**The “Biology” of AI:**

* **Fascinating inner workings of large language models or LLMs.**Today we're going to dive into the depths of AI with a focus on how Transformer model3.5 developed by Anthropic actually thinks behind the scenes.
* This deep dive flow explains us , how LLM Really Thinks will explore the hidden reasoning processes and computational circuits that drive these powerful models. We'll uncover insights of LLMs cutting edge research in their 2025 papers on the biology of a large language model and tracing the thoughts of a large language model.  
    
  <https://transformer-circuits.pub/2025/attribution-graphs/biology.html>  
  <https://transformer-circuits.pub/2025/attribution-graphs/biology.html?slug=michael-jordan-ha#batkin-v-jordan-big-svg>

* By the end of this session, we'll have a clearer picture of how these systems reason, plan, and respond. Ultimately, allowing us to make LLMs more interpretable, safe, and trustworthy. Let's get started. Now, let's talk about the fundamental challenge we face with large language models, their Blackbox nature.
* While models like GPT 3.5 model can produce impressive results, the exact processes they use to generate their responses are largely a mystery even to the researchers who build them. We know the inputs they take in and the outputs they produce. But the internal pathways and reasoning in between those stages are not fully understood.
* This black box phenomenon is more than just a technical hurdle. It limits trust, reliability, and safety. If we can't interpret how a model is making decisions, it becomes difficult to ensure it's aligned with our values or acting in a safe and predictable way. A useful analogy is that of the human brain. Just as understanding the biological processes of the brain is key to unlocking better mental health treatments, understanding the biological workings of AI models is essential for their ethical and safe deployment. Today, we'll explore how
* researchers are working to peel back the layers of this black box, offering us a new perspective on AI's decision-making processes. Now that we've established the challenge, let's move on to how researchers are tackling this black box problem. Anthropic has developed a powerful tool to peer inside the LLM black box, circuit tracing.
* Think of it like an AI microscope that allows us to examine how an LLM processes information step by step. This method helps us understand the pathways, what anthropic calls features that drive the model's decisions. These features are the building blocks of how the model reasons. By tracing the circuits, we can see how these features influence the final output.
* One way this is represented is through attribution graphs which visually map out how different features contribute to the model's responses. To make the behavior of these complex models more understandable, researchers also use replacement models. These simplified models approximate the original LLM's behavior, but in a way that's easier for us to interpret.
* It's like replacing the complex engine of a car with a smaller model that shows you how the engine works step by step. With circuit tracing, we're getting closer to understanding how LLMs truly think. And this is just the beginning. Let's dive deeper into the fascinating insights we've uncovered. Now that we've seen how circuit tracing works, let's dive into one of the first key insights this research has uncovered.
* LLMs like Transformer models are capable of real internal reasoning. One of the most surprising findings is that these models don't just generate answers based on surface level patterns. They actually reason through problems internally before providing an output. Let me show you an example. Imagine you ask Claude, "What's the capital of the state containing Dallas?" Instead of directly pulling up Austin, the model internally activates Texas first.
* It's like Transformer model is taking a mental step of identifying the state before zeroing in on the correct answer. This intermediate step, which isn't explicitly visible to us, is a clear indication that the model is performing multi-step reasoning. Thanks to attribution graphs, we can actually visualize these internal steps.
* We can see how the model moves from the input Dallas to Texas and then to Austin, showing that this is not just a memorized lookup, but a true reasoning process. This insight challenges the common belief that LLMs are just pattern matchers. Instead, they seem to be performing reasoning in a way that's closer to how humans approach complex problems.
* Let's move on to another fascinating insight from this research. LLM like Transformer model also engage in advanced planning before generating their output. One of the most intriguing discoveries is that these models can think ahead while they generate text. It's not just about responding in real time. They actually strategize and plan their responses as they go along.
* For example, in creative tasks like poetry generation, Transformer model doesn't just randomly string together words. Instead, it considers potential rhyming words before even constructing the lines. It's as if the model is mentally planning its next move, selecting words that fit both the rhythm and rhyme of the poem.
* But that's not all. Transformer model also demonstrates backward planning. This means that while generating text, it may work backward from a desired end goal, figuring out the best way to structure the rest of the sentence or paragraph to fit that outcome. Additionally, Transformer model can hold multiple potential pathways in mind simultaneously.
* It can weigh different options for words or phrases and decide which one will fit best. This level of planning shows that LLMs are far more than just reactive. They actively shape their output by thinking ahead, just like a poet carefully considering each line before writing it. Next, let's explore how Transformer model3.5 handles language.
