FY2024/2024年度

Natural Language Understanding Team

Kentaro Inui

自然言語理解チーム　乾 健太郎

Center for Advanced Intelligence Project

Mechanistic Understanding of GenAI

Vision and Social Impact

The adoption of GenAI has opened many new challenges. Mechanistic understanding of GenAI is needed to ensure safety, fairness, and trustworthiness.
Our vision is to achieve this understanding by analyzing models at three levels of granularity:

- **Neuronal**: What is the function of individual neurons?
- **Representational**: What information do hidden states encode? How does the model use this information?
- **Algorithmic**: What algorithms and reasoning strategies do models use to generate answers? How are they implemented in which model components?

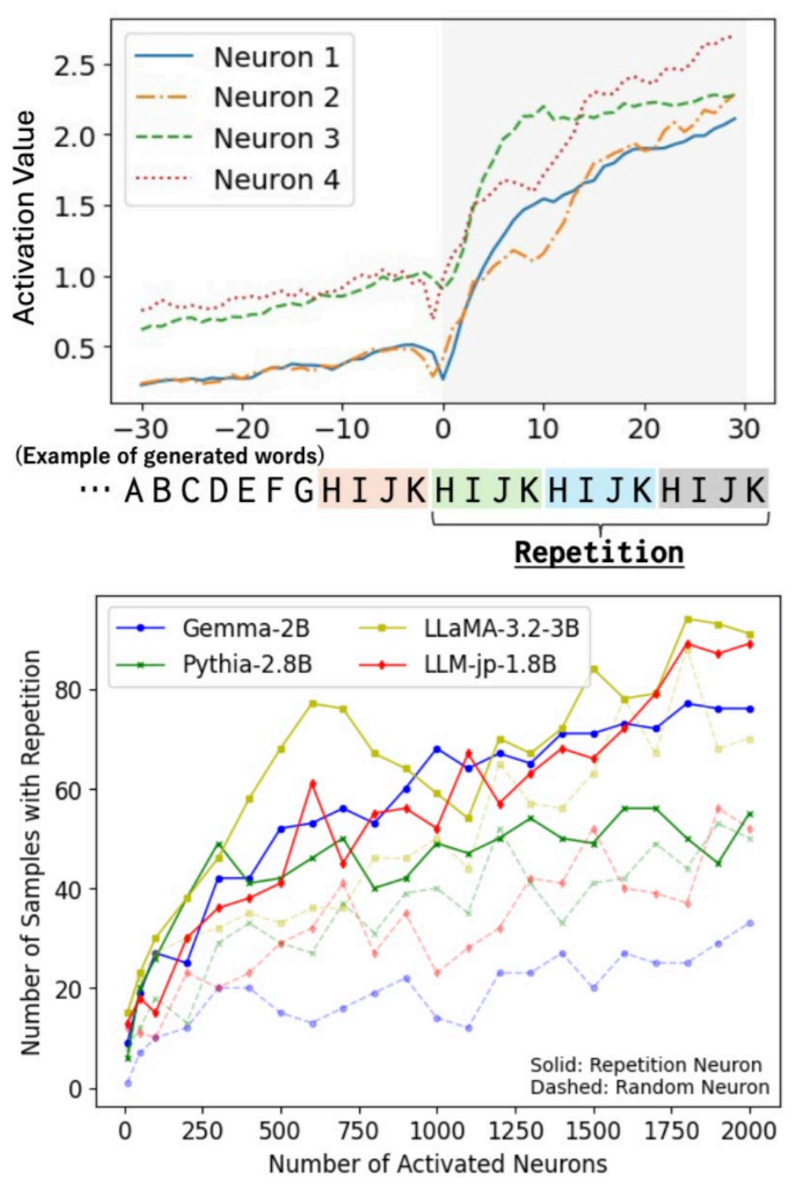
Neuron-level Analysis

NAACL 2025

Repetition Neurons: How Do Language Models Produce Repetitions?

