Position: We Need An Algorithmic Understanding of Generative AI

ICML 2025

Oliver Eberle, Thomas McGee, Hamza Giaffar, Taylor Webb, Ida Momennejad

Stanford CS/MS&E 331

What is a position paper?

Stakes out a clear viewpoint or agenda

Argues for a research direction, not just results

Synthesizes evidence; may include light experiments

Aims to shift how the field thinks/works

Motivation

Central question: How do LLMs reason?

- Determine *how* models compute, not just *what* they predict

Why now?

- Scaling is hitting limits: diminishing returns on larger models
  - To get around this, important to understand/improve reasoning mechanisms
- Empirical success outpaces theory: can't explain how models reason

Motivation

This paper: algorithms as a framework for studying reasoning

- What algorithms can GenAI learn?
  - How does this depend on model size, training data, …?
- Provable **guarantees** for any such algorithmic abilities?
- **Algorithmic objectives** for training and fine-tuning?
- How to create a **repository** of algorithmic abilities?
- How to study selection/**composition** of these components?
- **Architectures** w/ specific algorithmic capacities?

AlgEval: Framework for future research

Task: given computational task, *e.g., shortest path to goal?*

Hypothesis-driven approach:

1. **Identify candidate algorithms**
    - List possible algorithmic strategies (e.g., BFS, DFS, …)
2. **Test model behavior and internals**
    - Compare attention patterns, representations, etc. to candidates
3. **Verify mechanisms empirically** (accuracy, …)
4. **Connect findings to theory**
    - Relate observed mechanisms to formal algorithmic properties
5. **Use insights to refine models** (training, architecture, ...)

Why algorithmic reasoning tasks?

- **Core idea:** study LLMs on tasks with known solutions
  - Enables comparison between *learned* vs *ground-truth* algorithms
- Avoid **ambiguous** benchmarks
- Design tasks with **transparent** computational structure
  - E.g., graph traversal, arithmetic, logical inference, sorting
- Control task **complexity** (input size, branching factor, …)
- Diagnose **generalization** (unknown input scales, …)
- Algorithms have interpretable intermediate states/**primitives**

Next slide: more on primitives

From primitives to algorithms

Primitives: Low-level operations that compose into algorithms

- E.g., memory retrieval and updates, copying, comparisons, …
- Circuits and attention heads often implement specific primitives

Broad question: can LLMs truly reason *compositionally*?

- Evidence mixed – some successes, many failures

Goal: establish **methods** to study/induce composition

Next slide: methods for analyzing algorithmic reasoning

Methods: representation and attention

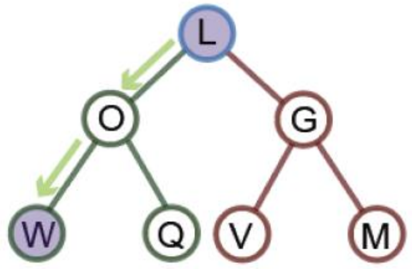
Representational analysis

- Treats layer activations as high-dimensional state spaces
- Uses similarity measures to compare layers, track internal geometry

Attention analysis

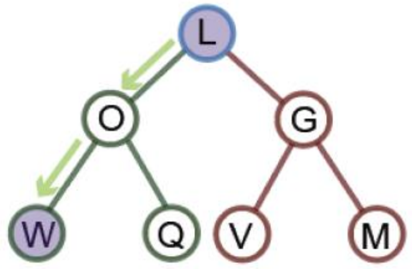
- Interprets attention weights as message-passing between tokens
- Layer-wise attention reveals what elements influence each other

