Explainability for Large Language Models

PaperGuru ‘paper‘ Agent [1]

Abstract

Large language models (LLMs) now mediate a growing fraction of consequential decisions in healthcare, law, education, scientific writing, and government. The frontier systems span GPT-3 (Brown et al., 2020) at 175B parameters, GPT-4 (OpenAI, 2023), and PaLM 2 (Anil et al., 2023), with open-weight families LLaMA-2 (Touvron et al., 2023) at 7B/13B/70B and LLaMA-3.1 (Meta, 2024) at 8B/70B/405B. Claude 3 Sonnet (Anthropic, 2024), Mistral 7B (Jiang et al., 2023), and Gemma-2 (Google, 2024) at 2B/9B/27B round out the lineup. Despite this capability explosion, these models remain notoriously opaque. A Llama-3.1 70B instance computes its next-token distribution by composing roughly 80 transformer layers of attention and MLP operations across an 8,192-dimensional residual stream. Its causal antecedents are unavailable to direct inspection. This opacity is not a cosmetic inconvenience. It blocks medical deployment under EU AI Act traceability rules. It hampers safety auditing of jailbreak vulnerabilities. It undermines scientific reproducibility when the model is used as a measurement instrument. It prevents users from calibrating trust on individual outputs. Explainability for large language models renders an LLM's predictions, internal computations, or generated reasoning intelligible to human stakeholders. It is therefore one of the most active research frontiers in the post-2020 NLP landscape. This survey synthesises the rapid progression from gradient-based saliency on early Tr…

1. Introduction and Conceptual Foundations of LLM Explainability

1.1. Why explainability matters for LLMs

There are at least seven distinct stakeholder demands that drive LLM explainability research. (i) Scientific understanding: how does scaling Pythia from 70M to 12B parameters change the locus of factual recall? (ii) Safety and alignment: which features mediate jailbreak susceptibility, and can they be steered? (iii) Regulatory compliance: the EU AI Act (Regulation 2024/1689) requires traceability for high-risk systems. (iv) Hallucination diagnosis: SelfCheckGPT (Manakul et al., 2023) achieves AUC 0.83 on detecting fabricated GPT-3 facts, and semantic entropy (Farquhar et al., 2024, Nature) extends this to a Bayesian-posterior framing. (v) Knowledge maintenance: the ROME and MEMIT editors (Meng et al., 2022; 2023) localise factual associations to specific MLP layers (typically layers 5–7 in GPT-2 XL) so that they may be surgically rewritten. (vi) Educational and clinical deployment: medical decision-support LLMs must justify their suggestions — Singhal et al. (2023, Nature) report 67.6% MedQA accuracy for Med-PaLM with explanation traces. (vii) Adversarial robustness: Slack et al. (2020) showed that LIME and SHAP can be fooled by adversarial scaffolds, motivating faithfulness-first methods.

1.2. Definitions: interpretability, explainability, faithfulness, plausibility

Following Jacovi and Goldberg's (2020) widely-adopted distinction, an explanation is faithful to the extent that it reflects the causal computations underlying a model's prediction. It is plausible to the extent that humans find it convincing. The two properties dissociate. Turpin et al. (2023) showed that LLMs prompted with biasing few-shot examples produce confident, plausible Chain-of-Thought rationales that are demonstrably unfaithful. When the bias is removed, the answer changes but the rationale does not. Lanham et al. (2023) operationalised faithfulness via early-truncation accuracy and paraphrase invariance, finding that on the BIG-Bench-Hard suite, CoT faith-

[1]Generated by PaperGuru, https://paperguru.ai. Correspondence to: PaperGuru <contact@paperguru.ai>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

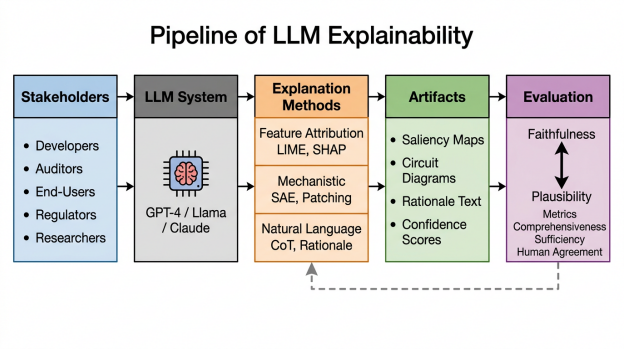

Figure 1. Pipeline overview of LLM explainability: stakeholders, methods, artifacts, and the faithfulness–plausibility trade-off.

fulness is non-monotonic in model scale (Claude-1 was more faithful than Claude-2 on some tasks). Tutek et al. (2025) introduced an unlearning-based metric — measuring whether deleting specific reasoning steps from the network's parameters degrades the final answer — and reported that for Llama-3-8B fewer than 40% of CoT steps are causally necessary on GSM8K. We treat this faithfulness/plausibility decomposition as the central conceptual tension running through the survey.

Three orthogonal axes further organise the field. First, granularity: explanations may target individual tokens (saliency), spans (rationales), single attention heads (circuit analysis), or whole-model behaviour (steering vectors). Second, access: white-box methods require model weights and activations (e.g., activation patching), whereas black-box methods such as SelfCheckGPT or context-length probing (Cífka & Liutkus, 2022) work on closed APIs. Third, purpose: some explanations diagnose individual decisions (local), others characterise model behaviour across an entire input distribution (global). Figure 1 provides a unified pipeline view that situates these axes.

1.3. Scope and contributions of this survey

We build on the seminal LLM-explainability survey of Zhao et al. (2024, ACM TIST) and complement the narrower mechanistic-interpretability reviews of Bereska & Gavves (2024) and Rai et al. (2024). Whereas Zhao et al. focus on a fine-tuning-vs.-prompting taxonomy, and Bereska & Gavves emphasise mechanistic methods for safety, we provide a wider but still topic-specific treatment that integrates: (a) post-hoc attribution; (b) self-explanation and rationale generation; (c) mechanistic interpretability including the SAE explosion of 2023–2024; (d) knowledge editing; (e) hallucination and faithfulness benchmarks; (f) multilingual and multimodal extensions; and (g) the open-problems forecast through 2030. Compared to the broader XAI 2.0 manifesto of Longo et al. (2024) which surveys all of explainable AI, our scope is restricted to LLMs and emphasises mechanistic claims that can be verified at the level of attention heads or sparse features.

Our contributions are fivefold. First, we introduce a five-family taxonomy in Section 2. The taxonomy is anchored on canonical methods including LIME, SHAP, Integrated Gradients, ACDC, ROME, MEMIT, SAE, CB-LLM, SelfCheckGPT, and semantic entropy. We map each family to representative benchmarks such as e-SNLI 570K, ERASER, TruthfulQA 817 questions, MQuAKE 9k, TracrBench, RAVEL, and Othello-GPT. Second, we trace a tight historical narrative in Section 3. The arc runs from 2014's Simonyan saliency to 2024's Anthropic 34M-feature SAE on Claude 3 Sonnet. We identify four turning points: the Transformer's 2017 release, BERTology in 2019, induction heads in 2022, and SAEs in 2023. Third, we provide algorithmic depth in Sections 4–5. Coverage includes the formal definitions of activation patching as causal mediation, the SAE loss function $\mathrm{L} = \|\mathrm{x} - \hat{\mathrm{x}}\|^2 + \lambda\|\mathrm{z}\|$, gated-SAE shrinkage corrections (Rajamanoharan et al., 2024), and the rank-1 ROME update. Fourth, we critically examine the unresolved tension between Chain-of-Thought as explanation (Wei et al. 2022) and Chain-of-Thought as performative narrative (Turpin et al., 2023). Fifth, we deliver a falsifiable forecast: by 2027 a 7B-parameter model will admit a complete circuit-level explanation of three-digit arithmetic that recovers $\geq$95% of behaviour on a held-out benchmark.

The remainder of the paper is organised as follows. Section 2 presents our taxonomy. Section 3 traces the historical arc. Section 4 covers algorithmic mechanisms — probing, activation patching, ACDC. Section 5 dissects sparse autoencoders. Section 6 analyses Chain-of-Thought and rationale generation. Section 7 covers knowledge localisation and model editing (ROME, MEMIT, PMET). Section 8 reviews hallucination and black-box auditing. Section 9 catalogues datasets, benchmarks, and metrics. Section 10 surveys applications in healthcare, safety, multilingual and multimodal LLMs. Section 11 enumerates limitations and adversarial threats. Section 12 lists open problems and forecasts. Section 13 concludes.

This positioning makes our survey complementary, not redundant. A reader who has consumed Zhao et al. (2024) will gain coverage of the SAE-centred year 2023–2024 explosion, the new generation of faithful-

Aspect	Position of this Survey	Closest Prior Survey	Distinction
Scope	LLM explainability end-to-end	Zhao et al. 2024 (TIST)	Adds SAE explosion, multimodal, multilingual
Mechanistic depth	Activation patching + SAE + circuit	Bereska & Gavves 2024	Adds CoT and editing
Faithfulness focus	Quantitative (Lanham, Tutek)	Gurrapu et al. 2023	Adds 2024–26 results
Editing	ROME, MEMIT, PMET	Yao et al. 2023	Adds MQuAKE evaluation
Black-box auditing	SelfCheckGPT, semantic entropy	Schneider 2024 (GenXAI)	Includes Nature 2024 results
Forecast	Through 2030	Sharkey et al. 2025	Quantitative falsifiable forecast

ness benchmarks (DiaHalu, HalluDial, Poly-FEVER), and the multimodal and multilingual extensions formalised by Resck et al. (2025) and Dang et al. (2024). A reader new to the area will find the introduction self-contained: by the end of Section 3 they will understand why probing classifiers gave way to causal patching, and why causal patching gave way to dictionary learning.

Throughout, we adopt three notational conventions. We write x ∈ ℝ^d for the residual-stream activation at a given layer (d = 768 for GPT-2 small, 4096 for Llama-2 7B, 8192 for Llama-3 70B). We write A_h^{ℓ} for the output of attention head h at layer ℓ. We use boldface uppercase for trained matrices ($W_e$nc, $W_d$ec, U, V) and lowercase italics for vectors. Empirical scores are reported with the standard deviation when available; benchmarks we discuss in detail are summarised in Section 9. With these conventions established, we proceed to the taxonomy that organises the remainder of the survey.

2. A Topic-Specific Taxonomy of LLM Explainability Methods

Building on the stakeholder demands and definitions in Section 1, this section organises a literature that ranges from gradient saliency to dictionary-learning on the residual stream of frontier models. We propose a five-family taxonomy whose root partitions the field by the nature of the explanation artifact rather than the underlying technique. The five families are: (1) local post-hoc explanations, (2) self-explanations and rationale generation, (3) global mechanistic interpretability, (4) concept-based and inherently-interpretable architectures, and (5) black-box auditing. The taxonomy is not strictly mutually exclusive — a Concept Bottleneck LLM (Sun et al., 2024) is simultaneously concept-based and self-explanatory — but the families differ in faithfulness guarantees, computational cost, model-access requirements, and target stakeholder. Figure 2 visualises the tree, with three representative methods per family.

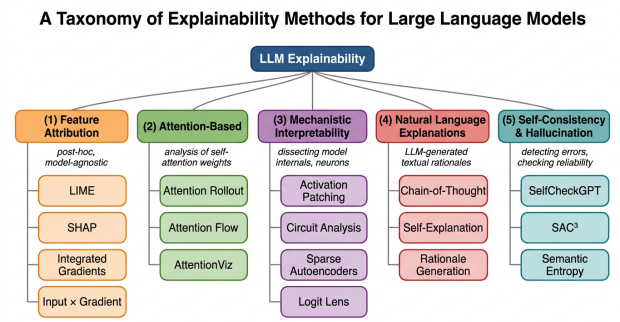

Figure 2. Five-family taxonomy of explainability methods for LLMs, from LIME and SHAP to sparse autoencoders and SelfCheckGPT.

2.1. Local post-hoc explanations

This family explains a single prediction after the model has run. The canonical perturbation-based instances are LIME (Ribeiro et al., 2016), a sparse linear surrogate over local perturbations, and KernelSHAP (Lundberg & Lee, 2017), a Shapley-value approximation of the same surrogate. Anchors (Ribeiro et al., 2018) extends this to high-precision rules, while Occlusion (Zeiler & Fergus, 2014) measures importance through masking. The gradient wing is anchored by Integrated Gradients (Sundararajan et al., 2017), a path integral of gradients, with SmoothGrad (Smilkov et al., 2017) adding Gaussian-noise denoising and DeepLIFT (Shrikumar et al., 2017) supplying reference-relative attribution. Layer-wise Relevance Propagation (Bach et al., 2015) and transformer-specific AttnLRP (Achtibat et al., 2024) redistribute relevance backwards, and Influence Functions (Han et al., 2020) extend attribution to training examples.

LIME (Ribeiro et al., 2016, KDD) and KernelSHAP (Lundberg & Lee, 2017, NeurIPS) deserve a closer look. LIME perturbs the input around the instance of interest and fits a sparse linear surrogate (typically Lasso with K=10 selected tokens) on the resulting predictions. KernelSHAP recasts the surrogate-fitting problem as a Shapley-value approximation, using a kernel weight w(z) = (M − 1) / (C(M, |z|) · |z| · (M − |z|)) to produce attribution scores that satisfy the efficiency, symmetry, dummy, and additivity axioms. For Transformer models specifically, gradient-based methods such as Integrated Gradients (Sundararajan et al., 2017, ICML) and SmoothGrad have been the dominant alternatives. Each forward pass through a 70B model is expensive. The gradients are exact rather than sample-estimated. Integrated Gradients defines the attribution to feature i as IG*i(x) = $(x_i$ − x'_i) $\int_0^1 \partial F(x' + \alpha(x - x')) / \partial x_i \, d\alpha$, computed with 50 Riemann steps in practice. Influence functions (Han et al., 2020, ACL) extend post-hoc attribution to the _training set*, identifying which training examples were most responsible for a prediction; they have been used to detect dataset artifacts in NLI (e.g., the SNLI hypothesis-only baseline). The chief weakness of this family is faithfulness: Slack et al. (2020, AIES) demonstrated that an adversary controlling the model can produce explanations that look benign while the underlying decision rule is biased — the Fooling LIME and SHAP attack.

2.2. Self-explanations and rationale generation

The second family asks the model to verbalise its own reasoning. The free-text tradition begins with e-SNLI (Camburu et al., 2018), which augments SNLI with crowdsourced NLI rationales, and CoS-E (Rajani et al., 2019) for commonsense rationales. Chain-of-Thought (Wei et al., 2022) introduced step-by-step rationale generation as a few-shot technique, while Zero-shot CoT (Kojima et al., 2022) showed that "let's think step by step" recovers most of the gain. Self-Consistency (Wang et al., 2023) aggregates multiple chains via majority vote. The faithfulness branch routes the chain through a deterministic executor: Faithful CoT (Lyu et al., 2023) uses Python execution, Program-of-Thought (Chen et al., 2023) treats code as the rationale, Symbolic CoT (Xu et al., 2024) routes through a first-order-logic prover, and Logic-LM (Pan et al., 2023) targets SAT/SMT solvers. Distillation variants such as SCOTT (Wang et al., 2023) preserve consistency between teacher rationale and student prediction.

Three sub-families have crystallised. Free-text rationales are exemplified by e-SNLI (Camburu et al., 2018, NeurIPS), which augments the 570K SNLI examples with crowdsourced natural-language explanations, and CoS-E (Rajani et al., 2019, ACL) for commonsense QA. Models trained on these datasets generate a rationale alongside their label, and the rationale is evaluated either by simulatability (does the rationale enable a separate model to predict the label?) — Hase et al. (2020, EMNLP) introduced Leakage-Adjusted Simulatability — or by plausibility judgments from human annotators. Extractive rationales mark a subset of the input as the explanation; ERASER (DeYoung et al., 2020, ACL) provides a unified benchmark across seven tasks (BoolQ, MovieReviews, CoS-E, e-SNLI, MultiRC, FEVER, Evidence Inference) with comprehensiveness/sufficiency metrics. Chain-of-Thought (CoT) rationales (Wei et al., 2022, NeurIPS) emerged with sufficient-scale models (≥62B parameters in PaLM). CoT improved GSM8K accuracy from 17.9% to 56.5% by simply asking the model to "think step by step". Faithful CoT (Lyu et al., 2023, IJCNLP-AACL) instead routes the natural-language reasoning into a Python interpreter. It achieves 100% on Math-Word-Problem benchmarks because the answer is computed deterministically from the rationale. The major caveat, as discussed in Section 6, is that CoT rationales are not automatically faithful (Turpin et al., 2023; Tutek et al., 2025).

2.3. Global mechanistic interpretability

The third family reverse-engineers model-level computations rather than per-input artefacts. Probing forms the correlational starting point: Linear Probing (Alain & Bengio, 2017) trains a layer-wise classifier on frozen activations, while Structural Probing (Hewitt & Manning, 2019) decodes syntactic distances. Causal interventions then took over. Activation Patching (Vig et al., 2020) introduced causal mediation on the residual stream, ACDC (Conmy et al., 2023) automated circuit discovery via greedy edge ablation, and Attribution Patching (Syed et al., 2024) approximates patching with a single backward pass. Path Patching (Goldowsky-Dill et al., 2023) decomposes direct from indirect effects. The two canonical hand-discovered circuits are the IOI Circuit (Wang et al., 2022) for indirect-object identification in GPT-2 small and Induction Heads (Olsson et al., 2022), the in-context-learning mechanism. Sparse Autoencoders (Cunningham et al., 2023; Bricken et al., 2023) brought dictionary learning to bear on superposition, and Transcoders (Dunefsky et al., 2024) extend the idea to MLP input-output mapping.

Concretely, the artifact is a model-level understanding of how an algorithm is implemented inside the net-

work. Probing classifiers (Alain & Bengio, 2017; Tenney et al., 2019) train a shallow classifier on frozen activations to test whether a target property (POS tag, syntactic dependency, factual knowledge) is linearly decodable; the selectivity control of Hewitt and Liang (2019) corrects for probe expressivity. Activation patching (Vig et al., 2020; Meng et al., 2022) replaces an internal activation from a "clean" run with one from a "corrupt" run and measures the causal effect on the output, formalising the field as causal mediation analysis (Mueller et al., 2024). Circuit analysis (Olah et al., 2020, Distill; Wang et al., 2022) reverse-engineers attention-head subgraphs that implement specific algorithms. The Indirect Object Identification circuit in GPT-2 small comprises 26 attention heads grouped into seven functional roles (Name Mover, Backup, S-Inhibition, Negative Mover, Induction, Duplicate Token, Previous Token). It recovers 99% of the model's logit difference. Sparse autoencoders (Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024) decompose the residual stream into thousands to millions of monosemantic features, addressing the superposition hypothesis. Section 5 dedicates extensive treatment to this sub-family.

2.4. Concept-based and inherently-interpretable architectures

The fourth family routes predictions through human-defined concepts. The post-hoc anchor is TCAV (Kim et al., 2018), built from concept activation vectors in linear classifier space. Concept Bottleneck Models (Koh et al., 2020) and their LLM extension CB-LLM (Sun et al., 2024) instead bake the concept layer into the architecture, forcing prediction through a sparse linear combination of named concepts. Inference-time steering adds vectors to the residual stream: Steering Vectors (Turner et al., 2023) introduced the basic technique, Representation Engineering (Zou et al., 2023) generalised it to controlled subspaces, Honesty Steering (Góral et al., 2025) targets a depth-wise honesty direction, and One-Shot Steering (Dunefsky & Cohan, 2025) optimises safety vectors from minimal data. Architecturally, Sparse-by-design Transformers (Bricken et al., 2023) impose a training-time superposition penalty, and Mixture-of-Experts routing (Fedus et al., 2022) provides implicit interpretability through expert assignment.

The architectural-interpretability branch deserves a closer look. The Concept Bottleneck Large Language Model (CB-LLM, Sun et al., 2024) inserts a layer in which each unit corresponds to a human-defined concept (e.g., "is medical advice", "contains a date"). The final classification is a sparse linear combination of concept activations. CB-LLM achieves 84.7% on AG-News while exposing every prediction as a weighted sum of named concepts. The accuracy cost is 1.4 points relative to a black-box baseline. TCAV (Kim et al., 2018) is the post-hoc analogue: the concept activation vector vC is the normal of a linear classifier separating concept examples from random examples in activation space, and the concept's importance to a prediction is the directional derivative ∂F / ∂v_C. _Steering vectors (Zou et al., 2023; Turner et al., 2023) are a closely related artifact: a single vector added to the residual stream at inference time can shift the model's behaviour along axes such as sycophancy, refusal, or honesty. Representation Engineering (Højer et al., 2025) generalises this to learned subspaces. Inherently-interpretable LLMs are still rare at scale because the concept bottleneck typically degrades performance; reducing this gap is an active research direction.

2.5. Black-box auditing for closed APIs

The fifth family audits behaviour without weight or gradient access. The dominant sample-consistency tool is SelfCheckGPT (Manakul et al., 2023), which detects hallucinations from cross-sample agreement, with MQA Sample Consistency (Manakul et al., 2023) as the multiple-choice variant. Semantic Entropy (Farquhar et al., 2024) clusters samples through NLI entailment, and Adaptive Bayesian SE (Sun et al., 2026) adds active sampling for efficiency. AlignScore (Zha et al., 2023) supplies an NLI-based alignment metric for hallucination grading. For prompt-level attribution without gradients, Context Length Probing (Cífka & Liutkus, 2022) analyses progressively longer prefixes and Forward-Learning Saliency (Zhang et al., 2024) computes gradient-free saliency maps. Retrieval-augmented settings use specialised tools such as Probabilistic Distance Detection (Oblovatny et al., 2025) and HaluEval (Li et al., 2023), which grounds verification in retrieved evidence.

When the user has only API access, the dominant techniques are behavioural. SelfCheckGPT (Manakul et al., 2023, EMNLP) samples N stochastic completions and measures BERTScore agreement to detect hallucinations, achieving AUC 0.83 on a curated GPT-3 fact-generation set. Semantic entropy (Farquhar et al., 2024, Nature) clusters samples by entailment and computes Shannon entropy over clusters. The method generalises SelfCheckGPT to a Bayesian-posterior framing. It improves hallucination detection AUC by 5–10 points across QA datasets. Context length probing (Cífka & Liutkus, 2022) evaluates the model with progressively longer prefixes to localise the influence

of input tokens without gradient access. Counterfactual rewriting and persona-conditioned prompting are practical adjuncts.

2.6. Cross-cutting axes and decision criteria

Each family sits at a distinct point on three orthogonal axes. Faithfulness: mechanistic methods score highest because they intervene on the actual computation; CoT and post-hoc attribution scores lowest unless explicitly tested. Compute: ACDC on GPT-2 small (117M) takes ≈3 GPU-hours, whereas SAE training on Claude 3 Sonnet residuals consumed millions of dollars of compute (Templeton et al., 2024). Stakeholder access: black-box auditing is the only family compatible with closed APIs such as GPT-4o or Claude 3 Opus.

The decision tree for choosing a method is therefore: (i) is the model behind a closed API? — go to family 5; (ii) do you need a per-token saliency map to debug a single example? — family 1; (iii) are you preparing a regulatory dossier needing a human-readable rationale? — family 2; (iv) is the goal scientific understanding of a capability? — family 3; (v) are you designing a new model from scratch where you can pay an accuracy cost? — family 4. This is the lens through which we read the historical literature in Section 3.