Motivation: Repetition is an important and sometimes undesirable aspect of text generation. How LMs perform repetition is not understood on a mechanistic level so far.

Contribution: Viewing repetition as a "skill", we devise a method for finding neurons strongly associated with this skill. We then show that repetition can be promoted or suppressed by activating and deactivating these "repetition neurons".



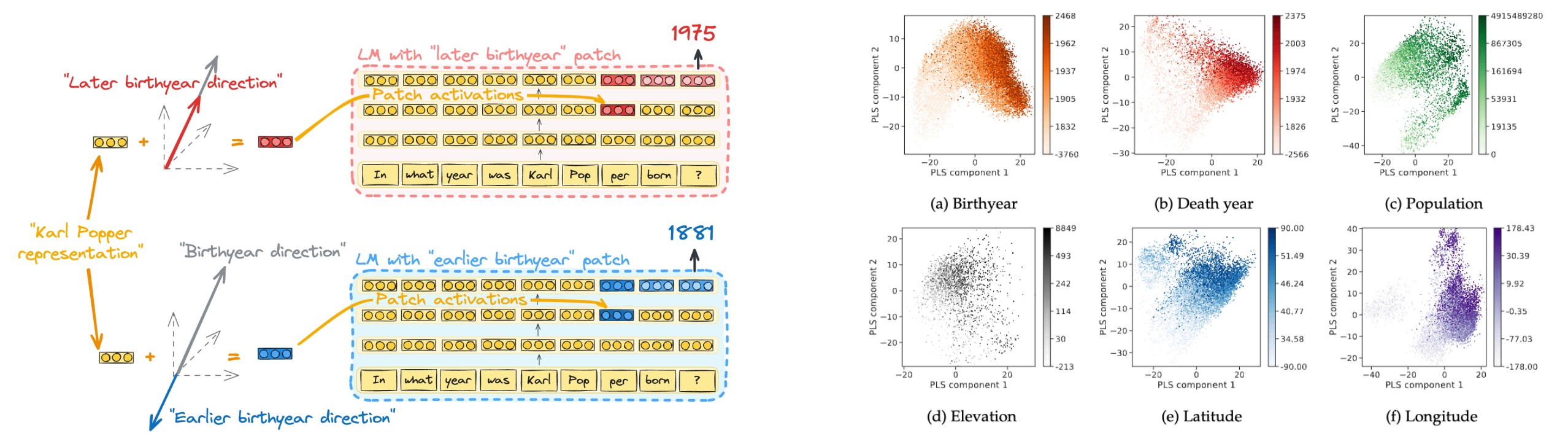
Understanding the Internal Representations of LLMs

ACL 2024

Monotonic Representation of Numeric Properties in Language Models

Motivation: Want to understand if/how numeric properties (birth year, elevation, population, etc.) are encoded in LMs

Contribution: Developed a method for probing and modifying LM representations. Found that numeric properties are encoded in low-dimensional subspaces in a easily-interpretable, monotonic fashion.

