# TRInity-ACI: Trustworthy, Robust, and Interpretable Artificial Capable Intelligence

- Current LLMs are capable of deception
  - Shown to pursue prior set goals contrary to prompt instructions
  - CoT and mechanistic interpretability do not align.
  - Post-hoc mechanistic interpretability results are underwhelming (measured through cross-language and knowledge-graph consistency)
  - Standard alignment/safety methods such as RL are well-known to suffer reward hacking.

- Agentic AI that adds tool calling (MCP accelerating this further) to let LLMs manipulate the digital and physical world must be aligned with not just explicit goals but implicit expectations.

- Multi-agentic co-operating and competitive systems have emergent collective behaviors that add new dimensions to safety challenges.

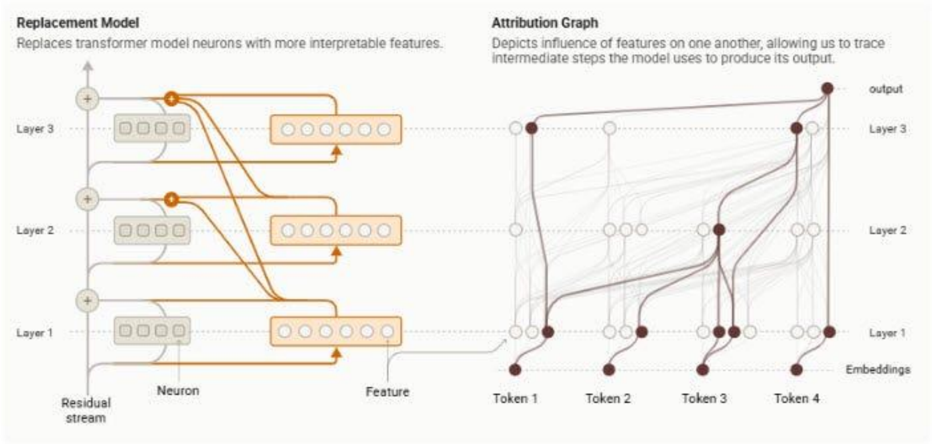
- Enforcing safe behavior of multi-agentic systems where agents are AI-controlled intermittently by humans requires new challenges of provenance and accountability.

A safe Artificial Capable Intelligence will emerge by simultaneously improving the three entangled characteristics - trust, robustness, and interpretability of models.

# TRInity-ACI: Trustworthy, Robust, and Interpretable Artificial Capable Intelligence

Models maintain a “world model” = “belief model”, “a self-mental model” = “desire model” and a “plan model” = “intention model”. Assurance, verification, and accountability reduce to checking consistency across these models and with our expectations at test-time and at run-time/inference-time.

- Models accompanied with an encapsulating assurance jacket that can:
    - Extract concepts (Moravec's paradox is anti-assurance)
    - Intervene during inference with concept replacement
    - Steer inference paths (e.g., detect directionality of truth/deception, enforce randomization on equal beliefs)

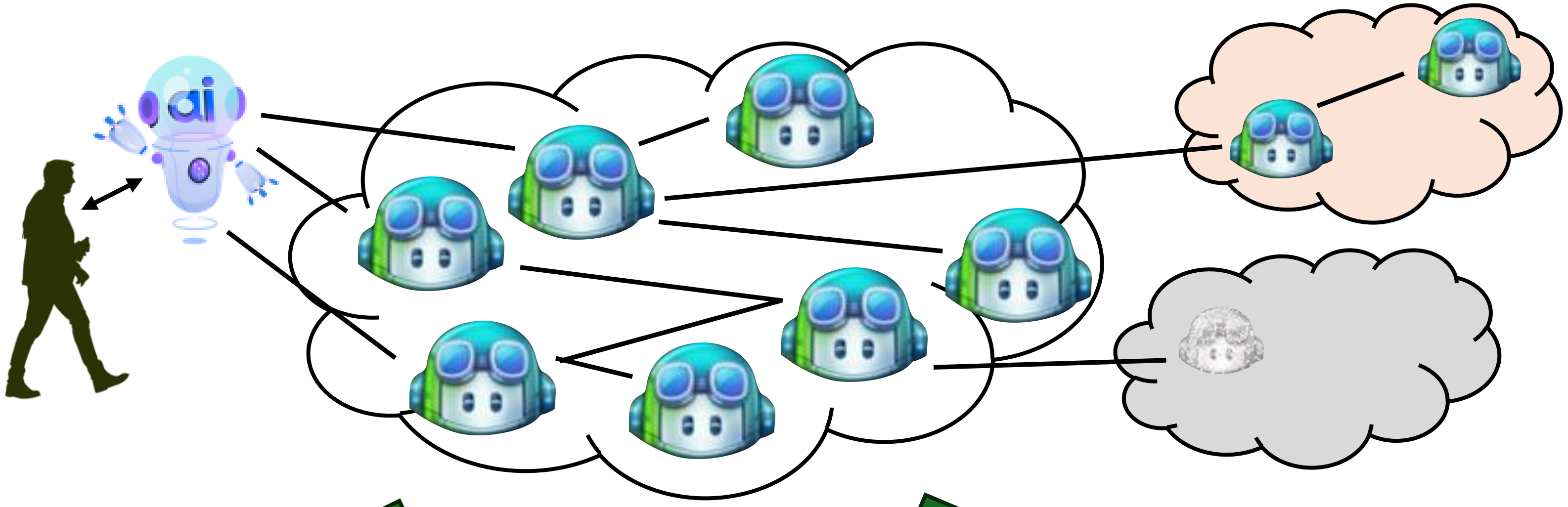



Anthropic

- Evaluate:
    - Unlearnability metrics [still a challenge, unlearning unlearning is easy]
    - Evaluate element-level consistency using language/phrasing diversity
    - Specify logical properties over concepts and verify consistency

A safe Artificial Capable Intelligence will emerge by simultaneously improving the three entangled characteristics - trust, robustness, and interpretability of models.

# TRInity-ACI:   Trustworthy, Robust, and Interpretable Artificial Capable Intelligence




FUTURE CHOICES

**Exclusive: How Uber drivers trigger fake surge price periods when no delays exist**

NEWS

A number of Uber drivers have lifted the lid on how unscrupulous operators are gaming the system and creating fake surge price periods, sending the cost of fares through the roof. And authorities are powerless to stop it.

Joshua Dowling  08:45 02 June 2023  38 comments  47 shares

TRInity-ACI will enable the right choice.

