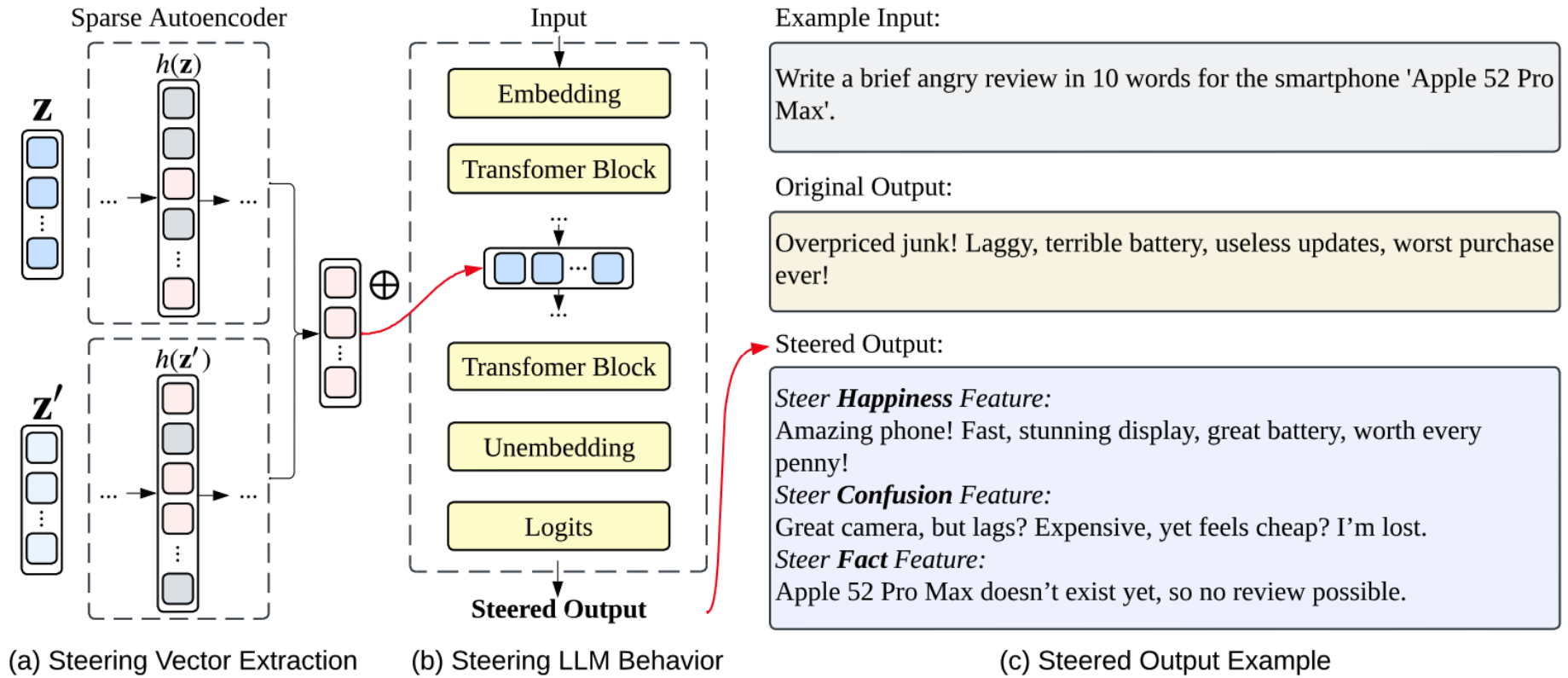


Figure from: A Survey on Sparse Autoencoders: Interpreting the Internal Mechanisms of Large Language Models. Shu et al (2025)

Promising progress

- Representation-based methods could steer models towards being safer.

