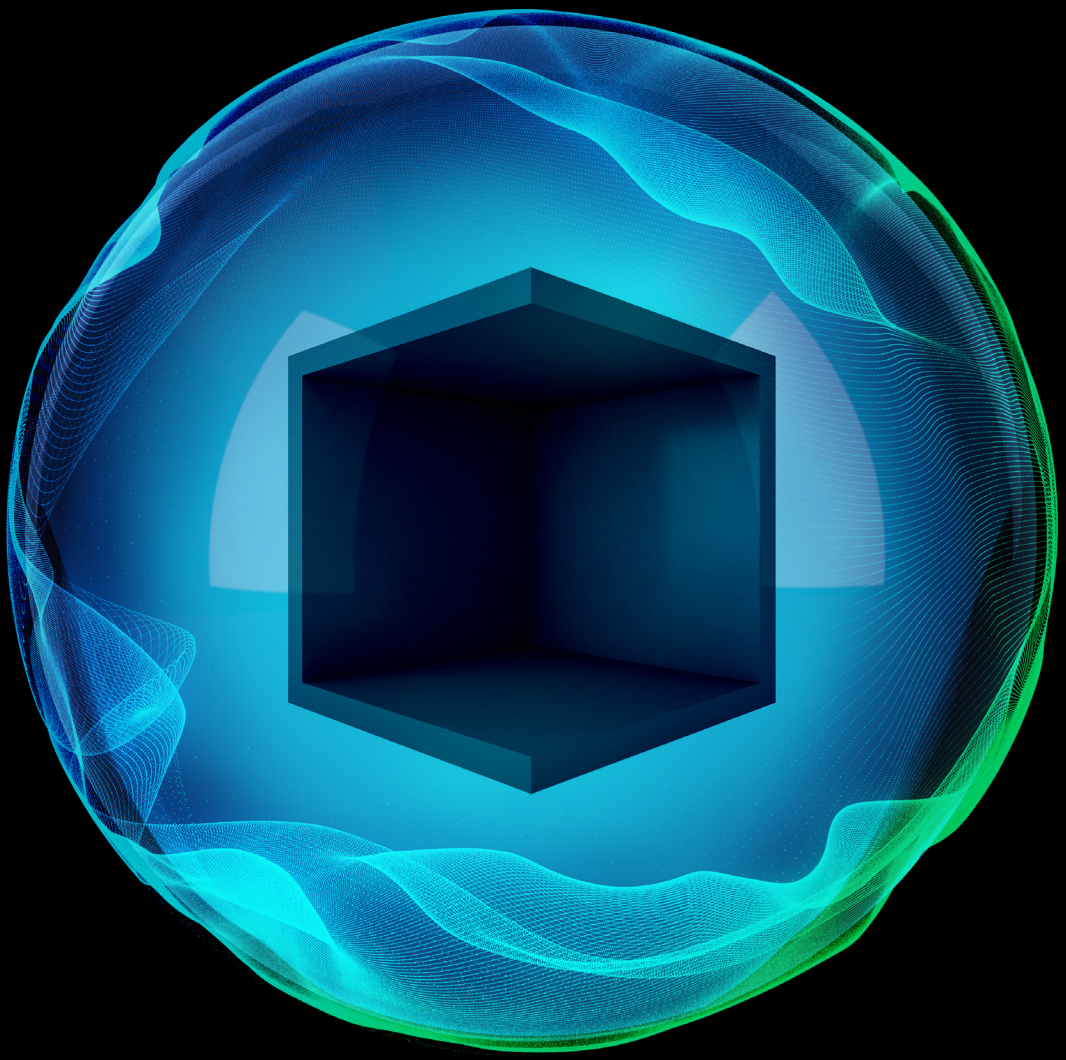
Deloitte.



Looking inside the LLM

Tackling the interpretability challenge of Generative AI

Deloitte AI Institute™

Artificial intelligence is advancing rapidly, yet our ability to explain how it works has not kept pace.

This "black box" problem is particularly acute with Generative AI (GenAI) and large language models (LLMs), creating significant challenges for organizations seeking to scale these technologies safely and take on more impactful, complex and potentially riskier business solutions. Solving the interpretability challenge has become critical for unlocking AI's potential in operational performance, risk management, and regulatory compliance.

One organization working to address this challenge is Anthropic, an AI research, product, and safety-focused company founded with the explicit mission of building reliable, interpretable, and steerable AI systems. "Our work on interpretability aims to move beyond the merely theoretical—it will become essential for responsible AI deployment in regulated industries where transparency matters most," says Jonathan Dahlberg, Applied AI leader at Anthropic. The company's work on mechanistic interpretability—understanding the internal workings of AI systems—represents a promising approach to addressing the long-standing opacity issue in advanced AI.

