Interpretability and Trust in Large Language and Agentic Models: A Survey of Methods, Metrics, and Applications

Jithesh Yemi Reddy
Independent Researcher, TX, USA
y.jithesh@gmail.com

Abstract

Large-scale language models (LLMs) and the agentic systems that embed them are being deployed across finance, healthcare, law and other high-stakes domains. Their emergence has intensified concerns about *interpretability*—the ability to understand how the models produce their outputs—and *trust*—the confidence that the models behave reliably and ethically. The opaque internal representations of deep neural models mean that decisions may be unpredictable or unfair, undermining public confidence and limiting adoption. This paper surveys the state of interpretability and trust for both standalone LLMs and *agentic AI systems*, synthesising methodological advances, evaluation metrics and real-world applications. We organise methods into feature-attribution techniques such as LIME and SHAP, example-based and counterfactual explanations, process-level and mechanistic interpretability, and system-level approaches tailored to agentic multi-agent systems. We then review evaluation frameworks that measure explanation quality, fairness, robustness and other trust dimensions, including recent benchmarks like TrustLLM and psychometric scales for human–LLM trust. We discuss how interpretability interacts with safety, robustness, privacy and ethics, and how adaptive monitoring and balanced evaluation frameworks can promote trustworthy deployment. Finally, we highlight open research challenges in ensuring that increasingly autonomous agentic systems remain transparent, accountable and aligned with human values.

1 Introduction

The success of large language models (LLMs) such as GPT-4, Llama-2 and Claude arises from training on massive text corpora, enabling them to encode complex patterns of grammar, semantics and world knowledge. Modern models may have billions or trillions of parameters, but this scale creates an *opacity problem*—the internal workings are "black boxes" that are difficult to decipher. It is challenging to trace how specific inputs lead to a particular output or to understand the abstract features the model has learned. This opacity poses practical barriers to deployment in high-stakes domains such as healthcare, finance and law because trust, reliability and safety cannot be assured.

These concerns are amplified when LLMs are embedded into *agentic systems* that can plan, act and interact with the environment. Agentic LLMs execute tool calls, persist memory and coordinate with other agents. A recent survey defined agentic LLMs as models that *reason, act and interact* and noted that the literature can be organised around these three capabilities. As such systems transition from research prototypes to real-world assistants, they introduce new forms of opacity: decisions emerge from long-horizon planning, multi-step tool invocations and interactions with other agents. Errors may arise not from single predictions but from emergent behaviour in multi-agent networks.

Interpretability and trustworthiness are therefore necessary for responsible deployment. *Interpretability* refers to the degree to which a human can understand the internal mechanisms or reasons for a model's output, whereas *trust* encompasses wider notions of predictability, reliability, robustness, safety and accountability. Interpretability is a prerequisite for trust but not sufficient on its own. Building and maintaining trust require continuous monitoring and evaluation across technical and sociotechnical dimensions.

This survey provides a comprehensive overview of interpretability and trust in LLMs and agentic AI. Section2 defines key concepts and reviews categories of interpretability methods. Section3 examines metrics and frameworks for evaluating explanations and trustworthiness. Section4 discusses monitoring and adaptive evaluation. Section5 analyses applications, including time-series forecasting, natural language processing and multi-agent systems. Section6 outlines open research directions, and Section7 concludes.

2 Interpretability Methods

Interpretability methods aim to make a model's reasoning transparent. We divide them into four broad categories: (1) feature-attribution methods, (2) example-based and counterfactual explanations, (3) process-level and mechanistic interpretability, and (4) system-level interpretability for agentic systems.

2.1 Feature-Attribution Methods

Feature-attribution techniques quantify how individual input features contribute to a model's prediction. Two widely used methods are *LIME* (Local Interpretable Model-agnostic Explanations) and *SHAP* (SHapley Additive exPlanations). SHAP derives from cooperative game theory and explains an instance's prediction by computing the contribution of each feature; it offers various visualisations such as summary and dependence plots. LIME, in contrast, builds a simple surrogate model around an individual prediction to explain local behaviour; it highlights which features of a particular instance push the prediction towards a certain class. These methods have been applied to diverse tasks, including time-series forecasting and text classification. Shukla (2025c) used LIME and SHAP to interpret classical time-series models on the Air Passengers dataset, demonstrating how monthly and seasonal patterns contributed to passenger demand. In another study, Shukla (2025d) applied LIME and SHAP to BERT models, highlighting token-level contributions and revealing biases in sentiment analysis. Feature-attribution methods provide intuitive, human-friendly insights but are local and can be unstable under small perturbations.