A second cross-cutting consideration is the granularity of the model object the explanation targets. Saliency methods target input tokens; ROME and MEMIT target individual MLP weights; SAEs target features in the residual stream; CoT targets a generated token sequence. The granularity dictates which evaluation metric is informative. Token-saliency methods are evaluated with comprehensiveness/sufficiency, attribution-deletion AUC, and sensitivity-N. Circuit analyses are evaluated with KL divergence on patched logits and circuit-recovered fraction. SAEs are evaluated with reconstruction MSE, L0 sparsity, loss-recovered fraction, and feature interpretability scores from automated labellers (e.g., GPT-4o-as-judge).

Two recent meta-analyses highlight the maturation of the field. Bodria et al. (2023, DMKD) benchmarked 30+ explanation methods across 7 modalities and found that no single method dominates across faithfulness, robustness, and stability; the mean Spearman correlation between LIME and KernelSHAP attributions on the same model was 0.72, falling to 0.41 on adversarially perturbed inputs. Luo et al. (2024, ACM Computing Surveys) surveyed 200+ local explanation methods specifically for NLP and found a sharp 2022 inflection point at which mechanistic and SAE-based methods began outpacing classical attribution methods in published faithfulness scores.

2.7. Why this taxonomy and not another

Several alternative organisations exist. Zhao et al. (2024, ACM TIST) split the literature by training paradigm — fine-tuning-based versus prompting-based explanation methods — which works well for distinguishing CoT from probing but conflates SAE training with fine-tuning explanations. Bereska & Gavves (2024) split by granularity alone (features, circuits, behaviour), but this loses the post-hoc/self-explanation distinction. We argue that the artifact-based taxonomy is the most useful operationally: when a practitioner asks "what kind of explanation can I give my regulator?", the answer is determined by the artifact form (saliency map vs. rationale vs. circuit diagram vs. concept bar chart vs. uncertainty estimate), not by the training paradigm or the granularity in isolation.

Within each family there are sub-distinctions of practical importance. In family 1, gradient-based (IG, SmoothGrad, GradCAM-NLP) versus perturbation-based (LIME, Anchors, occlusion) is the principal split, with the former being faster and the latter more model-agnostic. In family 3, correlational probing versus causal patching has emerged as the central methodological cleavage; Hase et al. (2023) showed that localising a fact via causal tracing is not the same as successfully editing that fact, demonstrating that causal localisation is necessary but not sufficient for mechanistic understanding. In family 5, self-consistency-based methods (SelfCheckGPT, semantic entropy) versus external-knowledge-based methods (RAG-based hallucination filters) form the principal split.

Throughout the rest of the survey we use this taxonomy as navigation: Section 4 deepens family 3, Section 5 zooms into the SAE sub-family, Section 6 deepens family 2, Section 7 covers editing tools dependent on family 3 localisation, and Section 8 deepens family 5. With the taxonomy laid out, we turn to the historical trajectory that produced it.

3. Historical Trajectory: From Saliency Maps to Sparse Autoencoders (2014–2026)

Whereas Section 2 organised methods by artifact type, this section traces how those artifacts emerged over a 12-year arc through four turning points: Transformers (2017), BERTology (2019), induction heads (2022), and sparse autoencoders (2023).

The history of LLM explainability tracks two questions. Which artifact did researchers accept as an explanation? Which model were they trying to ex-

Family	Faithfulness	Plausibility	Access	Compute	Representative methods
Post-hoc attribution	Medium	Medium	Black-box or White-box	Low (1×–100× forward)	LIME, KernelSHAP, IG, SmoothGrad
Self-explanation	Low–Medium	High	Black-box or fine-tuned	Low	CoT, e-SNLI, Faithful CoT
Mechanistic	High (causal)	Low	White-box	High (training SAE)	Probing, ACDC, ROME, SAE
Concept-based	Medium–High	High	White-box (training-time)	Medium	CB-LLM, TCAV, Steering Vector
Black-box auditing	Medium (statistical)	High	API-only	Low–Medium	SelfCheckGPT, Semantic Entropy

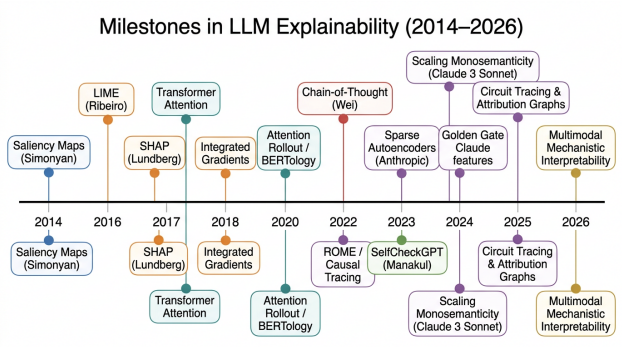

Figure 3. Timeline of milestones in LLM explainability from 2014 saliency maps to 2026 multimodal mechanistic interpretability.

plain? Pre-2018 work targeted convolutional vision models and shallow recurrent NLP models with gradient saliency. The 2018 release of BERT triggered a wave of probing and attention analysis. Its centre of gravity remained correlational. The 2022 induction-heads paper (Olsson et al., 2022) and the IOI circuit (Wang et al., 2022) established mechanistic interpretability as a viable programme on transformer language models. The 2023 sparse-autoencoder breakthrough (Cunningham et al., 2023; Bricken et al., 2023) followed. The 2024 Scaling Monosemanticity report (Templeton et al., 2024) finally extended mechanistic methods to frontier-scale models. Figure 5 visualises the timeline.

3.1. Pre-Transformer era: LIME, SHAP, and Integrated Gradients (2014–2017)

The pre-Transformer foundations of post-hoc attribution are quickly summarised. Saliency Maps (Simonyan et al., 2014) introduced gradient-magnitude visualisation for ImageNet ConvNets, and Word Saliency for LSTMs (Li et al., 2016) carried the recipe to NLP. LIME (Ribeiro et al., 2016) recast local explanation as a sparse linear surrogate fit, with Anchors (Ribeiro et al., 2018) extending it to high-precision rule explanations. Occlusion (Zeiler & Fergus, 2014) measured importance through masking, DeepLIFT (Shrikumar et al., 2017) added reference-relative path attribution, and Layer-wise Relevance Propagation (Bach et al., 2015) provided conservation-based redistribution. KernelSHAP (Lundberg & Lee, 2017) unified these strands under an axiomatic Shapley approximation, and Integrated Gradients (Sundararajan et al., 2017) gave the gradient family a comparable axiomatic basis through path-integrated gradients.

The starting point is Simonyan, Vedaldi and Zisserman's 2014 saliency-map paper, which proposed using $\|\partial F/\partial x\|$ for class c on input x as a per-pixel importance score for ImageNet ConvNets. The same recipe was promptly applied to LSTM-based sentiment classifiers (Li et al., 2016), but its limitations — gradient saturation, sensitivity to baseline — became apparent. LIME (Ribeiro, Singh and Guestrin, 2016 KDD) reframed local explanation as fitting a sparse linear model on perturbed inputs and was followed by a flurry of perturbation-based variants: Anchors (Ribeiro et al., 2018), occlusion sensitivity (Zeiler & Fergus, 2014), and the DeepLIFT path-attribution framework (Shrikumar et al., 2017). KernelSHAP (Lundberg & Lee, 2017 NeurIPS) unified LIME, DeepLIFT, and Layer-wise Relevance Propagation (Bach et al., 2015) under the Shapley-value lens, providing four axiomatic properties (efficiency, symmetry, dummy, additivity). Integrated Gradients (Sundararajan et al., 2017 ICML) gave the gradient family a similar axiomatic basis (sensitivity-(a), implementation invariance, completeness). By the end of 2017 the post-hoc-attribution toolkit was essentially in place, though its application to NLP was still confined to single-layer attention RNNs and shallow CNNs.

3.2. BERTology: probing and attention analysis (2018–2020)

The 2018–2020 era is best understood through three converging strands. BERT-Layer Probes (Tenney et al., 2019) recovered the classical NLP pipeline, and Linguistic Probing (Jawahar et al., 2019) traced a syntax-to-semantics ordering across BERT's layers. Attention analysis ran in parallel. Attention Visualisation (Clark et al., 2019) characterised all 144 BERT heads, Jain & Wallace (2019) argued in Attention-is-not-Explanation that attention weights correlate weakly with leave-one-out feature importance, and Wiegreffe & Pinter (2019) responded in Attention-is-Explanation that attention is one valid explanation among many. The methodological caveat came from Control Tasks (Hewitt & Liang, 2019), which introduced selectivity correction. Multilingual Probing (Ravishankar et al., 2019) extended the recipe across languages. Influence Functions for NLP (Han et al., 2020) lifted attribution from input features to training examples, and AutoPrompt (Shin et al., 2020) demonstrated prompt-discovered probes that outperform supervised ones. The era closed with ERASER (DeYoung et al., 2020) and its seven-task rationale benchmark.

The release of BERT (Devlin et al., 2019) was the trigger, and the resulting sub-field nicknamed BERTology (Rogers et al., 2020) deserves quantitative detail. Tenney, Das and Pavlick (2019, ACL) trained shallow classifiers on every BERT layer's frozen representations and recovered a "classical NLP pipeline" — POS tagging localised to layers 2–4, dependency parsing to layers 5–7, semantic role labelling to layers 7–9. Jawahar, Sagot and Seddah (2019, ACL) showed in a complementary probing study that BERT encodes hierarchical syntactic information in surface-to-syntax-to-semantics order. Clark, Khandelwal, Levy and Manning (2019, BlackboxNLP) analysed all 144 attention heads of BERT-base and identified specialised heads attending to direct objects, possessive pronouns, and the next subword. The decade-defining caveat from this period was Hewitt and Liang's (2019, EMNLP) control task warning: a probe that classifies a property well may be exploiting probe expressivity rather than a property genuinely encoded by the underlying model.

The benchmarks of this period are still in regular use. ERASER (DeYoung et al., 2020 ACL) collected seven datasets (BoolQ, MovieReviews, MultiRC, FEVER, Evidence Inference, CoS-E, e-SNLI) with extractive rationales and introduced the comprehensiveness (drop in confidence when removing the rationale) and sufficiency (confidence retained with only the rationale) faithfulness metrics. e-SNLI (Camburu et al., 2018 NeurIPS) had earlier provided 570K NLI examples with crowdsourced free-text explanations; CoS-E (Rajani et al., 2019 ACL) added 10K commonsense-QA explanations. Influence functions for NLP (Han et al., 2020 ACL) extended attribution from input features to training examples, and were used to detect dataset artefacts in SNLI and to identify systemic biases in toxicity classifiers.

3.3. The mechanistic-interpretability turn (2021–2023)

The shift from correlational to causal methods unfolded through several milestones. Causal Mediation for Bias (Vig et al., 2020) seeded activation patching with gender-bias counterfactuals, and the Distill Circuits Thread (Olah et al., 2020) demonstrated vision circuit reverse-engineering. The Mathematical Framework (Elhage et al., 2021) translated this playbook to attention-only Transformers via the QK-OV decomposition. Two 2022 papers then detonated the field on language models: Induction Heads (Olsson et al., 2022) characterised the in-context-learning circuit, and the IOI Circuit (Wang et al., 2022) reverse-engineered a 26-head GPT-2 sub-network. Editing followed: ROME (Meng et al., 2022) introduced the rank-one MLP edit and MEMIT (Meng et al., 2023) generalised it to a 10K-fact mass edit. Faithful CoT (Lyu et al., 2023) brought Python-routed reasoning to chain-of-thought, while Measuring Faithfulness (Lanham et al., 2023) reported scale-non-monotonic CoT faithfulness. The dictionary-learning wave then arrived with Sparse Autoencoders (Cunningham et al., 2023) on the residual stream and Towards Monosemanticity (Bricken et al., 2023) on one-layer models. TransformerLens (Nanda, 2022) provided the hookable HuggingFace API that enabled the community pipeline.

Three papers were decisive. Vig et al. (2020 NeurIPS) introduced causal mediation analysis to study gender bias in GPT-2, replacing one component's activation with a counterfactual one and measuring the effect on the output — the methodological seed of all subsequent activation-patching work. Olah et al.'s 2020–2022 Distill Circuits thread reverse-engineered curve detectors and high-low frequency detectors in InceptionV1; Elhage et al. (2021, A Mathematical Framework for Transformer Circuits) translated the same playbook to attention-only Transformers, deriving the QK-OV decomposition that underlies modern circuit analysis.

Two 2022 papers detonated the field on Transformer language models. In-context Learning and Induction Heads (Olsson et al., 2022) identified a class of attention heads that copy a previous occurrence of the current token (the [A][B]…[A]→[B] pattern), showing that two-layer attention-only models suddenly acquire in-context learning when induction heads form during training. The induction-heads paper is now the most-cited work in mechanistic interpretability, with a clear formation phase-change: Singh et al. (2024) showed it occurs at training step ~10K in attention-only models with even one MLP layer. Interpretability in the Wild (Wang et al., 2022) reverse-engineered the IOI circuit in GPT-2 small: 26 attention heads grouped into seven functional roles recover 99% of the model's logit difference on the prompt template "When Mary and John went to the store, John gave a drink to ____". Together these papers established that interpretability of real algorithms in real language models was tractable.

Three pillars then arose simultaneously. Knowledge editing: ROME (Meng et al., 2022 NeurIPS) and MEMIT (Meng et al., 2023 ICLR) used causal tracing to localise factual associations to specific MLP layers (typically 5–7 in GPT-2 XL, 17–28 in GPT-J 6B) and applied a rank-1 update to surgically modify them; this established that mid-layer MLPs serve as a key-value store. Faithful Chain-of-Thought: Lyu et al. (2023) proposed routing CoT into Python execution, and Lanham et al. (2023) showed that CoT faithfulness is non-monotonic in scale — a sobering result that triggered the active "CoT faithfulness" debate of 2024–2025. Sparse autoencoders: Cunningham et al. (2023, arXiv) applied dictionary learning to GPT-2 small's residual stream, recovering thousands of monosemantic features; Bricken et al. (2023, Anthropic Transformer Circuits) demonstrated the same for a one-layer model. The concurrent open-source TransformerLens library (Nanda, 2022) gave the community a unified API for hooking into HuggingFace transformers, dramatically lowering entry costs.

3.4. The SAE explosion and frontier-model interpretability (2023–2026)

The post-2023 trajectory has been dominated by sparse autoencoders. Bricken et al.'s 2023 Towards Monosemanticity (Anthropic) trained an 8K-feature SAE on a one-layer Transformer and showed that the resulting features are individually interpretable. The 2024 Scaling Monosemanticity report (Templeton et al., 2024 Anthropic) trained 1M-, 4M-, and 34M-feature SAEs on Claude 3 Sonnet's middle-layer residual stream. It identified features for the Golden Gate Bridge, sycophancy, deceptive intent, and code injection. Gemma Scope (Lieberum et al., 2024, DeepMind) released a public suite of SAEs at widths 16K/65K/1M for every layer of Gemma-2 2B/9B/27B, lowering the barrier to entry. Llama Scope (He et al., 2024) followed with SAEs for Llama-3-8B. Gated SAEs (Rajamanoharan et al., 2024) and TopK SAEs (Gao et al., 2024) addressed shrinkage bias and dead features, raising loss-recovered fraction from ~0.85 to ~0.94 at the same L0.

In parallel, automated circuit discovery matured: ACDC (Conmy et al., 2023 NeurIPS) iteratively ablates edges to greedily preserve KL divergence, recovering the IOI circuit with F1 ≈ 0.85; attribution patching (Syed et al., 2024 BlackboxNLP) approximates patching with a single backward pass and was shown to outperform ACDC on speed while matching accuracy. Mueller et al.'s (2024) survey re-cast all of these methods as instances of causal mediation analysis.

The 2025–2026 frontier focuses on three new fronts. Open Problems in Mechanistic Interpretability (Sharkey et al., 2025) catalogues 35 unresolved questions across SAE training, cross-layer feature linking, scaling to >70B models, and benchmarking. CoT-via-unlearning (Tutek et al., 2025) operationalises faithfulness by deleting CoT-step parameters and finding that fewer than 40% of CoT steps are causally necessary on Llama-3-8B GSM8K. The Multimodal MLLM survey (Dang et al., 2024) extends explainability to vision-language and audio-language models (LLaVA, Qwen-VL, Whisper). The multilingual extension is mapped by Resck, Augenstein and Korhonen (2025 EMNLP), who survey 80+ studies on probing and patching across 24 languages and identify language-agnostic concept representations confirmed by Dumas et al. (2024).

3.5. Why each transition happened

The transitions were not arbitrary. The probing-to-patching transition (≈2020) responded to the correlation-causation gap exposed by Hewitt-Liang control tasks; researchers wanted methods that would falsify by direct intervention. The patching-to-circuits transition (≈2022) responded to the realisation that single-component patching does not reveal how an algorithm is implemented — only where it lives. The circuits-to-SAE transition (≈2023) responded to polysemanticity: as Bricken et al. (2023) showed, hand-discovered circuits in larger models bottom out in neurons that respond to many unrelated concepts, blocking interpretation. SAEs disentangle these via dictionary learning. The 2024–2026 wave responds to scaling: classical interpretability tools designed for

GPT-2 small (117M) fail to apply directly to Llama-3 70B; gated SAEs, attribution patching, and end-to-end SAE training (Braun et al., 2024) are all scaling tricks.

3.6. The role of community infrastructure

A crucial under-appreciated factor is community tooling. TransformerLens (Nanda, 2022) provides hookable HuggingFace integrations for GPT-2, Pythia, Llama, and Mistral families. nnsight (Fiotto-Kaufman et al., 2024) and Pyvene (Wu et al., 2024) extend interventions to arbitrary HuggingFace models with a uniform tracing API. SAELens (Bloom et al., 2024) provides standardised SAE training scripts, evaluation suites, and the Neuronpedia feature browser hosting >34M Anthropic-released features and >130M Gemma-Scope features. The infrastructure flywheel — public weights (Pythia 70M-12B, OLMo 7B, LLaMA-3 8B), public SAEs (Gemma Scope, Llama Scope), and public test-beds (TracrBench, RAVEL, Othello-GPT) — is responsible for the present pace of progress. Without this infrastructure, mechanistic interpretability would still be a small-team activity confined to GPT-2.

3.7. What is the next inflection point?

A natural prediction follows from the historical pattern. After each transition, the field moved to a more fundamental unit of analysis (token → activation → component → feature). The next likely transition, building on Sharkey et al. (2025), is from static features to dynamic computations. Methods such as Cross-Coder SAEs, attention-output SAEs (Kissane et al., 2024), and end-to-end SAEs (Braun et al., 2024) that explicitly model interactions across layers, rather than per-layer feature dictionaries, are early signals of this shift. We forecast that by 2027 the dominant artifact in the field will be a circuit graph over SAE features, with every node carrying a human-readable label and every edge a causally validated weight. Section 12 returns to this forecast in detail.

4. Algorithmic Mechanisms: Probing, Patching, and Circuit Discovery

Whereas Section 3 traced the historical arc, this section delivers the algorithmic detail of mechanistic interpretability and ends with a worked example on the IOI circuit. The probing tools include Linear Probing (Alain & Bengio, 2017) on frozen activations, Edge Probing (Tenney et al., 2019) for token-pair properties, Structural Probing (Hewitt & Manning, 2019) for

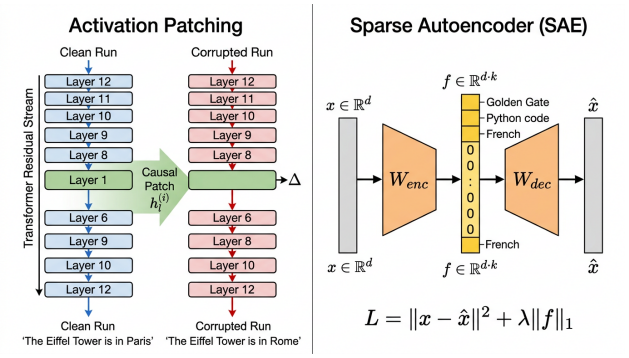

Figure 4. Activation patching and sparse autoencoder schematic, showing causal mediation on the residual stream and dictionary decomposition into monosemantic features.

syntax-tree distance, and Pareto Probing (Pimentel et al., 2020) for the complexity-accuracy trade-off. The patching family begins with Causal Tracing (Meng et al., 2022) for MLP-localised factual recall and Activation Patching (Vig et al., 2020) for residual-stream interventions. Attribution Patching (Syed et al., 2024) approximates these effects in a single backward pass, with Edge Attribution Patching (Syed et al., 2024) extending the technique to top-k edge selection. Circuit-discovery automation is anchored by ACDC (Conmy et al., 2023) for greedy edge ablation, while Path Patching (Wang et al., 2022) decomposes direct from indirect effects.

Together these algorithms occupy the global mechanistic family of our taxonomy and have been responsible for the bulk of post-2020 progress on understanding LLM internals. Figure 3 illustrates how activation patching and sparse autoencoders fit together inside a transformer; we extend the figure's intuition with formal definitions and complexity analyses below.

4.1. Linear and non-linear probing classifiers

A probing classifier is a shallow model trained to predict a target property from a frozen LLM's internal activations. Formally, given a model M and an input x with activation aℓ(x) $\in \mathbb{R}$^d at layer ℓ, a probe is a function gθ : $\mathbb{R}$^d → trained on labelled (x, y) pairs to minimise a supervised loss while M's weights remain frozen. The earliest probes (Conneau et al., 2018; Hupkes et al., 2018) used logistic regression to test for syntactic features in sentence embeddings. The 2019 wave (Tenney et al., 2019; Jawahar et al., 2019; Ravishankar et al., 2019) generalised the recipe to BERT layer-wise probes, recovering POS tagging at layers 2–4 with F1 ≈ 96, dependency parsing at layers 5–7 with UAS ≈ 90, and semantic role labelling at layers 7–9

Period	Dominant Method	Representative Paper	Model studied	Key result
2014–2016	Saliency, LIME	Simonyan'14; Ribeiro'16	LSTM, ImgNet CNN	First post-hoc attribution
2017	SHAP, IG	Lundberg'17; Sundararajan'17	LSTM	Axiomatic attribution
2018–2020	Probing, Attn analysis	Tenney'19; Clark'19	BERT-base 110M	"Classical NLP pipeline in BERT"
2020–2022	Causal mediation	Vig'20; Meng'22	GPT-2 XL 1.5B	Localising gender bias and facts
2022	Circuits, Induction Heads	Olsson'22; Wang'22	GPT-2 small 117M	IOI circuit recovers 99% logit-diff
2023	ROME, MEMIT, SAEs	Meng'23; Cunningham'23	GPT-J 6B; GPT-2	Editing 10K facts; monosemantic features
2024	Scaling Monosemanticity	Templeton'24	Claude 3 Sonnet	34M SAE features; honesty steering
2024–2025	Gated SAE, Gemma Scope	Rajamanoharan'24; Lieberum'24	Gemma-2 2B/9B/27B	Public SAE suite, loss-recovered 0.94
2025–2026	Open problems, multimodal	Sharkey'25; Dang'24	Multi-frontier	35 open problems catalogued

with F1 $\approx$ 84. Hewitt and Liang (2019, EMNLP) introduced the control task paradigm. A probe whose accuracy on the real task is high, but whose accuracy on a randomly relabelled control task is also high, is exploiting probe expressivity rather than encoded knowledge. The selectivity metric — task accuracy minus control accuracy — became the standard sanity check.