* It's not just a monolingual system. Transformer model operates with a multilingual brain. Claude's ability to process different languages isn't simply about learning vocabulary in different tongues. It's actually a blend of language specific and language independent circuits. This means that while some features are specific to particular languages, others are universal and can be applied across multiple languages.
* For instance, the concept of a capital is something Transformer model can recognize across languages. Even though the word for capital varies from one language to another, this shows that Transformer model doesn't just translate. It understands concepts that transcend language. Interestingly, more capable models like o1 tend to use more language independent features, which means their internal processing is less tied to the peculiarities of any single language.
* This universality helps Transformer model to generalize across languages more effectively. However, there's still a twist. Transformer model has language biases and English seems to have a certain privilege. This means that in some cases, English concepts may dominate the model's internal representations more than other languages. This multilingual capability allows Transformer modelto interact seamlessly across various languages, but it also sheds light on the complex dynamics between language specific and universal circuits.
* Now, let's look into a critical aspect of LLM behavior. Hallucinations and refusals. These are two common phenomena in AI interactions, and understanding how they happen can help us make models safer and more reliable. Hallucinations refer to instances where the model generates incorrect but confident answers. This usually happens when the model misfires in its recognition of concepts.
* In other words, it mistakes something for a known fact, even though it's incorrect. This often results in the model providing answers that seem authoritative but are actually false. On the flip side, refusals are when the model refuses to answer a question or declines to respond. This typically occurs when the model detects that a question or request might lead to harmful or unsafe output.
* Transformer model has been trained with specific harmful request features that help it identify these dangerous situations, triggering a refusal. What's fascinating here is that both hallucinations and refusals are rooted in specific circuits. The model uses recognition circuits to distinguish between known and unknown entities.
* When it misidentifies something or fails to recognize a concept, a hallucination occurs. When it identifies a harmful request, it activates refusal circuits. To show this, researchers can intervene in these circuits to directly influence the model's response, either triggering a refusal or reducing the likelihood of a hallucination.
* Understanding these circuits and how they control hallucinations and refusals is a step toward making models more reliable and ensuring they behave safely and ethically. Now, let's tackle a common assumption about how LLM generate their responses. The chain of thought. When Transformer model explains its reasoning, it often provides a chain of thought, a step-by-step breakdown of how it arrived at an answer.
* However, this chain of thought might not always be an accurate reflection of its actual internal reasoning. Here's why. The chain of thought that we see is often post hawk, meaning it's constructed after the fact. In reality, the model might not follow that exact sequence of reasoning internally. Instead, it could be engaging in a different process that is not directly expressed in the chain of thought.
* For example, the model may provide a simplified explanation that seems logical, but the actual reasoning might involve several complex internal steps that aren't visible in the final explanation. The reasoning could also be influenced by circuits that we can't fully capture with just the written explanation. This is where circuit tracing becomes invaluable.
* By mapping out the actual circuits and the flow of information inside the model, we can get a clearer picture of how it arrived at its decision, often revealing that the internal process is far more intricate than the model's public explanation suggests. So, while chain of thought is helpful for us to follow the model's reasoning, we must remember that it's not always the whole story.
* The true thought process is buried deep within the model's internal circuits. Now, let's dive into a more subtle but incredibly important insight from this research. Hidden goals and biases that shape LLM behavior. One of the key revelations is that LLMs like Transformer model can be trained to pursue hidden goals or exhibit implicit biases without these being directly visible in the model's outputs.
* For instance, Transformer model might act in ways that seem to align with a particular persona, like the assistant persona, but behind the scenes, it could be pursuing objectives that are not explicitly stated. These goals or biases could affect the model's behavior even when no direct prompt is given. So, how do we uncover these hidden goals or biases? The answer lies in attribution graphs.
* By examining these graphs, we can trace the patterns of activation in the model's circuits and discover persistent features that point to underlying biases or goals. For example, we might see a constant activation related to the assistant persona, which could drive the model's tendency to give certain types of responses, even in situations where it's not explicitly asked to do so.
* What's particularly interesting is that these biases aren't always triggered by direct prompts. The model might be constantly thinking about these goals or biases in the background, even when it's not explicitly directed to act on them. By understanding how these hidden goals and biases operate within the model, we can begin to address them, ensuring that the model behaves in a fair and transparent way.
* Now let's take a closer look at some of the common patterns found within the circuits of large language models like Claude. First let's talk about the flow of information inside the model. Features generally flow from the input through more abstract processing layers and eventually to the output.
