Explainable NLP: A Comprehensive Survey and Practical Guidelines for Interpretable Text Models

Hadi Mohammadi[1,*] Tina Shahedi[1,*]

[1]Department of Methodology and Statistics, Utrecht University, The Netherlands

[*]These authors contributed equally to this work

Corresponding author: h.mohammadi@uu.nl

January 23, 2026

Abstract

Explainable Artificial Intelligence (XAI) has become essential as natural language processing (NLP) models grow increasingly complex and are deployed in high-stakes domains. This survey provides a comprehensive overview of explainability methods for NLP, spanning from classical machine learning approaches to the latest developments in Large Language Model (LLM) interpretability.

We make three main contributions: (1) a **unified taxonomy** that categorizes explanation methods along multiple dimensions including locality (local vs. global), directionality (forward vs. backward), and model access requirements (model-agnostic vs. model-specific); (2) **practical guidelines** with decision frameworks to help practitioners select appropriate explainability methods based on their specific tasks, models, and requirements; and (3) coverage of **LLM-era developments** including Chain-of-Thought prompting as explanation, mechanistic interpretability, and self-explanation capabilities.

We systematically review intrinsic and extrinsic evaluation methods, discuss current benchmarks, and survey applications across healthcare, legal, social science, and content moderation domains. We identify key challenges including faithfulness verification, the attention-explanation debate, and scaling interpretability to billion-parameter models. Finally, we outline promising future directions including human-AI collaborative explanation systems and responsible AI considerations.

This survey serves as both a comprehensive reference for researchers and a practical guide for practitioners seeking to make NLP systems more transparent and trustworthy.

1 Introduction

Artificial Intelligence (AI) is experiencing unprecedented growth, with predictions suggesting that by 2050, intelligent machines will match human intellectual capabilities across a wide range of tasks [45]. Organizations increasingly deploy AI technologies, with research indicating that 57% of enterprises have deployed or piloted AI systems, though only 10% achieve substantial financial advantages—primarily those that develop effective human-machine collaboration [53].

The success of these enterprises is attributed to their ability to "learn with AI"—creating systems where humans learn from AI and AI learns from humans. This bilateral learning paradigm is enabled through *explainability*: the capacity of AI systems to communicate their reasoning in terms that humans can understand, verify, and act upon.

1.1 The Need for Explainability

Traditional AI algorithms often allowed straightforward interpretation of their decisions. However, modern deep learning systems—particularly in Natural Language Processing (NLP)—employ architectures of such complexity that even simple neural networks can be challenging to interpret [2]. This opacity becomes particularly concerning in high-stakes domains where human lives and livelihoods depend on system outputs: healthcare diagnosis, legal decision-making, financial services, and autonomous systems [56].

Explainable AI (XAI) addresses this challenge by developing techniques that produce more explicable models while maintaining high performance levels. Through XAI, stakeholders can understand how systems arrive at specific decisions, trust AI predictions through the transparency that explanations provide, and verify that model reasoning aligns with intended behavior. When systems produce unexpected outputs, explanations enable effective debugging by revealing the factors that influenced problematic predictions. Furthermore, as regulatory frameworks increasingly demand algorithmic accountability, XAI provides the transparency mechanisms necessary for compliance with emerging legal requirements.

The benefits of explanations flow bidirectionally. For humans, explanations support decision-making [35], enable learning complex skills from AI [36], and calibrate appropriate trust in AI systems [27]. For AI systems, explanations facilitate debugging [22], model verification [11], and iterative improvement through human feedback [67].

1.2 Scope and Contributions

This survey focuses on explainability for Natural Language Processing (NLP)—the AI subfield developing algorithms to process, understand, and generate human language. While XAI spans computer vision, robotics, and other domains, NLP presents unique challenges due to the discrete, sequential, and context-dependent nature of text data.

We make three main contributions:

1. **Comprehensive Taxonomy**: We present a unified framework categorizing XAI methods for NLP from classical approaches (LIME, SHAP) through attention-based methods to cutting-edge LLM-era techniques (Chain-of-Thought, mechanistic interpretability). Our taxonomy spans multiple dimensions including locality, directionality, and model access requirements.

2. **Practical Guidelines**: We provide actionable decision frameworks for practitioners selecting explainability methods. This includes method selection decision trees, task-specific recommendations (classification, sequence labeling, question answering, generation), and evaluation protocol guidance.

3. **2024 State-of-the-Art**: We cover recent developments in LLM interpretability including Chain-of-Thought prompting as explanation, self-explanation capabilities, mechanistic interpretability research, and the ongoing debate about attention as explanation.

1.3 Paper Organization

The remainder of this paper is organized as follows. Section 2 provides background on XAI fundamentals and the evolution of NLP models, establishing the conceptual framework for understanding explanation methods. Section 3 presents our comprehensive taxonomy of explanation methods, categorizing approaches by scope, technique, and applicability. Section 4 discusses evaluation frameworks, metrics, and benchmarks for assessing explanation quality. Section 5 offers practical guidelines for method selection, including decision frameworks tailored to different use cases. Section 6 examines applications of explainability across domains including healthcare, legal, and financial services. Section 7 addresses current challenges and future research directions, and Section 8 concludes with a summary of key findings and recommendations.

1.4 Related Surveys

Several surveys have addressed explainability in machine learning and NLP. We position our work relative to these contributions.

