From Models to Systems: A Survey of Explainability for Tool-Augmented Language Models and AI Agents

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly being used as part of complex agentic systems that orchestrate the use of external tools, such as retrieval mechanisms or code interpreters. In this survey, we argue that this development necessitates a rethinking of the goals of explainable artificial intelligence (XAI): Rather than focusing on providing users with explanations for *monolithic* machine learning models, we need *system-level* explanations that also provide information about which and how tools are used, as well as how external execution traces causally influence system behavior. We provide an overview of the existing methods in explainable AI and discuss the limitations of monolithic XAI methods in agentic contexts. Finally, we highlight open challenges in providing faithful explanations for LLM-based systems.

1 Introduction

Large language model (LLM)-based artificial intelligence (AI) systems continually grow in complexity and increasingly rely on sets of tools that interact with one another. This growing complexity makes it difficult for users to understand and predict the LLM's behavior accurately. At the same time, the shift from standalone models to multi-component systems introduces new forms of opacity at the system level. Traditional approaches in explainable artificial intelligence (XAI), such as input attribution or mechanistic interpretability, however, mainly assume monolithic, end-to-end differentiable models and do not address the explanation needs that arise from tool use, which is now central to modern LLM-based systems and AI agents. In this work, we survey existing explainability research from the perspective of tool-using LLM systems, that is, architectures that integrate external components into their inference process.

AI agents and augmented LLMs extend standalone generative models by integrating external information sources, tools, and structured interaction loops into inference. A first class of approaches focuses on *capability delegation*, in which LLMs rely on external components to perform well-defined subtasks. A canonical example is retrieval-augmented generation, which augments LLMs with mechanisms to query knowledge bases and condition generation on retrieved context (Lewis et al., 2020; Izacard and Grave, 2021). Augmentation mechanisms also include web search or external components like calculators or translation systems (Komeili et al., 2022; Thoppilan et al., 2022). Program-aided approaches couple LLMs with execution environments, such as Python interpreters, allowing models to generate programs while delegating computation to symbolic runtimes (Gao et al., 2023a). Works like Toolformer (Schick et al., 2023) move beyond manually specified delegation by enabling LLMs to learn when and how to select, invoke, and incorporate tools during generation. Recent architectures build on this idea through the framework of *agentic orchestration*, in which LLMs act as high-level controllers that plan tasks, select domain-specific models, and synthesize results (e.g., HuggingGPT : Shen et al., 2023). Subsequent systems scale tool use across large and diverse API spaces through parallel function execution (Kim et al., 2024), enhance robustness via self-reflection mechanisms (Du et al., 2024), or employ multiple LLMs through mixture-of-agents designs (Wang et al., 2025a). Efforts such as the Model Context Protocol (Parra and Spahr-Summers, 2024) standardize tool interfaces, enabling modular and interoperable agent ecosystems.

These developments increase functional expressivity but complicate explanation, as model outputs now emerge from interactions between the LLM, tools, and execution environments rather than a single forward pass, shifting the focus of explainability from *"why did the model output X?"* to *"how did delegation, orchestration, and interaction*

with external components affect X?". Augmented systems introduce qualitatively different explanation targets from those addressed by traditional XAI, raising critical questions from an explainability perspective: *Which tools are used and how do they work? How does the LLM invoke tools and use their output to generate responses? To what extent does the LLM rely on tools?*

In contrast to prior work, these questions concern decisions made at the *system level* rather than within a single model. Such explanations are particularly relevant given that users form mental models of LLM-based AI systems, holding beliefs about how they store information, reason, learn, and make decisions. Some of these beliefs are empirically reasonable given current system capabilities, while others reflect genuine misconceptions about model architecture or function. Users often believe that models access online information or external databases (Wang et al., 2025c), or that assistants "remember" and "learn" from prior interactions (Zamfirescu-Pereira et al., 2023), which may be true for systems that integrate retrieval, query persistent chat histories, or possess personalization layers. At the same time, users may incorrectly assume that models modify their own parameters or perform computations between sessions (O'Brien et al., 2025; Zamfirescu-Pereira et al., 2023). They may also overly trust intermediate outputs such as plans or rationales, interpreting them as indicators of correctness (Buçinca et al., 2021; He et al., 2025; Kim et al., 2025; Park et al., 2025). More generally, perceptions of factual reliability (Gessinger et al., 2025) or moral neutrality (Tolsdorf et al., 2025) are often unwarranted, as aligned models can still hallucinate, reflect biases, or reason inconsistently. Explanations for tool-using LLMs must therefore expose relevant system-level mechanisms and support the appropriate level of trust, rather than reinforcing surface-level intuitions.

This survey aims to provide an overview of research on explainability for tool-using LLMs and agents, and to identify open challenges in explaining them. In contrast to work focused on monolithic models, we adopt a system-level perspective on explainability for tool-augmented inference.

2 Background: Explainability in Non-Tool-Augmented Systems

Before discussing explainability methods that have been designed for tool-using systems, we briefly assess methods that have been developed for explaining the behavior of monolithic models, including LLMs without explicit tool use and simpler neural networks. These methods, while foundational, typically assume a single model or inference process, and therefore do not directly account for delegation, orchestration, and interaction with external components in tool-using LLM systems.

Early surveys on XAI focused on defining interpretability, motivating its necessity, and situating explanation within human and social contexts. Doshi-Velez and Kim (2017) emphasized the need for task-specific and human-grounded evaluation, while Miller (2019) grounded explainability in insights from the social sciences. Subsequent work proposed taxonomies and unifying frameworks for XAI (Murdoch et al., 2019; Arrieta et al., 2020; Vilone and Longo, 2021), surveyed complementary perspectives such as training data attribution (Hammoudeh and Lowd, 2024) and explanation evaluation (Zhou et al., 2021; Nauta et al., 2023).

With the advent of large transformer-based language models (Vaswani et al., 2017; Devlin et al., 2019), a parallel line of work examined explainability for NLP and LLMs. These surveys addressed issues such as faithfulness (Jacovi and Goldberg, 2020), reviewed attribution, probing, and mechanistic approaches (Luo et al., 2024; Lyu et al., 2024; Ferrando et al., 2024), and discussed how scale and instruction tuning challenge traditional interpretability notions (Zhao et al., 2024; Luo and Specia, 2024; Singh et al., 2024). Despite their breadth, these surveys largely conceptualize explainability at the level of standalone models and single inference passes, giving limited attention to tool use, orchestration, and multi-step system behavior.