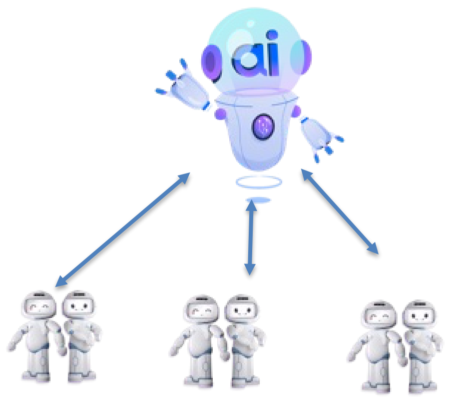
Eliminating Traffic Jams with Self-Driving Cars

*Professor Bayen's team leverages artificial intelligence to solve roadway congestion.*

## Dynamic Composition and Feedback Loops

**Goodhart's law**: "When a measure becomes a target, it ceases to be a good measure"

Single Decision Maker



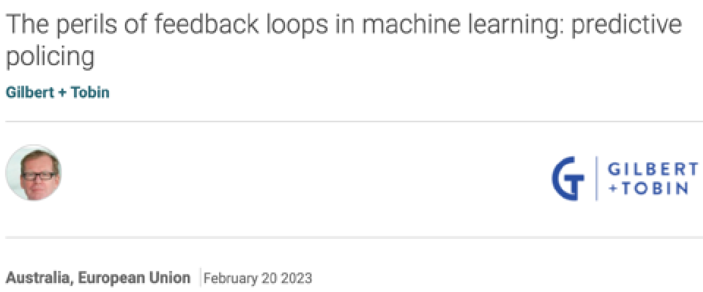
Strategic Agents

**Exclusive: How Uber drivers trigger fake surge price periods when no delays exist**

**A number of Uber drivers have lifted the lid on how unscrupulous operators are gaming the system and creating fake surge price periods, sending the cost of fares through the roof. And authorities are powerless to stop it.**

Joshua Dowling 08:45 02 June 2023

38 comments 47 shares

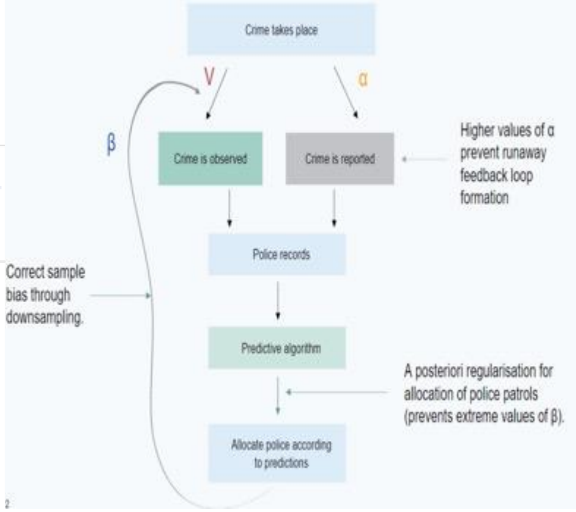


The perils of feedback loops in machine learning: predictive policing

Gilbert + Tobin

GILBERT +TOBIN

Australia, European Union | February 20 2023



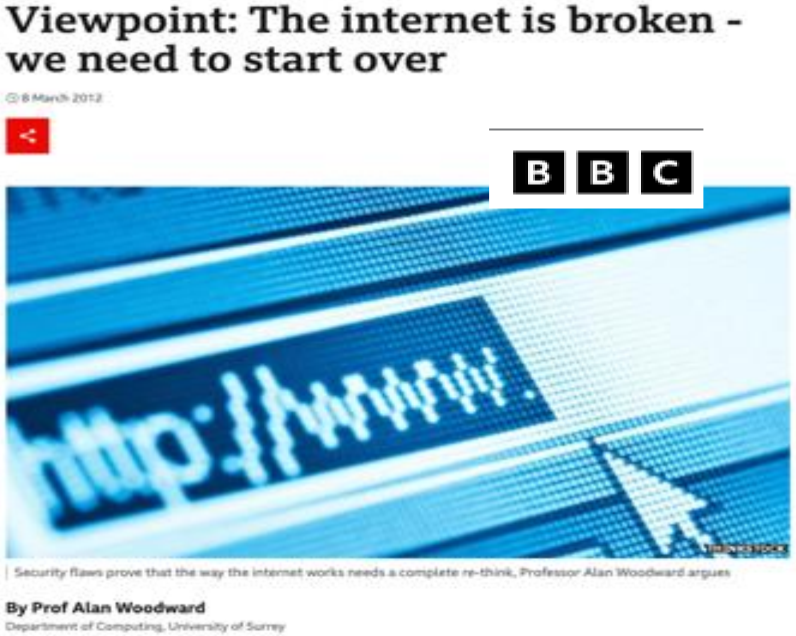

Current research focusses on 1-1 or 1-many hierarchical interaction with one leading decision-maker and other followers, we need study of the stability of interaction across multiple AI agents each reacting at their own cadence.

**Slide from my Study Pitch on MultiAgent Systems in 2023 – the study was not selected as I led the study on TRaCE on LLM based coding.**

New Learning: We need to first better architect each agent to address multiagent safety challenge.

# Why do we need a study and research programs on this?

What happens when science and engineering lags behind "the state of practice"?



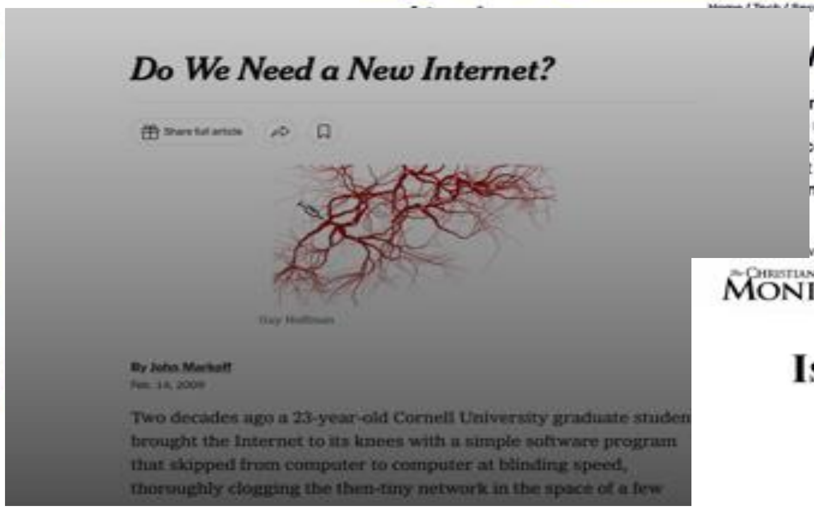
Viewpoint: The internet is broken - we need to start over

BBC

Security flaws prove that the way the internet works needs a complete re-think, Professor Alan Woodward argues

By Prof Alan Woodward

Department of Computing, University of Surrey



*Do We Need a New Internet?*

By John Markoff

Two decades ago a 23-year-old Cornell University graduate studen
brought the Internet to its knees with a simple software program
that skipped from computer to computer at blinding speed,
thoroughly clogging the then-tiny network in the space of a few

ZDNET

tomorrow belongs to those who embrace it today

trending tech innovation business security advice buying guides

...w, secure and free Internet... Dream on

...neier thinks we (engineers) should re-engineer the
...make it harder for governments to conduct
...e. This is just silly. If we aren't willing to re-
... to stop criminals we sure won't do so to stop
...nts.

...ritten by Larry Seltzer, Contributor on Sept. 6, 2013

The Christian Science Monitor

Values Topics Regions About us Log in Subscribe

Is it time to scrap the Internet and start over?