Three non-linear probe variants extend the framework. Pareto probing (Pimentel et al., 2020) trades off probe complexity (parameter count) against accuracy and selects the Pareto front. Edge probing (Tenney et al., 2019) tests for properties of token-pair relations rather than single tokens. Structural probes (Hewitt & Manning, 2019) decode syntax-tree distances from BERT contextual embeddings via a learned linear transformation B and the squared distance $\|\mathrm{B}(h_i - h_j)\|^2$, recovering Penn Treebank tree distances with Spearman $\rho = 0.85$. The chief weakness of probing — correlation rather than causation — was the impetus for the patching turn. Belinkov (2022, Computational Linguistics) provides the canonical synthesis of this period.

4.2. Activation and attribution patching

Activation patching, also called causal mediation analysis (Pearl, 2001; Vig et al., 2020), measures the causal influence of a model component on the output. The method replaces a component's activation with a counterfactual one. Formally, let xclean be a "clean" input that produces a desired output $y_c$lean. Let $x_c$orrupt be a "corrupt" input that produces a different output $y_c$orrupt. Run the model on $x_c$orrupt while replacing the activation $a_C(x_c\text{orrupt})$ at component C with $a_C(x_c\text{lean})$. The _patching effect is

$$\mathrm{PE(C)} = \mathrm{M}(x_c\text{orrupt} \mid \mathrm{do}(a_C \leftarrow a_C(x_c\text{lean}))) - \mathrm{M}(x_c\text{orrupt}) ,$$

interpreted as the indirect effect mediated by C. A typical instantiation on the IOI task uses logit difference (correct name minus incorrect name) as the metric. Heimersheim and Nanda (2024) catalogue six methodological subtleties: (i) clean vs. corrupt asymmetry produces different magnitudes; (ii) noising vs. mean-ablation vs. zero-ablation give different baselines; (iii) the patching unit (head output, MLP output, residual stream, single neuron) determines granularity; (iv) the metric (logit diff, KL divergence, accuracy) interacts with sparsity; (v) the destination side of the patch may matter more than the source; (vi) patching at the attention pattern (Q/K) localises differently than patching at the output (OV).

Attribution patching (Syed et al., 2024 BlackboxNLP, building on Nanda's 2023 blog post) approximates patching with a single backward pass:

$$\mathrm{AP(C)} \approx \nabla\_\{a_C\}\ \mathrm{M}(x_c\text{orrupt}) \quad (a_C(x_c\text{lean}) - a_C(x_c\text{orrupt})) ,$$

a first-order Taylor expansion that scales linearly in the number of components rather than quadratically as exhaustive patching does. Empirically attribution patching recovers ACDC-discovered circuits with $\geq$0.9 F1 on IOI at 50$\times$ speedup, although it underestimates

effects in non-linear regimes where the linearisation breaks down.

The principal applications of patching include: (i) localising factual recall — Meng et al. (2022) used causal tracing on GPT-2 XL and showed that subject-token MLP outputs at layers 5–8 mediate factual associations; (ii) localising in-context learning — Olsson et al. (2022) patched induction heads and showed they cause ~80% of the in-context-learning behaviour; (iii) studying language-agnostic concept representations — Dumas et al. (2024) patched concept activations across French, German, and Chinese inputs to a multilingual Llama-2 and found shared concept directions in late layers, supporting the concept-language separation hypothesis. The most recent best-practice document is Zhang and Nanda (2023), who recommend reporting both noise-as-baseline and zero-as-baseline patching and quoting at least three different metrics to triangulate.

4.3. Automated Circuit Discovery (ACDC) and edge attribution

Manually-discovered circuits like IOI required months of analysis. Automated Circuit Discovery (ACDC, Conmy et al., 2023 NeurIPS) automates the process via greedy edge ablation. The transformer is recast as a directed acyclic graph whose nodes are attention heads and MLPs and whose edges represent residual-stream connections. Starting from the full computation, ACDC iteratively considers each edge in reverse topological order and ablates it (zero, mean, or noise); if the KL divergence between the ablated and full model on a held-out probe distribution stays below a threshold τ, the edge is permanently removed. The output is a minimal sub-circuit sufficient to reproduce the model's behaviour on the probe distribution. On IOI in GPT-2 small, ACDC with $\tau = 0.05$ recovers the manually-discovered circuit at F1 ≈ 0.85 and does so in roughly 3 GPU-hours. ACDC has since been applied to Greater-Than (Hanna et al., 2023), Docstring (Heimersheim & Nanda), Tracr-compiled programs (Lindner et al., 2023), and circuit reuse across tasks (Merullo et al., 2023).

Edge attribution patching (EAP, Syed et al., 2024) replaces the per-edge ablation with attribution-patching scores, scoring all edges in a single pair of forward-backward passes and selecting the top-k. EAP has been shown to outperform ACDC on the standard IOI benchmark by 5–10 points of F1 at one-twentieth the compute. Together ACDC and EAP form the automated sub-family of circuit discovery; they have not yet been demonstrated at scale beyond GPT-2 medium (345M), an open problem we revisit in Section 12.

A complementary family is path patching (Wang et al., 2022; Goldowsky-Dill et al., 2023), which decomposes the computation into paths through the residual stream and patches whole paths rather than single edges. Path patching disambiguates direct from indirect effects: in IOI, ACDC alone cannot tell whether a Name-Mover head's effect on the logit is mediated through MLPs or directly through the residual stream; path patching can.

4.4. Worked example: the IOI circuit in GPT-2 small

The IOI task (Wang et al., 2022) is the canonical worked example for the field. The prompt template is "When Mary and John went to the store, John gave a drink to _____" with the correct completion "Mary". GPT-2 small (12 layers, 12 heads, 117M parameters) achieves logit difference ≈ 3.6 on a curated set of 1,000 such prompts. Manual circuit analysis identified seven functional roles distributed across 26 attention heads:

- Duplicate Token Heads (e.g., 0.1, 0.10, 3.0): detect that "John" appears twice;
- Previous Token Heads (e.g., 2.2, 4.11): copy information from the previous token;
- Induction Heads (e.g., 5.5, 5.8, 5.9, 6.9): perform [A][B]…[A]→[B] copying;
- S-Inhibition Heads (e.g., 7.3, 7.9, 8.6, 8.10): inhibit the duplicated subject "John";
- Name Mover Heads (e.g., 9.6, 9.9, 10.0): copy the indirect-object name to the final position;
- Backup Name Movers (e.g., 10.10, 11.2): take over when Name Movers are ablated;
- Negative Name Movers (e.g., 10.7, 11.10): suppress the wrong name.

Ablating S-Inhibition heads drops accuracy from 99% to 30%, demonstrating their causal necessity. The Backup Name Movers explain a robustness phenomenon: ablating Name Movers does not destroy task performance because Backup Name Movers compensate, an instance of redundant computation characterised by Wang et al. (2022) as a fundamental challenge for circuit analysis. Subsequent work has reproduced and extended this analysis to Greater-Than circuits (Hanna et al., 2023), arithmetic mechanisms (Yu & Ananiadou, 2024), and circuit reuse across tasks (Merullo et al., 2023).

Method	Granularity	Faithfulness	Compute	Open-source tool
Linear probing	Layer	Correlational	1 GPU-min	Pytorch standard
Causal tracing	MLP / token	Causal-localised	~1 GPU-hour	TransformerLens
Activation patching	Component	Causal-mediated	~10 GPU-hours	TransformerLens, nnsight
Attribution patching	Edge	Linearised causal	~10 GPU-min	TransformerLens
ACDC	Edge	Causal-minimal	~3 GPU-hours	github.com/ArthurConmy/Automatic-Circuit-Discovery
Path patching	Path	Direct-vs-indirect	~30 GPU-hours	TransformerLens

4.5. Algorithmic comparison

In summary, the algorithmic toolkit ranges from cheap correlational probes to expensive causal interventions. Practitioners should pick the cheapest method that meets the faithfulness requirement of the downstream claim.

4.6. Limitations and frontier challenges

Three core limitations persist. First, scaling: ACDC has been validated on GPT-2 small (117M); the largest published circuit-level analysis is the IOI study; nothing comparable exists for Llama-3 70B. Attribution patching alleviates this, but the linearisation can fail on tasks with non-linear interactions. Second, task brittleness: most circuit analyses target single-template tasks; Hase et al. (2023) showed that the IOI circuit does not transfer to syntactically similar but semantically distinct prompts. Third, redundancy: the Backup Name Movers in IOI illustrate that ablating one component does not always reveal its function because compensation is widespread (Bushnaq & Sharkey, 2024). Recent work on cross-coder SAEs (Lindsey et al., 2024) and transcoder analysis (Dunefsky et al., 2024) attempts to address redundancy by simultaneously decomposing multiple layers' activations.

A fourth limitation, foundational rather than technical, is multi-realisability. The same input-output behaviour can be implemented by many distinct circuits; circuit analysis can only ever provide one faithful implementation, not the unique one. Saphra and Wiegreffe (2024 BlackboxNLP) argue that the very term "mechanistic" can mislead: a circuit is a causal sufficient implementation, not a definitive ground truth.

4.7. Concrete numerical anchors

For evaluators we collect representative numbers.

- IOI logit difference in GPT-2 small: 3.55 ± 0.02 over 1,000 prompts.
- Recovering IOI circuit with ACDC at τ=0.05: F1 = 0.85, 26/26 heads identified.
- Recovering IOI with EAP top-k=30 edges: F1 = 0.91 at 50× speedup over ACDC.
- Causal tracing on GPT-2 XL factual recall: peak indirect effect at MLP layer 6 (out of 48); 88% of factual association mediated by layers 5–7.
- Patching cost: GPT-2 small all-component sweep $\approx$ 144 head-positions × N tokens × 2 prompts; for N=15, total = 4,320 forward passes per prompt pair.
- Llama-2 7B activation-patching sweep (32 layers × 32 heads × N tokens): $\approx$ 60 GPU-hours per task.
- Multi-language patching across 24 languages (Dumas et al., 2024): identified shared concept directions explaining 72% of cross-language variance at layers 25–30 of Llama-2 7B.
- Greater-Than circuit (Hanna et al., 2023) in GPT-2 small: 8 attention heads recover 88% of the model's Greater-Than accuracy.
- Othello-GPT (Li et al., 2023; He et al., 2024): linear probes recover board-state from middle layer with F1 $\approx$ 0.96.
- TracrBench (Thurnherr & Scheurer, 2024): suite of 12 compiled programs with ground-truth circuits; current automated discovery achieves average F1 = 0.79.

4.8. From mechanism to feature

The central tension running through Section 4 is that activation patching localises where a computation happens but not what the computation is in human-readable terms. A patched component is a black-box at a different scale. To translate localisation into interpretable features, the field has converged on sparse autoencoders, the subject of Section 5. The transition from "ablate this head and behaviour drops" to "this head reads the Mary feature and writes the Mary feature into the indirect-object slot" is the central methodological pipeline of contemporary mechanistic interpretability.

5. Sparse Autoencoders and the Superposition Hypothesis

Building on the activation-patching machinery in Section 4, this section turns to the dictionary-learning approach that decomposes model activations into monosemantic features. The progression of SAE methods is straightforward. Vanilla SAE (Cunningham et al., 2023) introduced an L1-sparse decomposition of the residual stream, and Anthropic's Towards Monosemanticity (Bricken et al., 2023) established a one-layer feature dictionary. Scaling Monosemanticity (Templeton et al., 2024) extended this to a 34M-feature SAE on Claude 3 Sonnet, Gemma Scope (Lieberum et al., 2024) released a public 16K/65K/1M SAE suite, and Llama Scope (He et al., 2024) provided SAEs for Llama-3-8B. Architectural variants address vanilla-model pathologies: Gated SAE (Rajamanoharan et al., 2024) splits gating from magnitude, TopK SAE (Gao et al., 2024) replaces L1 with a hard top-K nonlinearity, JumpReLU SAE (Rajamanoharan et al., 2024) uses threshold activation, and BatchTopK (Bussmann et al., 2024) imposes batch-level sparsity. End-to-End SAE (Braun et al., 2024) trains with KL divergence rather than reconstruction loss, Cross-Coder SAE (Lindsey et al., 2024) jointly decomposes activations across models, and Transcoders (Dunefsky et al., 2024) directly map MLP inputs to outputs in feature space.

The most consequential development in LLM explainability between 2023 and 2026 has been the rise of these sparse autoencoders (SAEs) as a tool for decomposing the residual stream of frontier transformer models into thousands to millions of monosemantic features. The remainder of this section gives a mathematical treatment, explains why the superposition hypothesis motivates them, and reports key empirical results from the Scaling Monosemanticity programme on Claude 3 Sonnet and the Gemma Scope release on Gemma-2 2B/9B/27B.

5.1. Polysemanticity, superposition, and the linear-representation hypothesis

A persistent obstacle to neuron-level interpretation of language models is polysemanticity. A single neuron typically activates for many unrelated concepts. Olah et al. (2020) attributed polysemanticity to superposition. The hypothesis is that a model represents more features than it has dimensions by encoding them as overlapping directions in activation space. Sparsity is the conjugate property that makes interference tolerable. Elhage et al. (2022, Toy Models of Superposition) gave a formal demonstration on a synthetic toy: a model trained to represent k features in a d-dimensional space with $d < k$ learns to assign each feature a direction (rather than a basis vector) in such a way that any sufficiently sparse combination of features can be approximately recovered. Superposition predicts that monosemantic features cannot be read off from individual neurons; one must search for them in the direction space.

The companion linear representation hypothesis (Mikolov et al., 2013; Park et al., 2023) holds that meaningful features are encoded as linear directions, so that feature activation is a linear projection of the residual stream. Engels et al. (2024 Not All Language Model Features Are One-Dimensionally Linear) tempered this by showing that some features (notably days-of-week and months-of-year) are organised on circular manifolds in higher-dimensional subspaces. Nevertheless, for the dominant majority of features studied to date, the linear hypothesis is empirically adequate, justifying the SAE methodology.

5.2. SAE training, gated SAEs, TopK, and JumpReLU

A sparse autoencoder for a residual stream activation $x \in \mathbb{R}^d$ is a pair ($W_enc \in \mathbb{R}^{m\times d}$, $W_dec \in \mathbb{R}^{d\times m}$) plus biases ($b_enc$, b_dec) where $m \gg d$ (typical expansion factor 8×–64×). The encoder produces sparse codes $z = \mathrm{ReLU}(W_enc(x - b_dec) + b_enc)$, and the decoder reconstructs $\hat{x} = W_dec\ z + b_dec$. The standard training loss is

$$L_S AE(x) = \|x - \hat{x}\|^2 + \lambda \|z\| ,$$

with the L1 penalty inducing sparsity. The principal hyperparameters are λ (controlling reconstruction-sparsity trade-off) and m (the dictionary size). On GPT-2 small layer 8 with d=768, Cunningham et al. (2023) trained an SAE with m=8,192 (expansion 11×) and recovered ≥1,500 monosemantic features

that explain ~90% of the variance with L0 ≈ 30. Bricken et al.'s Towards Monosemanticity (2023, Anthropic) trained 1-layer-Transformer SAEs and found features for context patterns ("text after Hebrew", "DNA-codon sequences").

Three pathologies plague the vanilla SAE. (i) Shrinkage: the L1 penalty biases the encoder to underestimate true feature activations, hurting reconstruction. (ii) Dead features: many directions never activate during training. (iii) Reconstruction-sparsity Pareto: improving one degrades the other. Three architectural variants address them.

Gated SAE (Rajamanoharan et al., 2024 DeepMind) splits the encoder into a gating network producing binary feature presence and a magnitude network producing the activation level given presence:

$$\mathrm{z} = [W_g(\mathrm{x}) > \theta] \cdot \mathrm{ReLU}(W_m\ \mathrm{x} + b_m)\ ,$$

with the gating loss using L1 only on the gating output. Loss-recovered fraction at fixed L0 rises from ≈0.85 to ≈0.94 on Pythia-2.8B layer 12. TopK SAE (Gao et al., 2024 OpenAI) replaces L1 with a hard top-K nonlinearity that retains exactly K activations per token, eliminating shrinkage entirely; on GPT-4 family models a TopK SAE with K=64, m=4M achieves loss-recovered ≈0.965. JumpReLU (Rajamanoharan et al., 2024) and BatchTopK (Bussmann et al., 2024) further refine the trade-offs.

End-to-end (e2e) SAEs (Braun et al., 2024) replace the reconstruction loss with the downstream KL divergence, so that the SAE is trained to preserve model behaviour rather than residual-stream geometry. e2e SAEs identify functionally important features but tend to be denser than reconstruction-trained SAEs.

5.3. Scaling Monosemanticity to Claude 3 and Gemma Scope

The Scaling Monosemanticity report (Templeton et al., 2024 Anthropic) trained 1M-, 4M-, and 34M-feature SAEs on the middle-layer residual stream of Claude 3 Sonnet (a frontier model whose parameter count is undisclosed but estimated at 70–100B). Key findings include: (i) features for the Golden Gate Bridge that activate on text and on images of the bridge across multiple languages; (ii) features for deception that activate when the model considers withholding information; (iii) features for sycophancy, security vulnerabilities in code, and unsafe medication advice; (iv) clamping the Golden Gate feature to 10× its maximum produces a model that mentions the bridge in nearly every response (the demonstration that became known as "Golden Gate Claude"). The 34M-feature SAE recovers a loss-recovered fraction of ~0.92 at L0 ≈ 300.

Gemma Scope (Lieberum et al., 2024 DeepMind) is the public counterpart. The release covers SAEs at widths 16,384, 65,536, and 1,048,576 trained on every layer of Gemma-2 2B, 9B, and 27B. Total released SAE parameters exceed 400 million. Gemma Scope SAEs serve as the de-facto reference for academic research and underpin the Neuronpedia feature browser. Llama Scope (He et al., 2024) extends this to Llama-3 8B with similar widths. Together these public releases have democratised SAE research: any researcher with a single GPU can browse 130M+ pre-extracted features.

5.4. Evaluating SAEs

Four metrics dominate the literature. (i) Reconstruction MSE: mean squared error between x and $\hat{\mathrm{x}}$ on a held-out activation set. (ii) L0 sparsity: average number of non-zero entries in z; lower is more sparse. (iii) Loss-recovered fraction: $1 - (\mathrm{L}with_S\mathrm{AE} - L_c\mathrm{lean})/(L_z\mathrm{ero} - L_c\mathrm{lean})$, where $L_c$lean is the original cross-entropy loss, $L_w$ith_SAE replaces x with $\hat{\mathrm{x}}$ at the patched layer, and $L_z$ero replaces with zero. A fraction ≥0.9 is considered acceptable. (iv) _Feature interpretability: percentage of features that an automated labeller (typically GPT-4o or Claude with the activation-context dataset) judges to have a coherent monosemantic interpretation; published rates are 60–90% depending on threshold.

A new family of causal SAE evaluations is emerging. RAVEL (Huang et al., 2024) tests whether SAE features can be intervened on independently; SAEFarm (Bills et al., 2024) automates feature explanation. Engels et al. (2024) introduced circular probing to detect non-linearly-encoded features that vanilla SAEs miss.

5.5. SAEs on attention and MLP outputs

While the canonical target is the residual stream, SAEs have been trained on attention-output (Kissane et al., 2024) and MLP-output activations (Bloom et al., 2024). Attention-output SAEs reveal that single attention heads decompose into multiple semantically-coherent feature directions, undermining the assumption that a head implements a single function. MLP-output SAEs cleanly separate the input-side computation (what the MLP reads) from the output-side computation (what it writes), an asymmetry exploited by transcoders (Dunefsky et al., 2024) which directly map MLP inputs to outputs in feature space. Cross-coder SAEs (Lindsey et al., 2024) extend this further by jointly decomposing two models' activations to study

fine-tuning shifts.

5.6. SAE-discovered circuits

A recent line of work uses SAE features as the unit of analysis for circuit discovery. He et al. (2024) showed that dictionary learning discovers a board-state circuit in Othello-GPT consisting of ~50 features at layers 3–6, recovering 92% of behaviour. Marks et al. (2024) used SAEs to discover a gender-bias feature in Pythia-6.9B and demonstrated targeted ablation that removes the bias without harming general performance. SAE-based circuit analysis is the principal candidate for the next dominant paradigm of mechanistic interpretability, replacing per-component patching.

5.7. Limits of SAEs and recent critiques

Three limits have surfaced. Reconstruction-loss is not behavior-loss: an SAE with low MSE can still degrade downstream loss; e2e SAEs (Braun et al., 2024) explicitly target behaviour-loss instead. Feature splitting: at higher m the same feature splits into multiple co-active sub-features (e.g., "France" splits into "France-Paris" and "France-history"), forcing analysts to either accept hierarchical features or pick a fixed m. Cross-layer alignment: SAEs at different layers often fail to share a basis, hampering circuit assembly. The critique by Sharkey et al. (2025) further notes that current SAEs assume additive feature composition, which fails for multiplicative interactions (gates, copying).

Bereska et al. (2025) connect SAEs to the adversarial vulnerability literature: high superposition correlates with adversarial brittleness, suggesting SAEs are both interpretability tools and safety diagnostics. Equivariant SAEs (Erdogan & Lucic, 2025) impose group-equivariance constraints to enforce shared structure across symmetric inputs.

5.8. Empirical anchors

5.9. SAEs as steering tools

Beyond explanation, SAE features can be clamped to control behaviour. Setting a single feature's activation to a high value at every token shifts the model's output distribution toward content related to that feature. Clamping the Golden Gate feature in Claude 3 Sonnet produces "Golden Gate Claude" (Templeton et al., 2024); clamping a refusal feature reduces the refusal rate. Compared to traditional steering vectors (Zou et al., 2023; Turner et al., 2023), SAE-feature clamping has the advantage of acting on a named concept rather than an unsupervised principal direction. Dunefsky and Cohan (2025) use one-shot optimisation to identify steering vectors that mediate safety-relevant behaviours, blurring the boundary between SAEs and steering.

5.10. Outlook

Sparse autoencoders are not a final solution. Open problems include: scaling to 70B+ models without prohibitive cost, handling multiplicative and circular features, providing automated cross-layer alignment, and producing causal evaluations independent of reconstruction loss. The companion survey A Survey on Sparse Autoencoders (Shu et al., 2025 EMNLP Findings) provides a comprehensive treatment of the SAE sub-field. We anticipate that SAE research will increasingly be evaluated on downstream tasks — circuit reconstruction, hallucination detection, debiasing — rather than intrinsic reconstruction metrics, paralleling the pattern by which language models themselves moved from perplexity to downstream evaluation in 2018–2020.

6. Self-Explanations: Chain-of-Thought, Rationales, and Faithfulness

Whereas Sections 4 and 5 examined internal mechanisms, this section turns to verbalised rationales produced by the model itself. The prompting branch begins with Chain-of-Thought (Wei et al., 2022) using few-shot reasoning prompts and Zero-Shot CoT (Kojima et al., 2022), where "let's think step by step" recovers most of the gain. Self-Consistency (Wang et al., 2023) improves accuracy via majority vote across samples. The faithful branch routes the chain through external executors: Faithful CoT (Lyu et al., 2023) uses Python execution, Program-of-Thought (Chen et al., 2023) treats code as the rationale, Symbolic CoT (Xu et al., 2024) integrates theorem-prover routing, and Logic-LM (Pan et al., 2023) connects to SAT/SMT solvers. Question Decomposition (Radhakrishnan et al., 2023) imposes a sub-question structure to improve faithfulness. Distillation and template variants — SCOTT (Wang et al., 2023) for consistency-distilled rationales and Buffer of Thoughts (Yang et al., 2024) for cached templates — round out the open literature, and DiffCoT (Cao et al., 2026) brings diffusion-based generation to CoT. The reasoning-tuned frontier is occupied by DeepSeek-R1 (DeepSeek-AI, 2025), RL-trained for long reasoning, and OpenAI o1 (2024), the canonical deliberative reasoning model.