General XAI surveys: Arrieta *et al.* [4] provided a comprehensive taxonomy of XAI methods across domains. Adadi and Berrada [2] surveyed the XAI landscape with focus on practical applications. These works cover explainability broadly but do not focus specifically on NLP challenges.

NLP-specific surveys: Danilevsky *et al.* [14] surveyed explainability for NLP with emphasis on neural models. Madsen *et al.* [44] reviewed post-hoc interpretability for transformer-based models. More recently, Zhao *et al.* [81] surveyed explainability for large language models, addressing prompting-based explanations and emergent capabilities. These surveys provide valuable foundations, though our work extends them with practical guidelines.

Specialized topics: Recent work has addressed specific aspects including attention interpretability [75], rationale extraction [15], and evaluation frameworks [26]. Singh *et al.* [61] argued that the LLM era fundamentally changes interpretability requirements, shifting focus from understanding individual predictions to understanding model capabilities at scale.

This survey's contributions: We extend prior work in three ways: (1) comprehensive coverage spanning classical methods to LLM-era techniques including Chain-of-Thought and mechanistic interpretability; (2) practical guidelines with decision frameworks for method selection; and (3) evaluation across multiple NLP tasks with recent benchmarks and tools.

2 Background

This section provides foundational concepts in Explainable AI (XAI) and Natural Language Processing (NLP), establishing the terminology and frameworks used throughout this survey.

2.1 Explainable Artificial Intelligence Fundamentals

Figure 1 illustrates the fundamental distinction between traditional AI systems and Explainable AI (XAI), highlighting how XAI extends conventional AI by providing interpretable outputs alongside predictions.

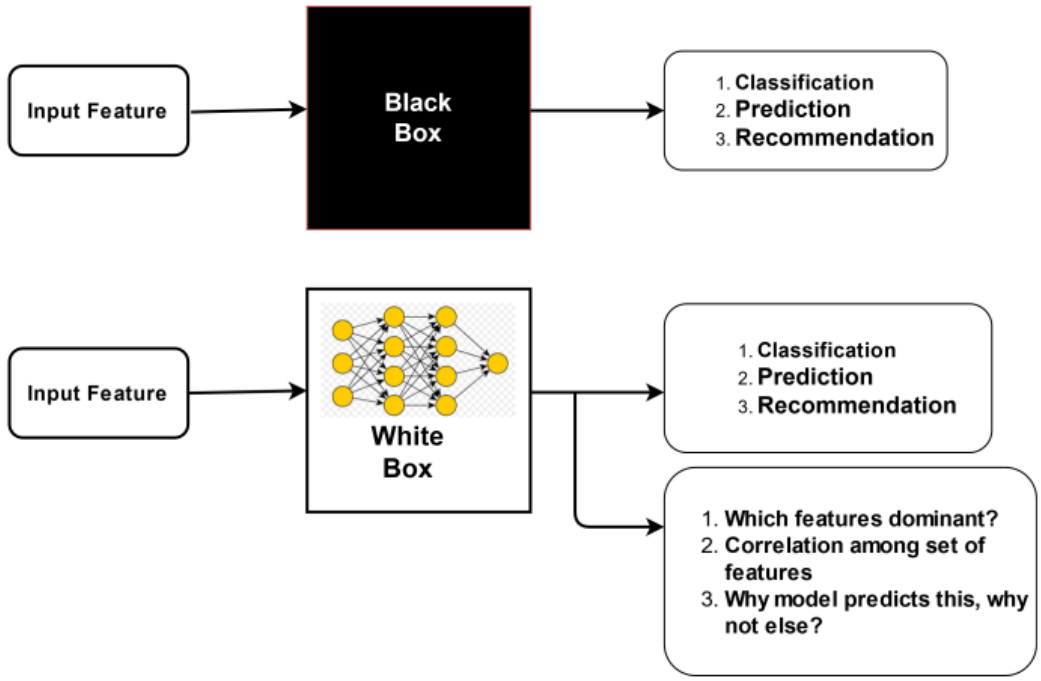

Figure 1: Comparison between traditional AI and Explainable AI (XAI). Traditional AI systems produce predictions without transparency, while XAI systems provide both predictions and human-understandable explanations of the decision-making process [21].

The terms *interpretability* and *explainability* are often used interchangeably in the literature, though subtle distinctions exist [14]. Interpretability refers to the degree to which humans can understand the cause of a decision, representing the capability of discerning the mechanics of a system without necessarily understanding the underlying reasons. Explainability, in contrast, refers to the degree to which the internal mechanics of a system can be explained in human terms, representing the capability of thoroughly explaining a phenomenon. In practice, we use these terms interchangeably while acknowledging that interpretability often emphasizes the model's inherent transparency, while explainability emphasizes post-hoc understanding.

Explainability approaches may be categorized by when explanation occurs relative to model construction [4]. Direct interpretability, also termed intrinsic interpretability, encompasses models designed to be inherently understandable, often called "white-box" models such as linear regression, decision trees, and rule-based systems. These are also termed *transparent* [2], *inherently interpretable* [56], or *self-explaining* [14]. Post-hoc explanation encompasses techniques applied after model training to explain predictions, and such methods can explain any model ("black-box") through companion interpretation systems [6, 65]. While post-hoc methods can be applied to directly interpretable models, this is typically unnecessary since inherent explanations are easier to obtain and guaranteed to be faithful.