Input attribution. Input attribution methods explain model outputs by assigning importance scores to input features, such as words or tokens. Prominent approaches include the perturbation-based methods LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), as well as the gradient-based integrated gradients (Sundararajan et al., 2017). These methods established attribution as a central paradigm for providing explanations. In contrast, the use of attention weights as faithful attribution signals remains contested (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019).

However, applying input attribution to LLMs can be challenging. The discrete nature of text complicates baseline selection and interpolation

paths, motivating adaptations such as discretized integrated gradients (Sanyal and Ren, 2021). Further, perturbation-based methods scale poorly to the long contexts typical of LLMs, can produce out-of-distribution inputs, and are not directly applicable to text generation. This motivated extensions of SHAP-style attribution methods for explaining generated responses (Paes et al., 2025). Overall, input attribution can highlight parts of an input that influenced a specific output, but it remains limited to explaining single input–output pairs and does not capture the multi-step decision-making, tool invocation, or orchestration processes that characterize tool-using LLMs.

Training data attribution. Training data attribution explains model behavior by identifying influential examples from the training corpus, for instance via influence functions (Koh and Liang, 2017), which estimate the contribution of individual training documents without retraining (e.g., Guo et al., 2021; Grosse et al., 2023; Choe et al., 2024). In practice, these methods face substantial scalability limitations for LLMs and are often restricted to smaller models or fine-tuning data. More fundamentally, training data attribution explains how models acquire knowledge during training, but provides little insight into runtime decision-making, tool selection, and multi-step interactions that characterize tool-using LLM systems.

Probing. Probing aims to determine what kinds of information are encoded in a model's internal representations by training a simple classifier—often linear—to predict a property of interest from those representations (Belinkov, 2022). It has been used to study a wide range of linguistic and semantic properties, from morphology and syntax (Belinkov et al., 2017; Hewitt and Manning, 2019) to more complex phenomena such as world knowledge (Li et al., 2023). While probing can reveal what information a model *could* in principle use, it does not necessarily explain what information the model actually causally relies on when making predictions (Elazar et al., 2021). Results are also sensitive to probe complexity, requiring careful controls (Hewitt and Liang, 2019; Voita and Titov, 2020). In the context of tool-using LLM systems, probing has seen limited adoption, as it provides little insight into runtime decision-making across tools, though it has been used to analyze whether retrieval-augmented models rely on parametric or contextual knowledge (Tighidet et al., 2024).

Causal intervention. Causal intervention methods aim to identify which internal components of a neural network are causally responsible for specific behaviors, and are often associated with work on mechanistic interpretability (e.g., Geiger et al., 2021; Hanna et al., 2023). One approach is activation patching (Vig et al., 2020), which replaces internal activations between minimally different inputs to localize components that drive changes in outputs. Related methods include causal tracing (Meng et al., 2022) and other localization techniques for neurons or attention heads (e.g., Goldowsky-Dill et al., 2023; Wang et al., 2023a).

By disentangling causal from correlational effects, these methods offer a principled way to study model internals, but practical challenges arise when applied to LLMs including difficulty in isolating minimal, faithful subnetworks that implement specific functionality (Mueller et al., 2025). Moreover, causal intervention techniques have not yet been systematically applied to tool-using systems, where behaviors emerge not only from model internal mechanisms but also from interactions between models, tools, and execution environments.

Self-explanations. Self-explanation refers to LLMs producing an output together with a natural-language rationale intended to explain why that output was generated (Rajani et al., 2019). While such rationales are often coherent and plausible, they may not faithfully reflect the causal factors that guided generation (Jacovi and Goldberg, 2020; Barez et al., 2025). Accordingly, explanations are commonly evaluated along three dimensions: *Plausibility* captures whether explanations appear coherent and convincing to humans (Lage et al., 2019; Jacovi and Goldberg, 2020). *Faithfulness* asks if the explanation aligns with the causal decision process of the model, which can be tested through causal interventions or counterfactual manipulations (Lyu et al., 2024).[1] *Simulatability* captures if explanations help users anticipate model behavior or failure modes (Doshi-Velez and Kim, 2017; Hase and Bansal, 2020). In tool-using LLM systems, self-explanations are particularly limited, as such rationales may omit, misrepresent, or hallucinate tool invocations, intermediate decisions, and interactions with external components, potentially leading to unwarranted trust in system behavior.

[1] We define *faithfulness* as the extent to which an explanation reflects the true causal factors influencing system behavior, rather than its ability to provide plausible justifications.

3 Explainability for Tool-using LMs

Explainability for tool-using LLMs spans a spectrum between two complementary perspectives: using tools to generate explanations of system behavior, and explaining why and how tools were invoked during inference. Many systems combine both directions, for example, by calling analysis tools to produce explanations while simultaneously exposing tool calls or reasoning traces. In this section, we cover approaches across this spectrum at increasing levels of abstraction.

3.1 RAG and Explainability

Retrieval-augmented generation enhances LLMs by providing mechanisms to retrieve passages from knowledge bases and to condition generation on the retrieved context. Existing work on explainability in the context of RAG focuses on how the retrieved context shapes model output rather than on the retrieval mechanism and its system-level decisions. Consequently, RAG occupies a boundary position between output-grounding techniques and tool-level explainability.

RAG for information tracing and grounding. RAG is widely used to ground model outputs in curated sources, supporting fact tracing, citation, and provenance attribution (Sankararaman et al., 2024; Zhu et al., 2025b,a; Xia et al., 2025; Qi et al., 2024; Armant et al., 2024). By pairing answers with retrieved documents, these systems make outputs easier to verify. However, standard RAG setups may provide relevant documents, but do not explain why documents were retrieved or how retrieval influenced the model's reasoning. This limits their usefulness to knowledge-based tasks where users can verify model outputs against the ground truth in the retrieved documents.