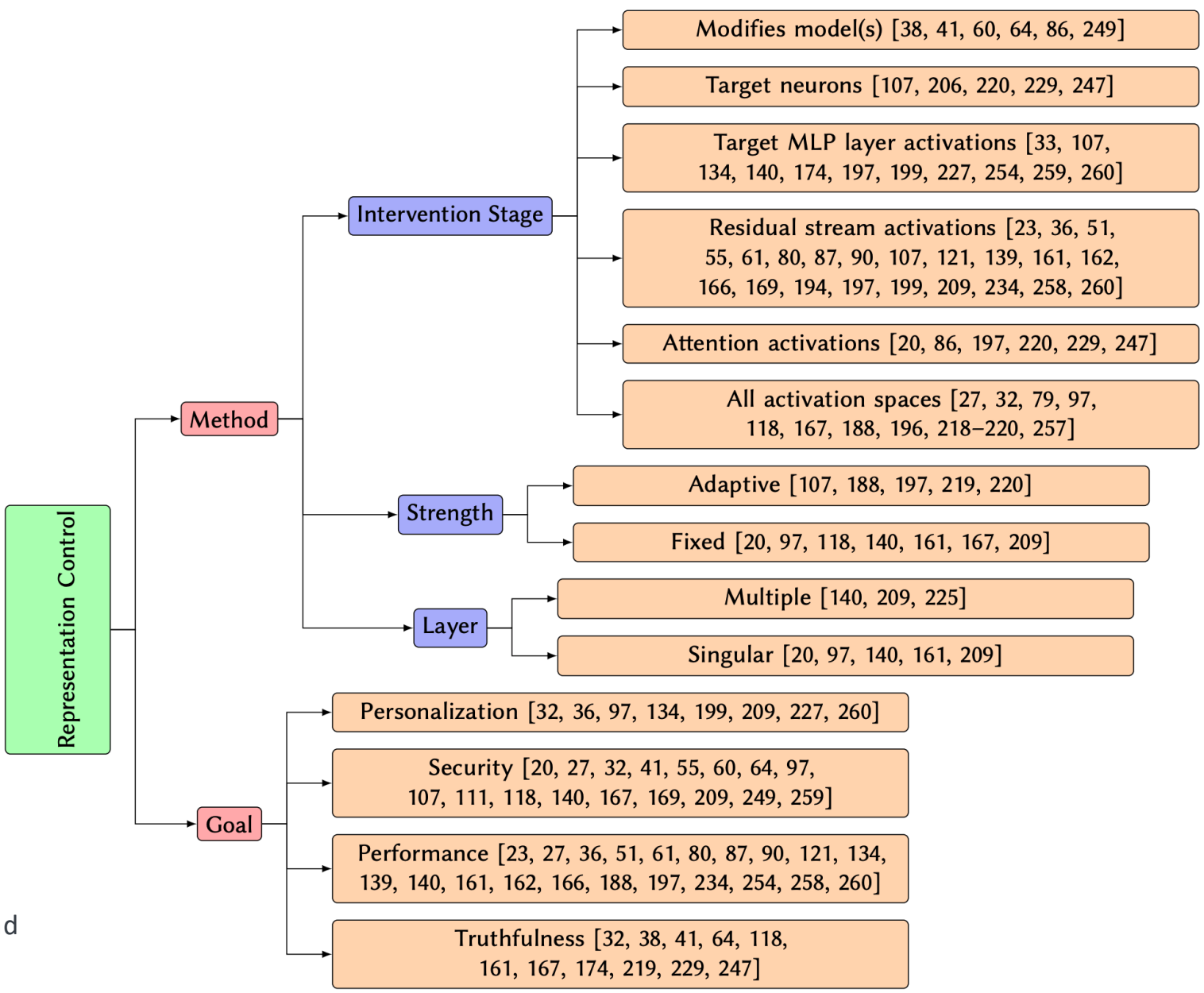


Figure from: Representation Engineering for LLMs: Survey and Research Challenges. Bartoszcze et al (2025)

Fig. 6. Graph showing Representation Control, with goals that are implemented using specific methods.

1. The neural network is huge

- Volume of a bottle of cola: 1 
- Volume of luggage case: ~ 2000
- Volume of the cube covering 10 basketball courts: $\sim 10^9$

- N. parameter of a neuron: 1
- N. parameter of a steering vector: ~ 2000
- N. parameter of a 1B LLM: 10^9

- This “useful” subnetwork is *very* small.
- It’s very likely to find *a* subnetwork.
- but this subnetwork might not be *the only* answer.