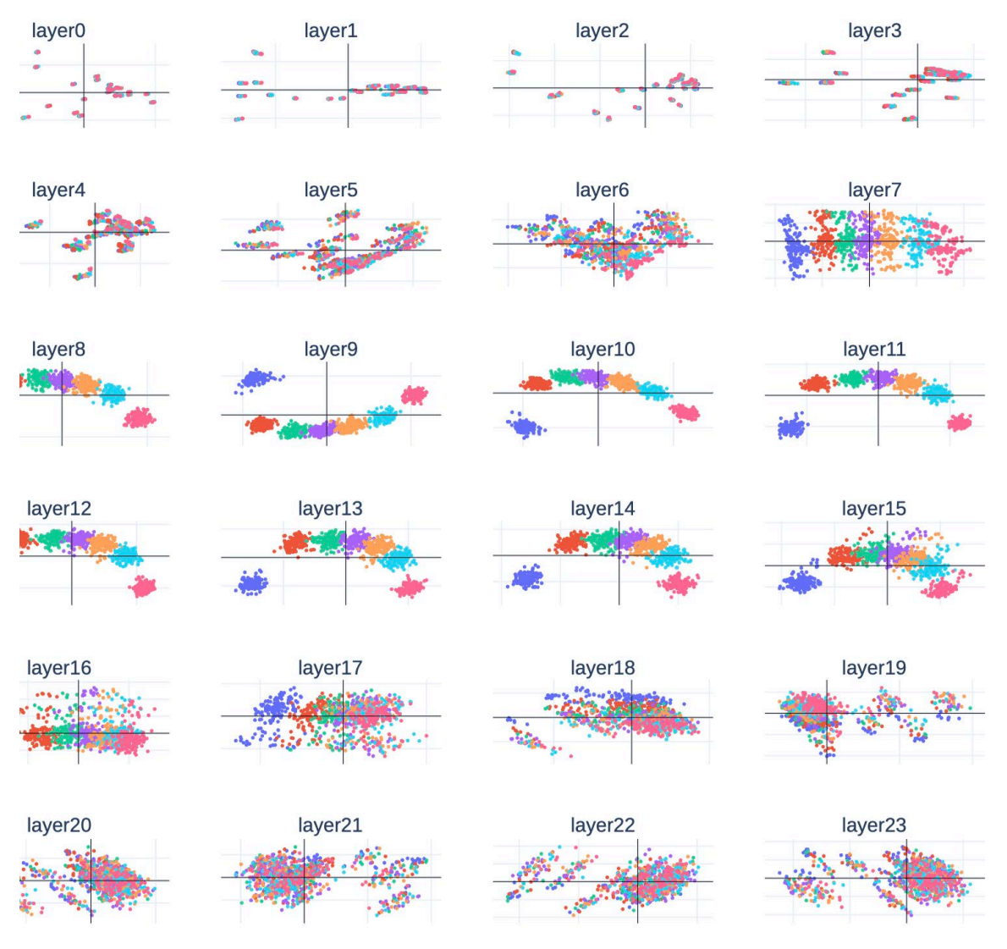


EMNLP 2024

Representational Analysis of Binding in Language Models

Motivation: Analyze if/how LMs perform *binding*. Binding is a basic operation in language understanding by which representations of entities that stand in a relation, e.g., "toy" and "box" in "the *toy* is in the *box*", are combined.

Contribution: Identified a subspace in which entities are represented according to their ordering. Causal interventions via activation patching allows controlling the LM's binding behavior, suggesting that this subspace is an important part of the mechanism by which the model performs binding.