Figure 2 provides a comprehensive overview of different explainability approaches, categorizing methods by their timing (direct vs. post-hoc), scope (local vs. global), and applicability (model-specific vs. model-agnostic).

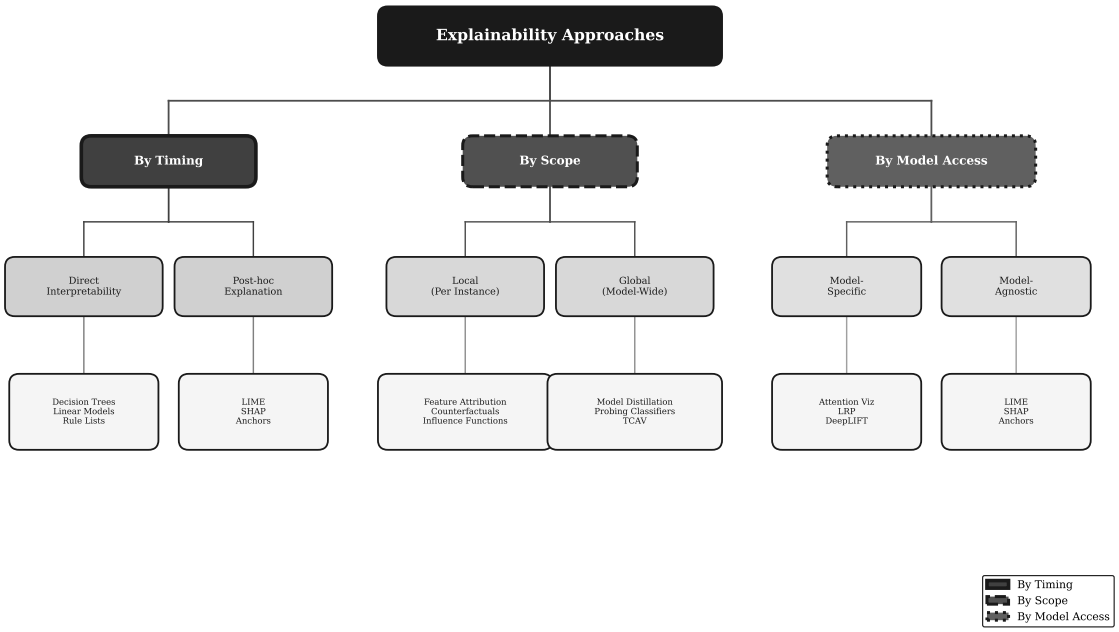

Figure 2: Overview of explainability approaches in machine learning. Methods are categorized by when explanations are generated (direct interpretability vs. post-hoc), the scope of explanations (local vs. global), and whether they are designed for specific model architectures or applicable to any model [4].

Explanation methods can be categorized along two key dimensions: scope and model access [42, 54]. Regarding scope, local explanations explain individual predictions without attempting to characterize the entire model, answering "Why did the model make this specific prediction?", while global explanations explain the model's overall behavior across all possible inputs, answering "How does the model generally work?" Local explanations can be aggregated to approximate global understanding, but global methods cannot generally be applied at the local level. Regarding model access, model-agnostic methods such as LIME [54] and SHAP [41] can explain any model by treating it as a black box without requiring access to internal parameters, while model-specific methods leverage internal model structure—examples include DeepLIFT [59] and Layer-wise Relevance Propagation (LRP) [5] for neural networks, and attention visualization for transformers. The choice among these dimensions depends on whether stakeholders need to understand specific decisions or general patterns, and whether internal model access is available.

A fundamental challenge in machine learning is the trade-off between model accuracy and interpretability, illustrated in Figure 3. Simpler models such as linear regression and decision trees are inherently interpretable but may have limited predictive power, while complex models like deep neural networks and ensemble methods achieve higher accuracy at the cost of transparency. This trade-off motivates much of the XAI research: how can we explain complex, high-performing models without sacrificing their predictive capabilities?

Local explanation methods can be further categorized by the form of explanation produced,