Case study: Graph navigation

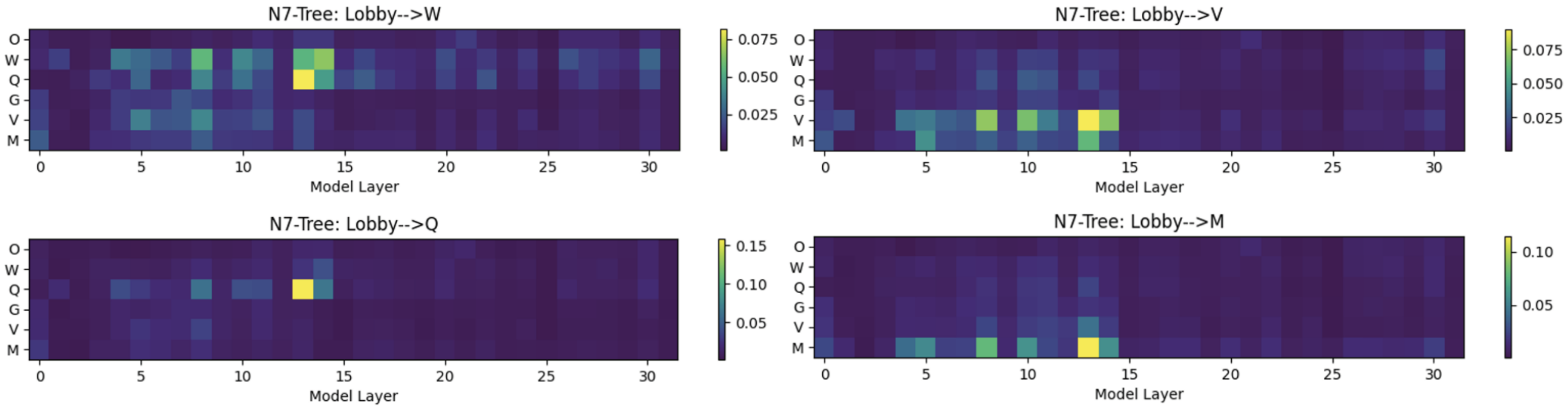


- **Task:** goal-directed navigation on a graph. Prompt:
  - Textual description of rooms (nodes) and connections (edges)
  - "Can you get to W from lobby?" → answer *Yes* or *No*
- Ground-truth algorithms for comparison:
  - Classical search methods e.g., BFS, DFS, and Dijkstra
- Hypothesis under test:
  - Each layer might correspond to one step in a search algorithm
  - Attention weights reveal which nodes are being "visited" at each step
- Models: Llama-3.1-8B and Llama-3.1-70B-Instruct