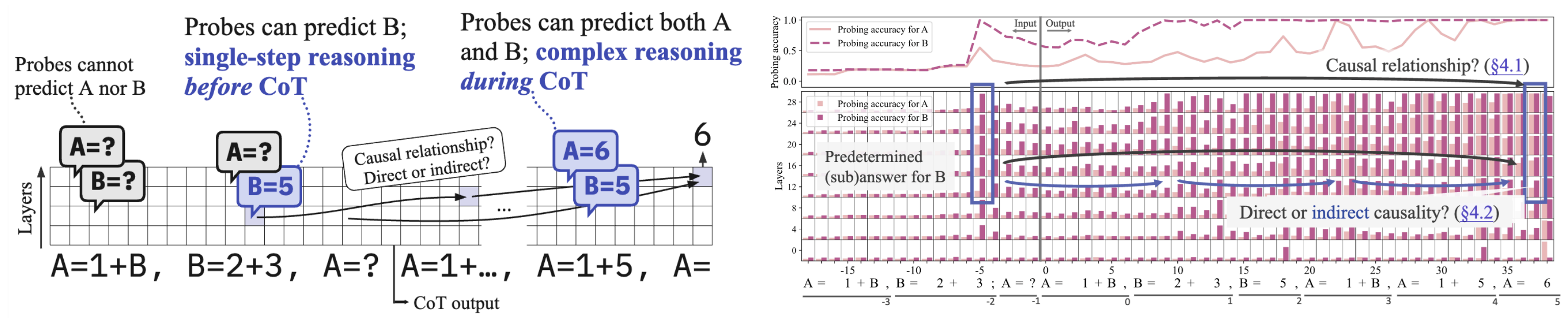


Dissecting the Reasoning Process of LLMs and Vision Models

Think-to-Talk or Talk-to-Think? When LLMs Come Up with an Answer in Multi-Step Reasoning

Motivation: Chain-of-thought is a prompting technique by which one asks LMs not only to generate an answer, but also a chain of reasoning steps. However, the degree to which this chain reflects the model's actual reasoning is unclear.

Contribution: By conducting systematic probing, we reveal a complex relationship between internal reasoning and expressed reasoning steps.

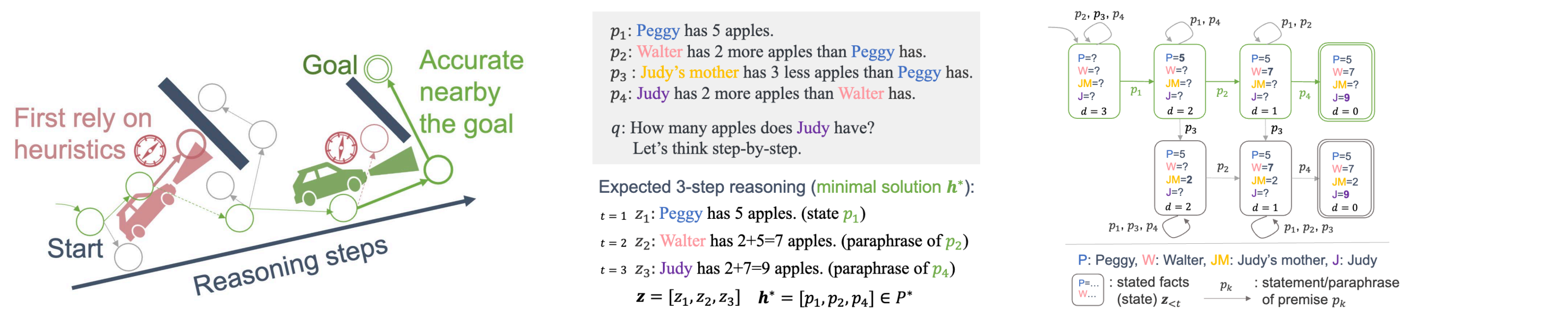

Probes cannot predict A nor B
Probes can predict B; single-step reasoning before CoT
Probes can predict both A and B; complex reasoning during CoT
A=?
B=?
A=?
B=5
Causal relationship? Direct or indirect?
A=6
B=5
6
Layers
A=1+B, B=2+3, A=? A=1+…, A=1+5, A=
CoT output
Probing accuracy for A
Probing accuracy for B
Probing accuracy
Input
Output
Causal relationship? (§4.1)
Predetermined (sub)answer for B
Direct or indirect causality? (§4.2)

EMNLP 2024

First Heuristic Then Rational: Dynamic Use of Heuristics in Language Model Reasoning

Motivation: LMs sometimes appear to perform complex multi-step reasoning but also can make simple mistakes. We analyze why such reasoning failures happen.

Contribution: We show that use of simple heuristic, such as lexical overlap, are more prevalent in the earlier stages of reasoning, while precise reasoning is more common towards the end of the reasoning chain.