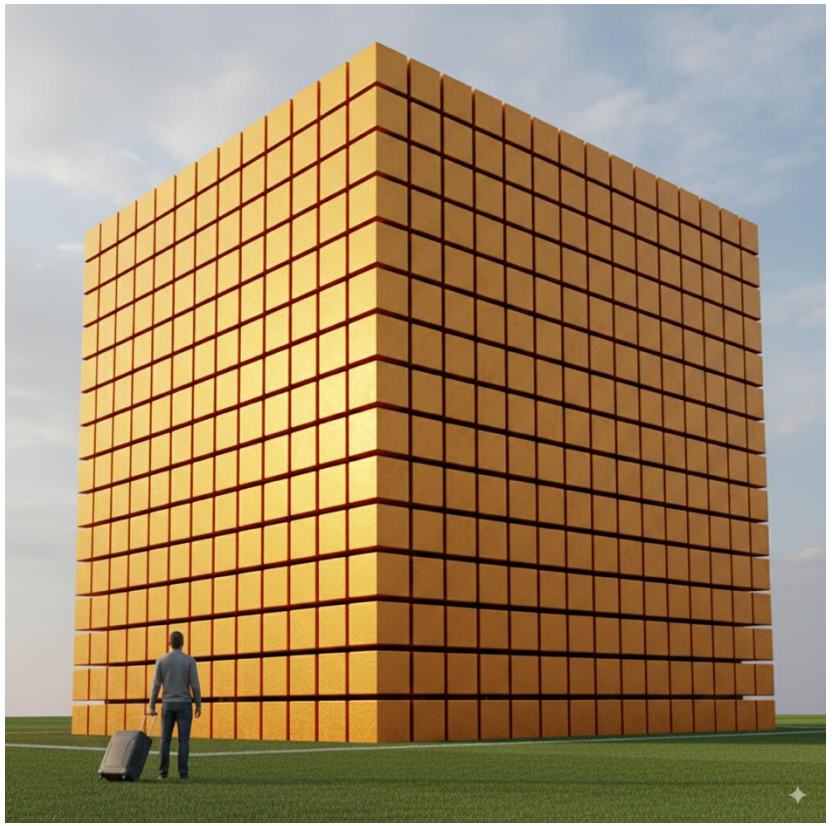
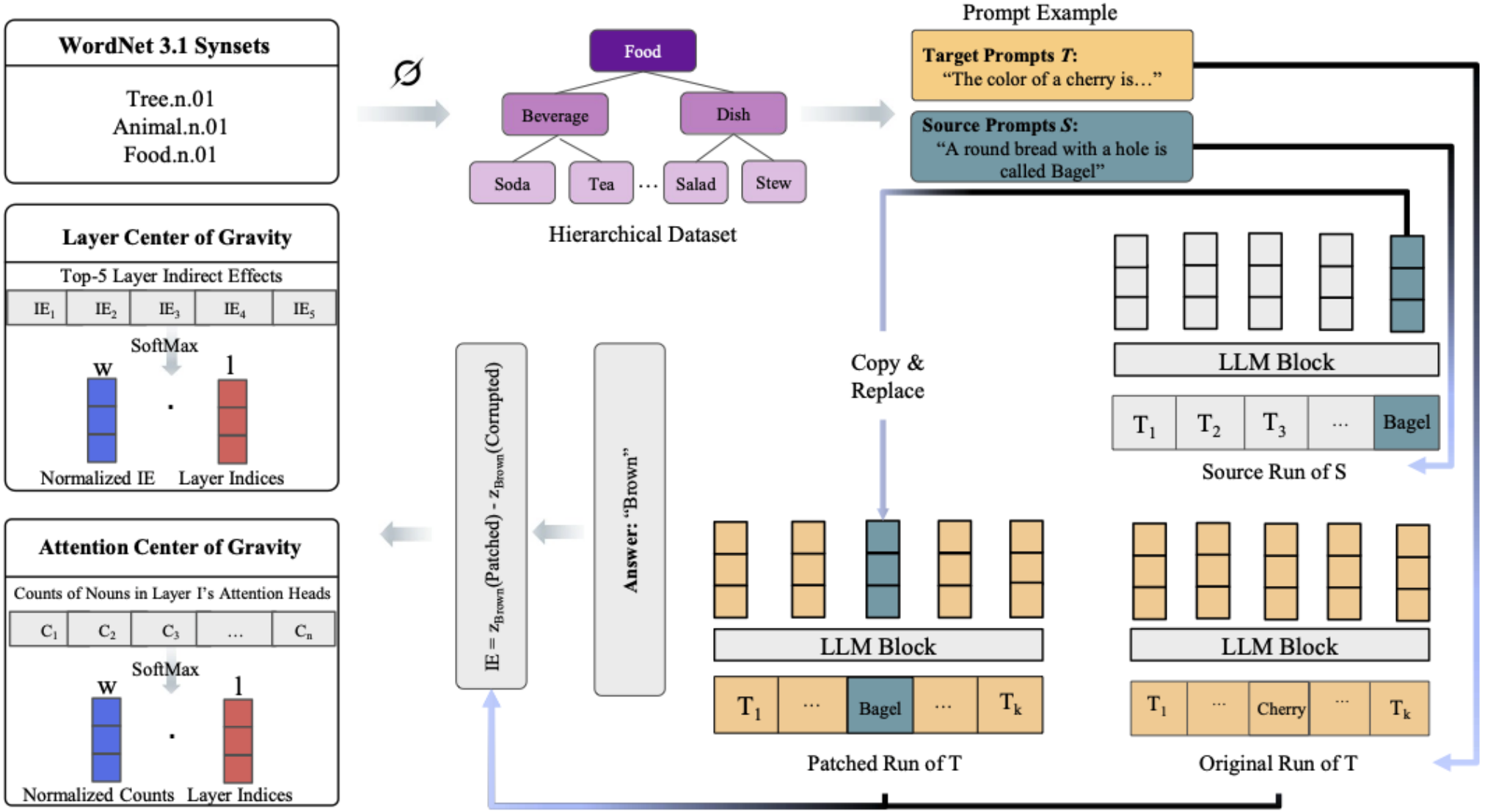