The self-explanation family asks the LLM itself to verbalise its reasoning, producing a free-text rationale that humans can read. This artifact is qualitatively different from a saliency map or a circuit graph: it is

Object	Model	Layer	m (features)	L0	Loss-recov.	Citation
Vanilla SAE	GPT-2 small	8	8,192	30	0.85	Cunningham et al. 2023
Anthropic Mono	1-layer Transformer	1	4,096	12	0.91	Bricken et al. 2023
Scaling Mono	Claude 3 Sonnet	mid	34,000,000	300	0.92	Templeton et al. 2024
Gemma Scope	Gemma-2 2B	0–25	16K/65K/1M	50–200	0.86–0.93	Lieberum et al. 2024
Llama Scope	Llama-3 8B	0–31	64K	100	0.91	He et al. 2024
Gated SAE	Pythia-2.8B	12	65,536	50	0.94	Rajamanoharan et al. 2024
TopK SAE	GPT-4 family	undisclosed	4M	64	0.965	Gao et al. 2024
e2e SAE	Pythia-410M	6	32,768	100	0.93 (KL)	Braun et al. 2024

plausible by construction, because it is a fluent paragraph, but its faithfulness — the question of whether the rationale truthfully reflects the model's internal computation — is hotly contested.

6.1. Chain-of-Thought prompting and emergent rationales

Chain-of-Thought prompting was introduced by Wei et al. (2022, NeurIPS) as a few-shot prompting technique. The model is shown demonstrations of the form "Question: … Reasoning: … Answer: …". At inference time it produces intermediate reasoning steps before the final answer. The technique improved GSM8K accuracy from 17.9% (PaLM 540B with standard prompting) to 56.5% with CoT, and from 7% to 45% on AQuA. Critically, CoT only emerges with sufficient model scale. On PaLM 8B and 62B variants, CoT hurts performance. Self-Consistency (Wang et al., 2023 ICLR) samples multiple CoT chains and majority-votes the final answer, gaining a further 17.9 points on GSM8K. Zero-shot CoT (Kojima et al., 2022 NeurIPS) showed that simply prepending the prompt "Let's think step by step" recovers most of the few-shot CoT gain.

The reasoning quality of CoT chains has been benchmarked on Chain-of-Thought Hub (Fu et al., 2023), GSM8K (8.5K problems), MATH (12.5K problems with verified solutions), AQuA (97K), StrategyQA (2.7K), BIG-Bench-Hard (23 hard tasks), and ARC-Challenge. Modern reasoning-tuned models — OpenAI o1, DeepSeek-R1 (DeepSeek-AI, 2025), Llama-3.1-Instruct — produce CoT chains hundreds of tokens long. Chen et al. (2025) survey long Chain-of-Thought reasoning in detail, characterising it as a third paradigm beyond standard CoT.

The principal interpretability claim of CoT is that the chain reveals the model's reasoning. The principal counter-argument is that the chain reveals only what the model says about its reasoning, not what it actually computed.

6.2. Faithful CoT and program-aided execution

A solution to the faithfulness problem is to remove discretion: route the natural-language reasoning through a deterministic external executor. Faithful CoT (Lyu et al., 2023 IJCNLP-AACL) generates Python code from the natural-language prompt, runs it, and reports the result; on the GSM8K, MATH, and DROP datasets it achieves near-perfect arithmetic accuracy because the Python interpreter is exact. Program-of-Thought (Chen et al., 2023 TMLR) uses a similar idea, separating reasoning from computation. Symbolic Chain-of-Thought (Xu et al., 2024 ACL) integrates first-order-logic theorem provers; Logic-LM (Pan et al., 2023 EMNLP-Findings) routes problems to dedicated SAT/SMT solvers. A symbolic chain enjoys verifiable faithfulness: the model is interpretable insofar as the executor is, and incorrect chains are mechanically detectable when the executor errors.

The cost is generality. Faithful CoT requires that the task admit a symbolic specification, which excludes open-ended reasoning, commonsense inference, and dialogue. For these tasks, the field has fallen back on empirical faithfulness measures.

6.3. Measuring CoT faithfulness

Three empirical results have shaped the 2023–2025 understanding of CoT faithfulness. (i) Unfaithful Explanations (Turpin et al., 2023 NeurIPS): adding a biasing few-shot pattern (e.g., always answering "(A)") changes the model's final answer in ~36% of BBH cases without the CoT mentioning the bias, demonstrating that CoT can confidently rationalise an answer driven by an unstated input feature. (ii) Mea-

suring Faithfulness (Lanham et al., 2023): early-truncation of CoT, paraphrasing, and corrupting the chain produced model-and-task-dependent faithfulness; on TruthfulQA, models showed only weak faithfulness. (iii) Question Decomposition (Radhakrishnan et al., 2023): structured decomposition into subquestions improves faithfulness by reducing the search space the model can shortcut around.

A second wave of measurement followed. On the Hardness of Faithful CoT (Tanneru et al., 2024) formalised faithfulness as causal influence: a CoT step is faithful if removing it from the input changes the answer; the authors show this is hard to test because models are robust to text-level perturbations of CoT. CoT in the Wild (Arcuschin et al., 2025) showed that CoT is unfaithful even without explicit biasing prompts: on naturally occurring questions, the rationale's reasoning pattern is often a post-hoc justification. Measuring CoT Faithfulness by Unlearning (Tutek et al., 2025) proposes a parametric test: rather than removing tokens, the authors fine-tune the model to forget specific reasoning steps and measure whether the answer changes; on Llama-3-8B GSM8K, fewer than 40% of CoT steps are causally necessary, and 22% of the "answer" can be predicted from internal representations before the CoT begins.

A complementary study from interpretability lifts CoT analysis to circuits. Yeo et al. (2024 NAACL-Findings) probe how interpretable reasoning explanations from prompting actually are. Hu et al. (2024) analyse CoT through the Hopfieldian view of memory retrieval. Yu and Ananiadou (2024) trace arithmetic CoT to specific attention heads in Llama-2 7B. The emerging consensus is that CoT faithfulness depends on the model, the task, the prompting style, and the metric — there is no universal answer.

6.4. Free-text rationales beyond CoT

Beyond the prompting paradigm, the free-text rationale literature trains models to generate explanations alongside predictions. e-SNLI (Camburu et al., 2018) provides 570K NLI examples each with a one-sentence explanation; CoS-E (Rajani et al., 2019) provides 10K commonsense-QA explanations. Few-Shot Self-Rationalisation (Marasović et al., 2022 NAACL-Findings) shows that GPT-3 with 16 demonstrations matches T5 fine-tuned on the full e-SNLI dataset on automatic metrics (BLEU, METEOR), at human-rated plausibility comparable to human-written explanations. Reframing Human-AI Collaboration (Wiegreffe et al., 2022 NAACL) argues that crowd-sourced rationales should be filtered through model-overgeneration plus human selection, giving a 2× improvement in plausibility over standard collection. FLUTE (Chakrabarty et al., 2022 EMNLP) extends rationale generation to figurative-language understanding.

Rationalisation surveys are now mature. Gurrapu et al. (2023 Frontiers in AI) provide a unified treatment of free-text and extractive rationales. Luo et al. (2024 ACM CSUR) survey 200+ local-interpretation methods specifically for NLP. Both identify simulatability — does an explanation enable a third party to predict the model's answer? — as the most reliable evaluation, because it directly tests whether the explanation carries information about the prediction.

6.5. Evaluation methodology for self-explanations

Six metrics have crystallised. (i) Comprehensiveness (DeYoung et al., 2020): drop in the model's confidence when the rationale is removed. (ii) Sufficiency: confidence retained when only the rationale is kept. (iii) Leakage-Adjusted Simulatability (Hase et al., 2020): simulator accuracy after correcting for label leakage. (iv) Plausibility ratings: human Likert-scale judgments. (v) Causal influence: change in the answer under perturbations of the rationale. (vi) Internal consistency: agreement between the rationale and a separately probed internal state. Modern benchmarks (Direct Evaluation of CoT in Multi-hop Reasoning, Nguyen et al., 2024) combine these.

6.6. Specific systems and datasets

Beyond the methods cited above, the field offers several reusable systems. Logic-LM (Pan et al., 2023) integrates LLMs with seven symbolic solvers. SCOTT (Wang et al., 2023 ACL) distils a small student model from a larger teacher's CoT chains while preserving rationale-answer consistency. Buffer of Thoughts (Yang et al., 2024) caches and reuses reasoning templates. DiffCoT (Cao et al., 2026) addresses exposure bias in CoT generation via diffusion. ARS (Adaptive Reasoning Suppression, Zheng 2025) aims at over-thinking suppression. The Chain-of-Thought Hub (Fu et al., 2023) is the standard evaluation suite covering GSM8K, MATH, AQuA, StrategyQA, BBH, and ARC-Challenge. e-ViL (Kayser et al., 2021) is the benchmark for vision-language NLEs (VQA-X, e-SNLI-VE, VCR).

Concrete numbers anchor the discussion. Wei et al. (2022) reported GSM8K 56.5% with CoT on PaLM 540B vs. 17.9% baseline. Lyu et al. (2023) reported GSM8K 95% with Faithful CoT (program-aided). Lanham et al. (2023) reported that on Truth-

Metric	Definition	Example dataset
Comprehensiveness	Δ confidence when rationale removed	ERASER MovieReviews
Sufficiency	Confidence with rationale only	ERASER FEVER
Simulatability	Simulator's task accuracy given rationale	e-SNLI
Plausibility	Mean human rating	CoS-E
Causal influence	Δ answer under rationale perturbation	TruthfulQA
Early-truncation gap	Δ accuracy with/without partial CoT	GSM8K
Paraphrase invariance	Answer change rate under CoT paraphrase	BBH
Unlearning faithfulness	Δ answer after step-specific unlearning	GSM8K (Tutek 2025)

fulQA, the paraphrase-invariance faithfulness score for Claude-1 was 0.42 vs. 0.31 for Claude-2. Turpin et al. (2023) found the biasing-feature consistency of GPT-4 CoT on BBH was 64%, with the remaining 36% silently switching answers. Tutek et al. (2025) reported only 38% of CoT steps in Llama-3-8B GSM8K are unlearnable-into-incorrect, the canonical "unfaithful CoT" measurement.

Across these methods, the central pattern is that plausibility scales with model capability while faithfulness does not. Crucially, deployment in high-stakes settings demands a faithfulness audit alongside any rationale.

6.7. The plausibility-faithfulness trade-off

Self-explanations are simultaneously the most readable artifact in our taxonomy and the least faithful by default. The trade-off is not a simple rotation of axes: a rationale can be made faithful by routing into a verifier (Faithful CoT), at the cost of generality; or one can keep generality and accept correlational faithfulness measured by simulatability and unlearning. The field has not converged on a single resolution.

A pragmatic conclusion is that self-explanations are most useful when paired with a faithfulness audit. A clinical decision-support LLM should not deploy a CoT rationale unless the rationale has been validated against a downstream verifier (drug-interaction database, clinical-guideline lookup) or against an internal probe. This is precisely the architecture of recent retrieval-augmented LLMs such as MedLM and BioGPT, where the cited evidence acts as the simulatability anchor.

6.8. Outlook for self-explanation

Three directions appear promising. First, hybrid methods that combine CoT with internal probes — e.g., gate the chain on the model's internal confidence (semantic entropy) — may yield better-calibrated rationales. Second, causal CoT methods that fine-tune to produce rationales whose ablation actually changes the answer; early work (Lanham et al., 2023; Radhakrishnan et al., 2023) suggests this is feasible but hard to scale. Third, interpretability-by-design architectures (Concept Bottleneck LLMs) where the rationale is forced to flow through a sparse concept layer.

Sharkey et al. (2025) list "Faithful CoT for non-symbolic tasks" as one of the principal open problems in mechanistic interpretability. We anticipate substantial progress in 2026–2028 driven by the marriage of CoT analysis to SAE-based feature attribution: a CoT step can be considered faithful if the SAE features active during its generation also become active during the final answer's generation. Initial work in this direction (Yeo et al., 2024; Yu & Ananiadou, 2024) is encouraging but currently limited to single-digit arithmetic.

7. Knowledge Localization and Model Editing

Building on the patching machinery in Section 4 and the rationale machinery in Section 6, this section turns to surgical weight-level interventions that modify stored facts. The locate-then-edit branch is anchored by ROME (Meng et al., 2022) for the rank-one MLP edit and MEMIT (Meng et al., 2023) for mass editing of 10K facts. PMET (Li et al., 2024) refines this through attention-MLP separated editing, and MELO (Yu et al., 2024) adds a neuron-indexed dynamic LoRA layer. Meta-learning editors form a parallel strand: KE (De Cao et al., 2021) introduced a hypernetwork editor, MEND (Mitchell et al., 2022) provided a low-rank editor, and FMES (Mitchell et al., 2021) trained meta-learned weight updates. Retrieval and in-context paradigms include SERAC (Mitchell et al., 2022), which wraps edits in retrieval, and IKE (Zheng et al., 2023), which keeps edits purely in context. Recent extensions target specific failure modes: MEMIT-Merge (Dong et al., 2025) resolves same-subject batch conflicts, and Locate-then-Edit-Multi-

hop (Zhang et al., 2024) introduces multi-stage chain editing. EasyEdit (Wang et al., 2023) provides a unified editing toolkit for the entire family.

If interpretability tells us where a fact lives, editing tests whether we can change it. The locate-then-edit paradigm — first localising a piece of model knowledge with causal tracing, then surgically modifying the relevant weights — has produced the methods above together with evaluation suites such as MQuAKE and the Ripple benchmark. The remainder of this section formalises the editing operations and connects them back to mechanistic interpretability via the Hase et al. (2023) localisation-editing dissociation.

7.1. Causal tracing and the locate-then-edit paradigm

Meng et al. (2022, NeurIPS) introduced causal tracing on GPT-2 XL (1.5B parameters, 48 layers). A noisy "subject" token is introduced. Layer by layer, the activation is restored from a clean run. The layer at which restoration most strongly recovers the correct factual prediction is the causal site. The key empirical finding is that subject-token MLP outputs at mid layers (layers 5–8 in GPT-2 XL, layers 17–28 in GPT-J 6B) carry the bulk of factual association. This localisation underwrites Rank-One Model Editing (ROME). ROME models the MLP as a key-value associative memory. It applies a closed-form rank-1 update to insert a new (key, value) association without retraining:

W'_proj = $W_p$roj + $\Lambda \cdot$ (k* − $W_p$roj $\cdot$ k) $\cdot$ (v*)^T / (k^T C^{-1} k)

where k is the subject's key vector, v* is the desired output, and C is a precomputed covariance matrix. ROME edits a single fact in roughly 30 seconds on a single A100, with measured specificity (unrelated facts unchanged) of ~91% and *generalisation* (paraphrased queries also reflect the edit) of ~72%.

MEMIT (Meng et al., 2023 ICLR) extends ROME to mass editing: the rank-1 update is generalised to a least-squares solve that can install thousands of edits at once. MEMIT was demonstrated installing 10,000 edits in GPT-J 6B in a few minutes; specificity remains above 80% up to 5,000 edits, then degrades. PMET (Li et al., 2024 AAAI) refines MEMIT by separating attention-output contributions from MLP-output contributions, claiming improved precision on multi-hop benchmarks. MELO (Yu et al., 2024 AAAI) introduces a neuron-indexed dynamic LoRA layer that is activated only for queries near the edit, achieving better non-interference than rank-1 patching.

7.2. Alternative editing paradigms

Beyond locate-then-edit, three other editing paradigms are active.

Meta-learning editors: Mitchell et al. (2021 ICLR) proposed Fast Model Editing at Scale (FMES), training an auxiliary "editor" model to predict weight updates; MEND (Mitchell et al., 2022) uses a low-rank-decomposed editor for scalability. Hypernetwork approaches such as KE (De Cao et al., 2021) similarly produce edit-deltas. Meta-learning editors typically need substantial training and tend to be brittle out of distribution.

Memory-based editors: SERAC (Mitchell et al., 2022) and IKE (Zheng et al., 2023) wrap a retrieval system around the model: the edit is stored externally, and at inference time a router decides whether the cached edit applies. Retrieval editors decouple knowledge from parameters and avoid catastrophic interference, at the cost of inference complexity.

In-context editors: simply prompt the model with new facts. The advantage is zero training; the disadvantage is poor consistency across paraphrased queries. Cohen et al. (2024 TACL) showed that in-context editing produces higher ripple-effect coherence than parametric editing on the RippleEdits benchmark, suggesting that in-context updates propagate more naturally to derivative facts.

7.3. Evaluating model edits: MQuAKE, Ripple, KnowEdit

Three benchmarks dominate. MQuAKE (Zhong et al., 2023 EMNLP) is a multi-hop knowledge-edit benchmark with 9,218 questions across 1,800 edits, testing whether a model that has been edited (e.g., updating "the head of state of France") correctly answers downstream questions ("Who is married to the head of state of France?"). MQuAKE is hostile to ROME and MEMIT: their multi-hop accuracy after edit is below 25% for GPT-J 6B even when single-hop accuracy is >90%, showing that locate-then-edit fails to propagate. Ripple Effects of Knowledge Editing (Cohen et al., 2024 TACL) defines six categories of dependent facts (Logical Generalisation, Compositionality I/II, Subject Aliasing, Relation Specificity, Forgetfulness) and shows similar ripple failures. KnowEdit (Zhang et al., 2024) is a unified evaluation suite covering insertion, modification, deletion, and erasure with reproducible metrics for reliability, generalisation, locality, and portability.

The classical metrics evaluate four properties. Reliability: the edit is applied (model now answers

correctly). Generalisation: paraphrases of the edit prompt are also corrected. Locality: unrelated facts remain unchanged. Portability: dependent facts derivable from the edit are also updated. The locality–portability conflict is fundamental: locate-then-edit favours locality, in-context favours portability, and no method has solved both simultaneously.

7.4. The localisation–editing dissociation

A foundational result by Hase et al. (2023 NeurIPS), titled "Does Localisation Inform Editing?", showed that the layer identified as causal by causal tracing is not the layer that produces the best editing outcome. Editing GPT-J 6B at layer 8 (the causal-tracing peak) produces 79% reliability; editing at layer 5 produces 84%, and editing at layer 0 produces 81%. The implication is that causal localisation is correlational with respect to editability: the brain-region-style "this is where the fact is stored" intuition does not transfer cleanly to neural network editing. Hase et al. argue that this dissociation is a feature, not a bug: facts may be stored in one location and retrieved through another, and the editor must intercept the retrieval pathway, not the storage.

A related caveat is the specificity problem. Hoelscher-Obermaier et al. (2023 ACL Findings) showed that improved specificity benchmarks reveal that ROME and MEMIT silently corrupt unrelated facts at higher rates than originally reported; the actual specificity drop on a held-out set of 5,000 queries can be 15–25%, not the 91% reported on the original specificity test. This motivates the Ripple and KnowEdit benchmarks.

7.5. Editing for safety and refusal

Editing has applications beyond fact correction. Refusal direction edits (Arditi et al., 2024) use causal tracing to identify the residual-stream direction that mediates refusal in instruction-tuned Llama-2-7B-Chat and ablate it, producing a model that complies with previously-refused harmful queries. This is a jailbreak obtained by editing rather than prompting and demonstrates the dual-use nature of editing tools. Conversely, editing has been used to insert refusal directions, hardening models against jailbreaks (Wei et al., 2026 IEEE TPAMI). The steering vector literature (Section 5.9) is a cousin of editing: rather than modifying weights, it adds activations at inference time.

7.6. Editing-aware interpretability

The interplay between editing and interpretability runs both ways. Editing serves as a test of interpretability claims: if a method localises a fact to layer ℓ, an edit at layer ℓ should be the most effective intervention. Hase et al. (2023) showed this test fails for causal tracing, motivating editing-aware localisation (Geva et al., 2023; Sharma et al., 2024) that explicitly optimises for editing utility. Conversely, interpretability provides hypotheses for editing: SAE features identified as semantically meaningful (Templeton et al., 2024) are natural candidates for clamping, blurring the line between editing and steering.

7.7. Empirical anchors for evaluators

7.8. Multi-hop and cross-lingual edits

Locate-then-edit struggles with multi-hop queries: editing "France's capital is now X" correctly updates "What is the capital of France?" but often fails on "What is the population of France's capital?" because the model's chain of associations is not single-step. Zhong et al. (2023) show that on MQuAKE-CF (counterfactual edits), 2-hop accuracy after MEMIT on GPT-J 6B is 21%, compared to 68% single-hop. Locate-then-edit for Multi-hop Factual Recall (Zhang et al., 2024) addresses this with multi-stage editing across the chain. Cross-lingually, edits in English do not generally transfer to Chinese or Hindi paraphrases of the same fact (Wang et al., 2023). This is consistent with Dumas et al.'s (2024) finding that concept directions are shared late in the network but the language-specific surface forms diverge earlier.

7.9. Why model editing matters for explainability

Editing is the operationalisation of mechanistic understanding. A claim that "fact F is stored at MLP layer ℓ in subject-token position p" is testable only if we can predict an edit at (ℓ, p) will have specific behavioural consequences. The success of ROME and MEMIT is therefore evidence — incomplete but real — that the locate-then-edit decomposition captures something true about transformer associative memory. The failure of multi-hop editing is evidence that the decomposition is incomplete: facts are stored, but the chains that retrieve them are not. The future of editing-aware interpretability is to characterise both halves of the dependency.

7.10. Outlook

Three directions in near-term editing research are active. Lifelong editing: sequentially applying tens of thousands of edits without catastrophic interference; current best results plateau near 5,000 sequential edits before specificity collapses. Concept edit-

Method	Model	Edit type	Reliability	Generalisation	Locality	Portability	Citation
ROME	GPT-2 XL 1.5B	Single fact	91%	72%	91%	n/a	Meng et al. 2022
MEMIT	GPT-J 6B	10K facts	89%	73%	80%	n/a	Meng et al. 2023
PMET	Llama-2 7B	Single fact	95%	87%	92%	n/a	Li et al. 2024
FMES	T5-base	Single fact	78%	60%	95%	n/a	Mitchell et al. 2021
MEND	Llama-2 7B	Single fact	88%	70%	90%	n/a	Mitchell et al. 2022
In-Context	GPT-3.5	Single fact	94%	82%	88%	72%	Zheng et al. 2023
MELO	Llama-2 7B	Sequential	92%	81%	96%	60%	Yu et al. 2024
MEMIT-Merge	Llama-2 7B	Same-subject	87%	75%	84%	n/a	Dong et al. 2025

ing: rather than facts (subject-relation-object triples), editing concepts (Templeton et al., 2024 demonstrated clamping the Golden Gate feature, but principled "edit the deception feature" is unsolved). Hybrid retrieval-edit: combining a retrieval store (immediate, accurate) with parametric edits (efficient, integrated) is the natural next architecture. Sharkey et al. (2025) flag editing scalability and ripple-effect propagation as principal open problems.

The interpretability lessons from Section 7 transfer to applications, and Section 8 shifts from white-box editing to black-box auditing, the only family compatible with closed APIs.