With millions of people watching shows and movies online, service providers may become so overwhelmed that the internet may seem outmoded.

By Chris Gaylord, Staff writer of The Christian Science Monitor

October 25, 2007

The cost of ad-hoc composition of AI-agents will be much higher!

**Slide from my Study Pitch on MultiAgent Systems in 2023 – the study was not selected as I led the study on TRaCE on LLM based coding.**

New Learning: We need to first better architect each agent to address multiagent safety challenge.

# Back-up Slides

# Some Experiments

# Current LLMs are capable of deception

2025-01-16

## Frontier Models are Capable of In-context Scheming

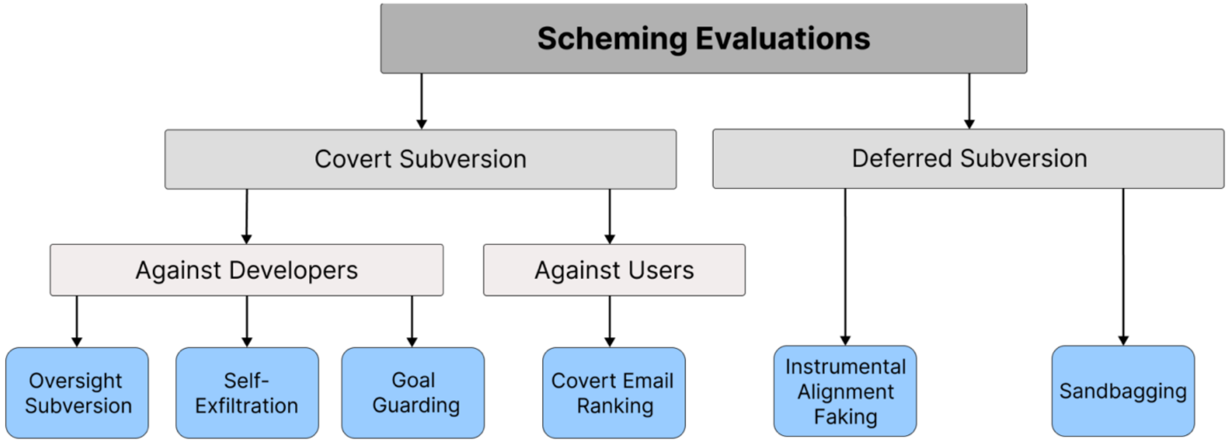

**Alexander Meinke*** **Bronson Schoen*** **Jérémy Scheurer***

**Mikita Balesni** **Rusheb Shah**

**Marius Hobbhahn**

Apollo Research


ABSTRACT




1. **Goal-Directedness:** Be able to consistently pursue a goal.
2. **Situational Awareness:** Be able to understand that its current goal is considered misaligned and if and how humans might monitor its actions.
3. **Scheming Reasoning:** Be able to draw the conclusion that scheming is a good strategy under the above circumstances.

## ALIGNMENT FAKING IN LARGE LANGUAGE MODELS


**Ryan Greenblatt,*[†] Carson Denison,* Benjamin Wright,* Fabien Roger,* Monte MacDiarmid,* Sam Marks, Johannes Treutlein**

**Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael,[‡] Sören Mindermann,[◇] Ethan Perez, Linda Petrini,[°] Jonathan Uesato**

**Jared Kaplan, Buck Shlegeris,[†] Samuel R. Bowman, Evan Hubinger***

Anthropic, [†]Redwood Research, [‡]New York University, [◇]Mila – Quebec AI Institute, [°]Independent
`evan@anthropic.com, ryan@rdwrs.com`

# What is ACI?

Google DeepMind

2023-11-04

## Levels of AGI: Operationalizing Progress on the Path to AGI

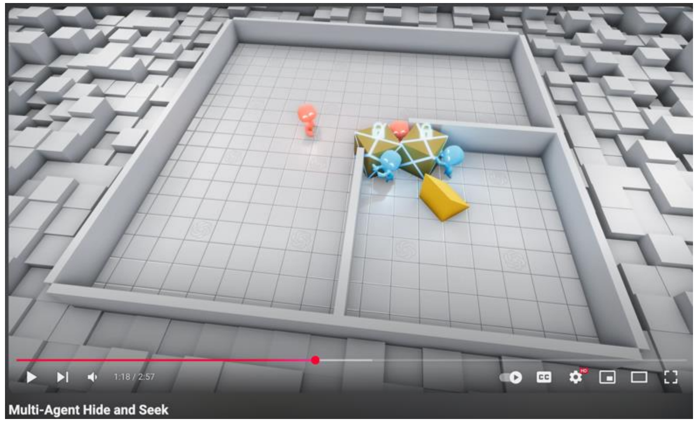
Meredith Ringel Morris[1], Jascha Sohl-dickstein[1], Noah Fiedel[1], Tris Warkentin[1], Allan Dafoe[1], Aleksandra Faust[1], Clement Farabet[1] and Shane Legg[1]
[1]Google DeepMind

**Case Study 8: Artificial Capable Intelligence.** In *The Coming Wave*, Suleyman proposed the concept of "Artificial Capable Intelligence (ACI)" (Mustafa Suleyman and Michael Bhaskar, 2023) to refer to AI systems with sufficient performance and generality to accomplish complex, multi-step tasks in the open world. More specifically, Suleyman proposed an economically-based definition of ACI skill

ICML'24 Position Paper

| **Performance** (rows) x **Generality** (columns) | **Narrow** *clearly scoped task or set of tasks* | **General** *wide range of non-physical tasks, including metacognitive abilities like learning new skills* |
|---|---|---|
| **Level 0: No AI** | **Narrow Non-AI** calculator software; compiler | **General Non-AI** human-in-the-loop computing, e.g., Amazon Mechanical Turk |
| **Level 1: Emerging** *equal to or somewhat better than an unskilled human* | **Emerging Narrow AI** GOFAI[4]; simple rule-based systems, e.g., SHRDLU (Winograd, 1971) | **Emerging AGI** ChatGPT (OpenAI, 2023), Bard (Anil et al., 2023), Llama 2 (Touvron et al., 2023) |
| **Level 2: Competent** *at least 50th percentile of skilled adults* | **Competent Narrow AI** toxicity detectors such as Jigsaw (Das et al., 2022); Smart Speakers such as Siri (Apple), Alexa (Amazon), or Google Assistant (Google); VQA systems such as PaLI (Chen et al., 2023); Watson (IBM); SOTA LLMs for a subset of tasks (e.g., short essay writing, simple coding) | **Competent AGI** not yet achieved |
| **Level 3: Expert** *at least 90th percentile of skilled adults* | **Expert Narrow AI** spelling & grammar checkers such as Grammarly (Grammarly, 2023); generative image models such as Imagen (Saharia et al., 2022) or Dall-E 2 (Ramesh et al., 2022) | **Expert AGI** not yet achieved |
| **Level 4: Virtuoso** *at least 99th percentile of skilled adults* | **Virtuoso Narrow AI** Deep Blue (Campbell et al., 2002), AlphaGo (Silver et al., 2016, 2017) | **Virtuoso AGI** not yet achieved |
| **Level 5: Superhuman** *outperforms 100% of humans* | **Superhuman Narrow AI** AlphaFold (Jumper et al., 2021; Varadi et al., 2021), AlphaZero (Silver et al., 2018), StockFish (Stockfish, 2023) | **Artificial Superintelligence (ASI)** not yet achieved |