RAG for generating explanations. RAG has also been used to generate explanations or rationales alongside predictions. Some approaches prompt RAG models to produce self-explanations conditioned on retrieved context (Garigliotti, 2025; Wei et al., 2025; Peng et al., 2025), while others adapt attribution methods to highlight influential parts of the combined prompt and context (Sudhi et al., 2024). A complementary line of work explains outputs in terms of retrieved sources by perturbing document sets or knowledge graphs to identify influential evidence (Rorseth et al., 2024; Balanos et al., 2025). Systems that use RAG to explain external processes rather than model behavior are out of scope in this survey (e.g., Minor and Kaucher, 2024; Li et al., 2025; Gao et al., 2023b).

Explaining and evaluating RAG models. Evaluation of RAG systems often draws on properties from interpretability literature, such as faithfulness, to capture the causal relationship between retrieved context and generated outputs. Prior work studies answer faithfulness and attribution groundedness (Gao et al., 2024; Qi et al., 2024; Song et al., 2025; Ye et al., 2024; Patel et al., 2024), and distinguishes between citation correctness and faithfulness (Wallat et al., 2025). Similar notions are reflected in benchmarks that assess context utilization, relevance, and answer completeness (Friel et al., 2025). Overall, however, RAG explainability primarily addresses grounding and justification of generated outputs. Most existing approaches explain answers instead of the retrieval and generation decisions, and thus provide only a partial explanation.

3.2 Tool-Level Transparency

Beyond information retrieval mechanisms, tool-using LLMs invoke external APIs or computational tools to perform reasoning and actions. At this level, explainability concerns how individual tool calls contribute to an agent's inference process, covering both the exposure of tool invocations themselves and explanations for why tools were used. Existing work spans from implicit transparency, to structured tool calls, to explicit, tool-supported explanation generation.

Transparent tool calls without explicit explanations. Structured tool calls that expose the tool name, input arguments, and outputs already provide a basic form of interpretability by making intermediate steps observable. Early systems such as ReAct (Yao et al., 2023b) interleave natural-language reasoning with external actions, allowing users to inspect which tools were invoked and with what effects. Toolformer (Schick et al., 2023) and HuggingGPT (Shen et al., 2023) generalize this paradigm by enabling LLMs to call diverse external tools through structured APIs. While these approaches do not generate dedicated explanations, their explicit tool-use traces make reasoning steps more transparent than end-to-end text generation.

Related work such as Gorilla (Patil et al., 2024) grounds tool calls in retrieved API documentation and evaluates correctness via syntactic matching,

further improving verifiability of tool usage. Similarly, standardization efforts like the Model Context Protocol(MCP) aim to improve transparency and interoperability of tool interfaces, facilitating inspection and auditing of individual tool calls.

Explanation generation with and for tool calls. Other systems explicitly generate explanations for tool usage or employ tools to construct explanations. SCRIBE (Fawzi et al., 2025) produces structured, step-by-step explanations grounded in tool outputs in educational settings, while ECHO (Vanbrabant et al., 2025) equips an LLM with access to model interfaces, data, and XAI utilities to orchestrate multi-step explanations of another model's behavior. Related approaches such as KnowThyself (Prasai et al., 2025) apply similar orchestration principles for model interpretability. Complementary work focuses on making intermediate reasoning artifacts explicit. Program-Aided Language Models (PAL) (Gao et al., 2023a) and Program-of-Thoughts (Chen et al., 2023) express reasoning as executable code, separating logical inference and computation from generation, enabling inspection of each step. Natural-language reasoning paradigms such as Chain-of-Thought (CoT) prompting (Wei et al., 2022), Tree-of-Thought (Yao et al., 2023a), and plan-based or self-reflective methods (Wang et al., 2023b; Shinn et al., 2023; Madaan et al., 2023) expose intermediate reasoning as structured text, improving transparency and often performance. However, different analyses show that such reasoning traces are frequently *unfaithful*: they describe plausible reasoning without necessarily reflecting the causal processes that produced the output (Turpin et al., 2023; Lanham et al., 2023; Paul et al., 2024). In tool-using settings, this limitation becomes more pronounced, as reasoning traces may plausibly justify tool use without corresponding to the actual tool calls executed or the causal role their outputs played in the final decision. Recent work on interactive reasoning (Pang et al., 2025) highlights this gap by enabling users to inspect, restructure, and intervene in CoT reasoning, suggesting that faithfulness in agentic contexts requires alignment not only between reasoning and output, but also between reasoning traces and the system's actions.

Overall, tool-level reasoning transparency improves interpretability by exposing or explaining individual tool invocations and intermediate decisions. However, these approaches typically focus on localized steps rather than entire agent trajectories. We discuss system-level methods that address this shortcoming in the next subsection.

3.3 System-Level Transparency

System-level transparency addresses explainability at the level of complete agent architectures and execution histories, rather than individual reasoning steps or isolated tool calls. In contrast to tool-level transparency, which focuses on local decisions within a single interaction, system-level approaches characterize component composition, information flow across tools, and agent behavior across entire trajectories or populations of runs.

Documentation. Structured documentation provides information on model capabilities, components, and intended use. Model Cards (Mitchell et al., 2019) and Foundation Model Transparency Reports (Bommasani et al., 2024) standardize disclosure of datasets, evaluation, risks, and architecture. While originally designed for monolithic models, these frameworks include some indicators relevant to tool-augmented systems–such as disclosure of model components and their integration–but do not support characterization of tools, tool dependencies, or runtime attribution.

Provenance-based approaches capture how outcomes arise from interactions among models, tools, and data. PROV-AGENT (Souza et al., 2025) extends the W3C PROV model[2] and integrates with the Model Context Protocol to record prompts, model responses, tool invocations, and their dependencies within unified provenance graphs. This enables reconstruction of agent workflows and attribution across human and machine-driven activities.

Auditing, verification, and observability. Beyond tracing individual executions, several systems support auditing and evaluation of agent behavior at scale. VeriLA (Sung et al., 2025) verifies multi-step agents by scoring intermediate subtasks and aggregating correctness over structured execution graphs. Observability platforms such as LangSmith Observability[3] and DeepEval[4] provide end-to-end traces and component-level evaluations, while visualization tools like Seaview (Bula et al., 2025), the OpenHands Trajectory Visualizer[5], and Agent Trajectory Explorer (Desmond et al., 2025) render

[2] w3.org/TR/prov-overview/
[3] langchain.com/langsmith/observability
[4] deepeval.com/docs/getting-started
[5] github.com/OpenHands/trajectory-visualizer

long agent trajectories in human-interpretable form for debugging and comparison. These approaches extend explainability from local reasoning steps to global system behavior, enabling analysis of how complex, tool-using LLM agents operate across time and interactions.