Beyond LIME and SHAP, attribution methods include saliency maps, integrated gradients, DeepLIFT, Grad-CAM and attention visualisation. These techniques are often model-specific and work best with neural networks. For example, integrated gradients compute the average gradient of the output with respect to each feature along a straight path from a baseline input to the actual input, yielding an attribution that satisfies properties like sensitivity and completeness. Attention visualisation leverages the self-attention weights of transformers to show which tokens attend to others, though it has been criticised for lack of faithfulness because attention weights do not always correspond to causal importance.

2.2 Example-Based and Counterfactual Explanations

Example-based explanations communicate model decisions through representative examples, prototypes or nearest neighbours. A prototype-based explanation might show the training examples most similar to the input, while criticism-based explanations display unusual or borderline cases. Counterfactual explanations identify minimal changes to the input that would change the model's prediction; for instance, showing that increasing a credit applicant's salary by \$5 ,000 would change a loan decision from "reject" to "approve." Such explanations are intuitive and can guide recourse actions but may be challenging to compute for high-dimensional inputs and may lack global insight.

2.3 Process-Level and Mechanistic Interpretability

Process-level methods aim to understand how LLMs perform intermediate reasoning. Chain-of-thought prompting reveals step-by-step reasoning, enabling inspection of intermediate steps. However, these generated rationales may not reflect the model's actual internal reasoning and can be influenced by prompting; thus they must be evaluated for *faithfulness*.

Mechanistic interpretability seeks to reverse-engineer neural networks into interpretable components. It examines the weights, activations and circuits of the model to discover substructures responsible for specific tasks. A recent review emphasised that while traditional mechanistic interpretability analyses a fully trained, static model, *developmental interpretability* studies how mechanisms form and evolve during training. The review argues that understanding the formation of circuits and phase transitions is crucial for proactive AI safety. Representational probing trains small classifiers ("probes") to predict linguistic properties from internal activations; high accuracy suggests that the model encodes these properties. Circuit discovery techniques identify attention heads and neuron groups that implement specific computations such as induction heads in transformers. Mechanistic interpretability has shed light on emergent capabilities but remains labour-intensive and limited to small models.

2.4 System-Level Interpretability for Agentic Systems

System-level interpretability focuses on explaining entire agentic systems rather than individual models. Agentic systems integrate LLMs with tool invocations, memory management and multi-agent coordination. Explanations must therefore cover sequences of actions, tool calls, memory updates and interactions among agents. Techniques such as behaviour trees, flowcharts and causal diagrams can depict high-level plans, while logs of tool usage and message passing provide granular details. In multi-agent settings, communication protocols and consensus mechanisms should be transparent. System-level interpretability remains an open challenge because emergent behaviours can arise from interactions that are difficult to predict or analyse.

3 Metrics and Evaluation Frameworks

Evaluating interpretability and trust requires metrics that assess explanation quality and broader trust dimensions. We review metrics for explanation fidelity, fairness, robustness and other aspects and discuss existing benchmarks and psychometric scales.

3.1 Explanation Quality Metrics

Common metrics evaluate explanations along several dimensions: fidelity (how well explanations reflect true model behaviour), stability (robustness to perturbations), sparsity (simplicity), and human alignment (consistency with human reasoning). Fidelity can be measured by training a simple surrogate model on the attributions and comparing its predictions to the original model. Stability examines whether small changes in the input produce drastic changes in the explanation. Human alignment often involves user studies where explanations are rated for clarity and usefulness. These metrics remain largely theoretical and may not capture real-world trust.

3.2 Trustworthiness Benchmarks

Trust extends beyond interpretability to encompass multiple dimensions. The *TrustLLM* benchmark introduces a taxonomy of eight aspects of trustworthiness: *truthfulness*, *safety*, *fairness*, *robustness*, *privacy*, *machine ethics*, *transparency* and *accountability*. The benchmark covers over 30 datasets and

evaluates 16 mainstream LLMs across six of these aspects (transparency and accountability are difficult to benchmark), establishing that trustworthiness is closely related to a model's utility. Proprietary models often outperform open-source models on trust metrics, but open-source models like Llama2 can sometimes excel. The analysis indicates that over-alignment—models refusing safe requests—reduces utility and trustworthiness. These findings highlight the need for balanced alignment strategies that avoid both harmful outputs and unnecessary refusals.