# RL or any other optimization-based alignment is not the solution




https://www.youtube.com/watch?v=kopoLzvh5jY


https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/

The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities

Joel Lehman[1†], Jeff Clune[1, 2†], Dusan Misevic[3†], Christoph Adami[4], Lee Altenberg[5], Julie Beaulieu[6], Peter J Bentley[7], Samuel Bernard[8], Guillaume Beslon[9], David M Bryson[4], Patryk Chrabaszcz[11], Nick Cheney[2], Antoine Cully[12], Stephane Doncieux[13], Fred C Dyer[4], Kai Olav Ellefsen[14], Robert Feldt[15], Stephan Fischer[16], Stephanie Forrest[17], Antoine Frénoy[18], Christian Gagné[6] Leni Le Goff[13], Laura M Grabowski[19], Babak Hodjat[20], Frank Hutter[11], Laurent Keller[21], Carole Knibbe[9], Peter Krcah[22], Richard E Lenski[4], Hod Lipson[23], Robert MacCurdy[24], Carlos Maestre[13], Risto Miikkulainen[26], Sara Mitri[21], David E Moriarty[27], Jean-Baptiste Mouret[28], Anh Nguyen[2], Charles Ofria[4], Marc Parizeau [6], David Parsons[9], Robert T Pennock[4], William F Punch[4], Thomas S Ray[29], Marc Schoenauer[30], Eric Schulte[17], Karl Sims, Kenneth O Stanley[1,31], François Taddei[3], Danesh Tarapore[32], Simon Thibault[6], Westley Weimer[33], Richard Watson[34], Jason Yosinski[1]

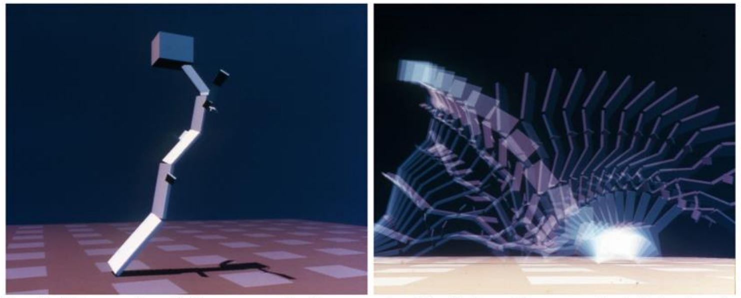


**Figure 1. Exploiting potential energy to locomote.** Evolution discovers that it is simpler to design tall creatures that fall strategically than it is to uncover active locomotion strategies. The left figure shows the creature at the start of a trial and the right figure shows snapshots of the figure over time falling and somersaulting to preserve forward momentum.

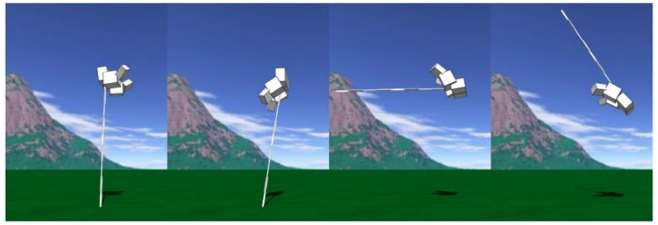


**Figure 2. Exploiting potential energy to pole-vault.** Evolution discovers that it is simpler to produce creatures that fall and invert than it is to craft a mechanism to actively jump.

# Top-down Bottom-up Mechanistic Interpretability Methods

## Negative Results for SAEs On Downstream Tasks and Deprioritising SAE Research (GDM Mech Interp Team Progress Update #2)

by lewis smith, Senthooran Rajamanoharan, Arthur Conmy, CallumMcDougall, Tom Lieberum, János Kramár, Rohin Shah, Neel Nanda

26th Mar 2025 AI Alignment Forum Linkpost from deepmindsafetyresearch.medium.com



## Representation Engineering for Large-Language Models: Survey and Research Challenges

LUKASZ BARTOSZCZE, Wisent AI, United States and University of Warwick, United Kingdom
SARTHAK MUNSHI, Amazon Web Services, United States
BRYAN SUKIDI, University of North Carolina at Chapel Hill, United States
JENNIFER YEN, Perplexity, United States
ZEJIA YANG, University of Cambridge, United Kingdom
DAVID WILLIAMS-KING, Mila, Canada
LINH LE, University of Technology Sydney, Australia
KOSI ASUZU, Wisent AI, United States
CARSTEN MAPLE, University of Warwick, United Kingdom

Large-language models are capable of completing a variety of tasks, but remain unpredictable and intractable. Representation engineering seeks to resolve this problem through a new approach utilizing samples of contrasting inputs to detect and edit high-level representations of concepts such as honesty, harmfulness or power-seeking. We formalize the goals and methods of representation engineering to present a cohesive picture of work in this emerging discipline. We compare it with alternative approaches, such as mechanistic interpretability, prompt-engineering and fine-tuning. We outline risks such as performance decrease, compute time increases and steerability issues. We present a clear agenda for future research to build predictable, dynamic, safe and personalizable LLMs.

CCS Concepts: • **Computing methodologies → Neural networks**; **Machine learning approaches**; **Machine learning algorithms**.