Explainability for computer systems and programs. Many challenges posed by tool-using LLM agents—such as attributing outcomes to intermediate steps, tools, and data—have long been studied in the context of program and system explainability. Work in software debugging framed explanations as causal accounts of system behavior, exemplified by the *Whyline* system (Ko and Myers, 2004, 2009), which allowed users to ask "why" and "why not" questions about program outputs, and by *delta debugging* (Zeller and Hildebrandt, 2025; Zeller, 2009), which isolates minimal failure-inducing inputs and state differences. Work in databases and workflow systems formalized explanation through *data provenance*, tracing how data items and transformations contribute to observed results (Buneman et al., 2001; Cheney et al., 2009; Freire et al., 2008). In parallel, research on *observability* in distributed systems developed techniques for reconstructing causal execution paths across components (Fonseca et al., 2007; Sigelman et al., 2010; Kaldor et al., 2017). These approaches established system-level causality, provenance, and observability as foundational principles for explaining complex, multi-component systems, providing important conceptual references for explainability in tool-using LLM agents.

3.4 Domain-Specific and Multimodal Systems

For some LLM agents, explainability is driven by how non-textual modalities and domain signals enter the decision loop. Explanations need to relate actions and tool calls to perceptual inputs (often vision), structured intermediates, or domain concepts. A key distinction from an explainability perspective is whether domain-specificity and modalities are handled by separate tools or integrated into a multimodal backbone:

Vision as a separate module or tool. Many agents route images through a dedicated perception component (e.g., detector, encoder, VQA), yielding intermediate outputs for downstream reasoning. Explanations focus on grounding decisions in visual evidence and exposing the perception–reasoning pipeline, for example, by linking answers to image regions or attributes (Yu and Ananiadou, 2025; Park et al., 2018). Recent work improves transparency with intermediate structures, such as task graphs (Nooralahzadeh et al., 2024), interpretable vision-language representations (Shaham et al., 2024), or flowchart-like attributions for multi-step visual reasoning (Suri et al., 2025).

Integrated multimodal backbones. Other systems utilize joint vision-language models. Here, explainability shifts from tool invocation to internal mechanisms of multimodal integration, such as cross-modal attention, circuits, or concept representations (Palit et al., 2023; Huo et al., 2024; Fang et al., 2024; Parekh et al., 2024; Bhalla et al., 2024; Gandelsman et al., 2025). These works do not address tool use directly, but they serve to demonstrate how explanation targets and faithfulness criteria shift when augmentation relies on internal representations instead of explicit tool calls.

Domain-specific settings. Explainability is also required in specialized domains, where transparency depends on linking behavior to domain knowledge and workflows. For example, in LLM-controlled robotics, interpretability requires grounding in a library of executable robot skills (Ichter et al., 2022). In medicine, clinical agents emphasize interpretable reasoning traces and tool use to enable expert oversight (Shimgekar et al., 2025; Shi et al., 2025).

4 Discussion

4.1 Does tool use improve interpretability?

Tool use makes LLMs more interpretable. Tool output is not the entire story of an LLM's behavior, but it constitutes an important and often informative part of it, as knowing which tool produced an intermediate result makes substantial portions of the reasoning process explicit. Because many tools operate according to comparatively transparent logic–such as arithmetic computation, database lookup, or API execution–they typically require less additional interpretation than purely statistical text generation. Moreover, intermediate tool outputs provide concrete artifacts that users can inspect, verify, and cross-check against their own knowledge or external sources, supporting stepwise validation of system behavior rather than relying solely on final answers. Empirical studies further suggest that mechanisms such as CoT prompting, RAG, and reference attribution can in-

crease user trust and effectiveness when interacting with chatbots (Wei et al., 2022; Menick et al., 2022; Zamfirescu-Pereira et al., 2023), although this trust is not always warranted and depends on the faithfulness of the exposed signals (Jacovi and Goldberg, 2020; Turpin et al., 2023; He et al., 2024).

Providing explanations is harder with tool use. At the same time, tool use introduces additional complexity that must itself be understood and explained. Whereas traditional language-only systems require explanations primarily for text generation, tool-calling agents require explanations not only for generated outputs but also for tool selection, invocation, and integration. Because tool calls are often easier to interpret than free-form generation, there is a risk that their visibility leads to an overemphasis on their role, creating unwarranted increases in user trust even in tasks where tools contribute little or no value. Given that explaining text generation remains an open and challenging problem, we caution against claims that tool use inherently makes LLMs more interpretable or trustworthy. Similar concerns have been raised in the context of RAG systems, where improved interpretability or groundedness is frequently cited as a motivation (e.g., He et al., 2024; Wang et al., 2024, 2025b), yet such claims are rarely validated outside dedicated and costly evaluation studies.

4.2 Open Challenges

Granularity. A central open challenge for explainability in tool-calling LLMs concerns the appropriate granularity and locality of explanations. Classical XAI methods were developed for relatively simple input-output mappings, where local explanations–such as attributing a sentiment classification to specific words–provides meaningful insight. In tool-calling systems, however, individual input tokens often carry limited explanatory value compared to factors external to the input, including retrieved context, tool outputs, execution results and orchestration choices. As a result, explanations that remain localized to surface inputs risk being misleading rather than merely incomplete, faithfully describing token-level correlations while omitting the dominant causal influences on the system's behavior. At the same time, shifting explanations toward system-level factors raises new challenges: how far back in an agent's execution history an explanation should extend, how to aggregate explanations across multiple steps and components, and how to balance fine-grained faithfulness against usability and cognitive load. Granularity is further intertwined with controllability, as overly coarse explanations do not support meaningful intervention, while overly fine-grained traces may expose levers that users cannot safely or effectively manipulate. Defining what constitutes a "local" explanation for an agentic system and how explanation granularity should adapt to task, user role, and risk remains an open problem.