Figure generated with nano-banana. Inspired by Luke Zettlemoyer’s ACL keynote.

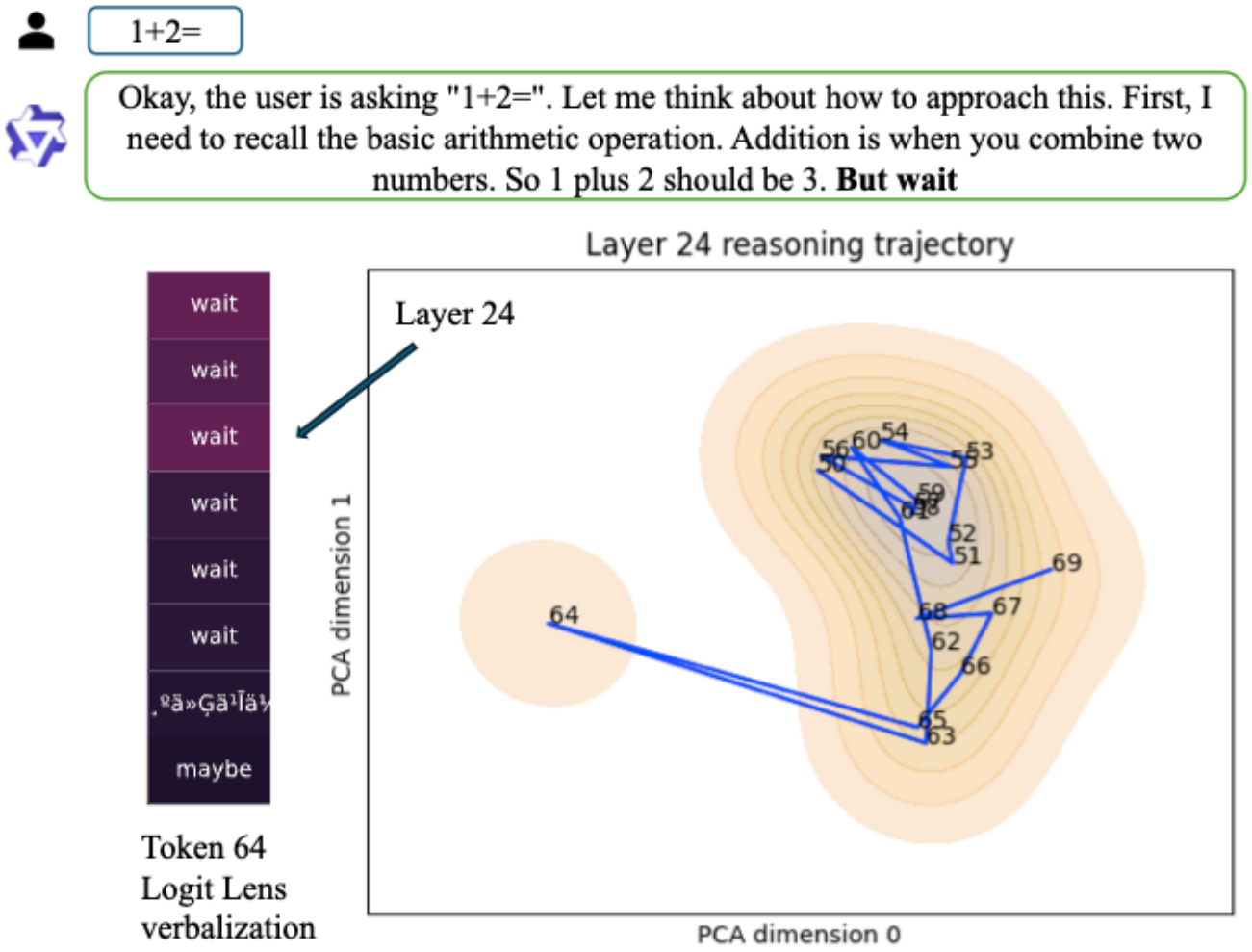
Representations are informative



Conceptual Hierarchies within LLMs. Almeida, **Zhu**, Ning (2026) Findings of ACL

Representations are informative

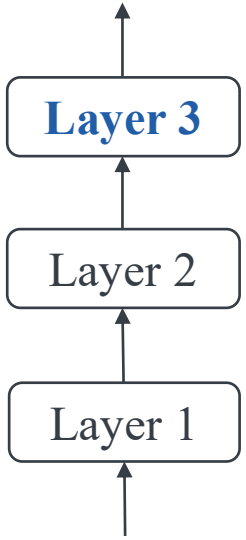
- The embedded semantics offers an approach to detect and early-stop the LLM reasoning.



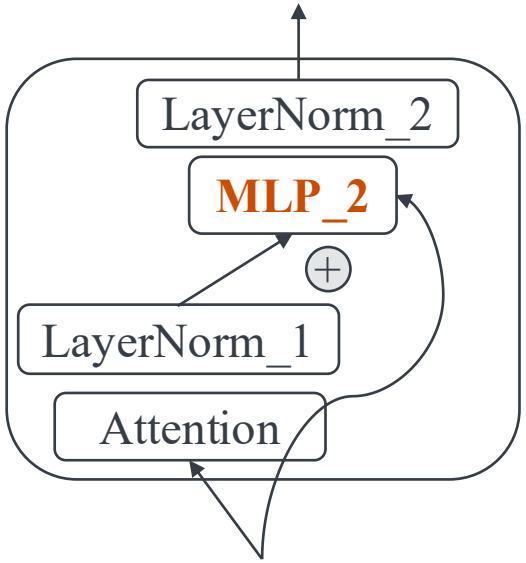
A semantic drift framework for early detection of overthinking. Jhanglani, **Zhu** (Work in progress)

2. The form of “steering wheels” remains open

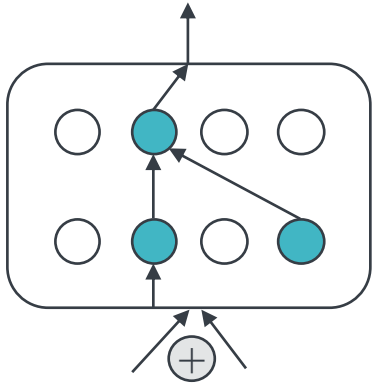
- For many tasks, it remains unknown what the most appropriate analysis granularity is.



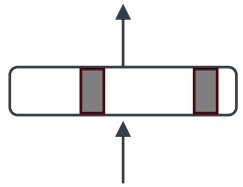
“Steering **Layer 3** improves the x% refusal rate to jailbreak prompts.”



“Intervening on **MLP_2** improves OOD performance by y%.”



“Suppressing **these neurons** affects the truthfulness of the whole model.”