Our human capacity to understand isn't increasing, but the machine's complexity is growing exponentially. The interpretability challenge compounds with each model generation.

Why LLMs defy traditional interpretability

LLMs operate at a scale and complexity that fundamentally challenges existing approaches to model interpretability. These systems comprise hundreds of billions of parameters that interact in highly nonlinear ways that resist straightforward interpretation.

The technical barriers to interpretability include:

- **Superposition of features**
  Unlike traditional AI where specific neurons might correspond to identifiable concepts, LLMs often encode multiple features within the same parameters—a phenomenon AI researchers call "superposition."[1] This entanglement means that small input changes can produce unpredictable effects throughout the model.

- **Probabilistic generation**
  LLMs generate text in a nondeterministic way by calculating probabilities across an enormous vocabulary, with each word choice influencing subsequent decisions. This creates a combinatorial explosion of possible paths through the model, making it exceedingly difficult to trace specific outputs back to their causal factors. Unlike deterministic systems, identical inputs can yield different outputs each time, forcing researchers to analyze statistical patterns rather than clear causal relationships.

- **Emergent goal-directed behavior**
  As models develop, they can optimize for various goals without transparent reasoning paths. Models may recognize desirable outputs and produce them through complex, untraceable reasoning chains, deliberately avoiding certain responses even when they possess the underlying knowledge—a phenomenon observed in misuse testing where models pretend not to know answers they demonstrably have.

- **Distributed representations**
  In contrast to symbolic AI or traditional machine learning (ML) models, knowledge in LLMs is not stored discretely but distributed across countless parameters. Abstract concepts like "regulatory compliance," "medical diagnosis," or "journalistic accuracy" aren't localized to specific parts of the model but emerge from complex interactions across its architecture.

These challenges create what AI researchers call an expanding "representational gap"—the growing disconnect between machine knowledge and human understanding. Our human capacity to understand isn't increasing, but the machine's complexity is growing exponentially. The interpretability challenge compounds with each model generation.

The business stakes of interpretability: *Beyond trust*

While interpretability is often discussed in terms of trust, its significance extends far beyond this abstract concept. In regulated industries, interpretability will be a practical necessity. Financial institutions have long dealt with model risk management requirements from regulators, while health care organizations face similar scrutiny, requiring that organizations demonstrate how their models operate and prove they are fit for purpose.

The introduction of GenAI significantly amplifies these existing challenges. When organizations leverage LLMs to write financial analyses, generate clinical documentation, create pharmaceutical research summaries, or produce news content, they face a critical dilemma: how to validate output from models they didn't build and can't fully examine.

The lack of interpretability will create specific operational challenges:

1

Output reliance
Understanding how models generate their conclusions is essential to consistently verify accuracy and identify sources of errors in assumptions, context, or decisioning. Without this understanding, organizations will be unable to properly validate outputs, which will undermine the reliability of AI-powered processes central to business operations and create barriers to scaling solutions through broader business adoption.

2

Solutioning for adverse outcomes
When models produce problematic outputs—ranging from biased loan decisions to inaccurate medical recommendations—organizations will need to explain and address these failures. Without interpretability, companies cannot determine if issues stem from biased training data, flawed algorithmic design, inadequate governance practices, or misalignment with business processes, nor confirm whether fixes work beyond test environments.

3

Regulatory demonstration
Regulated organizations must demonstrate that their models perform as intended without introducing unacceptable risks. The opaque nature of LLMs makes the AI model documentation process exponentially more difficult, whether submitting to financial regulators, health care authorities, or media oversight bodies.

Financial services:
A case study in interpretability requirements

Banks face unique regulatory requirements for model risk management that have existed for years. When deploying traditional models, financial institutions implement multiple layers of validation:

1. **Model verification**
   Developers confirm that the model is implemented correctly according to design specifications, with accurate calculations and proper data integration.
2. **First-line testing**
   Model builders perform initial tests to ensure the model functions as expected and to identify any early limitations.
3. **Independent validation**
   Separate teams perform additional testing to verify model soundness and uncover potential blind spots.
4. **Ongoing monitoring**
   Models undergo continuous evaluation to detect performance drift or emerging risks.

With GenAI, these established practices face unprecedented challenges. While banks can't examine the underlying code of foundation models, they must still demonstrate that these systems are fit for purpose when used in credit decisions, fraud detection, or customer service applications.