System-level characterizations. A central open challenge for explainability in tool-calling, agent-based systems concerns system-level characterizations, corresponding to global explanations of how a system works overall. Research on user mental models and misconceptions indicates substantial uncertainty about the high-level structure of complex AI systems, including which components they contain (e.g., session memory, internet access, retrieval modules, calculators, symbolic computation, background reasoning, or filtering mechanisms) and how these components interact. At the same time, system providers are often weakly incentivized to disclose such details, and deployed systems are fluid: components, capabilities, and orchestration strategies change frequently, rendering static descriptions quickly outdated.

Given the impact that architectural choices can have on system behavior and performance, accurately characterizing a system at a high level may contribute more to user understanding than increasingly fine-grained attribution methods applied to individual inputs. Local explanations–such as input attributions, calculation traces, or data attribution–can be difficult to interpret in isolation and often presuppose a correct understanding of the system's overall structure. Without this "big picture," such explanations risk being misunderstood or overgeneralized, whereas an accurate system-level characterization can already address many observed behaviors, limitations, and failure modes.

A further challenge lies in determining how such global characterizations should be constructed. Prior work on stakeholder needs for AI explanations provides a useful starting point, identifying recurring global questions such as what logic a system follows, how reliable its outputs are, and what data or resources it depends on (Liao et al., 2020). For tool-calling LLMs, this suggests shifting emphasis from exhaustive technical detail toward clarifying those aspects that users most fre-

quently misunderstand. One promising direction is the development of structured, standardized templates—analogous to model cards or foundation model transparency reports—that foreground tool use, external dependencies, and orchestration logic as first-class elements of system documentation. More broadly, these considerations point to the need for explanations at multiple, clearly distinguished levels: general knowledge about how AI systems work, high-level characterizations of specific deployed systems, more detailed global descriptions of salient features and behaviors, and local explanations addressing individual outputs.

Evaluation. Evaluation of explainability in tool-augmented systems remains underdeveloped, with existing approaches often narrow in scope, difficult to standardize, and challenging to compare across systems and tasks (Anjomshoae et al., 2019; Sado et al., 2023). A central challenge is that agent behavior is deeply entangled with human interaction: different users, prompts, and dialogue trajectories can steer an agent toward different tool uses, intermediate steps, and self-explanations, complicating controlled comparisons and reproducibility. Additionally, explanations in agentic systems concern long horizons and multiple components, so their quality cannot be assessed solely at the level of final outputs or individual steps. Provenance-oriented approaches suggest that richer and structured logging of actions, tool invocations, and information sources can improve auditability and support more systematic evaluation in such interactive settings (Kale et al., 2022), but how to evaluate the coherence and faithfulness of explanations across entire agent workflows remains an open problem.

Growing capabilities. The capabilities of AI agents are developing at a rapid pace. Recent evidence suggests that the length of software engineering tasks these agents can autonomously complete is doubling roughly every seven months (Kwa et al., 2025)—a trend which, if sustained, would imply that agents will soon be capable of executing workflows that currently require days or weeks of human effort. This growth reflects several interacting developments, including increasing model and data scale, expanding access to external tools and environments, and longer task horizons involving many sequential and interdependent decisions. While each of these trends raises new challenges for explainability, long-horizon task execution presents a particularly acute problem. In such settings, failures often emerge from the accumulation of many small decisions–tool selections, parameter choices, intermediate assumptions–rather than from a single identifiable error. As agents become more autonomous and deliver increasingly complete, ready-to-use outputs, opportunities for human intervention during execution diminish, shifting the burden toward post-hoc verification and auditing. Explainability methods must therefore operate at appropriate levels of abstraction that allow users to reconstruct outcomes, detect compounding errors, and assign responsibility across extended workflows. Whether existing XAI approaches can scale to these demands, or whether understanding of increasingly capable agentic systems will remain largely superficial, remains an open question.

5 Conclusion

In this survey, we argued that the ongoing shift of LLMs from standalone text generators to tool-using agents necessitates a rethinking of explainable AI. In such systems, understanding behavior requires more than explaining why a particular token was generated: it requires faithful accounts of how models delegate subtasks, orchestrate tool use, and integrate execution results over extended interactions. Our review of traditional explainability methods showed that most existing approaches were developed for monolithic models and single-prediction settings, and fall short of capturing multi-step, system-level processes that characterize tool-augmented LLMs. We surveyed work that addresses these gaps, including approaches tailored to specific forms of tool use such as retrieval-augmented generation and multimodal pipelines, methods that expose or explain individual tool calls, and system-level techniques that trace and audit provenance across complete agent trajectories.

Despite the potential interpretability benefits from explicit tool use, significant challenges remain. These include determining appropriate levels of explanatory granularity, developing robust and comparable evaluation practices, and aligning explanations with users' mental models of complex AI systems. Addressing these challenges calls for explainability approaches that operate at the system level, supporting faithful inspection of agent behavior across tools, data, and execution steps. Developing such approaches is a prerequisite for enabling meaningful oversight and appropriate trust in tool-using AI systems.

Limitations

This survey focuses on explainability for tool-calling large language models and agentic AI systems, and does not aim to provide a comprehensive overview of explainability research across all AI domains. As a result, related areas such as explainability for recommender systems, autonomous vehicles, control, or other not language-based systems are largely out of scope, except where they intersect with tool-augmented language model architectures.

In addition, our discussion emphasizes conceptual distinctions, design patterns, and representative approaches rather than exhaustive technical comparisons of individual methods. While we highlight strengths and limitations at a high level, the suitability of specific explainability techniques may depend on application context, system design, and user requirements in ways that are more nuanced than can be fully captured within the scope of this survey.

References

Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems.

Vincent Armant, Amira Mouakher, Felipe Vargas-Rojas, Danai Symeonidou, Joris Guérin, Isabelle Mougenot, and Jean-Christophe Desconnets. 2024. Can Knowledge Graphs and Retrieval-Augmented Generation be combined to Explain Query/Answer Relationships Truthfully? In *DAO-XAI 2024 Data meets Ontologies in Explainable AI co-located with the 27th European Conference on Artificial Intelligence (ECAI 2024)*, volume 3833, Santiago de Compostela, Spain.

Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58:82–115.

Georgios Balanos, Evangelos Chasanis, Konstantinos Skianis, and Evaggelia Pitoura. 2025. KGRAG-Ex: Explainable Retrieval-Augmented Generation with Knowledge Graph-based Perturbations. *arXiv preprint*. ArXiv:2507.08443 [cs].