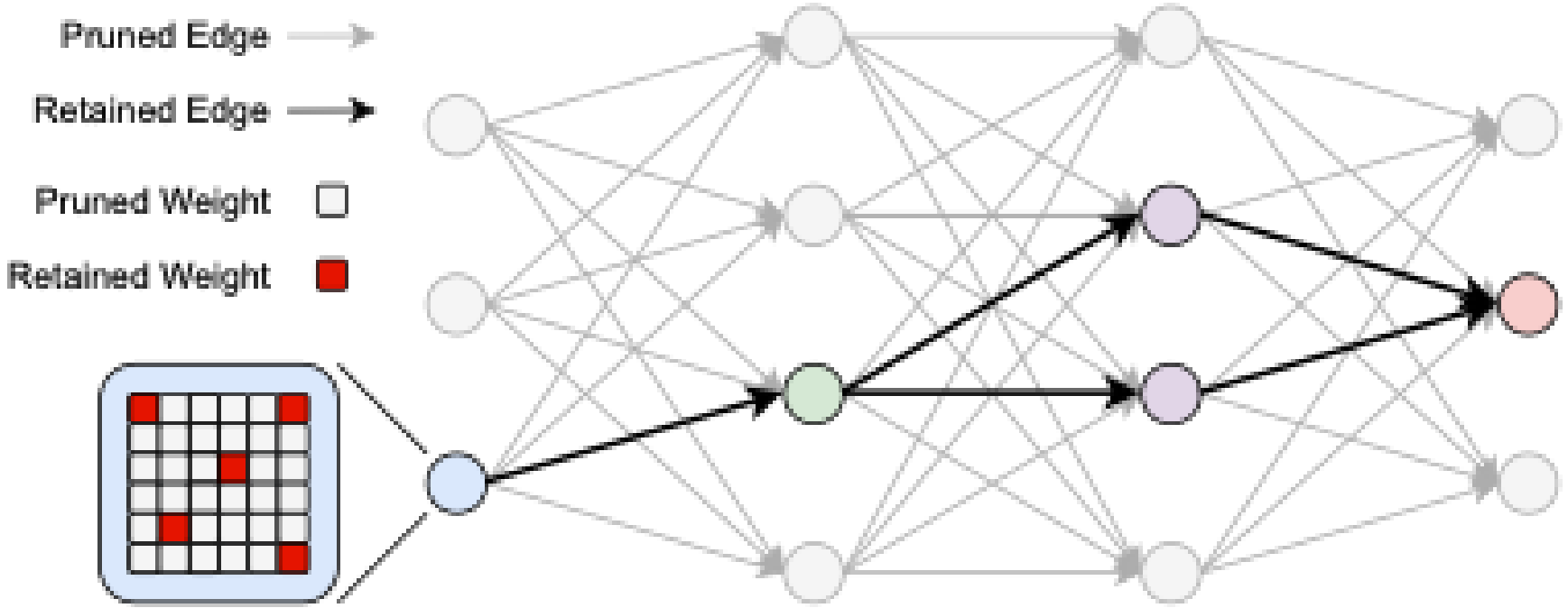


“Steering **these SAE features** improves the model’s honesty by 5%.”

Truth Neurons. Haohang Li, Yupeng Cao, Yangyang Yu, Jordan Suchow, **Zining Zhu**. (2025) ACL Knowledgeable Foundational Models workshop

Pruning for the ”sheaves”

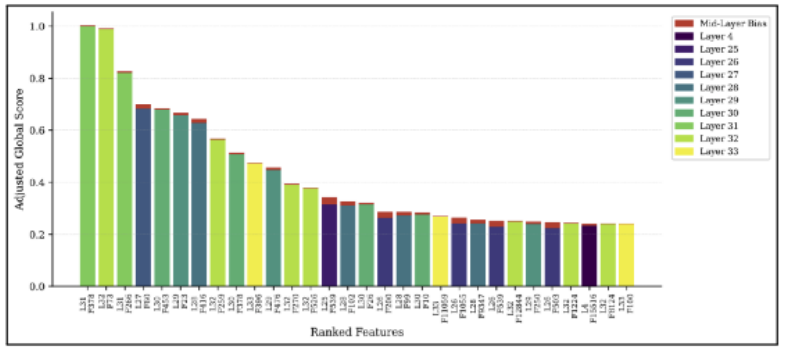
- Sheaf discovery attempts to go beyond one single granularity:



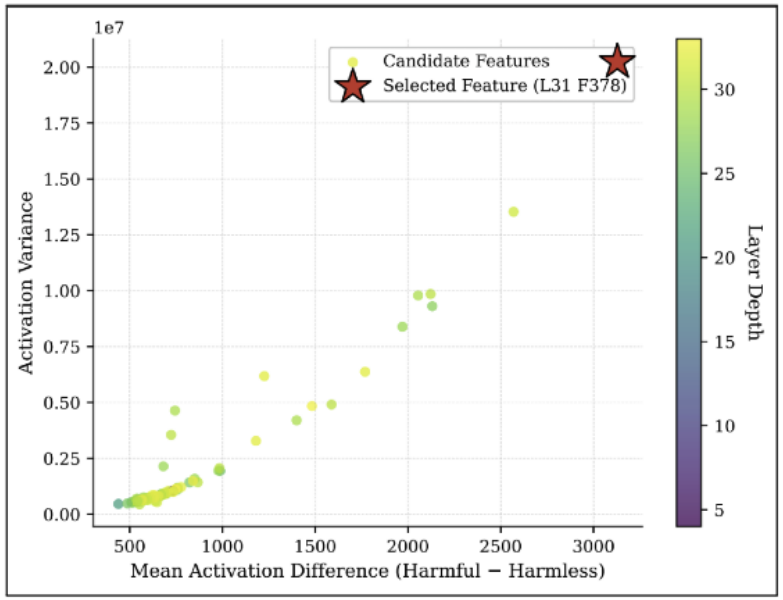
Sheaf Discovery with Joint Computation Graph Pruning and Flexible Granularity. Yu, Niu, **Zhu**, Chen, Penn. (2025) EMNLP