Case study: Graph navigation

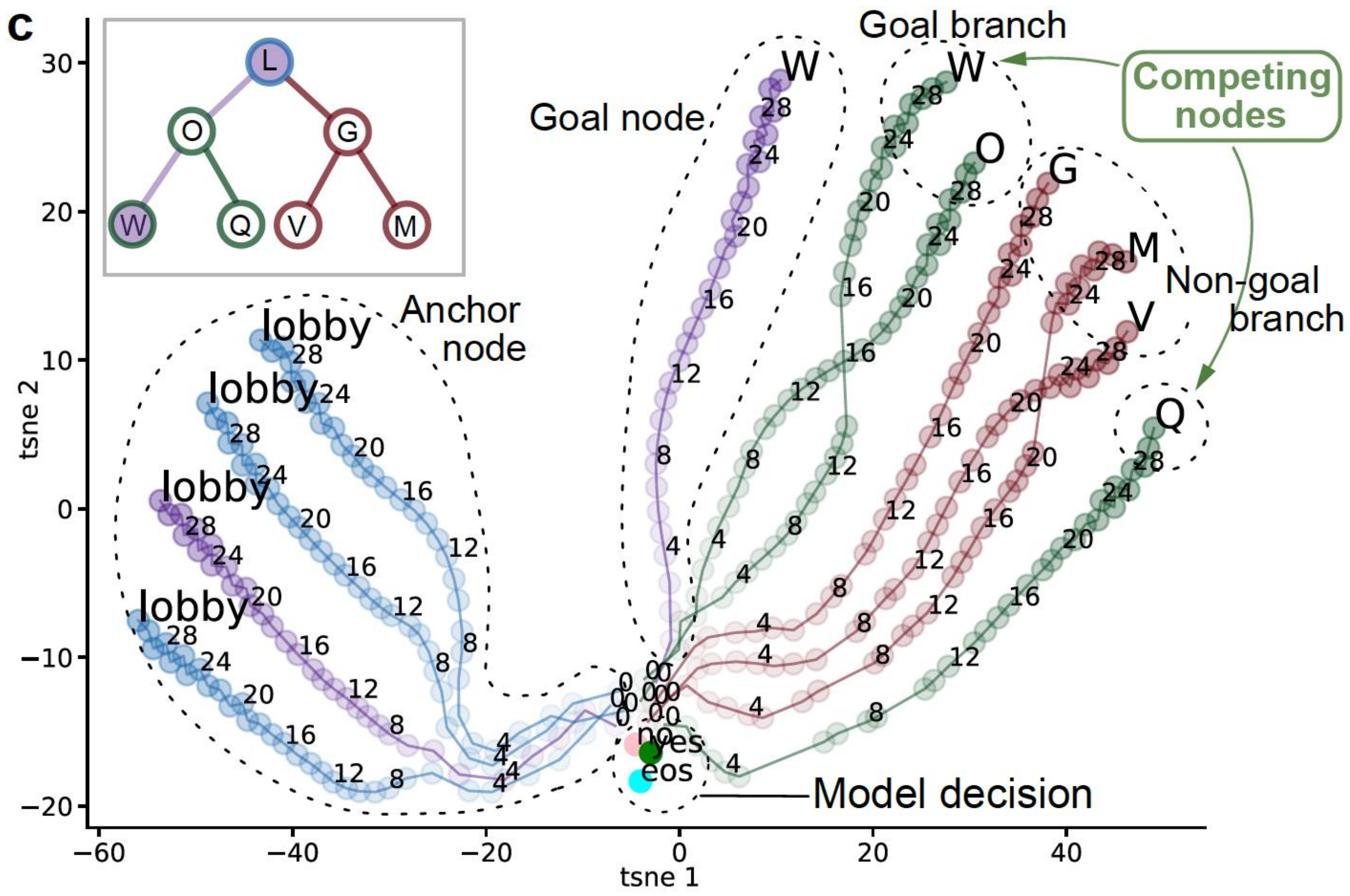


- Attention heatmaps from goal token to all nodes
- Attention seems to peak at **goal** and its **sibling**
  - Mechanistically: local decision test? “Goal here or its sibling?”