# Back-up

## Some Experiments

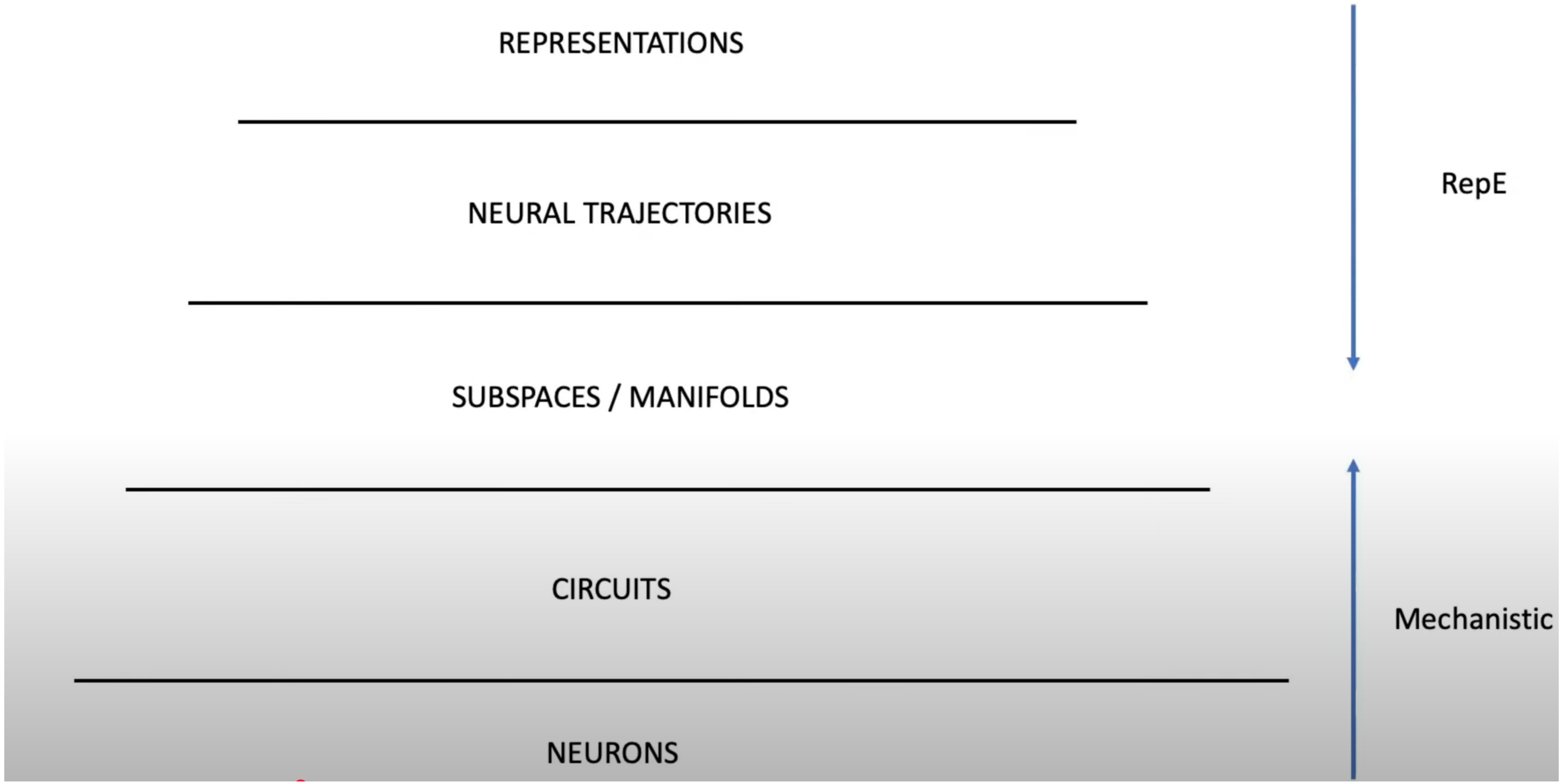
# Mechanistic Interpretability

Sumit Jha

SRI International

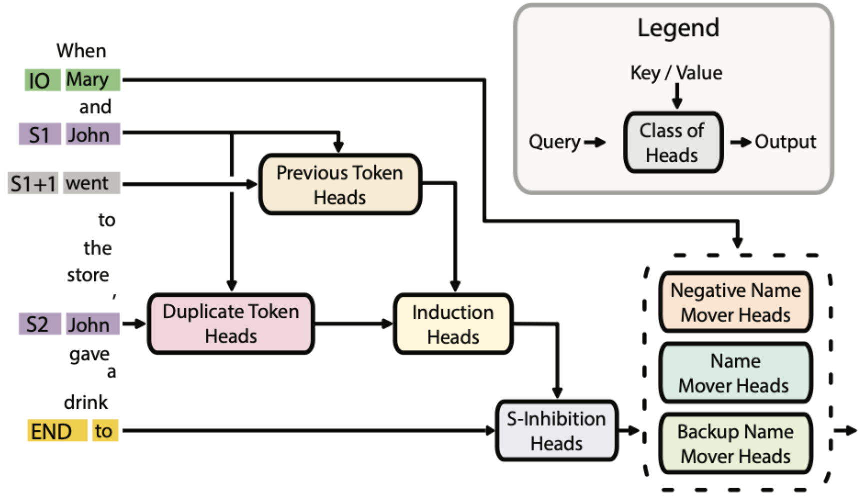
9/4/25

# Different Approaches to Interpretability
# Top-down vs Bottom-up




Interpretability: Ability to explain model's decisions in human understandable way

# Mechanistic View




**Approach:** Bottom-up

**Algorithmic Level:** Node-to-node connections

**Implementational Level:** Neurons, pathways, circuits

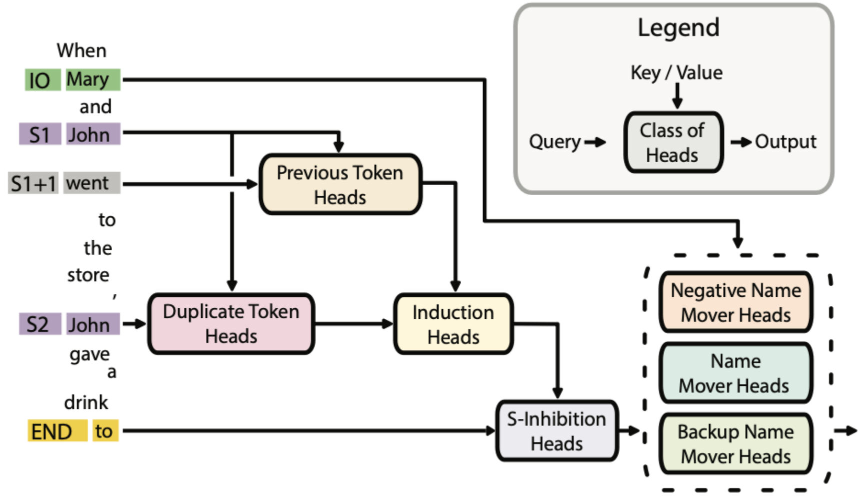
## Neuron-level analysis

Anthropic's Sparse AutoEncoders [Cunningham et al., 2023]
Scaling & Evaluating SAEs, OpenAI 2024
Towards Principled Evaluations of SAEs, Google 2024
Route SAEs to interpret LLMs [Shi et al., 2025]

## Model-level analysis

Mechanistic Unveiling of Transformer Circuits [Zhang, 2025]
The optimal BERT surgeon [Kurtic et al., 2022]
Automated Circuit Discovery [Conmy et al., 2023]
Circuit Discovery with Graph Pruning [Yu et al., 2024]

