

**Yogesh Haribhau Kulkarni** • YouAI Advisor (Helping organizations in their AI journeys) | PhD (Geometric Modeling) | Tech Colum...
now •






...

Decoding the Black Box: How Large Language Models Actually Work

Ever wonder what's happening inside the mind of a LLM (Large language models). They display impressive capabilities, but the mechanisms behind them remain mysterious. This is similar to biology, while evolution's basic principles are straightforward, the resulting biological mechanisms are quite intricate.

What if we could "poke" LLMs like scientists use microscopes to look inside cells? The [Anthropic](#) team has made fascinating discoveries doing exactly this!

Here's what they've observed:

-  Multi-step Reasoning: When answering "the capital of the state containing Dallas," the model creates an internal representation of "Texas" before arriving at "Austin"
-  Poetry Creation: Before writing each line of a poem, the model identifies potential rhyming words to use at line endings
-  Multilingual Processing: Uses both language-specific and abstract language-independent circuits
-  Guardrails: During fine-tuning, the model constructs a general-purpose "harmful requests" detector
-  Chain-of-thought: The model genuinely performs the computational steps it claims to be doing

How did they unlock these insights? By creating replacement models that reproduce the original transformer's activations using more interpretable components. I think of it like adding debug tokens to source code, but with "features" instead.

They produced attribution graphs, visual representations of the model's computational steps and grouped related nodes (super nodes) to create simplified depictions of what's happening inside these complex systems.

Fascinating research that helps us understand the "biology" of LLMs!

Read the full paper: <https://lnkd.in/dCQs-ZBk>

#MachineLearning #AI #LLMs #AIExplainability #DataScience #NeuralNetworks #AIResearch

#ComputationalThinking #mvpbuzz #gde Anthropic AI Google DeepMind GDG Pune GDG Cloud Pune



Like

Comment

Repost

Send