Case study: Graph navigation

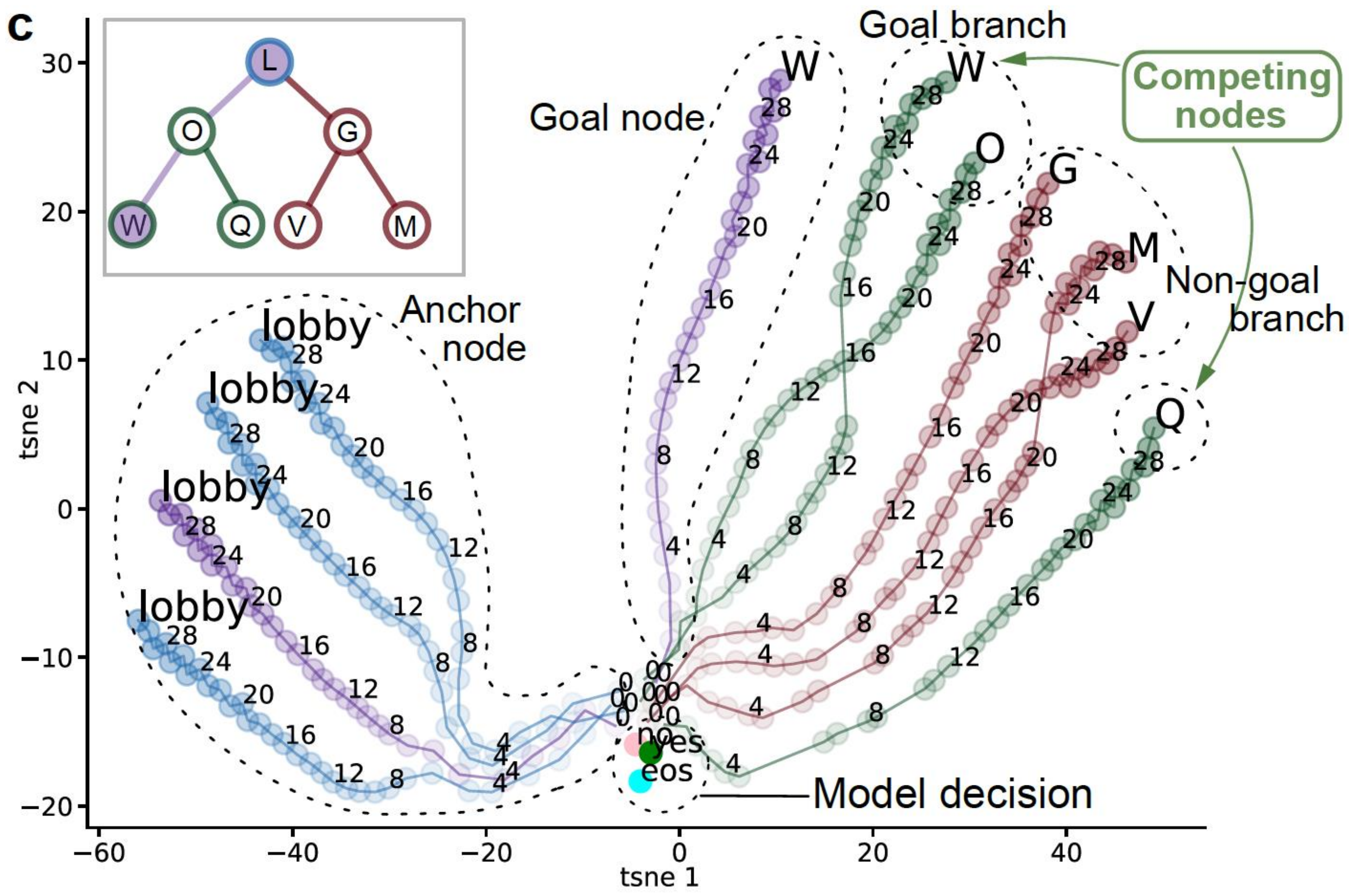
- 2D t-SNE of:
  - Room-token activations from all layers
  - Plus final eos token ("yes"/"no")
- Each color
  = room token
- Number next to point
  = layer index