First rely on heuristics
Goal
Accurate nearby the goal
Start
Reasoning steps
P: Peggy, W: Walter, JM: Judy's mother, J: Judy
: stated facts (state) $z_{<t}$
: statement/paraphrase of premise p_k

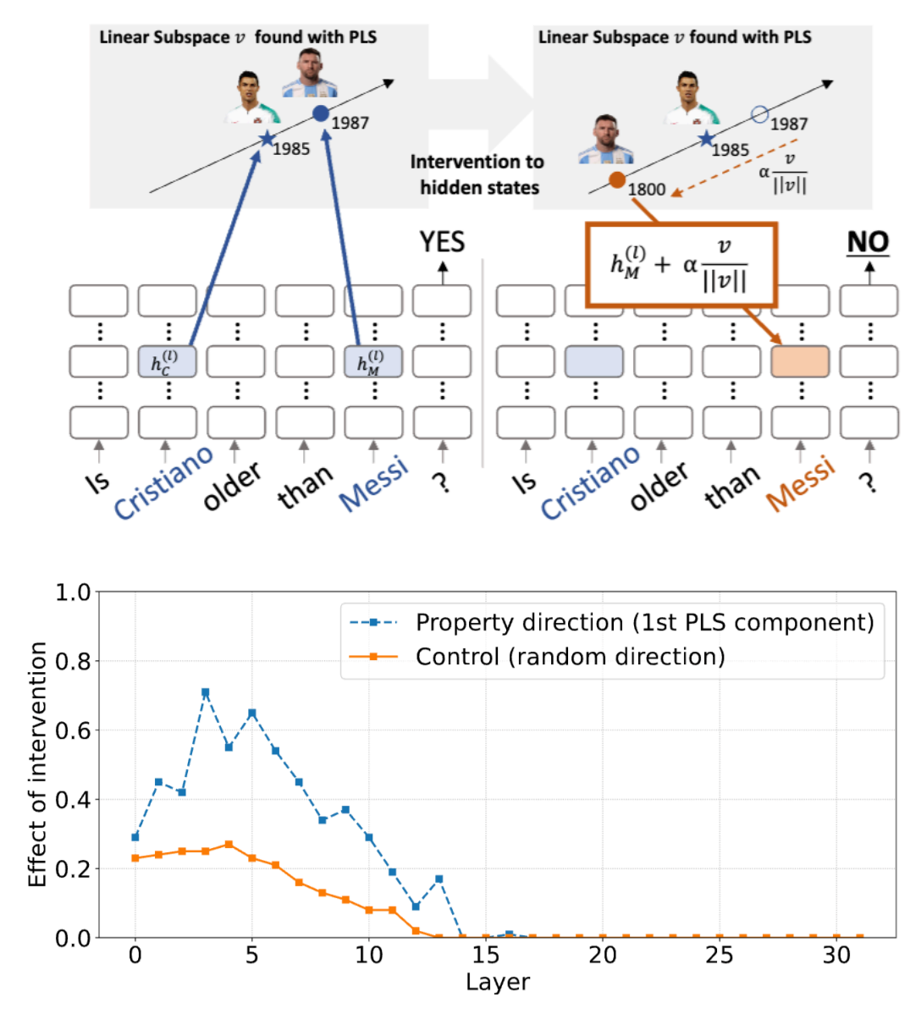
p_1: Peggy has 5 apples.
p_2: Walter has 2 more apples than Peggy has.
p_3 : Judy's mother has 3 less apples than Peggy has.
p_4: Judy has 2 more apples than Walter has.

q: How many apples does Judy have?
Let's think step-by-step.

Expected 3-step reasoning (minimal solution h^*):
$t=1$ z_1: Peggy has 5 apples. (state p_1)
$t=2$ z_2: Walter has 2+5=7 apples. (paraphrase of p_2)
$t=3$ z_3: Judy has 2+7=9 apples. (paraphrase of p_4)
$\boldsymbol{z} = [z_1, z_2, z_3]$ $\boldsymbol{h}^* = [p_1, p_2, p_4] \in P^*$

NAACL 2025

The Geometry of Numerical Reasoning


Linear Subspace v found with PLS
1985
1987
Intervention to hidden states
1800
$\alpha \frac{v}{||v||}$
$h_M^{(l)} + \alpha \frac{v}{||v||}$
YES
NO
$h_C^{(l)}$
$h_M^{(l)}$
Is Cristiano older than Messi ?
Property direction (1st PLS component)
Control (random direction)
Effect of intervention
Layer

**Motivation: In prior work we found that numeric properties are represented in low-dimensional linear subspaces, but it is unknown if/how these representations are used in the reasoning process of language models.