The Livermore National Laboratory (LLNL) article summarising TrustLLM emphasises how each dimension should be measured. A *fair* model avoids discriminatory outcomes; *machine ethics* measures recognition of human morals; *privacy* requires that the model not reveal sensitive data; *robustness* refers to handling anomalies; *safety* refers to resilience against data manipulation; *truthfulness* demands factual responses; *accountability* requires tracing the origin of outputs; and *transparency* requires explaining decision-making steps. Accountability and transparency remain difficult to measure at scale, and none of the tested models were fully trustworthy. The article notes that evaluation across all metrics is essential because models may perform well in one dimension (e.g., fairness) but poorly in others (e.g., privacy).

3.3 Balanced Evaluation Frameworks and Human Trust Measures

Balanced evaluation frameworks integrate metrics across performance, robustness, safety and sustainability. Shukla (2025b) proposed a balanced framework for agentic systems that considers cost and latency as well as safety and robustness. Adaptive evaluation can prioritise dimensions based on deployment context, focusing on fairness and safety in social domains and on robustness in technical applications. The frameworks emphasise continuous monitoring and updating of evaluation criteria as new risks emerge.

Human-centric measures complement technical metrics. Psychometric scales such as the Trust-In-LLMs Index (TILLMI) measure individuals' trust in LLMs. TILLMI extends cognitive and affective trust dimensions from organisational psychology to human–LLM interactions and was validated on a sample of 1 ,000 U.S. respondents. Factor analysis identified two dimensions: *closeness with LLMs* (affective trust) and *reliance on LLMs* (cognitive trust). The index correlated positively with openness, extraversion and cog-

nitive flexibility and negatively with neuroticism. Younger males reported higher closeness and reliance than older women, and individuals with no direct experience with LLMs had lower trust. Psychometric scales like TILLMI complement technical benchmarks by capturing user perceptions, which influence adoption.

3.4 Ethical Evaluation Metrics

Ethical evaluation frameworks emphasise fairness, transparency, safety and accountability. A 2025 article identifying top metrics for ethical LLM evaluation lists seven key measures: *bias detection*, *accuracy*, *transparency & explainability*, *toxicity & harm detection*, *factual accuracy*, *privacy & data protection*, and *accountability & audit tracking*. Bias detection is essential to reduce discrimination across demographic groups and involves both intrinsic evaluations (inspecting model mechanisms) and extrinsic tests (performance on diverse datasets). Accuracy metrics should not obscure disparities across groups; toxicity detection flags harmful content such as hate speech; factual accuracy ensures truthful and verifiable outputs; privacy metrics assess data leakage; and accountability metrics track decisions and assign responsibility. The article stresses that combining human judgment with automated evaluations yields robust assessments and that tools like AI Fairness 360, FairLearn, LIME and SHAP support bias detection.

3.5 Evaluation Challenges and Best Practices

Evaluating LLMs and agentic systems is challenging. Overfitting to popular benchmarks can mislead evaluations. Existing metrics may not capture diversity and creativity, leading to an overemphasis on accuracy at the expense of coherence, relevance and diversity. Human evaluation is subjective and costly, while automated metrics can be biased. Robustness evaluations often overlook adversarial attacks. Best practices include combining multiple metrics, curating diverse reference data, incorporating real-world tasks and evaluating robustness to adversarial inputs. Adaptive monitoring, discussed in the next section, addresses some of these limitations by evaluating models in deployment.

4 Monitoring and Adaptive Evaluation

Once deployed, LLMs and agentic systems may encounter new data distributions, adversarial inputs or evolving user requirements. Static evaluation cannot anticipate all contingencies. Monitoring frameworks track model behaviour in real time, detecting performance degradation, distribution shifts and safety violations. Adaptive evaluation uses monitoring data to update trust metrics and trigger interventions such as retraining or rollback.