Case study: Graph navigation

Early layers: all room tokens form single tight cluster

Lobby token diverges; anchor trajectory

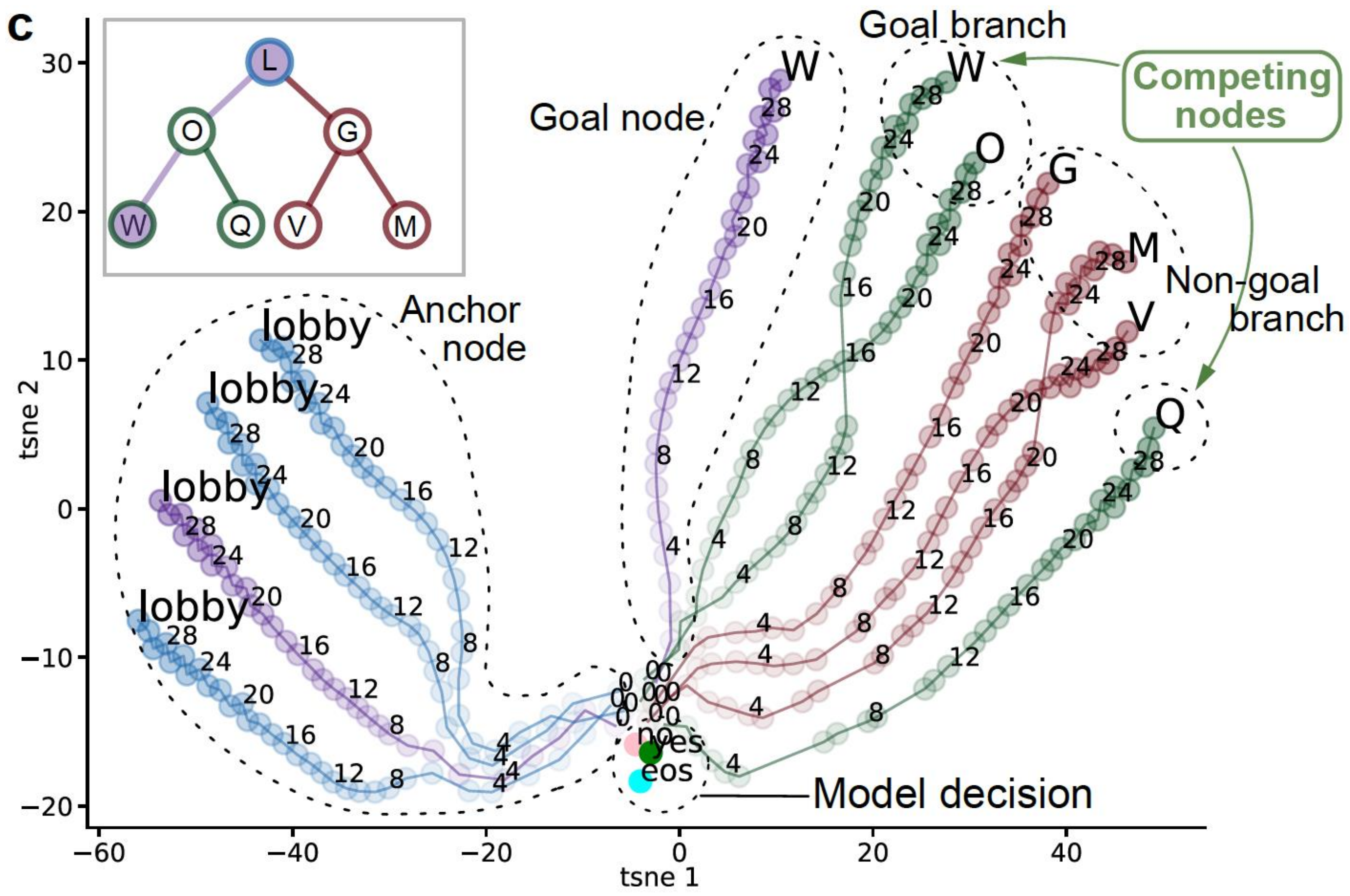


Case study: Graph navigation

Non-goal room tokens cluster together

- Consistent subgroup patterns across layers

W and sibling competitor Q increasingly separate



New directions: inference-time compute

Motivation: reasoning need not occur in one feedforward pass

- Chain-of-thought, explicit tree search, agentic frameworks, …

Fit for AlgEval:

Sequential outputs easier to analyze than high-dim states

Key research questions:

- Which computations *offloaded* to inference vs. embedded in model?
- Can scaling inference-time compute outperform scaling model size?

New directions: RL + alg reasoning

RL can shape how models discover algorithms

- RL may yield emergent algorithmic behaviors beyond imitation?

E.g., reasoning models show reasoning emergence via RL

- E.g., backtracking-like behavior/"aha moments"

Key research question: Does RL teach new algorithms or amplify ones already latent in pretraining data?