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, Adel Bibi, Robert Trager, Damiano Fornasiere, John Yan, Yanai Elazar, and Yoshua Bengio. 2025. Chain-of-thought is not explainability. Working paper, Oxford Martin AI Governance Initiative (AIGI), University of Oxford. Accessed 2025-12-16.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flávio P. Calmon, and Himabindu Lakkaraju. 2024. Interpreting CLIP with sparse linear concept embeddings (splice). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Betty Xiong, Sayash Kapoor, Nestor Maslej, Arvind Narayanan, and Percy Liang. 2024. Foundation model transparency reports. In *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES-24) - Full Archival Papers, October 21-23, 2024, San Jose, California, USA - Volume 1*, pages 181–195. AAAI Press.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in ai-assisted decision-making. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW1):188:1–188:21.

Timothy Bula, Saurabh Pujar, Luca Buratti, Mihaela A. Bornea, and Avirup Sil. 2025. Seaview: Software engineering agent visual interface for enhanced workflow. *CoRR*, abs/2504.08696.

Peter Buneman, Sanjeev Khanna, and Wang Chiew Tan. 2001. Why and where: A characterization of data provenance. In *Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings*, volume 1973 of *Lecture Notes in Computer Science*, pages 316–330. Springer.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Trans. Mach. Learn. Res.*, 2023.

James Cheney, Laura Chiticariu, and Wang Chiew Tan. 2009. Provenance in databases: Why, how, and where. *Found. Trends Databases*, 1(4):379–474.

Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya

Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, Jeff Schneider, Eduard Hovy, Roger Grosse, and Eric Xing. 2024. What is Your Data Worth to GPT? LLM-Scale Data Valuation with Influence Functions. *arXiv preprint*. ArXiv:2405.13954 [cs].

Michael Desmond, Ja Young Lee, Ibrahim Ibrahim, James M. Johnson, Avirup Sil, Justin MacNair, and Ruchir Puri. 2025. Agent trajectory explorer: Visualizing and providing feedback on agent trajectories. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 29634–29636. AAAI Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint*. ArXiv:1702.08608 [cs, stat].

Yu Du, Fangyun Wei, and Hongyang Zhang. 2024. Anytool: Self-reflective, hierarchical agents for large-scale API calls. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Junfeng Fang, Zac Bi, Ruipeng Wang, Houcheng Jiang, Yuan Gao, Kun Wang, An Zhang, Jie Shi, Xiang Wang, and Tat-Seng Chua. 2024. Towards neuron attributions in multi-modal large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Fares Fawzi, Vinitra Swamy, Dominik Glandorf, Tanya Nazaretsky, and Tanja Käser. 2025. SCRIBE: Structured chain reasoning for interactive behaviour explanations using tool calling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29273–29298, Suzhou, China. Association for Computational Linguistics.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models. *CoRR*, abs/2405.00208.

Rodrigo Fonseca, George Porter, Randy H. Katz, Scott Shenker, and Ion Stoica. 2007. X-trace: A pervasive network tracing framework. In *4th Symposium on Networked Systems Design and Implementation (NSDI 2007), April 11-13, 2007, Cambridge, Massachusetts, USA, Proceedings*. USENIX.

Juliana Freire, David Koop, Emanuele Santos, and Cláudio T. Silva. 2008. Provenance for computational tasks: A survey. *Comput. Sci. Eng.*, 10(3):11–21.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2025. RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems. *arXiv preprint*. ArXiv:2407.11005 [cs].

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. 2025. Interpreting the second-order effects of neurons in CLIP. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023a. PAL: program-aided language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR.

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023b. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *CoRR*, arXiv:2303.14524.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint*. ArXiv:2312.10997 [cs].

Darío Garigliotti. 2025. Self-explanatory Retrieval-Augmented Generation for SDG Evidence Identification. In *Advances in Conceptual Modeling*, pages 124–132, Cham. Springer Nature Switzerland.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9574–9586.

Iona Gessinger, Katie Seaborn, Madeleine Steeds, and Benjamin R. Cowan. 2025. Chatgpt and me: First-time and experienced users' perceptions of chatgpt's communicative ability as a dialogue partner. *Int. J. Hum. Comput. Stud.*, 194:103400.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing model behavior with path patching. *CoRR*, arXiv:2304.05969.

Roger B. Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger,

Kamile Lukosiute, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. 2023. Studying large language model generalization with influence functions. *CoRR*, arXiv:2308.03296.

Han Guo, Nazneen Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. 2021. FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10333–10350, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zayd Hammoudeh and Daniel Lowd. 2024. Training data influence analysis and estimation: a survey. *Mach. Learn.*, 113(5):2351–2403.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.

Gaole He, Gianluca Demartini, and Ujwal Gadiraju. 2025. Plan-then-execute: An empirical study of user trust and team performance when using LLM agents as A daily assistant. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, YokohamaJapan, 26 April 2025- 1 May 2025*, pages 414:1–414:22. ACM.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering. *Advances in Neural Information Processing Systems*, 37:132876–132907.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 6801–6816. Association for Computational Linguistics.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander Toshev, Vincent Vanhoucke, and 26 others. 2022. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pages 287–318. PMLR.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Kaldor, Jonathan Mace, Michal Bejda, Edison Gao, Wiktor Kuropatwa, Joe O'Neill, Kian Win Ong, Bill Schaller, Pingjia Shan, Brendan Viscomi, Vinod Venkataraman, Kaushik Veeraraghavan, and Yee Jiun Song. 2017. Canopy: An end-to-end performance tracing and analysis system. In *Proceedings of the 26th Symposium on Operating Systems Principles, Shanghai, China, October 28-31, 2017*, pages 34–50. ACM.

Amruta Kale, Tin Nguyen, Frederick Harris, Chenhao Li, Jiyin Zhang, and Xiaogang Ma. 2022. Provenance documentation to enable explainable and trustworthy ai: A literature review. *Data Intelligence*, 5:1–41.

Sehoon Kim, Suhong Moon, Ryan Tabrizi, Nicholas Lee, Michael W. Mahoney, Kurt Keutzer, and Amir Gholami. 2024. An LLM compiler for parallel function calling. In *Forty-first International Conference*

on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. OpenReview.net.

Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, YokohamaJapan, 26 April 2025- 1 May 2025*, pages 420:1–420:19. ACM.

Amy J. Ko and Brad A. Myers. 2004. Designing the whyline: a debugging interface for asking questions about program behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, page 151–158, New York, NY, USA. Association for Computing Machinery.

Amy J. Ko and Brad A. Myers. 2009. Finding causes of program output with the java whyline. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Boston, MA, USA, April 4-9, 2009*, pages 1569–1578. ACM.

Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1885–1894. PMLR. ISSN: 2640-3498.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8460–8478. Association for Computational Linguistics.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, and 6 others. 2025. Measuring AI ability to complete long tasks. *CoRR*, abs/2503.14499.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *CoRR*, abs/1902.00006.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *CoRR*, abs/2307.13702.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*.

Yuhan Li, Xinni Zhang, Linhao Luo, Heng Chang, Yuxiang Ren, Irwin King, and Jia Li. 2025. G-Refer: Graph Retrieval-Augmented Large Language Model for Explainable Recommendation. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, pages 240–251, New York, NY, USA. Association for Computing Machinery.

Q. Vera Liao, Daniel M. Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–15. ACM.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

Haoyan Luo and Lucia Specia. 2024. From understanding to utilization: A survey on explainability for large language models. *CoRR*, abs/2401.12874.

Siwen Luo, Hamish Ivison, Soyeon Caren Han, and Josiah Poon. 2024. Local interpretations for explainable natural language processing: A survey. *ACM Comput. Surv.*, 56(9):232:1–232:36.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, H. Francis Song, Martin J. Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. *CoRR*, abs/2203.11147.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38.

Mirjam Minor and Eduard Kaucher. 2024. Retrieval Augmented Generation with LLMs for Explaining Business Process Models. In *Case-Based Reasoning Research and Development*, pages 175–190, Cham. Springer Nature Switzerland.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 220–229. ACM.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. 2025. The quest for the right mediator: Surveying mechanistic interpretability for nlp through the lens of causal mediation analysis. *Computational Linguistics*, pages 1–48.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Comput. Surv.*, 55(13s):295:1–295:42.

Farhad Nooralahzadeh, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. 2024. Explainable multi-modal data exploration in natural language via LLM agent. *CoRR*, abs/2412.18428.

Gabrielle O'Brien, Antonio Pedro Santos Alves, Sebastian Baltes, Grischa Liebel, Mircea Lungu, and Marcos Kalinowski. 2025. User misconceptions of llm-based conversational programming assistants. *CoRR*, abs/2510.25662.

Lucas Monteiro Paes, Dennis Wei, Hyo Jin Do, Hendrik Strobelt, Ronny Luss, Amit Dhurandhar, Manish Nagireddy, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Werner Geyer, and Soumya Ghosh. 2025. Multi-level explanations for generative language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 32291–32317. Association for Computational Linguistics.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. 2023. Towards vision-language mechanistic interpretability: A causal tracing tool for BLIP. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 2848–2853. IEEE.

Rock Yuren Pang, K. J. Kevin Feng, Shangbin Feng, Chu Li, Weijia Shi, Yulia Tsvetkov, Jeffrey Heer, and Katharina Reinecke. 2025. Interactive reasoning: Visualizing and controlling chain-of-thought reasoning in large language models. *CoRR*, abs/2506.23678.

Jayneel Parekh, Pegah Khayatan, Mustafa Shukor, Alasdair Newson, and Matthieu Cord. 2024. A concept-based explainability framework for large multimodal models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8779–8788. Computer Vision Foundation / IEEE Computer Society.

Eunkyu Park, Wesley Hanwen Deng, Vasudha Varadarajan, Mingxi Yan, Gunhee Kim, Maarten Sap, and Motahhare Eslami. 2025. Critical or compliant? the double-edged sword of reasoning in chain-of-thought explanations. *Preprint*, arXiv:2511.12001.

David Soria Parra and Justin Spahr-Summers. 2024. Model Context Protocol (MCP).

Nilay Patel, Shivashankar Subramanian, Siddhant Garg, Pratyay Banerjee, and Amita Misra. 2024. Towards Improved Multi-Source Attribution for Long-Form Answer Generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3906–3919, Mexico City, Mexico. Association for Computational Linguistics.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2024. Gorilla: Large language model connected with massive apis. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought

reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032, Miami, Florida, USA. Association for Computational Linguistics.

Xiangyu Peng, Prafulla Kumar Choubey, Caiming Xiong, and Chien-Sheng Wu. 2025. Unanswerability Evaluation for Retrieval Augmented Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8452–8472, Vienna, Austria. Association for Computational Linguistics.

Suraj Prasai, Mengnan Du, Ying Zhang, and Fan Yang. 2025. Knowthyself: An agentic assistant for llm interpretability. *Preprint*, arXiv:2511.03878.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. 2024. Model Internals-based Answer Attribution for Trustworthy Retrieval-Augmented Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6037–6053, Miami, Florida, USA. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101. Association for Computational Linguistics.

Joel Rorseth, Parke Godfrey, Lukasz Golab, Divesh Srivastava, and Jaroslaw Szlichta. 2024. RAGE Against the Machine: Retrieval-Augmented LLM Explanations. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 5469–5472. ISSN: 2375-026X.

Fatai Sado, Chu Kiong Loo, Wei Shiung Liew, Matthias Kerzel, and Stefan Wermter. 2023. Explainable goal-driven agents and robots - A comprehensive review.

Hithesh Sankararaman, Mohammed Nasheed Yasin, Tanner Sorensen, Alessandro Di Bari, and Andreas Stolcke. 2024. Provenance: A Light-weight Fact-checker for Retrieval Augmented LLM Generation Output. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1305–1313, Miami, Florida, US. Association for Computational Linguistics.

Soumya Sanyal and Xiang Ren. 2021. Discretized integrated gradients for explaining language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10285–10299, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. 2024. A multimodal automated interpretability agent. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving AI tasks with chatgpt and its friends in hugging face. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Danli Shi, Xiaolan Chen, Bingjie Yan, Weiyi Zhang, Pusheng Xu, Jianchen Yang, Ruoyu Chen, Siyu Huang, Bowen Liu, Xinyuan Wu, Meng Xie, Ziyu Gao, Yue Wu, Senlin Lin, Kai Jin, Xia Gong, Yih-Chung Tham, Xiujuan Zhang, Li Dong, and 8 others. 2025. A multimodal ai agent for clinical decision support in ophthalmology.