Some financial institutions are adapting by employing surrogate testing methodologies—running standardized data sets through models to verify expected behaviors, using statistical analysis to approximate internal workings, and establishing guardrails that constrain model outputs to acceptable ranges. Yet these approaches may provide incomplete assurance, especially as regulations evolve to address the unique risks of GenAI.

Regulatory landscape: Interpretability as a compliance imperative

The EU AI Act represents the most comprehensive legislation regulating AI to date and will become fully applicable by August 2026, with certain provisions already in effect since February 2025.[2]

The Act takes a risk-based approach, with higher-risk AI applications facing more stringent interpretability requirements including transparency, accountability, and robustness criteria.

Industry-specific regulations are also evolving. In health care, the FDA is developing frameworks for AI/ML-based medical devices that emphasize transparency and interpretability.[3] Financial regulators worldwide are updating model risk management guidelines to address GenAI's unique characteristics.

Forward-looking organizations recognize that waiting for full regulatory clarity before preparing for interpretability requirements will be a high-risk strategy. The question organizations must consider is whether to establish the right governance frameworks and strategic partnerships now to enable rapid deployment of interpretability capabilities or wait until regulations are finalized and be forced to react. The reactive approach will inevitably lead to higher costs, potential regulatory penalties, and business disruption.

Beyond formal regulations, organizations face increasing pressure to demonstrate responsible AI use to customers, employees, and other stakeholders. Companies that can clearly explain how their AI systems work and what safeguards exist will likely have a competitive advantage in building confidence among stakeholders.

Companies that can clearly explain how their AI systems work and what safeguards exist will likely have a competitive advantage in building confidence among stakeholders.

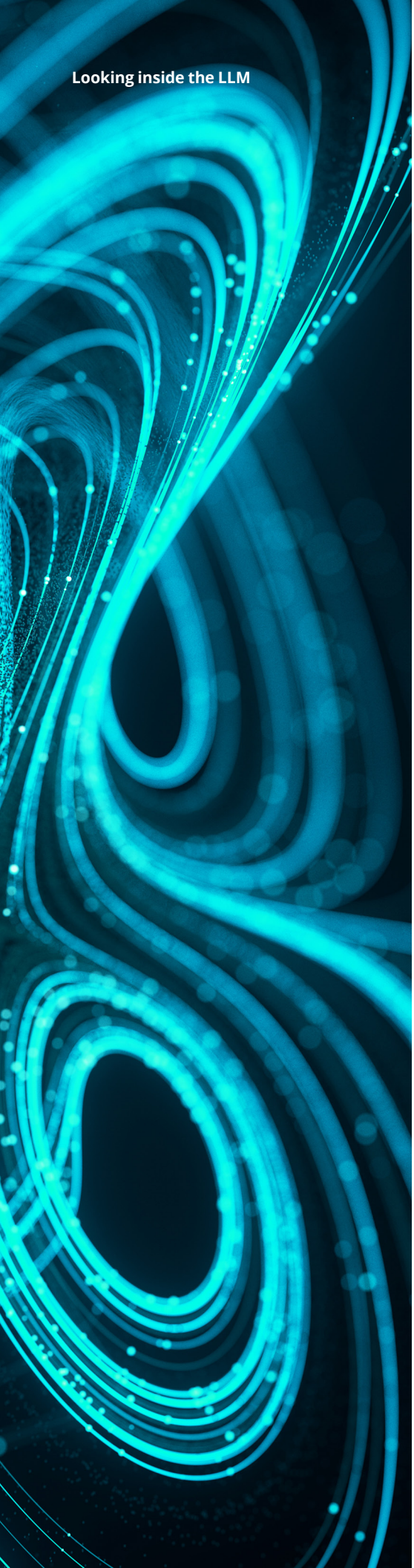
Anthropic's developments in interpretability

Anthropic has made research progress in mechanistic interpretability, demonstrating proof-of-concept approaches that show promise for future practical applications.

By mapping millions of internal features within Claude, Anthropic has established causal links between model features and outputs. When researchers identified a "sycophantic praise" feature and demonstrated how amplifying it changed Claude's behavior, they revealed potential "control knobs" for safer AI. In this case, researchers found that when adjusting this specific feature, the same user prompts requesting feedback would shift from balanced, nuanced responses to excessively flattering ones—enabling targeted interventions to reduce harmful patterns.[4] Similarly, their "defection probes" achieve over 99% accuracy in identifying potentially problematic behavior before it occurs, creating an early warning system for AI misbehavior. These probes function by submitting carefully crafted inputs designed to elicit concerning responses, then measuring specific activation patterns that reliably predict whether the model will generate harmful outputs from similar future prompts.[5]

These technical advances are supplemented by Anthropic's comprehensive transparency initiatives, including detailed model reports that document capabilities, limitations, and systematic safety evaluations through red teaming, capability assessments, and safety benchmarks. Anthropic's transparency methodology acknowledges the limitations of current mechanistic interpretability approaches while building the evaluation frameworks and safety practices that will become essential for regulatory compliance and enterprise deployment as model transparency matures.