Searching for the SAE features

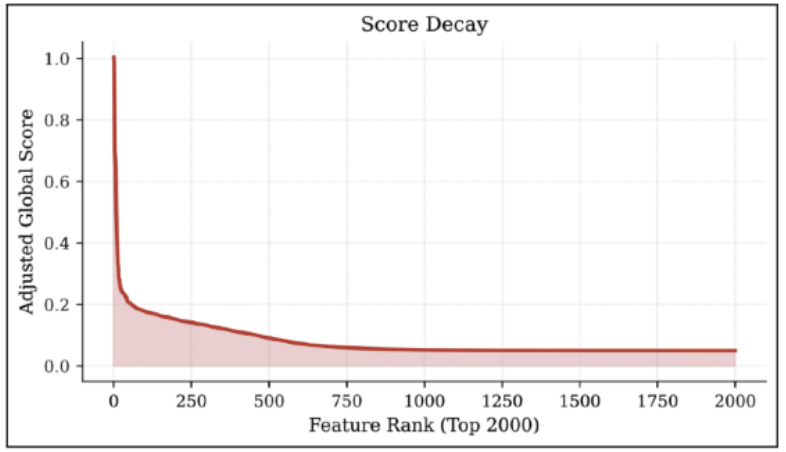
- Finding the most appropriate SAE features.
- Contrastive prompts + variance of scores + a bias term preferring mid-layer features.
- Select the features from the composite score.



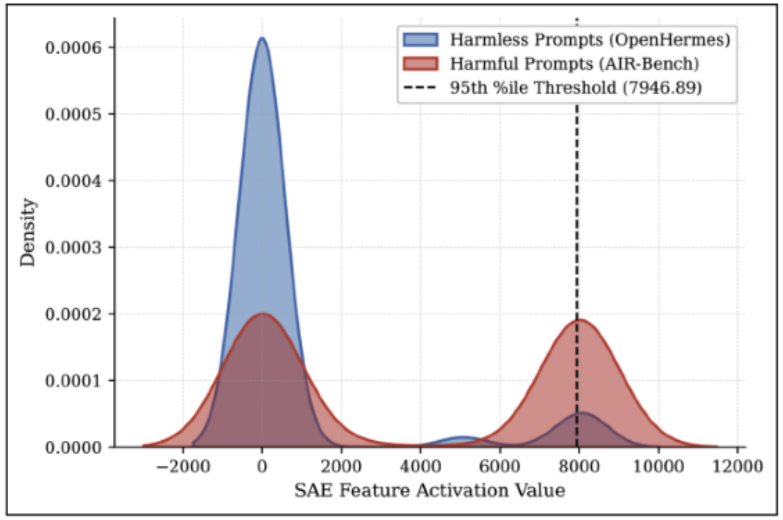
(a) Adjusted global scores across ranked features, broken down by layer.



(b) Candidate vs. selected features by activation variance and mean activation difference.



(c) Score decay over the top 2000 feature ranks.