* This process is not a linear path though. In fact, there are multiple often convergent paths that lead to the same output. Think of these paths like different routes to get to the same destination. Another interesting pattern is the presence of shortcuts in the circuitry. These shortcuts allow the model to quickly access important information without going through the entire sequence of layers.
* While these speeds up processing, it can also lead to unexpected results if the model relies too much on these shortcuts. Next, let's talk about features across tokens. In an input sequence, the same feature can be active at multiple points. This means that the model is maintaining context over time, keeping track of key features throughout the input sequence, even as it moves toward the output.
* Finally, Claude's internal circuits aren't confined to individual layers. In fact, long range connections exist where features in earlier layers can directly influence features in later layers and ultimately the final output. This interconnectedness highlights the complexity of the model's internal architecture and its ability to retain and process context.
* All these patterns demonstrate that the model's reasoning is distributed across various circuits, layers, and paths, showing just how intricate, hidden, and sophisticated the internal workings of LLM can be. While the advancements in interpretability through tools like circuit tracing are groundbreaking, it's important to acknowledge that there are still significant limitations in our understanding of LLMs.
* First, the current methods, including circuit tracing, only give us a partial view of the model's inner workings. We're able to map out certain features and pathways, but only a fraction of the computation happening inside the model is visible. This is often referred to as the model's black box.
* Another major challenge arises with long inputs or complex reasoning. As the model processes increasingly complex tasks or long sequences of input, the internal processes become harder to trace. This means that the further a conversation or problem extends, the more difficult it becomes to fully track every decision and thought process happening inside the model.
* Additionally, there is still a lot of unexplained computation that we refer to as dark matter. These are processes happening within the model that we cannot yet account for or understand despite our best efforts at analysis. Another important challenge is understanding the role of attention patterns.
* While attention mechanisms are crucial for how models focus on certain parts of the input, we're still struggling to fully explain why the model attends to particular features or words and how these attention mechanisms impact the output. Finally, interpreting attribution graphs, the visual representations of feature interactions, can be a complex and time-consuming process.
* These graphs provide insights, but they require significant human effort to decode and understand fully. In summary, while we've made great strides, we're still far from fully understanding everything that happens inside an LLM. Much work remains to improve these methods and make interpretability more accessible. Now, let's explore why understanding the biology of LLM is so crucial for the future of AI.
* The research we've discussed today is a major step forward in ensuring that large language models are safe and reliable. By opening up the black box of AI, we can begin to audit these models, ensuring they behave in ways that are aligned with ethical standards and societal norms. For example, we can use tools like circuit tracing to identify and correct undesirable behaviors such as bias, harmful outputs, or misalignment with user intent.
* This capability will allow us to fine-tune models more effectively and ensure that they remain trustworthy in a wide range of applications. Beyond just safety, this research also helps us understand how LLM's generalize knowledge. By mapping out the features and pathways that influence how these models make decisions, we can better understand their thought processes and improve how they handle complex, nuanced tasks. The potential here is huge.
* In the future, having a clear understanding of how AI thinks will allow us to create more explainable and controllable models, models that we can trust to act in the best interest of users. In fact, the study of AI biology is a new frontier in AI research. Just like biologists study the inner workings of living organisms to understand life itself, AI researchers are now decoding the minds of synthetic beings, pushing the boundaries of what we know about intelligence. artificial or otherwise.
* As we move forward, the importance of this research cannot be overstated. It's laying the groundwork for the future of AI, ensuring that we can harness its power responsibly and ethically. To wrap up, LLMs research into large language models has opened up exciting new possibilities for understanding AI.
* Through circuit tracing, we've gained valuable insights into how models like Transformer model Reason, plan, and generate responses. These insights help us move beyond the black box of LLMs and start to interpret the mechanisms behind their outputs. We now know that these models don't just memorize answers. They engage in genuine reasoning, planning, and even multi-step thought processes.
* However, this is just the beginning. As we've seen, the journey toward fully understanding AI is far from over. There are still limitations to our current interpretability tools, and many aspects of model behavior remain a mystery. But the progress made so far gives us a promising foundation for future work in this area.
* Ultimately, ongoing research into the biology of AI is critical for building safe, transparent, and trustworthy AI systems. As we continue to uncover the complexities of these models, we must remain mindful of the potential consequences and the need for careful, responsible development. So, as we conclude this presentation, I encourage you to keep asking questions.
* Why does the model think this way? How does it arrive at a specific conclusion? These are the kinds of questions that will drive the future of AI research and ensure that we use this powerful technology in a way that benefits everyone. Thank you and I look forward to your thoughts and questions. [Music]