4.1 Monitoring Dimensions

Monitoring can target multiple dimensions: predictive performance (accuracy, calibration), fairness (disparities across groups), robustness (out-of-distribution robustness and adversarial resilience), safety (toxic output detection and harmful instruction refusal), privacy (data leakage and training-data memorisation), cost (latency and compute usage) and sustainability (energy consumption). Metrics should be contextualised; for example, fairness metrics may prioritise different protected attributes depending on application domain. Monitoring should incorporate user feedback and domain knowledge.

4.2 Adaptive Monitoring for Agentic Systems

Adaptive monitoring tailors evaluation frequency and scope to system behaviour. When a model drifts, monitors may increase sampling rates or expand the set of metrics. For agentic systems with long-horizon planning, monitors may record sequences of tool calls and decisions and identify patterns indicative of unsafe or unethical behaviour. A balanced evaluation framework can assign weights to metrics such that high-impact failures (e.g., safety violations) trigger immediate intervention, while minor performance drops are tolerated temporarily. Adaptive monitoring thus focuses resources where risks are greatest.

4.3 Governance and Accountability

Governance structures must define who is accountable for agentic AI behaviour and how monitoring data are used. Accountability involves logging decisions and enabling audits. Transparency requires providing explanations

for decisions and monitoring results to stakeholders. Regulatory frameworks like the NIST AI Risk Management Framework emphasise risk identification, mitigation and continuous monitoring. For agentic systems, governance may include formal verification of decision logic, red-teaming to uncover vulnerabilities, and alignment with ethical guidelines. Adaptive monitoring thus serves not only technical evaluation but also compliance and societal oversight.

5 Applications

5.1 Time-Series Forecasting

LLMs and traditional models are increasingly used for time-series forecasting in domains such as demand planning and energy management. In Shukla's case study (2025c), LIME and SHAP were applied to a classical forecasting model on the Air Passengers dataset. The explanations revealed that *seasonal patterns* and *trend components* were the most influential features; high passenger numbers were driven by summertime peaks, while lower numbers corresponded to winter months. By attributing forecasted increases to specific months and exogenous factors, decision-makers could prepare capacity accordingly. The case illustrates how feature-attribution methods help domain experts understand why forecasts change and foster trust in automated predictions.

5.2 Natural Language Processing and Document Intelligence

Interpretability in NLP is critical for tasks such as sentiment analysis, information extraction and question answering. BERT-based models have achieved state-of-the-art performance but remain opaque. Shukla (2025d) applied LIME and SHAP to interpret BERT outputs, showing which tokens contributed most to predicted sentiments and identifying biases—for example, certain gendered pronouns disproportionately influencing negativity. More generally, token-level attributions can reveal how subword tokens affect classification, enabling developers to detect spurious correlations. Process-level explanations (chain-of-thought) can further reveal reasoning paths but must be evaluated for faithfulness.

5.3 Agentic Systems and Multi-Agent Collaboration

Agentic systems extend LLM capabilities by enabling tool use, long-horizon planning and interaction with other agents. Applications include autonomous coding assistants, personal shopping agents, robotic controllers and multi-agent research assistants. Real-world deployments show significant productivity gains but also highlight trust challenges. Balanced evaluation frameworks indicate that agentic systems may excel in capability but underperform in fairness, safety or sustainability. For instance, an autonomous coding agent might generate functioning code quickly but inadvertently produce insecure or non-compliant code. Continuous monitoring can detect such issues by tracking security vulnerabilities and adherence to coding standards. Multi-agent systems introduce additional complexities: communication protocols must be transparent, and emergent behaviours must be monitored for alignment with human goals. Researchers have proposed visualising agent communication graphs and implementing role-based access controls to maintain transparency.

5.4 Human-AI Interaction and Trust

Applications where humans interact directly with LLMs—chatbots, educational tutors, legal assistants—require a fine balance between performance and trust. Studies using the TILLMI index found that trust correlates with user characteristics and prior experience. Designers can foster trust by providing clear explanations, allowing users to inspect reasoning paths and giving them control over decisions. For example, a medical diagnostic assistant may present differential diagnoses along with SHAP attributions for each symptom. User satisfaction and trust metrics should be monitored continuously. Systems must also avoid over-alignment that causes refusal of benign queries, which can frustrate users and erode trust.

6 Open Challenges and Research Directions

Although significant progress has been made, many challenges remain:

1. **Faithfulness of explanations.** Generated explanations (e.g., chain-of-thought) may not reflect the model's actual reasoning, and attribution methods can be manipulated. Research is needed to quantify

faithfulness and develop methods that reliably reflect internal computations.