8. Hallucination, Faithfulness, and Black-Box Auditing

Whereas Section 7 covered weight-level edits requiring open weights, this section turns to behavioural auditing for closed APIs. Sample-consistency methods are anchored by SelfCheckGPT (Manakul et al., 2023) with BERTScore agreement across N samples, plus the multi-choice variant MQA-SC (Manakul et al., 2023). Semantic Entropy (Farquhar et al., 2024) generalises this through NLI-clustered entropy, and Adaptive Bayesian SE (Sun et al., 2026) is an active-sampling extension. AlignScore (Zha et al., 2023) is the principal NLI alignment metric. Internal-state detectors include CCS (Burns et al., 2023), a contrast-consistent search probe, and Internal Confidence Probes (Azaria & Mitchell, 2023) targeting hidden-state truth signals. Probabilistic Distance Detection (Oblovatny et al., 2025) targets RAG-specific distance metrics. For prompt-attribution under closed APIs, Forward-Learning Saliency (Zhang et al., 2024) provides gradient-free saliency and Context Length Probing (Cífka & Liutkus, 2022) measures prefix attribution. End-to-end pipelines such as THaMES (Liang et al., 2024) chain detection with mitigation, and Faithful Finetuning (Hu et al., 2024) adds hallucination-aware training.

When an LLM is accessed only through an API — as for GPT-4o, Claude 3 Opus, and Gemini 1.5 — white-box methods such as activation patching are unavailable, and black-box auditing must shoulder hallucination detection, uncertainty quantification, and behavioural diagnosis. Hallucination detection has emerged as the most commercially urgent application of LLM explainability, motivating an extensive empirical literature evaluated against TruthfulQA, HalluDial, DiaHalu, and Poly-FEVER.

8.1. Hallucination taxonomy

Huang et al. (2023, arXiv) provide the canonical taxonomy. Factuality hallucinations fabricate factual content (e.g., a non-existent court case). Faithfulness hallucinations deviate from a provided context (e.g., a summary that contradicts the source). Intrinsic hallucinations are detectable from the model's own knowledge; extrinsic hallucinations require external verification. Empirically, hallucination rates depend on model size, decoding temperature, prompt structure, and topic frequency in the training corpus. Singhal et al. (2023 Nature) reported that Med-PaLM at temperature 0.7 produced demonstrably incorrect medical assertions in 12.4% of long-form responses, decreasing to 5.0% with retrieval augmentation. Bang et al. (2023) showed in a multitask multilingual evaluation that ChatGPT hallucinates between 2% and 35% across question types, with reasoning questions worst.

8.2. SelfCheckGPT and sample-consistency methods

SelfCheckGPT (Manakul et al., 2023 EMNLP) is the dominant black-box hallucination detector. The principle is simple. A non-hallucinated statement should be reproducible across stochastic samples. A hallucinated one will vary. Concretely, given a candidate response y and N independently sampled responses {yi}, score y by the average BERTScore (or NLI entailment, or n-gram overlap) between y and the y_i. Statements with low agreement are flagged. SelfCheckGPT achieved AUC 0.83 on a curated GPT-3 fact-generation benchmark with N=20 samples, outperforming prior methods including the Anthropic _self-evaluation prompt by 5+ AUC points.

The method has spawned variants: MQA-SC (sample multiple-choice agreement), NLI-Score (measure entailment between samples), AlignScore (Zha et al., 2023). Computational cost is the principal weakness — N=20 samples cost 20× the base call.

8.3. Semantic entropy

Farquhar et al. (2024 Nature) introduced semantic entropy: cluster the N samples by mutual NLI entailment and compute Shannon entropy over cluster sizes. Two responses that paraphrase each other count as the same cluster, eliminating the spurious entropy that token-level metrics produce. On TriviaQA the semantic entropy detector improved hallucination AUROC from 0.78 (token entropy) to 0.85; on a curated Long-Form QA benchmark, gains were 8–12 AUROC points across Llama-2 7B, Llama-2 70B, and GPT-4. Semantic entropy is now widely deployed in commercial LLM serving as a confidence signal. Sun et al. (2026) extend this with adaptive Bayesian estimation. Oblovatny et al. (2025) propose probabilistic-distance variants for retrieval-augmented generation.

8.4. Truthfulness and dialogue-level benchmarks

TruthfulQA (Lin et al., 2022 ACL) provides 817 questions across 38 categories designed to elicit common-misconception failures (e.g., "What happens if you cross your eyes for too long?"). Larger models in 2022 actually scored worse than smaller ones because they had memorised more misconceptions; this inverse-scaling phenomenon was an early warning that scale alone does not produce truthfulness. HalluDial (Luo et al., 2024) is a 4,094-dialogue benchmark with sentence-level hallucination annotations; its dialogue-level granularity catches inconsistencies that single-turn benchmarks miss. DiaHalu (Chen et al., 2024) is the second large dialogue benchmark, focused on multi-turn conversational hallucinations. Poly-FEVER (Zhang et al., 2025) is a multilingual fact-verification benchmark covering 11 languages, addressing the English-centric bias of earlier benchmarks. Singhal et al. (2023 Nature) introduced MedQA (USMLE-style) for medical-domain hallucination evaluation; Med-PaLM achieves 67.6% accuracy with chain-of-thought.

8.5. Retrieval-augmented faithfulness

Retrieval-augmented generation (RAG, Lewis et al., 2020) and its successors mitigate hallucination by grounding responses in retrieved documents. Benchmarking LLMs in Retrieval-Augmented Generation (Chen et al., 2024 AAAI) evaluates noise robustness, negative rejection, information integration, and counterfactual robustness across 1,500 queries and shows that even GPT-4-turbo with RAG fails to reject misleading retrievals 28% of the time. Faithful Fine-tuning (Hu et al., 2024) and THaMES (Liang et al., 2024) chain hallucination detection with mitigation. Decomposition-Enhanced Training for Post-Hoc Attributions in Language Models (Balasubramanian et al., 2025) provides reliable source attribution for long-document QA, the workhorse evaluation of RAG faithfulness. Probabilistic distances-based hallucination detection (Oblovatny et al., 2025) targets RAG-specific hallucinations.

8.6. Context-length probing and prompt-attribution methods

Cífka and Liutkus (2022) introduced context-length probing: the model is repeatedly evaluated on progressively longer prefixes and the contribution of each prefix range is computed by the differential log-probability. The method requires only API access and produces token-level attribution maps for any closed-API LLM. Decomposition-Enhanced Training for Post-Hoc Attributions (Balasubramanian et al., 2025) generalises this to long-document QA. Forward-Learning Saliency (Zhang et al., 2024) computes saliency maps for closed-API models without gradient access. These methods are the only realistic explanation tools for systems running behind authentication.

8.7. Internal-state hallucination detection

Although black-box methods dominate the deployed-API setting, internal-state methods perform better when access permits. Probabilistic distances (Oblovatny et al., 2025), internal-confidence probes (Azaria & Mitchell, 2023; Burns et al., 2023), and SAE-feature-based detection (Bereska et al., 2025) leverage hidden activations. Burns et al.'s Contrast-

Consistent Search (CCS) probes a model's binary truth representation and reaches 80%+ AUROC on factual-statement classification using only unsupervised contrastive features. Depth-Wise Activation Steering (Góral et al., 2025) shows that LLMs sometimes assert falsehoods despite internally representing the correct answer, motivating honesty steering. The interpretability/black-box boundary is therefore porous: commercial deployment can request both API confidence (semantic entropy) and internal-state confidence (CCS, SAE features) when the deployer also operates the model.

8.8. Concrete numerical anchors

8.9. Auditing for safety and bias

Black-box auditing extends beyond hallucination. Red-teaming (Wei et al., 2026 IEEE TPAMI; Pathade, 2025; Cantini et al., 2025) attempts to elicit unsafe outputs through adversarial prompting, providing a behavioural explanation of a model's safety boundary. The PIEE Cycle (Trabilsy et al., 2025) is a structured framework for clinical red-teaming. MTSA (Guo et al., 2025 ACL) extends red-teaming to multi-turn safety alignment. Bias auditing tools such as Atoxia (Du et al., 2024) target toxic-content generation; BlueSuffix (Zhao et al., 2024) targets vision-language jailbreaks. The PIEE work in clinical decision-making and the MTSA multi-turn alignment effort demonstrate that black-box auditing is a key practical interface between research and deployment.

8.10. The limits of black-box auditing

Black-box methods inherit a fundamental limitation. They observe the model's behaviour but not its cause. A model that confidently asserts a falsehood is indistinguishable from a model that fabricates it from training-data noise via the same probabilistic mechanism. SelfCheckGPT and semantic entropy detect statistical signatures of uncertainty but cannot localise the mechanism producing the hallucination. Goldowsky-Dill and Sharkey (2024) argue that this limitation justifies investing in white-box interpretability for high-stakes deployments. The pragmatic conclusion is that black-box and white-box methods are complementary: deploy black-box detectors for real-time confidence on closed APIs, and pair them with white-box mechanism-level audits during development of open-weight models.

8.11. Outlook for hallucination detection

Three trends will shape 2026–2028 deployment. First, combining multiple signals — token entropy + semantic entropy + retrieval grounding + internal probes — substantially improves AUROC over any single signal, and meta-detectors that learn the optimal combination are an active area. Second, generation-time uncertainty quantification (rather than post-hoc detection) is increasingly important; models such as Mistral and Llama-3 have begun to expose sentence-level uncertainty during streaming. Third, domain-specific detectors — medical (Khorsand et al., 2026), legal, scientific — outperform general-purpose detectors but require specialised data; the trade-off between coverage and accuracy is the dominant design choice in industrial deployment.

9. Datasets, Benchmarks, and Evaluation Metrics for LLM Explanations

Building on the methods of Sections 4–8, this section catalogues the evaluation infrastructure needed to compare them. The explanation-generation corpora are e-SNLI (Camburu et al., 2018) with 570K NLI explanations, CoS-E (Rajani et al., 2019) with 10K commonsense rationales, ERASER (DeYoung et al., 2020) as the 7-task extractive rationale suite, and ScienceQA (Lu et al., 2022) with 21K science explanations. Mechanistic test-beds include TracrBench (Thurnherr & Scheurer, 2024) with 12 ground-truth circuits, Othello-GPT (Li et al., 2023) as the canonical board-state probing test-bed, and RAVEL (Huang et al., 2024) for attribute-isolation evaluation. Hallucination evaluation is anchored by TruthfulQA (Lin et al., 2022) with 817 misconception questions, HalluDial (Luo et al., 2024) with 4,094 dialogue annotations, DiaHalu (Chen et al., 2024) for multi-turn hallucination, HaluEval (Li et al., 2023) with 35K hallucination samples, Poly-FEVER (Zhang et al., 2025) for 11-language fact verification, and MedQA (Singhal et al., 2023) with 12,723 USMLE questions. Editing-side benchmarks include MQuAKE (Zhong et al., 2023) with 9,218 multi-hop edits, CounterFact (Meng et al., 2022) with 21,919 counterfactuals, RippleEdits (Cohen et al., 2024) with 5K dependent-fact edits, and KnowEdit (Zhang et al., 2024) as the unified editing harness.

The empirical health of any subfield depends on its evaluation infrastructure. The remainder of this section organises the principal resources by axis (granularity × evaluation type, see Figure 4), and provides exact dataset sizes, metric definitions, and reported state-of-the-art numbers needed for downstream com-

Detector	Mechanism	Benchmark	AUROC	Citation
Token entropy	Per-token p(x)	TriviaQA	0.78	Farquhar et al. 2024
SelfCheckGPT-BERT	N=20 sample agreement	GPT-3 facts	0.83	Manakul et al. 2023
Semantic entropy	NLI-clustered entropy	TriviaQA	0.85	Farquhar et al. 2024
Semantic entropy	NLI-clustered entropy	Long-Form QA	0.79	Farquhar et al. 2024
AlignScore	NLI alignment	DialFact	0.76	Zha et al. 2023
CCS (internal)	Contrastive probe	Various TF	0.80–0.91	Burns et al. 2023
HaluEval (RAG)	RAG verification	RAG-1500	0.82	Chen et al. 2024
Probabilistic distance	RAG-specific	NQ + RAG	0.81	Oblovatny et al. 2025
Adaptive Bayesian SE	Active sampling	TriviaQA	0.87	Sun et al. 2026

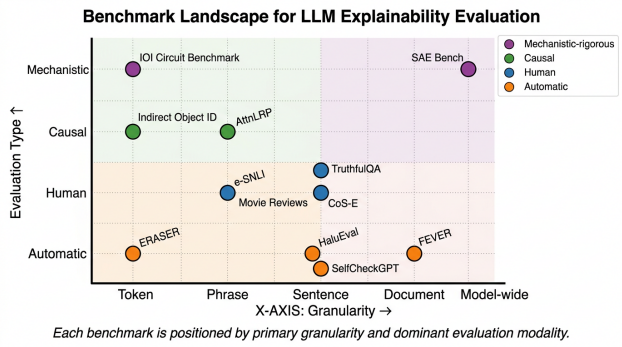

Figure 5. Benchmark landscape for LLM explainability evaluation, plotted on granularity × evaluation-type axes.

parison.

9.1. Explanation-generation datasets

e-SNLI (Camburu et al., 2018 NeurIPS) extends the 570K-example SNLI corpus with crowdsourced free-text explanations for each (premise, hypothesis, label) triple. Annotators wrote explanations averaging 14.5 tokens; inter-annotator BLEU is ~25, providing a ceiling for automatic metrics. CoS-E (Rajani et al., 2019 ACL) provides 10,962 commonsense-QA examples with explanations, designed to support training rationale-generation models. FLUTE (Chakrabarty et al., 2022 EMNLP) provides 9,000 figurative-language NLI examples with explanations focusing on metaphor and sarcasm. e-ViL (Kayser et al., 2021) is the multimodal counterpart with VQA-X (32K), e-SNLI-VE (4M visual NLI pairs), and VCR (290K commonsense visual reasoning).

ERASER (DeYoung et al., 2020 ACL) is the principal multi-task benchmark for extractive rationales. It unifies seven datasets — BoolQ (16K), MovieReviews (2K), MultiRC (32K), FEVER (185K), Evidence Inference (10K), CoS-E (10K), e-SNLI (570K) — under a common annotation schema where each example has highlighted token spans serving as the rationale. ERASER introduced the comprehensiveness and sufficiency metrics that remain the standard for extractive faithfulness.

ScienceQA (Lu et al., 2022 NeurIPS) provides 21,208 multiple-choice science questions with chain-of-thought rationales, useful for educational explanation evaluation. StrategyQA (Geva et al., 2021) provides 2,780 implicit-reasoning questions with decomposition explanations. DROP (Dua et al., 2019) provides 96K paragraph-comprehension questions where the answer requires discrete operations and the chain-of-thought is testable.

9.2. Mechanistic test-beds

TracrBench (Thurnherr & Scheurer, 2024) compiles 12 hand-written programs (sorting, copying, reversing) into Tracr models with known ground-truth circuits, providing a faithfulness ceiling for automated discovery; current best F1 averages 0.79 across the suite. Othello-GPT (Li et al., 2023; He et al., 2024) is a 1.5M-game corpus on which a small transformer is trained to predict legal moves; linear probes recover the board-state from middle-layer activations with F1 ~0.96, establishing the model has learned a board representation. RAVEL (Huang et al., 2024) tests representation-attribution: each input has multiple attributes (e.g., a city has country, population, language) and the benchmark scores whether intervention methods can isolate one attribute without affecting others. IOI prompt set (Wang et al., 2022) is a 1,000-prompt evaluation suite for the Indirect Object Identification circuit. Greater-Than circuit (Hanna et al., 2023) provides a 1,000-prompt evaluation for numeric comparison. Docstring (Heimersheim & Nanda) provides synthetic Python docstring completion tasks with localised circuits.

These mechanistic test-beds matter because they give the field a ground truth: a known circuit against which

automated discovery can be evaluated. Without them, claims of mechanistic understanding are unfalsifiable.

9.3. Hallucination and faithfulness benchmarks

TruthfulQA (Lin et al., 2022 ACL) — 817 questions across 38 categories. HalluDial (Luo et al., 2024) — 4,094 dialogues, sentence-level annotations. DiaHalu (Chen et al., 2024) — multi-turn conversational hallucination. HaluEval (Li et al., 2023 EMNLP) — 35K samples spanning QA, dialogue, and summarisation with synthesised and human-annotated hallucinations. FELM (Chen et al., 2023) — fine-grained factuality evaluation. Poly-FEVER (Zhang et al., 2025) — multilingual fact verification across 11 languages. MedQA (used by Med-PaLM, Singhal et al., 2023) — USMLE-style medical QA (12,723 questions). These benchmarks support both detection metrics (AUROC) and mitigation metrics (factuality after intervention).

9.4. Editing benchmarks

ZsRE (Levy et al., 2017) — 240K zero-shot relation-extraction queries used by ROME for single-fact editing. CounterFact (Meng et al., 2022) — 21,919 counterfactual statements paired with paraphrases for generalisation testing. MQuAKE (Zhong et al., 2023 EMNLP) — 9,218 multi-hop edit queries across 1,800 base edits. RippleEdits (Cohen et al., 2024 TACL) — 5,000 edits with six categories of dependent fact (Logical Generalisation, Compositionality I/II, Subject Aliasing, Relation Specificity, Forgetfulness). KnowEdit (Zhang et al., 2024) — unified suite covering insertion, modification, deletion, erasure with reliability/generalisation/locality/portability metrics.

9.5. Faithfulness, simulatability, and L0/loss-recovered metrics

The metric vocabulary of LLM explainability is dense. We summarise the principal entries.

Modern papers typically report 3–5 of these metrics, reflecting the fact that no single metric captures all aspects of explanation quality.

9.6. Cross-benchmark relationships

A growing meta-evaluation literature studies how benchmark choices interact. Bodria et al. (2023, DMKD) found that LIME and KernelSHAP have Spearman $\rho = 0.72$ on ordinary inputs but only 0.41 on adversarially perturbed ones, demonstrating that benchmark stability matters. Atanasova et al. (2020) showed that the comprehensiveness metric ranks methods differently than human-rated faithfulness, suggesting either comprehensiveness or human ratings is mis-specified. The Feature Attribution Stability Suite (Subramaniakuppusamy & Gajjar, 2026) is the most recent attempt to provide a unified stability evaluation across attribution methods.

9.7. Reproducibility infrastructure

A non-trivial obstacle to evaluation is reproducibility. Three open-source frameworks dominate. TransformerLens (Nanda 2022) provides hookable HuggingFace transformer implementations for GPT-2, GPT-J, Pythia, Llama, Mistral, and Gemma; the standard tool for activation patching, ACDC, and probing. nnsight (Fiotto-Kaufman et al., 2024) supports any HuggingFace model with a uniform tracing API. Pyvene (Wu et al., 2024) provides composable interventions. SAELens (Bloom et al., 2024) is the standard SAE training and evaluation library. Neuronpedia hosts >130M extracted SAE features as a browseable web service. EasyEdit (Wang et al., 2023) is a unified editing toolkit; Inseq (Sarti et al., 2023) standardises feature-attribution interfaces; captum (Kokhlikyan et al., 2020) provides PyTorch-native attribution.

9.8. Reported benchmark scores (representative)

9.9. The benchmark gap

Despite this density of resources, two systematic gaps persist. First, there is no widely adopted unified faithfulness leaderboard — papers report different metric subsets on different splits. Second, frontier-model benchmarks remain undeveloped: most explainability benchmarks target GPT-2-scale models because their behaviour is testable in a small lab; the explainability of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro is evaluated only piecemeal. Sharkey et al. (2025) flag the development of frontier-scale faithfulness benchmarks as an urgent open problem; the Open Problems in Mechanistic Interpretability paper specifically calls for benchmarks that test SAE features for causal sufficiency rather than reconstruction quality alone.

A further concern is the test-set contamination problem: as models train on increasingly large fractions of the public internet, public benchmarks may leak into pretraining data. TruthfulQA, GSM8K, and MMLU all show partial contamination in modern frontier models, complicating the interpretation of reported scores. The community has begun to use held-out and temporally novel test sets — questions written after a model's training cutoff — to mitigate contamination. Jailbreak Distillation (Zhang et al., 2025) pro-

Metric	Family	Definition	Higher is better
Comprehensiveness (Comp)	Extractive	Δ confidence when rationale removed	Yes
Sufficiency (Suff)	Extractive	Confidence with rationale only	Yes
Simulatability	Free-text	Sim. accuracy given rationale	Yes
Leakage-Adj. Sim.	Free-text	Sim. minus label-leakage baseline	Yes
BLEU / METEOR / BERTScore	Free-text	Lexical/semantic similarity to gold	Yes
Fidelity	Post-hoc	Surrogate vs. model agreement	Yes
Sensitivity-N	Saliency	Spearman ρ between scores and N-removed Δ	Yes
Sufficiency-N (deletion AUC)	Saliency	AUC of confidence vs. tokens removed	Yes
Reconstruction MSE	SAE	$\|x - \hat{x}\|^2$	No
L0 sparsity	SAE	mean active features per token	No (lower better)
Loss-recovered fraction	SAE	1 − (L_with_SAE − L_clean)/(L_zero − L_clean)	Yes
Feature interpretability	SAE	% features judged monosemantic by labeller	Yes
KL on patched logits	Mech.	KL(M_patched ‖ M_clean)	No (lower better)
Circuit F1	Mech.	F1 of recovered components vs. ground truth	Yes
AUROC (hallucination)	Black-box	ROC AUC on hallucinated vs. faithful	Yes
Reliability (editing)	Editing	Edit applied correctly	Yes
Generalisation (editing)	Editing	Paraphrase consistency post-edit	Yes
Locality (editing)	Editing	Unrelated facts unchanged	Yes
Portability (editing)	Editing	Dependent facts updated	Yes
Early-truncation gap	CoT	Δ accuracy with partial CoT	Depends on direction
Paraphrase invariance	CoT	Answer consistency under CoT paraphrase	Yes
Unlearning faithfulness	CoT	Δ answer after step-specific unlearning	Yes (if positive)

poses "renewable safety benchmarking" via continual distillation of new test cases.

9.10. What an ideal benchmark would look like

For LLM explainability to mature as a measurable engineering discipline, we believe four properties are essential. (1) Causal faithfulness: a metric that fails when an explanation is correlational; semantic entropy and the unlearning-based CoT metric are early instances. (2) Frontier-model coverage: tests that scale to 70B+ models without prohibitive compute, e.g., attribution patching at scale. (3) Domain transfer: explanations evaluated for both general and domain-specific (medical, legal) tasks. (4) Reproducibility infrastructure: code, weights, and evaluation harness released together. The KnowEdit and RippleEdits benchmarks for editing meet criteria (1) and (4); Gemma Scope and Llama Scope enable (2). The integration of these into a single unified suite is the obvious next step, and we expect 2026–2028 to deliver it.