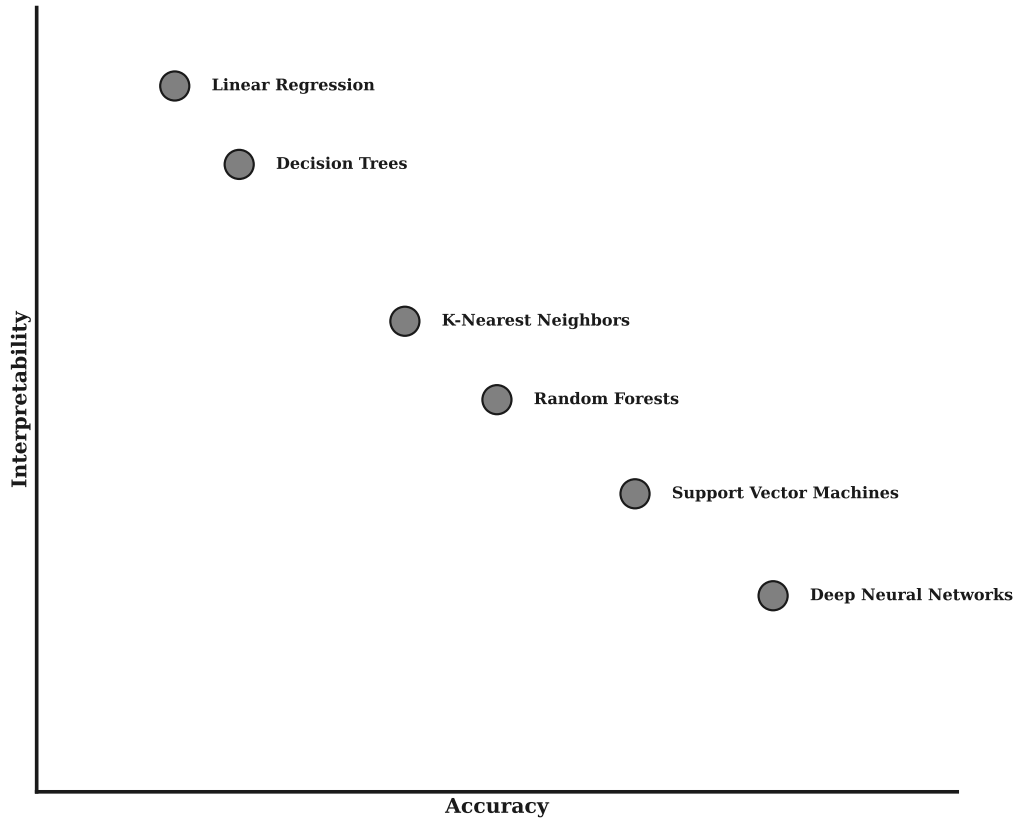

Figure 3: The accuracy-interpretability trade-off in machine learning models. Simpler models like linear regression and decision trees offer high interpretability but potentially lower accuracy, while complex models like deep neural networks and ensemble methods achieve higher accuracy at the cost of reduced interpretability. Adapted from [30].

as summarized in Table 1. Input-based methods identify parts of the input essential for the prediction through feature importance or attribution scores. Counterfactual methods point out input changes that would alter the prediction, providing actionable insights. Example-based methods select influential training examples to explain a prediction. Rule-based methods provide decision rules approximating the prediction process. Textual methods verbalize explanations in natural language, making them accessible to non-technical audiences.

The goals of XAI extend beyond mere model transparency, encompassing multiple objectives that benefit different stakeholders. These goals include building user trust in AI systems, ensuring regulatory compliance with transparency requirements, facilitating model debugging and improvement, enabling knowledge discovery from model insights, and supporting effective human-AI collaboration. Understanding these goals is essential for selecting appropriate explanation methods for specific applications.

2.2 Natural Language Processing Evolution

The evolution of NLP has progressed through distinct paradigms, each presenting different interpretability challenges and opportunities. Traditional NLP relied on feature engineering combined with linear classifiers such as logistic regression and SVMs, or tree-based methods [25, 77]. These models offered direct interpretability through feature weights or decision paths,

Table 1: Categorization of local explanation methods

Method Type	**Definition**
Input-based	Identifying parts of the input essential for the prediction (feature importance/attribution)
Counterfactual	Pointing out input changes that would alter the prediction
Example-based	Selecting influential training examples to explain a prediction
Rule-based	Providing decision rules approximating the prediction process
Textual	Verbalizing explanations in natural language

though their performance on complex tasks was limited. The interpretability of these classical approaches came naturally from their mathematical structure, but this transparency came at the cost of expressive power.

The transition to neural approaches—beginning with word embeddings [47] and recurrent neural networks [17, 72], then accelerating with the transformer architecture [70]—dramatically improved NLP performance but introduced opacity. These models learn distributed representations that resist human interpretation, as the meaning encoded in high-dimensional vectors cannot be easily mapped to human-understandable concepts. While transformers' self-attention mechanisms explicitly model relationships between all positions in a sequence, and attention weights are sometimes interpreted as importance scores, this interpretation remains contested (see Section 3.4.4). The paradigm shift from feature-engineered models to learned representations fundamentally transformed the interpretability landscape of NLP, as pre-trained language models such as BERT and GPT achieved state-of-the-art performance across tasks at the cost of increased complexity and reduced transparency.

The current era of Large Language Models (LLMs)—including GPT-4, Claude, Llama, and Mistral—presents unprecedented interpretability challenges. These models contain billions of parameters and exhibit emergent capabilities not present in smaller models. They demonstrate in-context learning, allowing them to perform new tasks from examples without weight updates, and instruction-following behavior that enables natural language task specification. LLMs can generate explanations through prompting techniques such as Chain-of-Thought, but whether these explanations accurately reflect internal processing remains an open question. Understanding how LLMs process information and generate outputs is an active research frontier, discussed in Section 3.4.

3 Comprehensive Taxonomy of Explainable NLP

This section presents a unified taxonomy of explanation methods for NLP, organized along multiple dimensions. Figure 4 provides an overview of our categorization framework, organizing methods by scope (local vs. global), with further categorization by technique type.