# Mechanistic View




**Approach:** Bottom-up

**Algorithmic Level:** Node-to-node connections

**Implementational Level:** Neurons, pathways, circuits

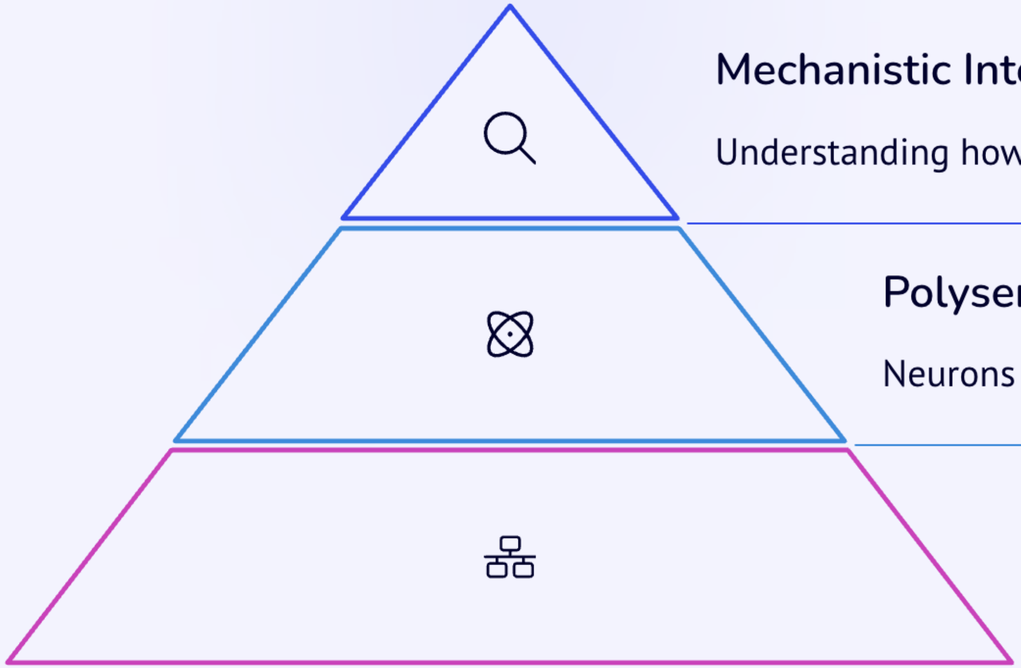
## Neuron-level analysis

Anthropic's Sparse AutoEncoders [Cunningham et al., 2023]
Scaling & Evaluating SAEs, OpenAI 2024
Towards Principled Evaluations of SAEs, Google 2024
Route SAEs to interpret LLMs [Shi et al., 2025]

## Model-level analysis

Mechanistic Unveiling of Transformer Circuits [Zhang, 2025]
The optimal BERT surgeon [Kurtic et al., 2022]
Automated Circuit Discovery [Conmy et al., 2023]
Circuit Discovery with Graph Pruning [Yu et al., 2024]

# Sparse Autoencoders



### Mechanistic Interpretability

Understanding how neural networks calculate outputs

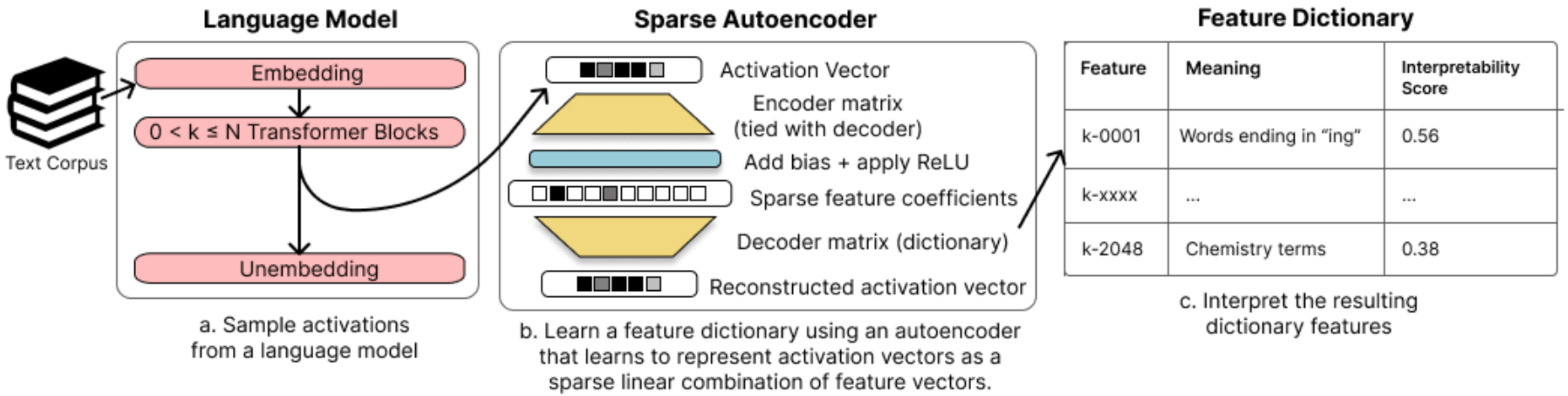
### Polysemanticity Challenge

Neurons activate for multiple unrelated features

### Superposition Hypothesis

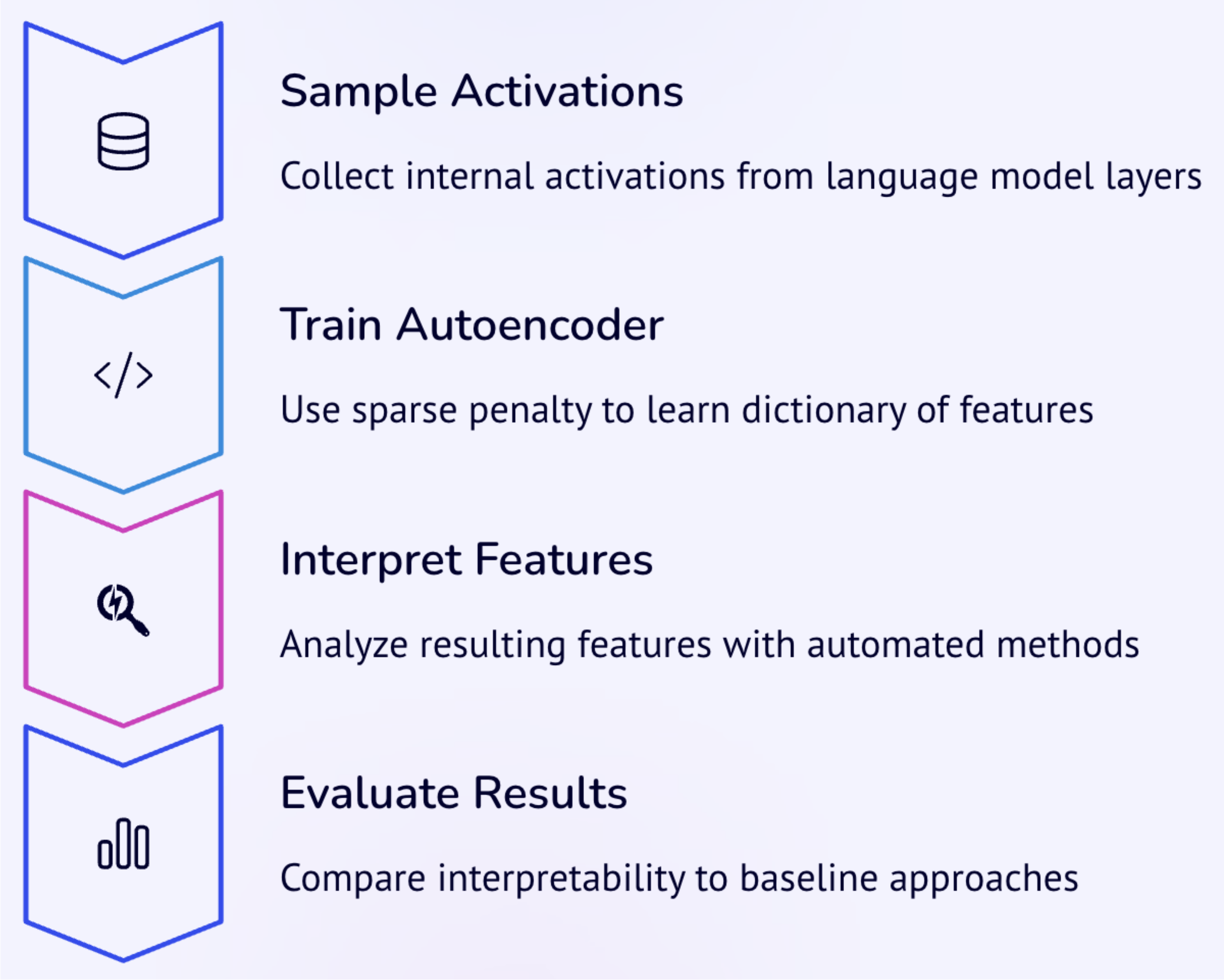
Networks learn more features than dimensions