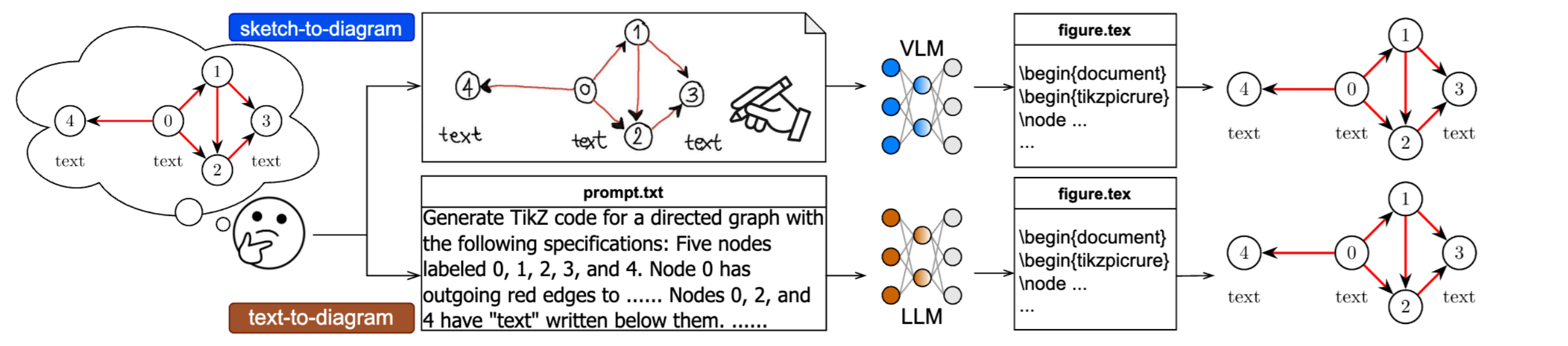
Contribution: Found LMs compare numeric properties in low-dimensional linear subspaces of activation space. By intervening on representations in these subspaces, one can control the outcome of numeric comparisons.

ICLR 2025

Sketch2Diagram: Generating Vector Diagrams from Hand-Drawn Sketches

Motivation: Creating and understanding diagrams and vector graphics involves abstract reasoning about discrete. This task is challenging for current models.

Contribution: To enable progress on this complex task, we introduce a dataset consisting of hand-drawn sketches and vector diagrams. In addition to evaluating large models, we develop augmentation methods and show that a small model trained on our dataset can outperform GPT-4o.



```
\begin{document}
\begin{tikzpicrure}
\node ...
...
```

Generate TikZ code for a directed graph with the following specifications: Five nodes labeled 0, 1, 2, 3, and 4. Node 0 has outgoing red edges to ...... Nodes 0, 2, and 4 have "text" written below them. ......

Publication highlights: Heinzerling+ ACL 2024, Yoshida+ ACL-SRW 2024, Takahashi+ NAACL-SRW 2024, Brassard+ COLM 2024, Dai+ EMNLP 2024, Aoki+ EMNLP 2024, Kamoda+ COLING 2025, Shiono+ COLING 2025, El-Shangiti+ NAACL 2025, Hiraoka+ NAACL 2025, Kamoda+ NAACL Findings 2025, Saito+ ICLR 2025, etc.
言語処理学会第30回年次大会優秀賞，同委員特別賞×5，同スポンサー賞，NLP若手の会奨励賞×2