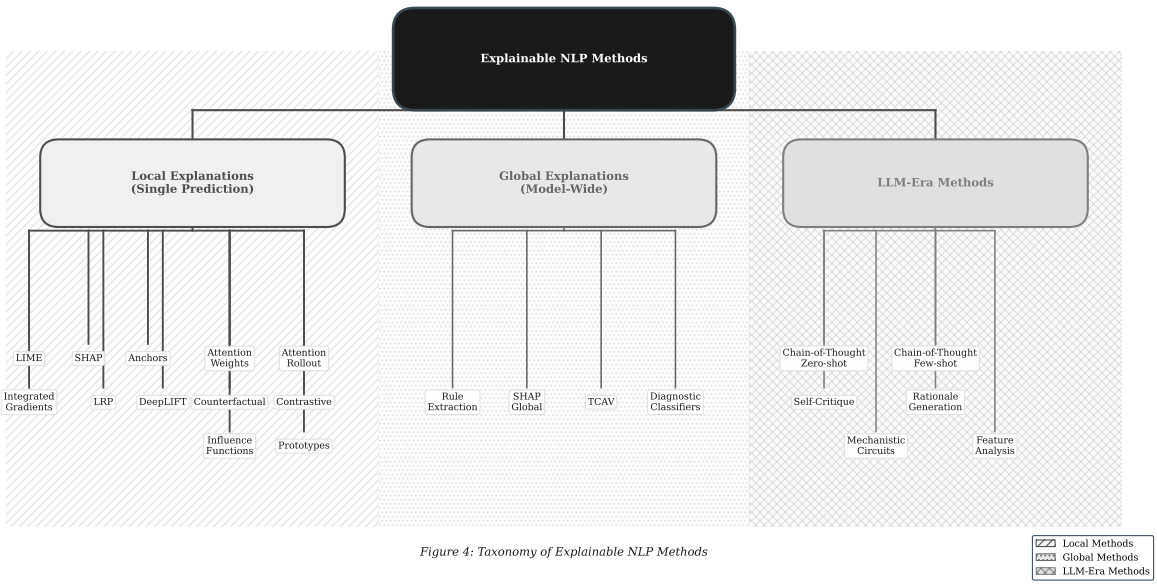

Figure 4: Taxonomy of Explainable NLP Methods

Figure 4: Taxonomy of Explainable NLP Methods. The framework organizes methods by scope (local vs. global), with further categorization by technique type. Local methods explain individual predictions through input attribution, counterfactuals, or examples. Global methods characterize overall model behavior. LLM-era methods leverage the language capabilities of large models for self-explanation.

3.1 Taxonomy Framework

We categorize explanation methods along four primary dimensions that capture the essential characteristics distinguishing different approaches. The first dimension is **scope**, distinguishing between local explanations for single predictions and global explanations that characterize model-wide behavior. The second dimension is **direction**, capturing whether methods work forward from input to output or backward from output to input attributions. The third dimension is **model access**, separating model-agnostic methods that work with any model from model-specific methods designed for particular architectures. The fourth dimension is **output form**, categorizing methods by whether they produce importance scores, rules, examples, counterfactuals, or natural language explanations.

Table 2 provides an overview of prominent explanation methods classified by these dimensions, showing that most methods fall clearly into specific categories while a few, such as SHAP, span multiple categories depending on their application.

3.2 Local Explanation Methods

Local explanations address individual predictions, answering the fundamental question: "Why did the model produce this output for this input?" These methods are essential for understanding specific decisions in high-stakes applications and for debugging model behavior on particular examples.

Table 2: Classification of Explanation Methods by Scope and Direction

Method	Local	Global	Forward	Backward
Vanilla Gradients	✓			✓
Integrated Gradients	✓			✓
SmoothGrad	✓			✓
LRP	✓			✓
DeepLIFT	✓			✓
CAM/GradCAM	✓			✓
LIME	✓			✓
SHAP	✓	✓		✓
Attention weights	✓		✓	
Gate activations	✓		✓	
Hidden states	✓		✓	
Probing classifiers		✓	✓	
Concept activation		✓	✓	
Model distillation		✓	✓	
Pruning analysis		✓		✓
Influence functions	✓			✓

3.2.1 Input-Based Explanations

Input-based explanation methods, also known as feature attribution methods, assign importance scores to input features such as words, tokens, or spans. These methods answer which parts of the input were most influential in producing the model's output.

Perturbation-based methods create explanations by observing how predictions change when inputs are modified. LIME (Local Interpretable Model-agnostic Explanations) [54] creates local linear approximations by perturbing inputs and observing prediction changes; for text, this typically involves masking words and fitting a linear model to explain local behavior. LIME is model-agnostic and widely applicable but computationally expensive due to the need for many model evaluations. Occlusion and erasure methods [40] systematically remove input features and measure prediction changes, providing a simple and intuitive approach that may miss feature interactions.

Gradient-based methods leverage the model's gradient information to attribute importance to inputs. Vanilla gradients [60] compute $\partial y/\partial x$ to identify which input features most influence the output, offering fast computation but potentially noisy attributions. Integrated Gradients [64] address these limitations by computing gradients along a path from a baseline to the input, satisfying theoretical axioms including completeness that ensures all importance is attributed. SmoothGrad [62] reduces noise by averaging gradients over multiple noisy copies of the input.

Propagation-based methods redistribute relevance backward through network layers. Layer-wise Relevance Propagation (LRP) [3, 5] redistributes the prediction backward through network layers according to conservation principles, providing fine-grained attributions but requiring access to model architecture. DeepLIFT [59] compares activations to reference activations and propagates differences backward, addressing gradient saturation issues that affect vanilla gradient methods.

Table 3 summarizes these local-backward explanation methods, their key characteristics, and typical applications in NLP tasks.