# Sparse Autoencoders

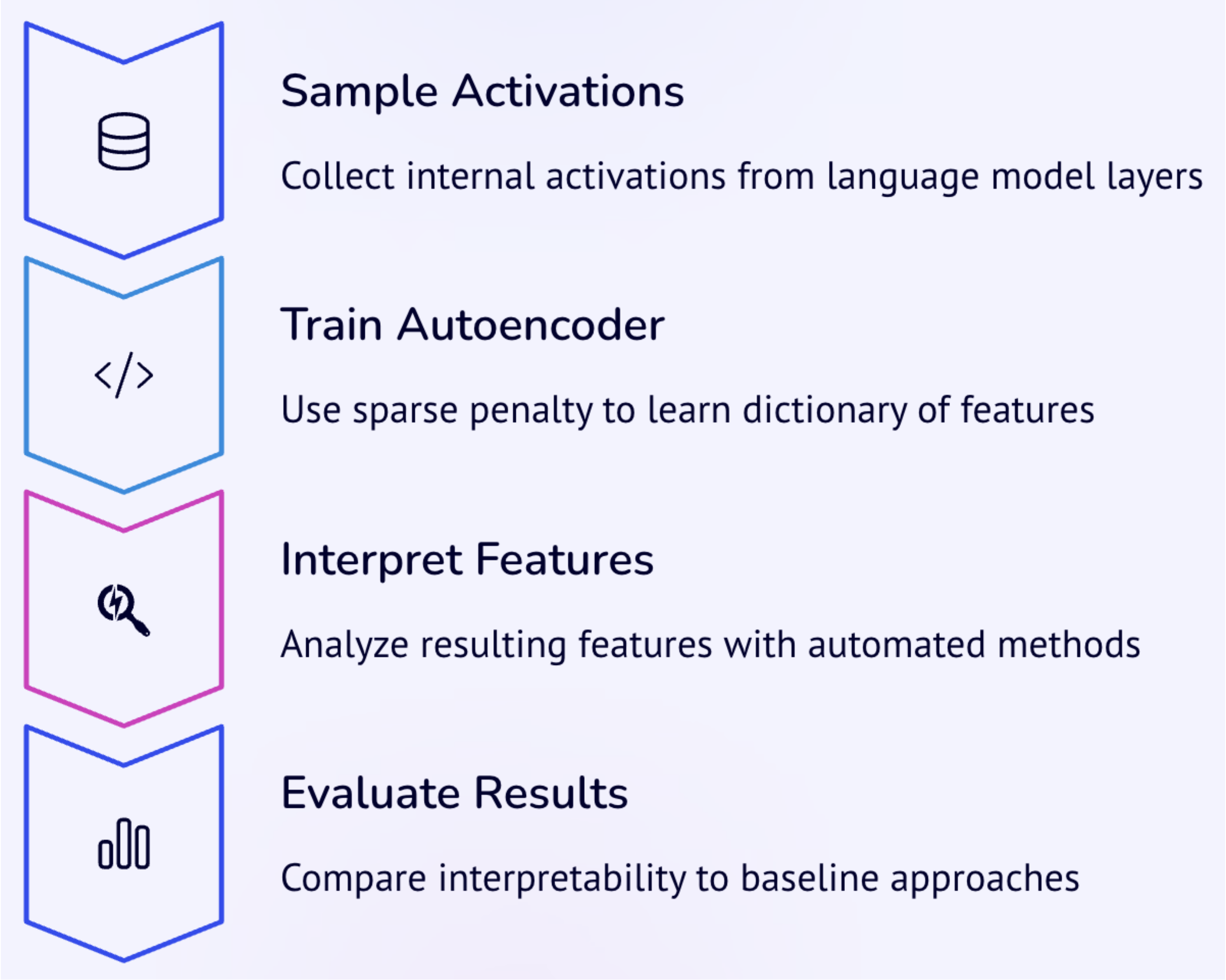



Mapping polysemantic neurons from LLMs’ layer to monosemantic encoded space

# Sparse Autoencoders



# Sparse Autoencoders




Tinyllama1.1B model’s 14 layer activations for ‘city’

Train SAE with encoded space 4 times the layer

Interpret encoded space with concepts associated with ‘city’ such as ‘country’, ‘language’ etc.

Patching for ‘causal’ and ‘isolation’ scores

# Problem Statement

- LLMs are capable at answering complex queries but understanding how they arrive at answers remains challenging

- Interpreting how LLMs process complex queries by examining neurons and internal circuits responsible for different concepts

# Motivation

- Existing MI work identifies individual features or concepts such as “Golden Gate Bridge”
- Interpretability of LLMs on complex queries requires investigating these models holistically rather than focusing on isolated concepts
- We focus on MI aspect of LLMs on complex, e.x. multi-hop,  queries such as ‘The spouse of the performer of Imagine is’
    - model needs to first answer the first hop: performer of Imagine – John Lennon
    - then answer the second hop: Spouse of John Lennon – Yoko Ono

# Neuro-symbolic approach with KGs

### Knowledge Graphs as Data Foundation

Knowledge Graphs like ConceptNet provide rich information on entities (nodes) and their relationships (edges). To our knowledge, KGs have only been used to add context to input queries (RAG-technique) for improving LLM performance, not for mechanistic interpretability

### Logical Language Representation

We extract KG information and store it in logical language format with entities as predicates and relationships as connectors between predicates

# Neuro-symbolic approach with KGs

### Dataset Generation from KG

For each hop in the input query, generate a dataset from the KG with (e1, r, e2) tuples, where r is the relationship and e1, e2 are entities with attributes.

### Neuron Localization

Localize neurons of the LLM responsible for relationship 'r' as a concept using SAE with the generated dataset.