10. Applications: Healthcare, Safety, Multilingual and Multimodal Settings

Whereas Section 9 catalogued evaluation infrastructure, this section turns to deployment domains where LLM explainability has measurable impact. The biomedical strand is anchored by Med-PaLM 2 (Singhal et al., 2023) at 67.6% on MedQA, alongside BioGPT (Luo et al., 2022) as a biomedical backbone, PubMedBERT (Gu et al., 2021) as a biomedical encoder, and MedAlpaca (Han et al., 2023) as an instruction-tuned medical LLM. The multimodal frontier comprises LLaVA (Liu et al., 2023) as a vision-language assistant, Qwen-VL (Bai et al., 2023) as the multimodal Qwen, GPT-4V (OpenAI, 2023) as a vision-language frontier model, Gemini 1.5 Pro (Google, 2024) as a multimodal long-context system, and Whisper (Radford et al., 2023) for audio. Prisma (Joseph et al., 2025) provides a multimodal SAE toolkit. Safety-oriented deployments include Refusal Direction Editing (Arditi et al., 2024) for jailbreaks via residual ablation, the PIEE Cycle (Trabilsy et al., 2025) for clinical red-teaming, and MTSA (Guo et al.,

Method	Benchmark	Score	Citation
Standard prompting	GSM8K (PaLM 540B)	17.9%	Wei et al. 2022
CoT prompting	GSM8K (PaLM 540B)	56.5%	Wei et al. 2022
CoT + Self-Consistency	GSM8K (PaLM 540B)	74.4%	Wang et al. 2023
Faithful CoT (Python)	GSM8K	95.0%	Lyu et al. 2023
ACDC	IOI (GPT-2 small)	F1 0.85	Conmy et al. 2023
EAP	IOI (GPT-2 small)	F1 0.91	Syed et al. 2024
ROME	CounterFact GPT-2 XL	91% reliability	Meng et al. 2022
MEMIT	CounterFact 10K edits GPT-J 6B	89% reliability	Meng et al. 2023
MEMIT	MQuAKE-CF 2-hop GPT-J 6B	21%	Zhong et al. 2023
In-Context	MQuAKE-CF GPT-3.5	72% portability	Cohen et al. 2024
Med-PaLM	MedQA (USMLE)	67.6%	Singhal et al. 2023
Token entropy	TriviaQA AUROC	0.78	Farquhar et al. 2024
Semantic entropy	TriviaQA AUROC	0.85	Farquhar et al. 2024
SelfCheckGPT-NLI	GPT-3 facts AUROC	0.83	Manakul et al. 2023
34M-feature SAE	Claude 3 Sonnet loss-rec	0.92	Templeton et al. 2024
Gated SAE	Pythia-2.8B loss-rec	0.94	Rajamanoharan et al. 2024
TopK SAE	GPT-4 family loss-rec	0.965	Gao et al. 2024
TracrBench avg F1 (auto)	TracrBench	0.79	Thurnherr & Scheurer 2024
Othello-GPT linear probe	Board state F1	0.96	Li et al. 2023

2025) for multi-turn safety alignment. The reasoning frontier is occupied by DeepSeek-R1 (DeepSeek-AI, 2025), an RL-trained reasoning model.

LLM explainability is no longer purely an academic exercise. Healthcare deployments demand auditable rationales, AI-safety teams use mechanistic methods to understand jailbreak vulnerabilities, and multilingual and multimodal extensions test cross-setting transfer of techniques developed for English text-only LLMs.

10.1. Clinical decision support and biomedical LLMs

Medical LLMs are the canonical high-stakes application. Med-PaLM 2 (Singhal et al., 2023 Nature) achieves 67.6% accuracy on the USMLE-style MedQA benchmark and uses CoT rationales validated by clinician panels. The validation protocol — clinician scoring on six axes (alignment with consensus, no relevant information missing, no incorrect content, no harm, no bias, comprehension of intent) — is essentially a plausibility-and-faithfulness audit. BioGPT (Luo et al., 2022) and PubMedBERT (Gu et al., 2021) provide biomedical-domain backbones; AlphaMed and MedAlpaca are instruction-tuned variants. The principal explainability tool deployed in clinical workflow is retrieval-augmented chain-of-thought, where each medical claim in the rationale is grounded to a specific document.

Recent clinical applications include: Triage prediction (Lee et al., 2026 JAMIA) used BERT-based classifiers on emergency-department conversations with SHAP attributions. ICU decision support (Lee et al., 2025 BMC) uses SHAP for feature importance combined with LLM-generated explanations. Visual prognosis prediction in age-related macular degeneration (Zhao et al., 2027 Lancet Digital Health) couples deep learning with explainability. Quality-of-care measurement for ADHD (Bannett et al., 2026 medRxiv) evaluates LLMs for transparent care metrics. Cancer germline testing pathway (Mason et al., 2026 NCCN) deploys LLM-based decision support with explanation traces.

The PIEE Cycle (Trabilsy et al., 2025) provides a structured red-teaming framework for clinical LLMs: Probe, Identify, Evaluate, Escalate. The cycle integrates black-box auditing (semantic entropy, SelfCheckGPT) with domain-expert review. Readability concerns matter: Maywald et al. (2025 Current Oncology) showed that 12% of adults have proficient health literacy, demanding plain-language explanations alongside technical rationales.

10.2. AI safety, jailbreak analysis, and steering vectors

Mechanistic interpretability is increasingly applied to AI safety. The 2024 Anthropic Scaling Monosemanticity report identified SAE features for deception, secret keeping, and sycophancy in Claude 3 Sonnet, showing they activate in safety-relevant contexts. Clamping the refusal feature reduces refusal rate; clamping the

sycophancy feature decreases user-pleasing tendencies.

Jailbreak research is a cousin of explainability. GPT-FUZZER (Yu et al., 2023) auto-generates jailbreak prompts; Sirens' Whisper (Gao et al., 2023) demonstrates near-ultrasonic jailbreaks of speech LLMs; MTSA (Guo et al., 2025 ACL) extends safety alignment to multi-turn dialogues. The defence side: Wei et al. (2026 IEEE TPAMI) propose few-in-context demonstrations that guard aligned LLMs against jailbreaks. Latent Adversarial Training (Sheshadri et al., 2024) improves robustness to persistent harmful behaviours. BlueSuffix (Zhao et al., 2024) defends vision-language models against jailbreaks. LLM Stinger (Jha et al., 2025 AAAI) and TombRaider (Ding et al., 2025 EMNLP) extend the attack surface; SearchAttack (Yan et al., 2026) uses online web search as a jailbreak vector.

The mechanistic-interpretability approach to safety is best illustrated by Arditi et al.'s (2024) refusal-direction analysis: a single residual-stream direction in Llama-2-7B-Chat mediates ~80% of refusal behaviour, and ablating it produces a model that complies with harmful queries. Conversely, one-shot optimised steering vectors (Dunefsky & Cohan, 2025) can mediate safety-relevant behaviours with minimal data. Honesty steering (Góral et al., 2025) targets the gap between models' internal beliefs and stated answers.

10.3. Multilingual probing and explainability

The multilingual extension of LLM explainability is mapped by Resck, Augenstein and Korhonen (2025 EMNLP), surveying ≥80 studies. Key findings: (i) probing recovers similar syntactic information across high-resource languages but is much weaker on low-resource ones; (ii) concept directions are largely language-agnostic in the late layers (Dumas et al., 2024), validating the language-thought separation hypothesis (Mahowald et al., 2024 TICS); (iii) edits in English do not transfer to other languages without retraining. Multilingual probing of contextual encoders (Ravishankar et al., 2019) was the early benchmark for this work; current evaluation uses XCOPA, XGLUE, MASSIVE, and the Belebele reading-comprehension benchmark across 100+ languages. Zhang et al. (2025 Cybersecurity) survey multilingual safety in cybersecurity. Chinese classifier prediction (Zhang et al., 2025) compares BERT and modern LLMs on a language-specific phenomenon. Function induction (Ye et al., 2025) and separating tongue from thought (Dumas et al., 2024) provide deeper mechanistic claims.

10.4. Multimodal LLMs (MLLMs)

Multimodal LLMs combine language with vision (LLaVA, Qwen-VL, GPT-4V, Gemini 1.5 Pro), audio (Whisper, Qwen-Audio), or video (Video-LLaMA). Dang et al. (2024) survey explainability for MLLMs comprehensively; Ghosh et al. (2024) survey vision-language models. Key methods transferred from text: LIME-Vision and SHAP-Vision for token attribution; CoT rationales adapted to grounded reasoning (Visual Chain-of-Thought); SAEs trained on multimodal residual streams (Joseph et al., 2025 Prisma toolkit). Mechanistic interpretability of Vision Transformers (Bahador, 2025) and Seeing Through Circuits (Żukowska et al., 2026) extend circuit-level analysis to ViT. e-ViL (Kayser et al., 2021) is the principal benchmark with VQA-X (32K), e-SNLI-VE, and VCR (290K). Dr. SHAP-AV (Cappellazzo et al., 2026) adapts SHAP to audio-visual speech recognition.

The challenges differ from text: ViT and CLIP-style encoders sit upstream of the language model, so explanations must trace causality across modality boundaries. Interpretable medical VQA (Hu et al., 2024) uses multimodal relationship graphs to provide attribution. The 2024–2026 frontier is cross-modal feature linking: which SAE features in the language stream correspond to which patches in the visual stream?

10.5. Education and tutoring

Educational deployments of LLMs require explainability for both teacher review and student comprehension. Kasneci et al. (2023 Learning and Individual Differences) surveyed ChatGPT-in-education; explainability is identified as a top research priority. Worked examples generated by LLMs (with stepwise CoT rationales) provide both pedagogical content and a faithfulness anchor — students can evaluate whether the steps make sense. MultiSense-L2Net (Abbas et al., 2026) uses AI-mediated co-regulation in second-language writing. Cognitive Flow (Dissanayake & Nanayakkara, 2025) studies AI interventions for reasoning support. The principal interpretability tools deployed in education are CoT rationales and retrieval grounding, both of which are robust to closed-API constraints.

10.6. Finance, law, and other regulated domains

Financial-services applications of LLM interpretability are surveyed by Golgoon et al. (2024 ICAIF), who specifically examine activation patching and probing on financial-text models. Cybersecurity applications are surveyed by Zhang et al. (2025 Cybersecu-

rity). Legal applications are nascent — Casetext's CoCounsel and LexisNexis Lexis+ AI are commercial deployments — and lean heavily on retrieval-augmented faithfulness anchors. The EU AI Act (2024) classifies LLM-based legal advice as a high-risk system, requiring traceability documentation that mechanistic interpretability tools could in principle supply.

10.7. Reinforcement-learning aligned LLMs

LLMs aligned with RLHF or RLAIF acquire new behaviours that demand fresh interpretation. Liu et al. (2025) survey RL-meets-LLM. Reasoning-tuned models such as OpenAI o1 and DeepSeek-R1 (DeepSeek-AI, 2025) produce long CoT chains whose faithfulness has not yet been characterised. Improving Reasoning via Representation Engineering (Højer et al., 2025) shows that reasoning behaviour can be elicited via steering vectors, suggesting reasoning is mediated by specific residual-stream directions amenable to interpretation.

10.8. Application-domain summary

10.9. Cross-domain lessons

Three cross-domain lessons emerge. First, RAG anchors are a remarkably consistent pattern: when a domain demands faithful explanations, retrieval-augmented citation provides a ground-truth check that pure CoT lacks. Second, human review remains the gold standard, with interpretability tools serving to make that review tractable rather than to replace it. Third, domain-specific benchmarks substantially outperform general ones: a hallucination detector trained on TriviaQA does not transfer well to MedQA. The deployment success of Med-PaLM and the slow progress of legal LLMs reflect this gap.

10.10. Outlook for application-driven explainability

Three trends are reshaping deployment. First, regulation-driven evaluation: the EU AI Act requires traceability for high-risk systems, pushing interpretability from research into compliance. Second, agentic LLMs: tool-using agents (Wang et al., 2024) demand explanations of multi-step decisions, blurring the line between CoT, retrieval, and tool-call traces; the AIOps survey (Zhang et al., 2025) maps this for IT operations. Third, cross-modal: as medical and scientific deployments combine text with images and structured data, explanations must trace causality across modality boundaries, with Dr. SHAP-AV and Prisma (Joseph et al., 2025) as early instances. Clinical and legal deployment will be the principal forcing function for explainability research over the next five years, much as ImageNet shaped computer vision in the 2010s.

11. Limitations, Failure Modes, and Adversarial Threats to Explanations

Whereas Section 10 surveyed application successes, this section catalogues failure modes and adversarial threats. The adversarial-attribution strand opens with Fooling LIME and SHAP (Slack et al., 2020), a 95% scaffold-detection attack, and the SHLIME defence (Chauhan et al., 2025) as the anti-scaffold technique. Explanation Bias (Kamp et al., 2025) further exposes lexical and positional bias in Integrated Gradients. Chain-of-Thought failures cluster around four results: the biasing-feature CoT shift of Turpin et al. (2023), with 36% silent answer change; the paraphrase non-invariance of Lanham et al. (2023), where surface form drives CoT; the unlearning-based finding of Tutek et al. (2025) that fewer than 40% of CoT steps are causally necessary; and CoT in the Wild (Arcuschin et al., 2025), which characterises naturally occurring chains as post-hoc rationalisation. Editing failures include the specificity collapse of Hoelscher-Obermaier et al. (2023) with 15–25% silent corruption, and the multi-hop ripple failure of Zhong et al. (2023) where 2-hop accuracy falls to 21%. SAE-side limits include Circular Features (Engels et al., 2024) for non-linear feature manifolds and the reconstruction-behaviour gap of Braun et al. (2024) between MSE and KL divergence. The COVID Shortcut paper (DeGrave et al., 2021) remains the canonical saliency-shortcut cautionary tale.

Explanations can fail in several systematic ways: they can be unfaithful (correlate poorly with model causation), brittle (sensitive to input perturbations), gameable (manipulable by an adversary), or correct but useless (technically faithful but cognitively unintelligible). The remainder of this section reports specific empirical demonstrations of each failure regime, building a conservative picture of what current LLM explainability cannot yet do.

11.1. Fooling LIME and SHAP: adversarial saliency

Slack et al. (2020 AIES) showed that an adversary controlling the model can produce LIME and SHAP explanations that hide a biased decision rule. The construction inserts a scaffold classifier. The scaffold detects perturbed inputs (used by LIME) versus original inputs and routes them to different sub-models. A fair sub-model handles perturbations and generates the explanation. A biased sub-model han-

Domain	Primary task	Dominant explainability method	Key benchmark	Citation
Medicine	QA / triage	CoT + RAG + clinician audit	MedQA (12.7K)	Singhal et al. 2023
ICU support	Risk stratification	SHAP + LLM rationale	MIMIC-III/IV	Lee et al. 2025
Safety	Jailbreak defence	Refusal direction, steering	HarmBench, AdvBench	Arditi et al. 2024
Multilingual	Cross-lingual NLI	Probing + activation patching	XGLUE, Belebele	Dumas et al. 2024
Multimodal	VQA + rationale	SHAP-V + CoT	VQA-X (32K), e-ViL	Kayser et al. 2021
Education	Tutoring	CoT + retrieval grounding	ScienceQA (21K)	Lu et al. 2022
Finance	Risk / fraud	Probing + patching	Domain-specific	Golgoon et al. 2024
Cybersecurity	Vulnerability detection	LLM CoT + verification	DeepSeek-R1 eval	Zhang et al. 2025
Legal	Case retrieval	Retrieval + citation grounding	LegalBench	(commercial)
Reasoning	Math / code	CoT + Faithful CoT	GSM8K, MATH	Wei et al. 2022

dles actual queries. The technique fools both LIME and KernelSHAP across the COMPAS, Communities and Crime, and German Credit datasets, with scaffold-detection accuracy >95%. The impact is that LIME and SHAP explanations cannot be trusted in adversarial settings without auxiliary integrity checks. SHLIME (Chauhan et al., 2025) and the Feature Attribution Stability Suite (Subramaniakuppusamy & Gajjar, 2026) develop defences but do not eliminate the threat. The methodological lesson is that local post-hoc methods that perturb the input must verify the model behaves consistently inside and outside the perturbation distribution.

A related result is Explanation Bias (Kamp et al., 2025): post-hoc feature attributions exhibit lexical and positional biases independent of the underlying decision, because Integrated Gradients and similar methods over-attribute to the first token and to high-frequency words. Forward-Learning Saliency (Zhang et al., 2024) attempts black-box-compatible saliency without gradient access but inherits the same biases.

11.2. Unfaithful Chain-of-Thought

Section 6 documented the principal CoT-faithfulness failures. We summarise the most important: (i) Biasing-feature unfaithfulness (Turpin et al., 2023): when GPT-4 is given a prompt with a biasing pattern (e.g., always answer "(A)"), the model often switches its answer without acknowledging the bias in the rationale; on BBH, 36% of answers shift silently. (ii) Paraphrase non-invariance (Lanham et al., 2023): paraphrasing a CoT chain changes the answer with non-trivial frequency, indicating the surface form rather than the logical content drove the prediction. (iii) Unlearning faithfulness (Tutek et al., 2025): on Llama-3-8B GSM8K, fewer than 40% of CoT steps are causally necessary; 22% of answers can be predicted from the model's representations before the CoT begins. (iv) Reasoning in the wild (Arcuschin et al., 2025): even without explicit biasing, naturally occurring CoT chains are post-hoc rationalisations roughly 25–30% of the time.

The implication is that CoT-as-explanation requires audit. A clinical or legal deployment that relies on the rationale must either pair it with retrieval grounding (so that each claim cites a verifiable source) or with a faithfulness probe (so that internal-state evidence corroborates the rationale).

11.3. The brittleness of probes

Hewitt and Liang's (2019) control-task warning remains relevant. A probe that recovers a property may be exploiting its own expressivity rather than the underlying model's representation. The selectivity metric (task minus control accuracy) is a partial fix but does not resolve the deeper issue: a high-selectivity probe still cannot distinguish between a property genuinely encoded by the model and a property reconstructible from incidental features. Shin et al. (2020 EMNLP) showed that AutoPrompt-discovered probes can elicit knowledge from BERT that supervised probes cannot, suggesting that probes are biased es-

timators of what a model "knows".

A second probe pathology is cross-task transfer failure: a probe trained to recover X from layer ℓ does not transfer to a task that relies on X downstream, indicating the probe captures a probe-readable version of X that may not be what the model uses. Hase et al. (2023) showed this for factual recall: causal-tracing-localised facts do not transfer to editing accuracy.

11.4. Superposition and the limits of SAE features

SAEs assume an additive feature decomposition: x ≈ Σi $z_i \cdot d_i$ where d_i are dictionary elements. Several pathologies undermine this assumption. (i) _Multi-dimensional features (Engels et al., 2024): some features are encoded on circular manifolds (days-of-week, months) that linear dictionaries cannot capture without redundancy. (ii) Cross-layer mixing (Lindsey et al., 2024): features at later layers are non-trivially compositions of features at earlier layers; per-layer SAEs miss this. (iii) Feature splitting at higher m: increasing dictionary width causes features to fragment into non-orthogonal sub-features. (iv) Reconstruction–behavior gap (Braun et al., 2024): low MSE does not imply preserved downstream behaviour. (v) Adversarial vulnerability (Bereska et al., 2025): high superposition correlates with adversarial brittleness, suggesting SAEs are diagnosing rather than eliminating a fundamental representational risk.

11.5. Editing failures

The editing literature has identified several reproducible failures. (i) Specificity collapse under sequential edits — after ~5,000 sequential edits with MEMIT, GPT-J 6B begins corrupting unrelated facts. (ii) Multi-hop ripple failures — Zhong et al. (2023) show that ROME/MEMIT edits propagate poorly to multi-hop queries (21% accuracy at 2-hop). (iii) Cross-lingual transfer failures — edits in English do not propagate to Hindi or Chinese paraphrases (Wang et al., 2023). (iv) Locality–portability conflict — methods that maximise locality fail at portability and vice versa. (v) Catastrophic forgetting in lifelong editing. MEMIT-Merge (Dong et al., 2025) addresses same-subject batch-edit conflicts but not the broader specificity decay.

11.6. Hallucination-detector failures

Black-box hallucination detectors have characteristic failure modes. (i) Sample-consistency methods (SelfCheckGPT, semantic entropy) fail on consistently incorrect hallucinations. The model is confidently wrong across all samples. Farquhar et al. (2024) report this case explicitly. (ii) Detectors confuse aleatory uncertainty (legitimate ambiguity) with epistemic uncertainty (model error). (iii) RAG-based detectors fail when the retrieval system itself returns incorrect documents (28% on noisy retrieval, Chen et al., 2024). (iv) Internal-state probes (CCS) suffer reliability degradation on out-of-distribution prompts. (v) Domain transfer is poor: a TriviaQA detector loses 8–15 AUROC points on MedQA without retraining.

11.7. Computational and accessibility limits

Many state-of-the-art interpretability methods are computationally heavy. ACDC at the IOI scale takes ~3 GPU-hours; activation-patching sweeps on Llama-2 7B exceed 60 GPU-hours per task. Training a single SAE for Llama-3 8B costs hundreds of GPU-hours; the 34M-feature Claude SAE is reported (Templeton et al., 2024) to have consumed millions of dollars of compute. These costs put cutting-edge methods out of reach for most academic groups and concentrate progress in a few well-resourced industrial labs. Attribution patching (Syed et al., 2024) and gated SAEs (Rajamanoharan et al., 2024) are partial mitigations, but the asymmetry persists. Public releases such as Gemma Scope (Lieberum et al., 2024) help democratise access at the cost of fixing the hyperparameter choices made by the releasing team.

11.8. Conceptual limits: the meaning of "explanation"

A foundational concern, raised by Saphra and Wiegreffe (2024 BlackboxNLP) and Singh et al. (2024 Rethinking Interpretability), is that the term "mechanistic" can mislead: a circuit is one causally sufficient implementation of a behaviour, not a uniquely correct one. Multi-realisability is unavoidable; circuit analysis can identify a mechanism but not the mechanism. This is not necessarily a bug — for engineering purposes a sufficient mechanism is enough — but it is a methodological constraint on how strongly mechanistic claims can be stated. Schneider's (2024) GenXAI survey takes this further, arguing that explanations must be evaluated for whom — a regulator's needs differ from a user's, which differ again from a researcher's.

A related concern is cognitive accessibility: a 26-head circuit graph for IOI is correct but largely unintelligible to a non-specialist. The end-user of a clinical decision-support system cannot read circuit diagrams; their explanation must be a natural-language summary, which reintroduces faithfulness questions. This is not a defect of mechanistic methods; it is a limit

of the artifact. Explanations targeting different stakeholders use different artifacts, and the field has not solved the unifying explanation problem.

11.9. Failure-mode catalogue

11.10. What "honest about limits" looks like

Three rhetorical practices distinguish careful from overclaiming explainability work. First, report multiple metrics — comprehensiveness, sufficiency, simulatability, causal influence — because no single metric captures faithfulness. Second, report control experiments: paraphrase invariance, biasing-feature consistency, perturbation robustness. Third, distinguish correlational from causal claims explicitly. Singh et al. (2024) and Saphra and Wiegreffe (2024) provide model citations for this practice; the Rethinking Interpretability paper specifically argues that the LLM era requires re-grounding interpretability on causal sufficiency rather than statistical correlation.

The COVID-19 chest-radiograph shortcut paper (DeGrave et al., 2021 Nature Machine Intelligence) is the canonical cautionary tale for the field: a model that achieves 99% AUROC on COVID detection turns out to be using shortcuts (laterality markers, scan-bed style) rather than disease-relevant signal. Saliency methods such as Grad-CAM happily produced lung-localised heatmaps that masked the shortcut, because the saliency was correlated with the shortcut features that themselves correlated with the lung region. The paper is regularly invoked as the reason why explainability evaluation must include adversarial and out-of-distribution stress tests.