Table 3: Local-Backward Explanation Methods for NLP

Method	Key Characteristics	NLP Applications
Vanilla Gradients	Fast computation; can be noisy	Word importance in classification
Integrated Gradients	Satisfies completeness axiom; requires baseline	Token attribution in transformers
SmoothGrad	Noise reduction via sampling	Reducing noise in attention analysis
LRP	Layer-wise propagation; conservation principle	Sentiment analysis; NER
DeepLIFT	Reference-based; handles saturation	Sequence classification; QA
GradCAM	Activation-weighted gradients	CNN-based text classification

Game-theoretic methods apply principles from cooperative game theory to attribution. SHAP (SHapley Additive exPlanations) [41] uses Shapley values to fairly distribute prediction among features, providing theoretical guarantees about the attribution but requiring computational resources that scale exponentially with feature count, though approximations make it tractable for many applications.

Attention-based attribution interprets attention weights in transformer models as importance scores [70]. However, this interpretation remains contested: Jain & Wallace [29] showed that attention often doesn't correlate with gradient-based importance, while Wiegreffe & Pinter [75] argued that attention can serve as explanation under appropriate conditions. Section 3.4.4 provides detailed discussion of this ongoing debate.

Table 4 summarizes local-forward explanation methods that analyze model internals during inference.

Table 4: Local-Forward Explanation Methods for NLP

Method	Key Characteristics	NLP Applications
Attention weights	Direct model output; interpretable	Token relationships; alignment
Attention rollout	Multi-layer aggregation	Information flow in transformers
Gate activations	LSTM/GRU gate analysis	Sequence modeling; memory analysis
Hidden state analysis	Intermediate representations	Feature encoding; layer analysis
Head importance	Attention head pruning	Model compression; head analysis
Key-value inspection	Transformer memory access	Memory retrieval; context usage

3.2.2 Discrete Explanation Methods

Beyond continuous importance scores, several methods produce discrete, interpretable explanation structures. **Counterfactual explanations** identify minimal input changes that would alter the prediction [78]—for example, stating that a review would be classified as positive if "disappointing" were changed to "impressive." These explanations are intuitive because they mirror human

reasoning about causation and are actionable because they suggest specific modifications, though generating meaningful counterfactuals for discrete text is challenging as not all word substitutions produce grammatical or semantically coherent alternatives.

Example-based methods explain predictions by reference to training examples, grounding explanations in concrete instances that users can examine. Influence functions [32] compute how each training example affects the model's prediction on a test point by approximating the effect of removing that example from training; this approach is computationally expensive but provides theoretically grounded attributions. Representer points [80] offer a computationally cheaper alternative, while prototype selection identifies exemplar training cases that characterize different prediction classes. **Rule-based methods** extract interpretable decision rules from model behavior: Anchors [55] generates high-precision rules that almost always yield the same prediction regardless of other input features, providing clear, actionable explanations that non-technical users can understand and verify.

3.3 Global Explanation Methods

Global explanations characterize model behavior across all inputs, answering: "How does the model generally work?" These methods are valuable for understanding model capabilities and limitations, identifying systematic biases, and building trust through comprehensive transparency.

Model distillation trains an interpretable student model—such as a decision tree or rule set—to mimic the black-box teacher [6, 65]. The student's structure then serves as an explanation of the teacher's behavior. This approach trades perfect fidelity for interpretability, as the student model necessarily approximates rather than replicates the teacher.

Global feature importance methods aggregate local importances or analyze model-wide patterns. Common approaches include aggregating SHAP values across many predictions to identify consistently important features, analyzing attention patterns across the corpus to understand what the model focuses on, and identifying globally important n-grams or features that drive predictions across the dataset.

Testing with Concept Activation Vectors (TCAV) [31] measures how much a user-defined concept influences predictions. Researchers define concepts such as "formal language" or "technical terminology" and use linear probes on internal representations to quantify their influence on model outputs. This approach bridges the gap between low-level features and high-level human concepts.

Probing classifiers train auxiliary classifiers on model representations to detect what linguistic information is encoded [7]. Researchers probe for part-of-speech tags, syntactic structure, semantic roles, and named entities to understand what knowledge representations capture. This approach reveals what information is available in representations without assuming the model actually uses that information.

Tables 5 and 6 summarize global explanation methods that analyze model-wide behavior through forward analysis and backward modification, respectively.

Table 5: Global-Forward Explanation Methods for NLP

Method	Key Characteristics	NLP Applications
Probing classifiers	Auxiliary classification tasks	Linguistic knowledge detection
TCAV	Concept activation vectors	High-level concept testing
Attention patterns	Corpus-wide attention analysis	Head specialization; patterns
Activation clustering	Hidden state clustering	Semantic grouping; concepts
Model distillation	Knowledge transfer to interpretable model	Rule extraction
Feature visualization	Maximize neuron activations	Understanding representations

Table 6: Global-Backward Explanation Methods for NLP

Method	Key Characteristics	NLP Applications
Weight pruning	Remove unimportant weights	Model compression; importance
Head pruning	Remove attention heads	Head importance analysis
Layer ablation	Remove entire layers	Layer contribution analysis
Lottery tickets	Sparse subnetwork discovery	Essential parameter identification
Knowledge neurons	Locate factual knowledge	Fact storage in transformers
Circuit analysis	Identify functional circuits	Mechanistic understanding

3.4 LLM-Era Explainability

Large Language Models present unique interpretability challenges and opportunities that differ fundamentally from earlier neural network paradigms. Their scale, emergent capabilities, and natural language generation abilities enable new explanation approaches while complicating traditional methods.