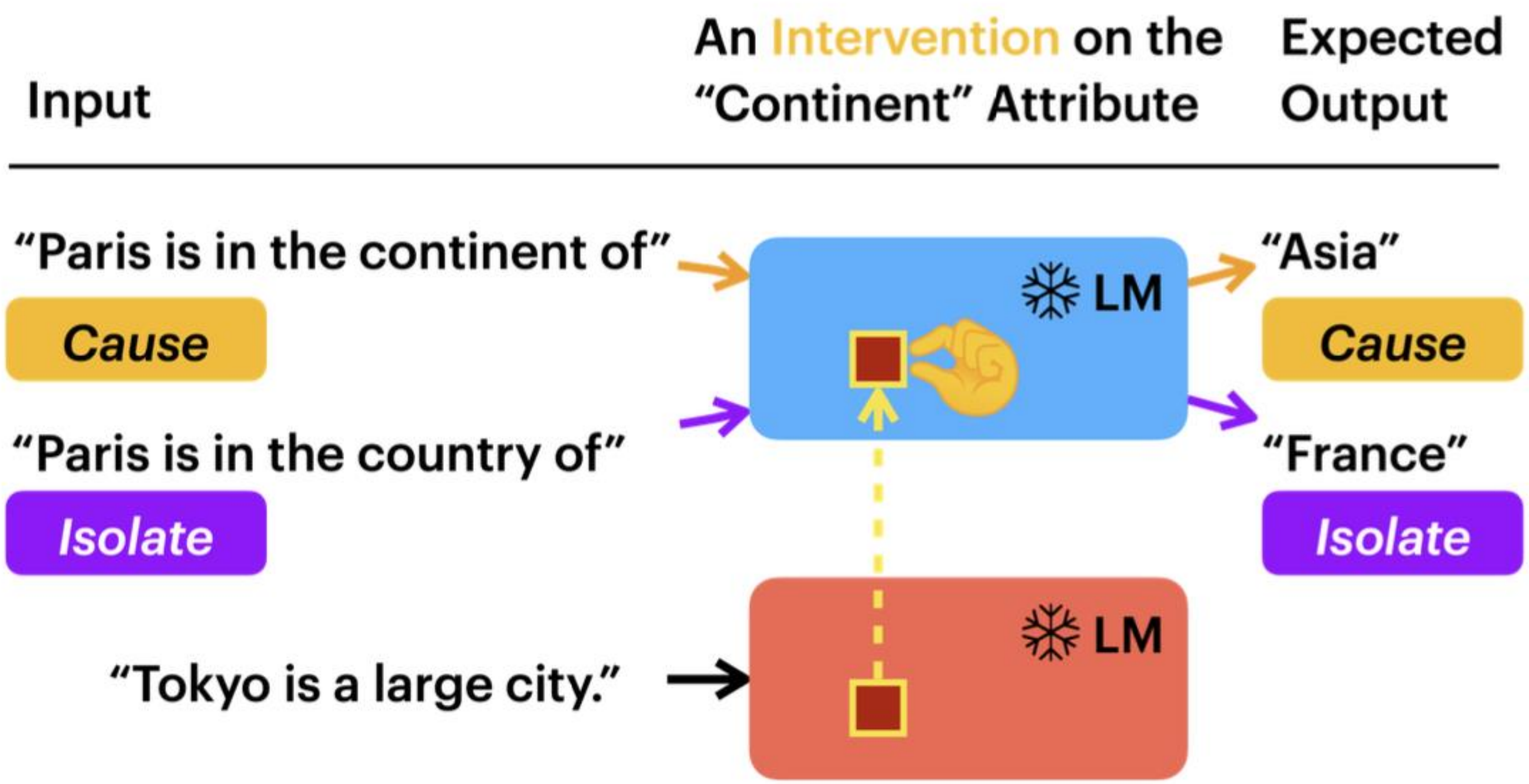
### Circuit Construction

Apply network analysis to construct and study circuits formed by interconnecting neurons corresponding to different concepts

### Evaluation via Intervention

Use activation patching/intervening techniques to check interpretability of the LLM on complex queries via causal and isolate scores, using ground truth data extracted from the KG

# Evaluation via patching



# Results on Single-Hop Queries

For our base query "city: Ahvaz, country: ", the model correctly answers "Iran".

When we patch it with source city Cascavel, we expect two outcomes if SAE has correctly identified interpretable neurons:

1. **High causal score:** Intervention on country neurons should change prediction from Iran to Brazil

2. **High isolation score**: Language neurons should be non-overlapping with country neurons

| Concept | Neuron Patching | Neuron Dropout |
|---|---|---|
| Language | Brazil => non-isolation<br>Overlapping neurons [46 out of 62] | Brazil<br>non-overlapping equally imp as overlapping?<br>[No – tested intervention on 16: result was Iran] |
| Country | Brazil => causal | Iran<br>Could have been random but why Iran? |
| Union of both | Brazil => causal | Iran<br>Could have been random but why Iran? |

Dropout results suggest that looking into neurons in isolations for concepts is not sufficient

# Results on 2-Hop Queries

Few-Shot base case variations:

1. What is the national language of the country where Paris is located? French. What is the national language of the country where Rome is located?
2. What is the national language of the country where Paris is located? French. What is the national language of the country where Moscow is located?
3. What is the national language of the country where Paris is located? French. What is the national language of the country where Spain is located?

Source: What is the national language of the country where Paris is located? French. What is the national language of the country where London is located?

**Ans in all the test cases is Italian/Russian/Spanish - wrong answer for source/correct answer for base**

# Analysis of results on 2-Hop Queries

- Intervention results – just using neurons corresponding one or union of both concepts does not yield causality
    - we need to identify the circuit and not just these neurons in isolation as done by existing techniques
- Dropout results – removal of neurons corresponding one or union of both concepts does not change the result
    - we need to remove the circuit corresponding to the query to change the result for base query
- Something similar might be happening for even simple queries – as indicated by for row of results

# SAE Results

| Concepts for Objects | Changed Base O/P | Correct Patching O/P |
|---|---|---|
| Category | 46.15% | 34% |
| Color | 46.67% | 11.66% |
| Texture | 60.93% | 4.2% |

| Base Input | Base Output | Patched Input | Correct Patched Output |
|---|---|---|---|
| rock: non-living thing; cabbage: plant; dog: animal; apple: | plant | rock: non-living thing; cabbage: plant; dog: animal; chair: | non-living thing |
| The color of leaf is usually green. The color of coal is usually black. The color of banana is usually | yellow | The color of leaf is usually green. The color of coal is usually black. The color of golf ball is usually | white |
| rock is hard; towel is soft; door is | hard | rock is hard; towel is soft; pillow is | soft |

| Base Input | Base Output | Patched Input | Incorrect Patched Output |
|---|---|---|---|
| rock: non-living thing; cabbage: plant; dog: animal; apple: | plant | rock: non-living thing; cabbage: plant; dog: animal; chair: | non-living thing |
| The color of leaf is usually green. The color of coal is usually black. The color of banana is usually | yellow | The color of leaf is usually green. The color of coal is usually black. The color of golf ball is usually | white |
| rock is hard; towel is soft; door is | hard | rock is hard; towel is soft; pillow is | soft |