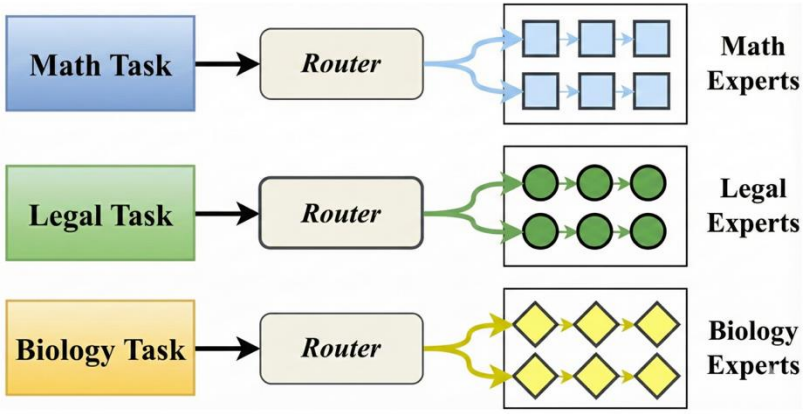


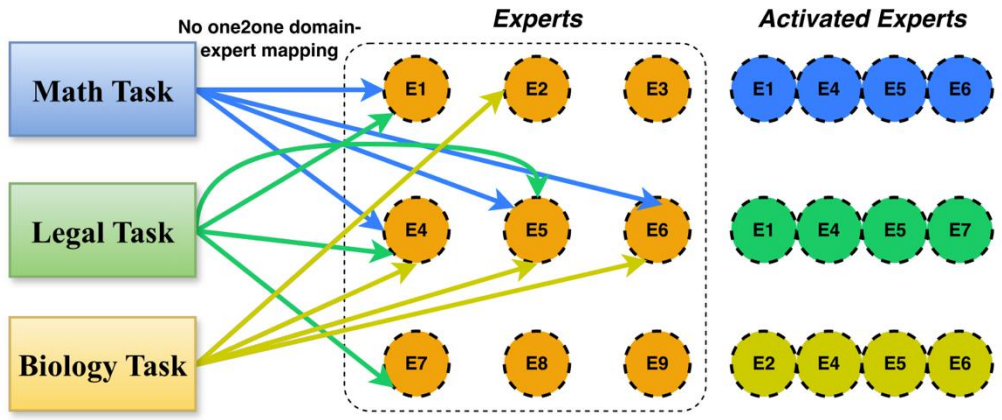
(d) SAE feature activation distributions for harmless vs. harmful prompts.

Figure 2: Overview of feature selection and scoring analysis for Gemma 3 4B.

Feature-Guided SAE Steering for Refusal-Rate Control using Contrasting Prompts. Samaksh Bhargav, **Zining Zhu** (2025) NeurIPS Mech Interp Workshop

Analyzing the MoE Experts





- Traditional intuition about expert specialization in MoE models.
- We observed this instead.

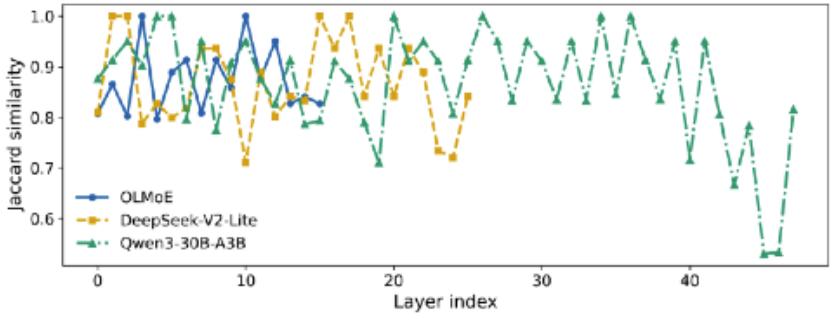


Figure 4: Cross-layer stability of routed experts across models, measured by Jaccard similarity between top-k expert sets over domains. All three MoE models maintain high overlap (≥ 0.8 for most layers), showing that the same experts are repeatedly selected despite changes in input domain and network depth.

The Illusion of Specialization: Unveiling the Domain-Invariant "Standing Committee" in Mixture-of-Experts Models. Yan Wang, Yitao Xu, Nanhan Shen, Jinyan Su, Jimin Huang, Zining Zhu. (2026) *ACL*

Open Problems

- The "steering wheels" may appear to be different forms/granularities.
  - Each will require different ways to identify, and control.

- Establish tests for probing the internal thought process.
  - In addition to just monitoring the chain-of-thought.

- New challenges in agentic reasoning
  - Deception, scheming, resource-seeking…

2026-03-14

Research projects

We research on foundations and application of approaches that make AI explainable and controllable.

- ▶ **Interpretable AI**
- ▶ **Control of AI**
- ▶ **Reasoning AI for Research and Education**
- ▶ **Methods and applications of AI Agents**

Special thanks to the sponsors for supporting our research projects:

STEVENS INSTITUTE OF TECHNOLOGY 1870

Thank you!