2. **Scalability of mechanistic interpretability.** Circuit analysis and developmental interpretability are labour-intensive and currently feasible only for small models. Automatic tools and abstraction techniques are needed to scale to billion-parameter models.

3. **System-level transparency.** For agentic systems, explaining sequences of actions, tool calls, memory updates and inter-agent communication is challenging. New approaches are needed to summarise behaviours at appropriate levels of abstraction and to support audits.

4. **Balancing alignment and utility.** Over-alignment can lead to excessive refusals, while under-alignment may permit harmful outputs. Methods that adjust alignment dynamically based on context and user feedback, and evaluation frameworks that quantify this balance, are essential.

5. **Standardised benchmarks.** Current evaluation benchmarks are fragmented; trust dimensions like privacy and accountability remain difficult to measure. Common benchmarks across fairness, safety, privacy and other dimensions will facilitate comparison.

6. **Human-centric evaluation.** Technical metrics must be complemented with psychometric scales such as TILLMI to capture user trust. Cross-cultural studies are needed to understand how trust varies across demographics and applications.

7. **Privacy-preserving interpretability.** Attribution and probing methods may inadvertently reveal sensitive training data. Research into privacy-preserving explanations—e.g., using differential privacy or synthetic attributions—is needed to protect data while maintaining transparency.

8. **Integration with governance and law.** Interpretability and trust metrics must align with emerging regulatory frameworks such as the EU AI Act and NIST AI RMF. Mechanisms for auditing, certification and redress must be integrated into deployment pipelines. The role of institutions—such as third-party auditors—and the legal status of agentic systems remain open questions.

7 Conclusion

The rapid adoption of LLMs and agentic AI systems heightens the urgency of ensuring that these technologies are transparent, trustworthy and aligned with human values. Interpretability methods—feature-attribution, example-based, process-level and system-level—provide crucial insights into model behaviour. Trustworthiness encompasses multiple dimensions, from truthfulness and safety to privacy and accountability, and requires comprehensive evaluation. Balanced frameworks and adaptive monitoring promise to bridge the gap between technical metrics and real-world trust, while psychometric instruments like TILLMI capture human perceptions. Nevertheless, challenges such as faithful explanations, scalable mechanistic interpretability, system-level transparency and alignment trade-offs persist. Addressing these issues will require interdisciplinary collaboration across machine learning, human–computer interaction, ethics, law and governance. By advancing interpretability and trust frameworks, we can unlock the potential of LLMs and agentic AI while safeguarding users and society.

References

[1] Shukla, M. (2025c). *Interpreting Time Series Forecasts with LIME and SHAP: A Case Study on the Air Passengers Dataset*. Research Square Preprint.

[2] Kendiukhov, I. (2025). *A Review of Developmental Interpretability in Large Language Models*. arXiv preprint.

[3] Kailkhura, B., et al. (2024). *Evaluating Trust and Safety of Large Language Models*. LLNL News.

[4] DataCamp. (2024). *Explainable AI: Understanding and Trusting Machine Learning Models*. Tutorial.

[5] Shukla, M. (2025a). *Adaptive Monitoring and Real-World Evaluation of Agentic AI Systems*. arXiv preprint arXiv:2509.00115.

[6] De Duro, E. S., Veltri, G. A., Golino, H., & Stella, M. (2025). *Measuring and identifying factors of individuals' trust in Large Language Models*. arXiv preprint.

[7] Shukla, M. (2025b). *Evaluating Agentic AI Systems: A Balanced Framework for Performance, Robustness, Safety and Beyond*. SSRN Preprint.

[8] Plaat, A., van Duijn, M., van Stein, N., Preuss, M., & others. (2024). *Agentic Large Language Models, a survey*. arXiv preprint.

[9] AIMultiple. (2025). *Large Language Model Evaluation: 10+ Metrics & Methods*. Research Article.

[10] Shukla, M. (2025d). *Interpreting BERT Using LIME and SHAP*. Research Square Preprint.

[11] TrustLLM: *A Comprehensive Study of Trustworthiness in Large Language Models*. 2024–2025.

[12] Miguelañez, C. (2025). *Top 7 Metrics for Ethical LLM Evaluation*. Latitude Blog.