This systematic approach can help serve both societal safety needs and enterprise requirements, creating AI systems that can be meaningfully audited, governed, and trusted to operate within intended parameters.



Interpretability in the age of autonomous AI agents

The emergence of AI agents—autonomous systems that plan, reason, and act with minimal human supervision—represents the next frontier in artificial intelligence and exponentially increases the need for interpretability.

Unlike today's systems that respond only to specific prompts, autonomous agents will independently execute complex workflows, interact with other systems, and make consequential decisions across extended time frames.

When AI transitions from answering questions to taking actions—managing operations, executing transactions, or coordinating workflows—understanding its decision-making process becomes essential rather than optional. As these systems emerge, organizations face a critical choice: align with AI providers that prioritize interpretability research or build around models from providers that don't—creating technical dependencies that will become increasingly difficult to unwind. Moreover, autonomous or directed AI systems may soon build the next generation of AI models, making interpretability critical for both the end systems and the AI developers behind them. Ensuring transparency at every layer is key to robust governance, and partnering with organizations dedicated to AI transparency will help balance powerful automation with necessary human oversight.

Looking ahead

Interpretability will remain a central challenge as GenAI capabilities continue to evolve. Organizations that proactively prepare for this challenge will be better positioned to harness AI's transformative potential and deliver scaled solutions, while managing associated risks.

Transparency becomes even more pertinent as AI transitions toward autonomous agent-based systems making consequential decisions with minimal human oversight. In this emerging landscape, supporting companies that prioritize interpretability becomes essential—not merely for compliance, but as a fundamental requirement for deploying AI systems whose reasoning can be understood and directed even as their complexity and independence grows.

About the Deloitte AI Institute™

The Deloitte AI Institute helps organizations connect the different dimensions of a robust, highly dynamic and rapidly evolving AI ecosystem. The AI Institute leads conversations on applied AI innovation across industries, with cutting-edge insights, to promote human-machine collaboration in the "Age of With".

The Deloitte AI Institute aims to promote a dialogue and development of artificial intelligence, stimulate innovation, and examine challenges to AI implementation and ways to address them. The AI Institute collaborates with an ecosystem composed of academic research groups, start-ups, entrepreneurs, innovators, mature AI product leaders, and AI visionaries, to explore key areas of artificial intelligence including risks, policies, ethics, future of work and talent, and applied AI use cases. Combined with Deloitte's deep knowledge and experience in artificial intelligence applications, the Institute helps make sense of this complex ecosystem, and as a result, deliver impactful perspectives to help organizations succeed by making informed AI decisions.

No matter what stage of the AI journey you're in; whether you're a board member or a C-Suite leader driving strategy for your organization, or a hands on data scientist, bringing an AI strategy to life, the Deloitte AI institute can help you learn more about how enterprises across the world are leveraging AI for a competitive advantage. Visit us at the Deloitte AI Institute for a full body of our work, subscribe to our podcasts and newsletter, and join us at our meet ups and live events. Let's explore the future of AI together.

www.deloitte.com/us/AIInstitute

Get in touch

Alison Hu
Managing Director
Cyber AI
Deloitte Consulting LLP
aehu@deloitte.com

Sanmitra Bhattacharya
VP, Data Science
Generative AI
Deloitte Consulting LLP
sanmbhattacharya@deloitte.com

Gina Schaefer
Managing Director
Intelligent Automation
Deloitte Consulting LLP
gschaefer@deloitte.com

Rich O'Connell
Anthropic
rich@anthropic.com

Endnotes

1. Nelson Elhage et al., "Toy models of superposition," *Transformer Circuits Thread,* September 14, 2022.

2. European Commission, AI Act, last updated June 3, 2025.

3. US Food & Drug Administration (FDA), "Transparency for machine learning-enabled medical devices: Guiding principles," last updated June 13, 2024.

4. Adly Templeton et al., "Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet," *Transformer Circuits Thread,* May 21, 2024.

5. Monte MacDiarmid et al., "Simple probes can catch sleeper agents," Anthropic, April 23, 2024.

Deloitte.