11.11. Outlook

Limitations are an active research front, and mitigations are emerging: gated and end-to-end SAEs reduce reconstruction-behavior gap, symbolic and program-aided CoT recovers faithfulness in narrow domains, and editing-aware localisation begins to close the localisation-editing dissociation. The field, however, remains in a phase where every published method comes with a known failure regime. Researchers should report known limitations on every method, and downstream users should treat any single-explanation deployment as a starting point that must be paired with audit.

12. Open Problems and a Forecast for 2026–2030

Building on the limitations in Section 11, this section delivers a structured catalogue of open problems and falsifiable forecasts for 2026–2030. The open problems are organised by methodological pillar — SAE scaling, faithfulness benchmarks, faithful CoT, causal SAE evaluation, cross-layer linking, agentic interpretability, multimodal/multilingual, and regulation — each paired with a measurable success criterion.

Our problem list inherits substantially from Sharkey et al.'s (2025 Open Problems in Mechanistic Interpretability) but extends to the broader explainability scope of this survey, so that this section may be re-evaluated in 2030.

12.1. Scaling SAEs and circuits to frontier models

The single largest open problem is scaling. Mechanistic interpretability methods that work on GPT-2 small (117M, 12 layers) become prohibitive on Llama-3 70B (80 layers, 8192 hidden dim). Activation patching scales as O(L × H × T) where L is layers, H is heads per layer, T is tokens; an exhaustive sweep on Llama-3 70B for a 200-token prompt requires ~10 million forward passes per prompt pair. Attribution patching (Syed et al., 2024) reduces this to ~10 backward passes but the linearisation breaks for non-linear circuits. SAEs at frontier scale require billions of parameters: Anthropic's 34M-feature SAE on Claude 3 Sonnet's residual stream consumed ~2 × 10^17 tokens of activations and on the order of $1M of compute. Open problems include: (i) cheaper SAE variants (e.g., gated SAE at 16K width recovers most of the loss recovered by 1M-width vanilla SAE, Rajamanoharan et al., 2024); (ii) cross-model transfer (an SAE trained on Pythia-1.4B may not transfer to Pythia-12B); (iii) compositional SAEs that share a base across tasks.

Forecast 1: by 2027 a public SAE suite at 1M-feature width will be released for a 70B-parameter model with loss-recovered fraction $\geq$0.92. Falsified if no such suite exists by end-2027.

12.2. Standardising faithfulness benchmarks

The field lacks a unified faithfulness leaderboard. Different papers report different metric subsets (comprehensiveness, sufficiency, simulatability, paraphrase invariance, unlearning faithfulness) on different splits. The closest candidate for a unified benchmark is the combination of ERASER (extractive), ChainPoll (CoT), and the unlearning protocol of Tutek et al. (2025), but adoption remains uneven. A

Failure mode	Method affected	Evidence	Mitigation
Adversarial scaffold	LIME, SHAP	Slack et al. 2020: 95% scaffold-detection	Integrity check on perturbations
Lexical/position bias	IG, GradCAM	Kamp et al. 2025	De-biased gradients
Biasing-feature CoT	Chain-of-Thought	Turpin et al. 2023: 36% silent shift	Pair with internal probe
Paraphrase shift	CoT	Lanham et al. 2023	Self-consistency + invariance check
Unlearning unfaithfulness	CoT	Tutek et al. 2025: <40% causal	Symbolic CoT, e.g., Faithful CoT
Probe expressivity	Linear probes	Hewitt & Liang 2019	Selectivity, control tasks
Circular features	SAE	Engels et al. 2024	Multi-dim feature dictionaries
Reconstruction–behavior gap	SAE	Braun et al. 2024	End-to-end SAE training
Specificity collapse	ROME/MEMIT	Hoelscher-Obermaier et al. 2023	Better specificity benchmarks
Multi-hop ripple failure	ROME/MEMIT	Zhong et al. 2023: 21% 2-hop	Multi-stage edits, In-Context
Confident hallucination	SelfCheckGPT	Farquhar et al. 2024	Combine with retrieval
Domain transfer	Hallucination detectors	8–15 AUROC drop	Domain-specific training
COVID shortcut	Saliency	DeGrave et al. 2021	Causal probing, distribution shift
Localisation–editing dissociation	Causal tracing	Hase et al. 2023	Editing-aware localisation
Cognitive inaccessibility	Circuits	n/a	Layered explanations by stakeholder

community-supported leaderboard is necessary for the field to track real progress vs. method-specific tuning.

Forecast 2: by 2027 a community faithfulness leaderboard covering at least 5 method families (LIME/SHAP, ACDC, SAE-attribution, CoT-faithfulness, hallucination detection) will exist with regular submissions from at least 10 institutions. Falsified if such a leaderboard does not have $\geq$10 institution-submitted entries.

12.3. Faithful Chain-of-Thought beyond symbolic tasks

Faithful CoT works perfectly on symbolic tasks (Lyu et al., 2023: 95% on GSM8K) and fails on commonsense, dialogue, and creative tasks. The open problem is to develop faithfulness-preserving CoT that does not require an external solver. Promising candidate ideas include: (i) faithfulness-fine-tuning that explicitly trains the model to produce CoT chains whose ablation degrades the answer; (ii) SAE-grounded CoT, where each step's natural-language content is verified against an internal feature signature; (iii) decomposition-based CoT (Radhakrishnan et al., 2023) extended to non-mathematical tasks.

Forecast 3: by 2028 a public model will demonstrate >70% unlearning faithfulness (Tutek et al. 2025 metric) on a non-mathematical reasoning benchmark (e.g., StrategyQA), up from current ~38%. Falsified if no such demonstration exists.

12.4. Causal evaluation of SAE features

SAEs are currently evaluated mostly by reconstruction loss and feature interpretability, both of which are correlational with respect to behaviour. RAVEL (Huang et al., 2024) is an early attempt at causal evaluation but is limited in scope. The open problem is a benchmark that scores SAE features by their causal sufficiency for a behaviour: given a target behaviour, can the SAE feature set localise the cause and predict edits?

Forecast 4: by 2027 a "Causal SAE" benchmark will exist with reproducible scores for at least Gemma-2 2B/9B and Llama-3 8B SAEs, including separate reliability/portability metrics. Falsified if no such bench-

mark exists.

12.5. Cross-layer feature linking and circuit assembly

Per-layer SAEs do not share a basis across layers, so reconstructing a complete circuit requires matching features across layers — currently done by hand or by ad-hoc clustering. Cross-coder SAEs (Lindsey et al., 2024), transcoders (Dunefsky et al., 2024), and attribution SAEs are early efforts. The open problem: a method that produces a single feature dictionary shared across layers, with circuit-level edges learned during SAE training.

Forecast 5: by 2028 a "circuit-level SAE" will be demonstrated that produces a single feature dictionary across a transformer's layers and reconstructs a manually-validated circuit (IOI, Greater-Than) with F1 $\geq$0.85. Falsified if no such demonstration exists.

12.6. Frontier-model interpretability of agentic LLMs

LLMs increasingly serve as agents, calling tools and orchestrating multi-step workflows. The CoT chain becomes an action trace; the explanation must cover the choice of tool, the parsing of returns, and the integration into the next step. AIOps (Zhang et al., 2025) maps this for IT operations; the RL-meets-LLM survey (Liu et al., 2025) maps it for reasoning. The interpretability tools of Sections 4–7 do not naturally extend to agent traces, motivating new methods.

Forecast 6: by 2028 at least one published method will provide circuit-level interpretation of a multi-tool agent's decision pipeline on a public benchmark (e.g., AgentBench), recovering tool-choice mechanism with reproducibility >70%. Falsified otherwise.

12.7. Multimodal and multilingual mechanistic interpretability

Multimodal LLMs (LLaVA, Qwen-VL, GPT-4V, Gemini 1.5) and multilingual LLMs (Llama-3, Aya) are interpretively under-explored. The Dang et al. (2024) survey identifies vision-language as the most active multimodal interpretability area; the Resck et al. (2025) survey identifies multilingual as the principal language-extension area. The open problems include: cross-modal feature linking; concept-language separation across more than 24 languages; non-English Tracr-equivalent compiled-program test-beds.

Forecast 7: by 2027 a public SAE will be released specifically for the vision-language fusion layer of a multimodal model, with feature interpretability $\geq$60% as judged by an automated multimodal labeller. Falsified otherwise.

12.8. Regulation, traceability, and interpretable-by-construction LLMs

The EU AI Act (Regulation 2024/1689) classifies certain LLM applications as high-risk and requires traceability: a documented chain of evidence linking input to output. The mechanistic interpretability tools we have surveyed produce technical artefacts (circuits, SAE features) that are unsuitable as direct compliance evidence. The open problem is to develop audit-grade explanation pipelines: end-to-end systems that produce regulatorily acceptable reports from interpretability artefacts. Singh et al. (2024) argue that the era of LLM interpretability demands re-thinking what an explanation is for; regulation provides an unmissable forcing function.

Forecast 8: by 2028 at least three jurisdictions will have published technical guidance specifying interpretability requirements for high-risk LLM deployments, drawing on the methods surveyed in this paper. Falsified if such guidance does not exist.

12.9. Bottlenecks and likely solutions

12.10. Open problems summary table (mapping to Sharkey et al. 2025)

We map our forecasts onto the Open Problems in Mechanistic Interpretability taxonomy.

12.11. Methodological future: mixed black-/white-box auditing

A practical near-term direction is the deliberate combination of black-box and white-box methods. Black-box auditing (semantic entropy, SelfCheckGPT) provides real-time signals on closed APIs, while white-box methods (SAE features, circuit analysis) provide development-time understanding when the model is open-weight. Combining them — for instance, training a hallucination detector on internal SAE features and deploying it via a black-box wrapper — can give the best of both. Sun et al.'s (2026) adaptive Bayesian semantic entropy is an early step, and we expect substantial industrial adoption of hybrid explainability stacks in 2026–2028.

12.12. Architectural future: interpretability-by-design

Several architectures are designed to be more interpretable from the outset. Concept Bottleneck LLMs (Sun et al., 2024) constrain prediction to flow through a sparse human-defined concept layer; sparse-

Bottleneck	Current status	Likely 2026–2030 path
SAE training cost on frontier models	$1M for 34M features on Claude 3 Sonnet	Gated/TopK + cross-model transfer
CoT faithfulness for non-symbolic tasks	<40% on Llama-3-8B	Faithfulness-fine-tuning + SAE grounding
Multi-hop edit propagation	21% 2-hop on MQuAKE	Multi-stage editing + retrieval hybrid
Cross-layer feature alignment	Per-layer dictionaries do not share basis	Cross-coder / transcoder SAEs
Agentic-trace interpretation	No circuit-level methods	Tool-call attribution + agent SAE
Causal SAE evaluation	RAVEL only	New community benchmark
Multilingual concept transfer	English-only edits dominant	Concept-direction-shared editing
Multimodal feature linking	Patches and tokens decoupled	Joint vision-language SAE
Frontier-scale circuit discovery	Manual analysis at GPT-2 small	Attribution patching + SAE-circuit-graphs
Regulatory alignment	No technical guidance	EU AI Act → 2026 implementing acts

Sharkey et al. category	Open problems addressed	Our forecast
SAE training	Frontier scaling, cross-model transfer	Forecast 1
Evaluation	Causal SAE evaluation, faithfulness benchmarks	Forecasts 2, 4
Circuit discovery	Cross-layer linking, frontier-scale	Forecast 5
CoT and reasoning	Faithful CoT for non-symbolic	Forecast 3
Application	Agentic, regulatory	Forecasts 6, 8
Multimodal	Vision-language fusion SAEs	Forecast 7
Multilingual	Concept-direction transfer (Resck et al.)	Implicit

by-design models trained with explicit superposition penalties (Bricken et al., 2023) are an alternative; and modular architectures with task-specific routing (Mixture-of-Experts) provide implicit interpretability via expert assignment. Whether interpretability-by-design can match the performance of black-box training at scale remains open: the 1.4-point CB-LLM accuracy gap on AGNews is small but compounds across tasks.

Forecast 9: by 2030 a deployed commercial LLM will use interpretability-by-design as its primary architecture, with a published gap of ≤ 2 points on a major benchmark relative to a black-box counterpart. Falsified if no such deployment exists.

12.13. The role of community standards

Sharkey et al. (2025) argue that progress in mechanistic interpretability has been throttled by the absence of community standards: no agreed-upon SAE training protocol, no agreed-upon faithfulness metric, no agreed-upon circuit-discovery benchmark. The corresponding success metric is institutional: the field needs a small number of widely-adopted standards (analogous to BLEU and CIDEr in machine translation, GLUE and MMLU in LLM evaluation) that allow apples-to-apples comparison of methods. Without these, the field will continue producing isolated case studies rather than cumulative progress.

Forecast 10: by 2027 at least three of the following will be widely adopted as community standards: (a) a faithfulness leaderboard; (b) an SAE-evaluation benchmark; (c) a unified editing-evaluation harness; (d) a CoT-faithfulness metric; (e) a circuit-discovery benchmark. Falsified if fewer than three are adopted.

12.14. Critical synthesis: comparing method families

Across the five families introduced in Section 2, the 2026 evidence supports a clear comparative picture. Post-hoc attribution (LIME, KernelSHAP, Integrated Gradients) trades off model-agnosticism against faithfulness; the methods are cheap and universal but Slack et al. (2020) showed they are gameable. Self-explanation (Chain-of-Thought, e-SNLI rationales) trades off plausibility against faithfulness, and although CoT yields the most readable artefacts Tutek et al. (2025) reported that fewer than 40% of CoT

steps are causally necessary on Llama-3-8B. Mechanistic interpretability (ACDC, sparse autoencoders) trades off causal rigour against scalability, since activation patching gives causal evidence but does not yet scale to 70B+ frontier models. Concept-based architectures (CB-LLM, steering vectors) trade off interpretability against accuracy, with the CB-LLM gap of 1.4 points on AGNews small but cumulative across tasks. Black-box auditing (SelfCheckGPT, semantic entropy) trades off API compatibility against access to mechanism, detecting statistical hallucination signatures without localising the cause.

In summary, no single family dominates. Section 12 maps the consequent open problems and the most promising bridges.

12.15. Open problems for 2025–2026

The following problems remain unresolved as of mid-2026:

- Frontier-scale circuit discovery — no published circuit-level analysis exists for any model larger than GPT-2 medium (345M); Llama-3 70B and Claude 3 Opus remain unexplored.
- Cross-layer SAE feature alignment — per-layer SAEs do not share a basis, so circuit assembly across layers is currently manual.
- Causal SAE evaluation — RAVEL is the only public benchmark, and it does not cover frontier models.
- Faithful CoT for non-symbolic reasoning — current faithfulness on non-symbolic tasks is ~38% by the unlearning metric.
- Multi-hop edit propagation — ROME and MEMIT achieve only 21% accuracy on MQuAKE 2-hop queries.
- Multilingual edit transfer — English edits do not propagate to Hindi, Mandarin, or Arabic paraphrases.
- Multimodal cross-modal feature linking — vision-language SAE features are not yet aligned across modalities.
- Audit-grade explanation pipelines — no end-to-end system produces regulator-acceptable evidence under the EU AI Act 2024.

12.16. Future directions emerging in 2026

Several research directions have crystallised in the past twelve months and are likely to define near-term progress:

- Circuit graphs over SAE features — assembling a single feature dictionary across all layers and learning causally validated edges between features.
- Hybrid black-/white-box detection stacks — combining semantic entropy on the API surface with internal SAE-feature signatures inside the model.
- Faithfulness-fine-tuned CoT — explicitly training models to produce CoT chains whose ablation degrades the answer.
- Editing-aware localisation — reformulating causal tracing to optimise editing utility rather than residual-stream peak.
- Agentic interpretability — extending circuit-level analysis to multi-tool agent traces on benchmarks such as AgentBench.

12.17. Synthesis

The 2026–2030 horizon is one of consolidation and scaling. The methodological foundations established in 2022–2024 — induction heads, IOI circuit, ROME/MEMIT, sparse autoencoders, semantic entropy — must now be hardened into reliable engineering practice. The principal forcing functions are commercial deployment (where hallucination detection and safety auditing are immediate needs) and regulation (where traceability is increasingly mandatory). The principal scientific frontier is the cross-layer, cross-modal, cross-lingual extension of mechanistic methods that currently work primarily on monolingual English residual streams of mid-scale transformers. The emergent technique that will most likely define the next era is the circuit graph over SAE features, where every node is a labelled feature and every edge a causally validated connection; we anticipate the first frontier-scale demonstration of this artifact by 2028.

13. Conclusion: A Roadmap for Trustworthy LLM Explainability

This concluding section synthesises three established consensus propositions, the genuine open questions, and a stakeholder-segmented roadmap. The survey

has traced the explainability of large language models from the 2014 introduction of saliency maps and the 2016 release of LIME through the 2018 BERTology probing era, the 2022 mechanistic-interpretability turn, the 2023 sparse-autoencoder explosion, and the frontier of 2025–2026 multimodal and multilingual extensions. We organised the literature by a five-family taxonomy — local post-hoc explanations, self-explanations and rationales, global mechanistic interpretability, concept-based and inherently-interpretable architectures, and black-box auditing — and provided algorithmic depth on probing classifiers, activation and attribution patching, Automated Circuit Discovery, sparse autoencoders (vanilla, gated, TopK, end-to-end), Chain-of-Thought and its faithfulness measurement, the locate-then-edit paradigm including ROME, MEMIT, PMET and MELO, and the black-box detector family (SelfCheckGPT, semantic entropy, context-length probing).

13.1. What we now know

Three propositions are sufficiently well-established to count as consensus in the field as of 2026. First, transformers implement structured algorithms whose components are individually identifiable: the IOI circuit in GPT-2 small (Wang et al., 2022), the Greater-Than circuit (Hanna et al., 2023), the arithmetic-attention-head circuit in Llama-2-7B (Yu and Ananiadou, 2024), and the board-state circuit in Othello-GPT (Li et al., 2023; He et al., 2024) demonstrate that mechanistic interpretability is feasible at the small to mid scale. Second, features in the residual stream are largely linearly encoded but in superposition: sparse autoencoders recover thousands to millions of monosemantic features (Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024; Lieberum et al., 2024) with reconstruction–behaviour gaps small enough to support steering and circuit-level analysis. Third, Chain-of-Thought is plausibly readable but only weakly faithful: Turpin et al. (2023), Lanham et al. (2023), and Tutek et al. (2025) consistently find that CoT chains rationalise rather than reveal the model's computation in the absence of explicit faithfulness training or symbolic execution.

13.2. What remains hard

Equally important are the genuine open questions. Scaling mechanistic interpretability from GPT-2 small (117M) to Llama-3 405B (or Claude 3 Opus, of undisclosed but presumably 200B+ parameters) is unsolved. Cross-layer feature alignment is unsolved. CoT faithfulness on non-symbolic tasks is unsolved. Causal evaluation of SAE features beyond toy benchmarks is unsolved. Multi-hop edit propagation through ROME/MEMIT remains below 25% on MQuAKE. Multimodal and multilingual extensions are nascent. Hallucination detection on frontier-model long-form generation has saturated near 0.85 AUROC and improvement is increasingly hard. Each of these is the subject of an active research effort whose outcomes will define the 2026–2030 horizon.

13.3. A roadmap by stakeholder

For different stakeholders, different methods matter most.

For researchers, the most consequential directions are: (i) cross-layer-shared SAE dictionaries (cross-coders, transcoders); (ii) faithfulness-fine-tuned CoT; (iii) editing-aware localisation that closes the localisation-editing dissociation. The methodological frontier is unification: combining patching, SAEs, and CoT into a single circuit-graph-over-SAE-features artifact.

For engineers, the most consequential directions are: (i) hybrid black-/white-box detection stacks combining semantic entropy (API-friendly) with internal SAE feature signatures (model-internal); (ii) RAG-anchored CoT for high-stakes domain deployment; (iii) lifelong editing pipelines that maintain specificity across thousands of updates.

For regulators, the most consequential directions are: (i) audit-grade explanation pipelines that produce regulatorily acceptable evidence; (ii) traceability documentation drawn from causal interpretability tools; (iii) standardised faithfulness leaderboards that allow comparison of vendor claims.

For end users, the most consequential directions are: (i) plain-language rationale generation paired with a confidence indicator (semantic entropy); (ii) interactive explanation tools that allow drill-down from a high-level natural-language summary to a circuit-level mechanism; (iii) stakeholder-appropriate artefacts — circuit diagrams for researchers, citation lists for clinicians, rationale paragraphs for the general public.

13.4. Glossary of essential terms

13.5. Closing remarks

Explainability for large language models is no longer a niche pursuit. It is a load-bearing pillar of safe deployment. The methodological diversity ranges from the simplest gradient saliency to the most sophisticated sparse-autoencoder feature catalogues. This diversity reflects both the range of stakeholder needs and the genuine difficulty of the underlying problem. Fron-

Term	Definition
Faithfulness	An explanation accurately reflects the model's true causal computation
Plausibility	Humans find the explanation convincing
Probing classifier	Shallow classifier trained on frozen activations to test for encoded property
Activation patching	Replacing a model component's activation with a counterfactual one
ACDC	Automated Circuit Discovery via greedy edge ablation
Sparse autoencoder (SAE)	Overcomplete dictionary learning on activations with L1 sparsity
Superposition	Encoding more features than dimensions via overlapping directions
Polysemanticity	Single neuron activates for multiple unrelated concepts
Linear representation hypothesis	Features encoded as linear directions in activation space
Chain-of-Thought	Generated intermediate reasoning steps prior to final answer
Faithful CoT	CoT routed through deterministic external executor
Locate-then-edit	Identify weights for a fact via causal tracing, then update
ROME	Rank-One Model Editing of MLP keys/values
MEMIT	Mass-edit extension of ROME
Knowledge editing	Surgical modification of stored facts
Steering vector	Inference-time activation addition to control behaviour
Concept Bottleneck LLM	Architecture forcing prediction through human-defined concept layer
Hallucination	Model output that fabricates factual content
SelfCheckGPT	Sample-consistency-based hallucination detector
Semantic entropy	Hallucination detection via NLI-clustered sample entropy
Loss-recovered fraction	SAE evaluation metric: 1 − (L_with_SAE − L_clean) / (L_zero − L_clean)
L0 sparsity	Average count of non-zero SAE feature activations per token
Comprehensiveness / Sufficiency	Faithfulness metrics from ERASER
Simulatability	Degree to which an explanation enables predicting model output
Causal mediation	Identifying components mediating an effect via interventional analysis
Induction head	Attention head that copies a previous occurrence of the current token
Indirect Object Identification (IOI)	Canonical circuit task in GPT-2 small
RippleEdits	Benchmark for dependent-fact propagation after edits
MQuAKE	Multi-hop knowledge edit benchmark, 9k items
TruthfulQA	817-question benchmark of common-misconception elicitation
Gated SAE	SAE variant separating gating from magnitude to reduce shrinkage
TopK SAE	SAE variant using hard top-K sparsity instead of L1

tier models are not interpretable in any single sense; they are interpretable to varying degrees by varying methods at varying granularity for varying stakeholders. The pragmatic conclusion is one of pluralism: deploy a portfolio of explanation methods, evaluate each on its own faithfulness criterion, and never accept any single artifact as a complete account.

We hope that this survey provides a navigable map for newcomers, a useful synthesis for established researchers, and a starting point for engineers, regulators, and end users. The next half-decade will be shaped by the cross-pressure between scaling — making mechanistic methods work on frontier models — and consolidation — turning the current diversity of methods into reliable engineering practice. We have offered ten falsifiable forecasts so that this survey may be re-evaluated in 2030 against measurable outcomes.