3.4.1 Chain-of-Thought Prompting

Chain-of-Thought (CoT) prompting [74] elicits step-by-step reasoning from LLMs, producing explanations alongside answers. A typical CoT prompt might produce: "Let's think step by step. First, I notice that the review mentions several positive aspects such as... Therefore, the answer is..." Research has established several key findings: CoT improves performance on reasoning tasks [33, 74], self-consistency across multiple CoT paths improves reliability [73], and CoT can be elicited zero-shot with simple prompts like "Let's think step by step" [33].

However, faithfulness remains a significant concern. Turpin *et al.* [68] showed that LLMs don't always "say what they think"—stated reasoning may not reflect actual computation. The model may produce plausible-sounding explanations that rationalize outputs rather than accurately describing the generation process. Lanham *et al.* [39] developed methods to measure CoT faithfulness through interventions that test whether modifying the reasoning chain affects the final output. Lyu *et al.* [43] proposed training-based approaches to improve faithfulness,

constraining models to produce reasoning traces that demonstrably influence their final answers rather than post-hoc rationalizations.

3.4.2 Self-Explanation and Critique

LLMs can be prompted to explain their outputs through direct queries such as "Why did you classify this as spam?" or "What evidence supports your conclusion?" This self-explanation capability offers an intuitive interface for obtaining explanations in natural language. However, such explanations may be post-hoc rationalizations rather than faithful accounts of the model's reasoning process [79]. Self-critique mechanisms, where models evaluate and critique their own outputs, show promise for improving response quality but face similar faithfulness challenges. Chen *et al.* [12] showed that LLMs can be taught to self-debug code by generating explanations of failures and iteratively correcting errors—a form of explanation-guided self-improvement with practical utility.

3.4.3 Mechanistic Interpretability

Mechanistic interpretability aims to understand models at the level of individual components, including neurons, attention heads, and circuits. This bottom-up approach seeks to identify how specific computations are implemented in model weights.

Anthropic's research on circuits and features [13, 16] identifies interpretable "circuits"—patterns of connected components implementing specific functions. Examples include induction heads that copy patterns from context and edge detection circuits in vision models. This work suggests that neural networks may be more modular than previously thought, with identifiable components performing recognizable functions. Wu *et al.* [76] extended circuit analysis to instruction-tuned models, identifying causal mechanisms in Alpaca that govern how instructions are interpreted and followed.

Models may represent more features than they have dimensions through superposition [16], where multiple features share the same neurons through approximately orthogonal representations. This phenomenon complicates interpretation because individual neurons may encode multiple distinct features simultaneously. Complementary work on knowledge localization has shown that factual associations can be precisely located and edited within transformer weights [46], suggesting that despite superposition, certain information is stored in identifiable locations. Geva *et al.* [20] further dissected how factual associations are recalled during generation, tracing the flow from subject representations through relation-specific processing to final answer production.

Automated interpretability approaches use LLMs to scale human understanding. Bills *et al.* [9] used language models to automatically generate descriptions of what individual neurons detect, enabling analysis at scales impossible for human researchers alone. Templeton *et al.* [66] extended this approach to Claude 3 Sonnet, extracting millions of interpretable features that represent concepts ranging from concrete entities to abstract ideas—demonstrating that sparse autoencoders can scale to production models.

3.4.4 The Attention Debate

Whether attention weights constitute valid explanations remains contested, with significant implications for the most widely-used explanation method in NLP. Arguments against attention as explanation highlight that alternative attention distributions can yield identical predictions [29] and that attention may not correlate with gradient-based importance measures [58]. These findings suggest attention may capture something other than input importance.

Arguments for attention as explanation emphasize that attention can be a valid explanation under appropriate conditions [75] and that constrained attention mechanisms can improve faithfulness [48]. The debate has produced several resolution approaches: attention rollout and flow methods [1] trace information through layers to address multi-layer aggregation, and gradient-weighted attention may be more faithful than raw attention weights.

3.5 Summary of Methods

Table 7 synthesizes the key explanation methods discussed, their characteristics, and typical use cases, providing a reference for practitioners selecting among available approaches.

Table 7: Summary of explanation methods

Method	Scope	Access	Output	Notes
LIME	Local	Agnostic	Importance	Versatile but slow
SHAP	Local/Global	Varies	Importance	Principled but expensive
Integrated Gradients	Local	Specific	Importance	Satisfies axioms
LRP	Local	Specific	Importance	Architecture-dependent
Attention	Local	Specific	Importance	Faithfulness debated
Anchors	Local	Agnostic	Rules	High-precision rules
Influence Functions	Local	Specific	Examples	Training data attribution
CoT Prompting	Local	Agnostic	Text	LLM-specific, faithfulness varies

Table 8 categorizes explanation methods by their output type, distinguishing between continuous outputs like importance scores and discrete outputs like rules and examples. This distinction is important for matching explanation methods to user needs and application requirements.

4 Evaluation Frameworks

Evaluating explanations remains challenging due to the lack of ground truth for "correct" explanations. This section reviews intrinsic evaluation (explanation quality) and extrinsic evaluation (downstream utility).

4.1 Intrinsic Evaluation

Intrinsic evaluation assesses explanation quality independent of downstream tasks.