Soorya Ram Shimgekar, Shayan Vassef, Abhay Goyal, Navin Kumar, and Koustuv Saha. 2025. Agentic ai framework for end-to-end medical data inference. *Preprint*, arXiv:2507.18115.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Benjamin H Sigelman, Luiz André Barroso, Mike Burrows, Pat Stephenson, Manoj Plakal, Donald Beaver, Saul Jaspan, and Chandan Shanbhag. 2010. Dapper, a large-scale distributed systems tracing infrastructure.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *CoRR*, abs/2402.01761.

Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. 2025. Measuring and enhancing trustworthiness of llms in RAG through grounded attributions and learning to refuse. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Renan Souza, Amal Gueroudji, Stephen DeWitt, Daniel Rosendo, Tirthankar Ghosal, Robert B. Ross, Prasanna Balaprakash, and Rafael Ferreira da Silva. 2025. PROV-AGENT: unified provenance for tracking AI agent interactions in agentic workflows. In *IEEE International Conference on eScience, eScience 2025, Chicago, IL, USA, September 15-18, 2025*, pages 467–473. IEEE.

Viju Sudhi, Sinchana Ramakanth Bhat, Max Rudat, and Roman Teucher. 2024. RAG-Ex: A Generic Framework for Explaining Retrieval Augmented Generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 2776–2780, New York, NY, USA. Association for Computing Machinery.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Yoo Yeon Sung, Hannah Kim, and Dan Zhang. 2025. Verila: A human-centered evaluation framework for interpretable verification of LLM agent failures. *CoRR*, abs/2503.12651.

Manan Suri, Puneet Mathur, Nedim Lipka, Franck Dernoncourt, Ryan A. Rossi, Vivek Gupta, and Dinesh Manocha. 2025. Follow the flow: Fine-grained flowchart attribution with neurosymbolic agents. *CoRR*, abs/2506.01344.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, and 38 others. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Zineddine Tighidet, Jiali Mei, Benjamin Piwowarski, and Patrick Gallinari. 2024. Probing language models on their knowledge source. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 604–614, Miami, Florida, US. Association for Computational Linguistics.

Jan Tolsdorf, Alan F. Luo, Monica Kodwani, Junho Eum, Mahmood Sharif, Michelle L. Mazurek, and Adam J. Aviv. 2025. Safety perceptions of generative AI conversational agents: Uncovering perceptual differences in trust, risk, and fairness. In *Twenty-First Symposium on Usable Privacy and Security, SOUPS 2025, Seattle, WA, USA, August 10-12, 2025*, pages 93–112. USENIX Association.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Sebe Vanbrabant, Gilles Eerlings, Gustavo Alberto Rovelo Ruiz, and Davy Vanacken. 2025. ECHO: Enhancing Conversational Explainable AI through Tool-Augmented Language Models. *Proc. ACM Hum.-Comput. Interact.*, 9(4):EICS014:1–EICS014:33.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401.

Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion*, 76:89–106.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness is not Faithfulness in Retrieval Augmented Generation Attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pages 22–32, New York, NY, USA. Association for Computing Machinery.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2025a. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023a. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023b. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2609–2634. Association for Computational Linguistics.

Nengbo Wang, Xiaotian Han, Jagdip Singh, Jing Ma, and Vipin Chaudhary. 2025b. CausalRAG: Integrating Causal Graphs into Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22680–22693, Vienna, Austria. Association for Computational Linguistics.

Xingyi Wang, Xiaozheng Wang, Sunyup Park, and Yaxing Yao. 2025c. Mental models of generative AI chatbot ecosystems. In *Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI 2025, Cagliari, Italy, March 24-27, 2025*, pages 1016–1031. ACM.

Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Yanbin Lu, Xiaojiang Huang, and Yingzhen Yang. 2024. RecMind: Large Language Model Powered Agent For Recommendation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4351–4364, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2025. Instructrag: Instructing retrieval-augmented generation via self-synthesized rationales. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Sirui Xia, Xintao Wang, Jiaqing Liang, Yifei Zhang, Weikang Zhou, Jiaji Deng, Fei Yu, and Yanghua Xiao. 2025. Ground Every Sentence: Improving Retrieval-Augmented LLMs with Interleaved Reference-Claim Generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 969–988, Albuquerque, New Mexico. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xi Ye, Ruoxi Sun, Sercan Arik, and Tomas Pfister. 2024. Effective Large Language Model Adaptation for Improved Grounding and Citation Generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6237–6251, Mexico City, Mexico. Association for Computational Linguistics.

Zeping Yu and Sophia Ananiadou. 2025. Understanding multimodal llms: the mechanistic interpretability of llava in visual question answering. *Preprint*, arXiv:2411.10950.

J. D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: How non-ai experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 437:1–437:21. ACM.

Andreas Zeller. 2009. *Why Programs Fail - A Guide to Systematic Debugging, 2nd Edition*. Academic Press.

Andreas Zeller and Ralf Hildebrandt. 2025. Simplifying and isolating failure-inducing input: A retrospective on delta debugging. *IEEE Trans. Software Eng.*, 51(3):820–824.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2):20:1–20:38.

Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5).

Junnan Zhu, Min Xiao, Yining Wang, Feifei Zhai, Yu Zhou, and Chengqing Zong. 2025a. TROVE: A Challenge for Fine-Grained Text Provenance via Source Sentence Tracing and Relationship Classification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11755–11771, Vienna, Austria. Association for Computational Linguistics.

Yuqicheng Zhu, Nico Potyka, Daniel Hernández, Yuan He, Zifeng Ding, Bo Xiong, Dongzhuoran Zhou,

Evgeny Kharlamov, and Steffen Staab. 2025b. ArgRAG: Explainable Retrieval Augmented Generation using Quantitative Bipolar Argumentation. In *Proceedings of The 19th International Conference on Neurosymbolic Learning and Reasoning*, pages 697–718. PMLR. ISSN: 2640-3498.