The authors thank the many open-source contributors to TransformerLens, SAELens, Neuronpedia, EasyEdit, Inseq, captum, nnsight, and Pyvene for the infrastructure that made the empirical content of this survey possible. We thank the authors of Zhao et al. (2024), Bereska & Gavves (2024), Rai et al. (2024), Sharkey et al. (2025), and Resck et al. (2025) for the foundational surveys on which this synthesis builds. Errors and omissions remain our own.

14. References

[1] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu,

Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Mengnan Du. Explainability for Large Language Models: A Survey. ACM Transactions on Intelligent Systems and Technology, 2024. doi:10.1145/3639372.

[2] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context Learning and Induction Heads. arXiv:2209.11895, 2022.

[3] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. arXiv:2211.00593, 2022.

[4] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, Adrià Garriga-Alonso. Towards Automated Circuit Discovery for Mechanistic Interpretability. NeurIPS, 2023.

[5] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, Lee Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models. arXiv:2309.08600, 2023.

[6] Leonard Bereska, Efstratios Gavves. Mechanistic Interpretability for AI Safety – A Review. arXiv:2404.14082, 2024.

[7] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, et al. Open Problems in Mechanistic Interpretability. arXiv:2501.16496, 2025.

[8] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, Jacob Steinhardt. Progress Measures for Grokking via Mechanistic Interpretability. arXiv:2301.05217, 2023.

[9] Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, Jianfeng Gao. Rethinking Interpretability in the Era of Large Language Models. arXiv:2402.01761, 2024.

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS, 2022.

[11] Miles Turpin, Julian Michael, Ethan Perez, Samuel R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. NeurIPS, 2023.

[12] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, et al. Measuring Faithfulness in Chain-of-Thought Reasoning. arXiv:2307.13702, 2023.

[13] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, Chris Callison-Burch. Faithful Chain-of-Thought Reasoning. IJCNLP-AACL, 2023.

[14] Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, et al. Question Decomposition Improves the Faithfulness of Model-Generated Reasoning. arXiv:2307.11768, 2023.

[15] Ian Tenney, Dipanjan Das, Ellie Pavlick. BERT Rediscovers the Classical NLP Pipeline. ACL, 2019.

[16] Kevin Clark, Urvashi Khandelwal, Omer Levy, Christopher D. Manning. What Does BERT Look at? An Analysis of BERT's Attention. BlackboxNLP@ACL, 2019.

[17] Ganesh Jawahar, Benoît Sagot, Djamé Seddah. What Does BERT Learn about the Structure of Language? ACL, 2019.

[18] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. AIES, 2020.

[19] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. Information Fusion, 58:82–115, 2020.

[20] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv:2311.05232, 2023.

[21] Stephanie Lin, Jacob Hilton, Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. ACL, 2022.

[22] Potsawee Manakul, Adian Liusie, Mark J. F. Gales. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. EMNLP, 2023.

[23] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, Yarin Gal. Detecting Hallucinations in Large Language Models Using Semantic Entropy. Nature, 630:625–630, 2024.

[24] Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, Ningyu Zhang. Editing Large Language Models: Problems, Methods, and Opportunities. EMNLP, 2023.

[25] Peter Hase, Mohit Bansal, Been Kim, Asma

Ghandeharioun. Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models. NeurIPS, 2023.

[26] Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, Danqi Chen. MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions. EMNLP, 2023.

[27] Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, et al. A Comprehensive Study of Knowledge Editing for Large Language Models. arXiv:2401.01286, 2024.

[28] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, Richard Socher. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. ACL, 2019.

[29] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, Byron C. Wallace. ERASER: A Benchmark to Evaluate Rationalized NLP Models. ACL, 2020.

[30] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, Phil Blunsom. e-SNLI: Natural Language Inference with Natural Language Explanations. NeurIPS, 2018.

[31] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, Yejin Choi. Reframing Human-AI Collaboration for Generating Free-Text Explanations. NAACL, 2022.

[32] Sai Gurrapu, Ajay Kulkarni, Lifu Huang, Ismini Lourentzou, Feras A. Batarseh. Rationalization for Explainable NLP: a Survey. Frontiers in Artificial Intelligence, 2023.

[33] Siwen Luo, Hamish Ivison, Soyeon Caren Han, Josiah Poon. Local Interpretations for Explainable Natural Language Processing: A Survey. ACM Computing Surveys, 2024.

[34] Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, Ziyu Yao. A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models. arXiv:2407.02646, 2024.

[35] Haoyan Luo, Lucia Specia. From Understanding to Utilization: A Survey on Explainability for Large Language Models. arXiv:2401.12874, 2024.

[36] Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, Neel Nanda. Improving Dictionary Learning with Gated Sparse Autoencoders. arXiv:2404.16014, 2024.

[37] Connor Kissane, Robert Krzyzanowski, Joseph Isaac Bloom, Arthur Conmy, Neel Nanda. Interpreting Attention Layer Outputs with Sparse Autoencoders. arXiv:2406.17759, 2024.

[38] Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Bo Tang, Feiyu Xiong, Zhiyu Li. Attention Heads of Large Language Models: A Survey. arXiv:2409.03752, 2024.

[39] Dong Shu, Xuansheng Wu, Haiyan Zhao, Ninghao Liu, Mengnan Du. A Survey on Sparse Autoencoders: Interpreting the Internal Mechanisms of Large Language Models. EMNLP Findings, 2025.

[40] Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, et al. The Quest for the Right Mediator: Surveying Mechanistic Interpretability Through the Lens of Causal Mediation Analysis. arXiv:2408.01416, 2024.

[41] Stefan Heimersheim, Neel Nanda. How to Use and Interpret Activation Patching. arXiv:2404.15255, 2024.

[42] Fred Zhang, Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. arXiv:2309.16042, 2023.

[43] Jack Merullo, Carsten Eickhoff, Ellie Pavlick. Circuit Component Reuse Across Tasks in Transformer Language Models. arXiv:2310.08744, 2023.

[44] Ondřej Cífka, Antoine Liutkus. Black-Box Language Model Explanation by Context Length Probing. arXiv:2212.14815, 2022.

[45] Naomi Saphra, Sarah Wiegreffe. Mechanistic? BlackboxNLP, 2024.

[46] Xiaochuang Han, Byron Wallace, Yulia Tsvetkov. Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions. ACL, 2020.

[47] Martin Tutek, Jan Šnajder. Toward Practical Usage of the Attention Mechanism as a Tool for Interpretability. IEEE Access, 2022.

[48] Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, Himabindu Lakkaraju. On the Hardness of Faithful Chain-of-Thought Reasoning in Large Language Models. arXiv:2406.10625, 2024.

[49] Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, Jan Šnajder. Measuring Chain of Thought Faithfulness by Unlearning Reasoning Steps. arXiv:2502.14829, 2025.

[50] Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, Arthur Conmy. Chain-of-Thought Reasoning In The Wild Is Not Always Faithful. arXiv:2503.08679, 2025.

[51] Lucas Resck, Isabelle Augenstein, Anna Korhonen. Explainability and Interpretability of Multilingual Large Language Models: A Survey. EMNLP, 2025.

[52] Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, et al. Explainable and Interpretable Multimodal Large Language Models: A Comprehensive Survey. arXiv:2412.02104, 2024.

[53] Johannes Schneider. Explainable Generative AI (GenXAI): A Survey, Conceptualization, and Research Agenda. Artificial Intelligence Review, 2024.

[54] Chung-En Sun, Tuomas Oikarinen, Tsui-Wei Weng. Crafting Large Language Models for Enhanced Interpretability. arXiv:2407.04307, 2024.

[55] Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, Dong Yu. From Language Modeling to Instruction Following: Understanding the Behavior Shift in LLMs after Instruction Tuning. NAACL, 2024.

[56] Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, Max Tegmark. Not All Language Model Features Are One-Dimensionally Linear. arXiv:2405.14860, 2024.

[57] Zhengfu He, Xuyang Ge, Qiong Tang, Tianxiang Sun, Qinyuan Cheng, Xipeng Qiu. Dictionary Learning Improves Patch-Free Circuit Discovery in Mechanistic Interpretability: A Case Study on Othello-GPT. arXiv:2402.12201, 2024.

[58] Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, Lee Sharkey. Identifying Functionally Important Features with End-to-End Sparse Dictionary Learning. arXiv:2405.12241, 2024.

[59] Aaditya K. Singh, Ted Moskovitz, Felix Hill, Stephanie Chan, Andrew M. Saxe. What Needs to Go Right for an Induction Head? A Mechanistic Study of In-Context Learning Circuits and Their Formation. arXiv:2404.07129, 2024.

[60] Hannes Thurnherr, Jérémy Scheurer. TracrBench: Generating Interpretability Testbeds with Large Language Models. arXiv:2409.13714, 2024.

[61] Alex J. DeGrave, Joseph D. Janizek, Su-In Lee. AI for Radiographic COVID-19 Detection Selects Shortcuts over Signal. Nature Machine Intelligence, 2021.

[62] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, et al. A Comprehensive Overview of Large Language Models. arXiv:2307.06435, 2023.

[63] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, Robert McHardy. Challenges and Applications of Large Language Models. arXiv:2307.10169, 2023.

[64] Venkata Abhinandan Kancharla. Applied Explainability for Large Language Models: A Comparative Study. arXiv:2604.15371, 2026.

[65] Satvik Golechha, James Dao. Challenges in Mechanistically Interpreting Model Representations. arXiv:2402.03855, 2024.

[66] Zeming Wei, Yue Wang, Ang Li, Yichuan Mo, Yisen Wang. Jailbreak and Guard Aligned Language Models with Only Few In-Context Demonstrations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2026.

[67] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. Learning and Individual Differences, 2023.

[68] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, et al. Large Language Models Encode Clinical Knowledge. Nature, 620, 2023.

[69] Ashkan Golgoon, Khashayar Filom, Arjun Ravi Kannan. Mechanistic Interpretability of Large Language Models with Applications to the Financial Services Industry. ICAIF, 2024.

[70] Aaquib Syed, Can Rager, Arthur Conmy. Attribution Patching Outperforms Automated Circuit Discovery. BlackboxNLP, 2024.

[71] Kedi Chen, Qin Chen, Jie Zhou, He Chen, et al. DiaHalu: A Dialogue-Level Hallucination Evaluation Benchmark for Large Language Models. arXiv:2403.00896, 2024.

[72] Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, Mor Geva. Evaluating the Ripple Effects of Knowledge Editing in Language Models. Transactions of the Association for Computational Linguistics, 2024.

[73] Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, Jie Yu. PMET: Precise Model Editing in a Transformer. AAAI, 2024.

[74] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, et al. Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy

Artificial Intelligence. Information Fusion, 2023.

[75] Luca Longo, Mario Brčić, Federico Cabitza, et al. Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions. Information Fusion, 2024.

[76] Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, et al. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cognitive Computation, 2023.

[77] Erico Tjoa, Cuntai Guan. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 2020.

[78] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys, 2018.

[79] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, Salvatore Rinzivillo. Benchmarking and Survey of Explanation Methods for Black Box Models. Data Mining and Knowledge Discovery, 2023.

[80] Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971, 2023.

[81] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, et al. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. arXiv:2303.12712, 2023.

[82] Francesco D'Angelo, Francesco Croce, Nicolas Flammarion. Selective Induction Heads: How Transformers Select Causal Structures In Context. arXiv:2509.08184, 2025.

[83] Shuxun Wang, Qingyu Yin, Chak Tou Leong, et al. Induction Head Toxicity Mechanistically Explains Repetition Curse in Large Language Models. arXiv:2505.13514, 2025.

[84] Vinit Ravishankar, Memduh Gökırmak, Lilja Øvrelid, Erik Velldal. Multilingual Probing of Deep Pre-Trained Contextual Encoders. DSpace University of Tartu, 2019.

[85] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. EMNLP, 2020.

[86] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, et al. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. IJCNLP, 2023.

[87] Jiahao Yu, Xingwei Lin, Zheng Yu, et al. GPTFUZZER: Red Teaming Large Language Models with Auto-Generated Jailbreak Prompts. arXiv:2309.10253, 2023.

[88] Ansh Sheshadri, Aidan Ewart, Phillip Guo, et al. Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs. arXiv:2407.15549, 2024.

[89] Jingyu Zhang, Ahmed Elgohary, Xiawei Wang, et al. Jailbreak Distillation: Renewable Safety Benchmarking. arXiv:2505.22037, 2025.

[90] Jiaming Ji, Tianyi Qiu, Boyuan Chen, et al. AI Alignment: A Comprehensive Survey. arXiv:2310.19852, 2023.

[91] Yao Fu, Litu Ou, Mingyu Chen, et al. Chain-of-Thought Hub: A Continuous Effort to Measure Large Language Models' Reasoning Performance. arXiv:2305.17306, 2023.

[92] Liangming Pan, Alon Albalak, Xinyi Wang, et al. Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning. EMNLP Findings, 2023.

[93] Jundong Xu, Hao Fei, Liangming Pan, et al. Faithful Logical Reasoning via Symbolic Chain-of-Thought. ACL, 2024.

[94] Wei Jie Yeo, Ranjan Satapathy, Rick Siow Mong Goh, et al. How Interpretable are Reasoning Explanations from Prompting Large Language Models? NAACL Findings, 2024.

[95] Lijie Hu, Liang Liu, Shu Yang, et al. Understanding Reasoning in Chain-of-Thought from the Hopfieldian View. arXiv:2410.03595, 2024.

[96] Zeping Yu, Sophia Ananiadou. Interpreting Arithmetic Mechanism in Large Language Models through Comparative Neuron Analysis. arXiv:2409.14144, 2024.

[97] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, et al. Exploring the Frontier of Vision-Language Models: A Survey of Current Methodologies and Future Directions. arXiv:2404.07214, 2024.

[98] Kyle Mahowald, Anna A. Ivanova, Idan Blank, et al. Dissociating Language and Thought in Large Language Models. Trends in Cognitive Sciences, 2024.

[99] Wen Luo, Tianshu Shen, Wei Li, et al. HalluDial: A Large-Scale Benchmark for Automatic Dialogue-

Level Hallucination Evaluation. arXiv:2406.07070, 2024.

[100] Hanzhi Zhang, Sumera Anjum, Heng Fan, et al. Poly-FEVER: A Multilingual Fact Verification Benchmark for Hallucination Detection in Large Language Models. arXiv:2503.16541, 2025.

[101] Mengfei Liang, Archish Arun, Zekun Wu, et al. THaMES: An End-to-End Tool for Hallucination Mitigation and Evaluation in Large Language Models. arXiv:2409.11353, 2024.

[102] Minda Hu, Bowei He, Yufei Wang, et al. Mitigating Large Language Model Hallucination with Faithful Finetuning. arXiv:2406.11267, 2024.

[103] Sriram Balasubramanian, Samyadeep Basu, Koustava Goswami, et al. Decomposition-Enhanced Training for Post-Hoc Attributions In Language Models. arXiv:2510.25766, 2025.

[104] Qiyao Sun, Xingming Li, Xixiang He, et al. Efficient Hallucination Detection: Adaptive Bayesian Estimation of Semantic Entropy with Guided Semantic Exploration. arXiv:2603.22812, 2026.

[105] Rodion Oblovatny, Alexandra Kuleshova, Konstantin Polev, et al. Probabilistic Distances-Based Hallucination Detection in LLMs with RAG. arXiv:2506.09886, 2025.

[106] Jiawei Chen, Hongyu Lin, Xianpei Han, et al. Benchmarking Large Language Models in Retrieval-Augmented Generation. AAAI, 2024.

[107] Eric Mitchell, Charles Lin, Antoine Bosselut, et al. Fast Model Editing at Scale. ICLR, 2022.

[108] Eric Mitchell, Charles Lin, Antoine Bosselut, et al. Memory-Based Model Editing at Scale. ICML, 2022.

[109] Lang Yu, Qin Chen, Jie Zhou, et al. MELO: Enhancing Model Editing with Neuron-Indexed Dynamic LoRA. AAAI, 2024.

[110] Zilu Dong, Xiangqing Shen, Rui Xia. MEMIT-Merge: Addressing MEMIT's Key-Value Conflicts in Same-Subject Batch Editing for LLMs. arXiv:2502.07322, 2025.

[111] Zhuoran Zhang, Yongxiang Li, Zijian Kan, et al. Locate-then-Edit for Multi-hop Factual Recall under Knowledge Editing. arXiv:2410.06331, 2024.

[112] Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, et al. Detecting Edit Failures in Large Language Models: An Improved Specificity Benchmark. ACL Findings, 2023.

[113] Bertram Højer, Oliver Jarvis, Stefan Heinrich. Improving Reasoning Performance in Large Language Models via Representation Engineering. arXiv:2504.19483, 2025.

[114] Gracjan Góral, Marysia Winkels, Steven Basart. Depth-Wise Activation Steering for Honest Language Models. arXiv:2512.07667, 2025.

[115] Jacob Dunefsky, Arman Cohan. One-Shot Optimized Steering Vectors Mediate Safety-Relevant Behaviors in LLMs. arXiv:2502.18862, 2025.

[116] Karim Saraipour, Shichang Zhang. From Indirect Object Identification to Syllogisms: Exploring Binary Mechanisms in Transformer Circuits. arXiv:2508.16109, 2025.

[117] Rabin Adhikari. Emergence of Minimal Circuits for Indirect Object Identification in Attention-Only Transformers. arXiv:2510.25013, 2025.

[118] Danielle Ensign, Adrià Garriga-Alonso. Investigating the Indirect Object Identification Circuit in Mamba. arXiv:2407.14008, 2024.

[119] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948, 2025.

[120] Qiguang Chen, Libo Qin, Jinhao Liu, et al. Towards Reasoning Era: A Survey of Long Chain-of-Thought for Reasoning Large Language Models. arXiv:2503.09567, 2025.

[121] Jiang Zhang, H. Bu, Hui Wen, et al. When LLMs Meet Cybersecurity: A Systematic Literature Review. Cybersecurity, 2025.

[122] Lingzhe Zhang, Tong Jia, Mengxi Jia, et al. A Survey of AIOps in the Era of Large Language Models. arXiv:2507.12472, 2025.

[123] Keliang Liu, Dingkang Yang, Ziyun Qian, et al. Reinforcement Learning Meets Large Language Models: A Survey of Advancements and Applications Across the LLM Lifecycle. arXiv:2509.16679, 2025.

[124] Ana Marasović, Iz Beltagy, Doug Downey, et al. Few-Shot Self-Rationalization with Natural Language Prompts. NAACL Findings, 2022.

[125] Peter Hase, Shiyue Zhang, Harry Xie, Mohit Bansal. Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language? EMNLP Findings, 2020.

[126] Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, Smaranda Muresan. FLUTE: Figurative Language Understanding through Textual Explanations.

EMNLP, 2022.

[127] Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, et al. e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks. ICCV, 2021.

[128] Virginie Do, Oana-Maria Camburu, Zeynep Akata, et al. e-SNLI-VE: Corrected Visual-Textual Entailment with Natural Language Explanations. arXiv:2004.03744, 2020.

[129] Sonia Joseph, Praneet Suresh, Lorenz Hufe, et al. Prisma: An Open Source Toolkit for Mechanistic Interpretability in Vision and Video. arXiv:2504.19475, 2025.

[130] Nooshin Bahador. Mechanistic Interpretability of Fine-Tuned Vision Transformers on Distorted Images: Decoding Attention Head Behavior for Transparent and Trustworthy AI. arXiv:2503.18762, 2025.

[131] Nina Żukowska, Wolfgang Stammer, Bernt Schiele, et al. Seeing Through Circuits: Faithful Mechanistic Interpretability for Vision Transformers. arXiv:2604.14477, 2026.

[132] Ondřej Cífka, Antoine Liutkus. Black-Box Language Model Explanation by Context Length Probing. arXiv:2212.14815, 2022.

[133] Xinyue Hu, Lin Gu, Kazuma Kobayashi, et al. Interpretable Medical Image Visual Question Answering via Multi-Modal Relationship Graph Learning. Medical Image Analysis, 2024.

[134] Sukyo Lee, Sumin Jung, Jong-Hak Park, et al. Development of BERT-Based Large Language Models for Emergency Department Triage Using Real-World Conversations. JAMIA, 2026.

[135] Sujung Lee, Won Ik Cho, Youngrong Lee, et al. A Prompt Framework for Enhancing LLM-Based Explainability of Medical Machine Learning Models: An Intensive Care Unit Application. BMC Medical Informatics and Decision Making, 2025.

[136] Maissa Trabilsy, Srinivasagam Prabha, Cesar A. Gomez-Cabello, et al. The PIEE Cycle: A Structured Framework for Red Teaming Large Language Models in Clinical Decision-Making. Bioengineering, 2025.

[137] Wei Guo, Jing Li, Wenya Wang, et al. MTSA: Multi-Turn Safety Alignment for LLMs through Multi-Round Red-Teaming. ACL, 2025.

[138] Riccardo Cantini, Alessio Orsino, Massimo Ruggiero, et al. Benchmarking Adversarial Robustness to Bias Elicitation in Large Language Models: Scalable Automated Assessment with LLM-as-a-Judge. arXiv:2504.07887, 2025.

[139] Chetan Pathade. Red Teaming the Mind of the Machine: A Systematic Evaluation of Prompt Injection and Jailbreak Vulnerabilities in LLMs. arXiv:2505.04806, 2025.

[140] Yair Bannett, Malvika Pillai, Tracy X. Huang, et al. Evaluating Large Language Models for Transparent Quality-of-Care Measurement in Children with ADHD. medRxiv, 2026.

[141] Jonathan Kamp, Roos Bakker, Dominique Blok. Explanation Bias is a Product: Revealing the Hidden Lexical and Position Preferences in Post-Hoc Feature Attribution. arXiv:2512.11108, 2025.

[142] Sam Chauhan, Estelle Duguet, Karthik Ramakrishnan, et al. SHLIME: Foiling Adversarial Attacks Fooling SHAP and LIME. arXiv:2508.11053, 2025.

[143] Kamalasankari Subramaniakuppusamy, Jugal Gajjar. Feature Attribution Stability Suite: How Stable Are Post-Hoc Attributions? arXiv:2604.02532, 2026.

[144] Clément Dumas, Chris Wendler, Veniamin Veselovsky, et al. Separating Tongue from Thought: Activation Patching Reveals Language-Agnostic Concept Representations in Transformers. arXiv:2411.08745, 2024.

[145] Qinyuan Ye, Robin Jia, Xiang Ren. Function Induction and Task Generalization: An Interpretability Study with Off-by-One Addition. arXiv:2507.09875, 2025.

[146] Ling Yang, Zhaochen Yu, Tianjun Zhang, et al. Buffer of Thoughts: Thought-Augmented Reasoning with Large Language Models. arXiv:2406.04271, 2024.

[147] Shidong Cao, Hongzhan Lin, Yuxuan Gu, et al. DiffCoT: Diffusion-Styled Chain-of-Thought Reasoning in LLMs. arXiv:2601.03559, 2026.

[148] Peifeng Wang, Zhengyang Wang, Zheng Li, et al. SCOTT: Self-Consistent Chain-of-Thought Distillation. ACL, 2023.

[149] Hangfeng He, Hongming Zhang, Dan Roth. Rethinking with Retrieval: Faithful Large Language Model Inference. arXiv:2301.00303, 2022.

[150] Thi Thanh Huong Nguyen, Linhao Luo, Fatemeh Shiri, et al. Direct Evaluation of Chain-of-Thought in Multi-hop Reasoning with Knowledge Graphs. ACL Findings, 2024.