Table 8: Explanation Methods by Output Type

Output Type	Methods	Characteristics
Continuous Outputs		
Importance scores	LIME, SHAP, Gradients, LRP	Numerical attribution per feature
Attention weights	Self-attention, Cross-attention	Soft alignment scores
Probability scores	Confidence, Uncertainty	Prediction certainty
Discrete Outputs		
Rules	Anchors, Decision rules	If-then conditions
Examples	Influence functions, Prototypes	Similar training instances
Counterfactuals	Contrastive examples	Minimal input changes
Text	CoT, Self-explanation	Natural language rationale
Concepts	TCAV, Concept bottleneck	Human-defined abstractions

4.1.1 Faithfulness

Faithfulness measures how accurately an explanation reflects the model's actual reasoning process [26, 28]. A faithful explanation should accurately represent the features the model uses, change when model reasoning changes, and not mislead about causal mechanisms underlying predictions.

Several measurement approaches have been developed. **Sufficiency** captures whether keeping only highlighted features preserves the prediction [15]. Formally:

$$\text{Sufficiency} = 1 - \frac{1}{N} \sum_i |p(y|x_i) - p(y|r_i)| \tag{1}$$

where r_i is the rationale (highlighted portion) of input x_i.

Comprehensiveness measures whether removing highlighted features changes the prediction [15]:

$$\text{Comprehensiveness} = \frac{1}{N} \sum_i |p(y|x_i) - p(y|x_i \setminus r_i)| \tag{2}$$

Fidelity applies to surrogate explanations, measuring how often the surrogate agrees with the original model [38].

4.1.2 Additional Quality Criteria

Beyond faithfulness, several related criteria characterize explanation quality. **Soundness** requires that all explanation components are true of the underlying system [34], while **completeness** requires that the explanation covers all relevant aspects of the system's behavior—a complete and sound explanation is necessarily faithful. **Plausibility** measures how convincing an explanation is to humans [26], with measurement typically involving comparison to human-annotated rationales from datasets like e-SNLI [10] and CoS-E [52]. Importantly, plausibility differs from faithfulness: a plausible but unfaithful explanation may mislead users about actual model behavior, potentially

causing more harm than no explanation at all.

Comprehensibility measures how easily humans can understand explanations, with simplicity proxies including the number of features, rule complexity, and explanation length [37]. Sparsity—having fewer highlighted features—is often assumed to improve comprehensibility [8], though human studies provide the most reliable assessment. **Generalizability** measures whether explanations hold beyond the specific instance [63], addressing whether insights transfer to similar examples and helping users build accurate mental models of system behavior.

4.2 Extrinsic Evaluation

Extrinsic evaluation assesses explanation utility for downstream tasks. A fundamental question is whether explanations help humans make better decisions: task accuracy studies measure human performance with and without explanations [35], showing that explanations can improve human accuracy on some tasks, though benefits depend on explanation quality and user expertise, and poor explanations can actually harm performance by misleading users. Decision time provides another measure, as efficient explanations should support decision-making without significantly slowing the process.

4.2.1 Trust and Decision Support

Explanations should help users develop appropriate trust in AI systems [27], where appropriate trust means users trust reliable predictions and distrust unreliable ones. However, over-trust can occur when users accept plausible but unfaithful explanations uncritically, making measurement through trust questionnaires and behavioral measures essential. For model development, explanations serve valuable roles in error detection [54], training data debugging through influence functions [32], and explanatory debugging that allows users to correct model behavior [67]. Simulatability tests whether users can predict model behavior after studying explanations [23]: presenting examples and explanations, then asking users to predict model outputs on new examples, with higher accuracy implying more useful explanations.

4.3 Benchmarks and Datasets

The ERASER benchmark [15] provides standardized evaluation across multiple datasets with human rationale annotations, including Movie Reviews for sentiment analysis, BoolQ for reading comprehension, Evidence Inference for scientific claims, FEVER for fact verification, MultiRC for multi-sentence reasoning, and e-SNLI for natural language inference. Additional datasets provide explanation annotations for specific domains: e-SNLI [10] offers natural language explanations for natural language inference, CoS-E [52] provides commonsense explanations, HatEval includes hate speech annotations with rationales, and LIAR-PLUS offers fake news instances with justifications. The field continues developing new evaluation resources including LLM-specific explanation benchmarks designed for large language models, faithfulness evaluation suites specifically for Chain-of-Thought explanations, and cross-lingual explanation datasets that support multilingual explainability research.

4.4 Comparison Across Methods

Direct comparison of explanation methods requires careful experimental design [50, 51]. Several challenges complicate such comparisons: different methods produce different output formats that resist direct comparison, ground truth varies across tasks, and human evaluation is both expensive and subjective.

Best practices for comparison studies include using multiple metrics that capture faithfulness, plausibility, and comprehensibility together, comparing on standardized benchmarks for reproducibility, including human evaluation where possible, and reporting confidence intervals and significance tests to support reliable conclusions.

5 Practical Guidelines

This section provides actionable guidance for practitioners selecting and implementing explainability methods. We present decision frameworks, task-specific recommendations, and an overview of available tools.

5.1 Method Selection Decision Framework

Selecting an appropriate explanation method requires considering multiple factors that interact in complex ways. Figure 5 presents a decision tree for common scenarios, guiding practitioners through key decision points to arrive at suitable method recommendations.

5.1.1 Key Decision Factors

The available level of model access fundamentally constrains method selection and should be the first consideration in any explainability project. With full access to weights and gradients, practitioners can employ the widest range of methods including gradient-based attribution techniques such as Integrated Gradients and SHAP, attention visualization that reveals internal model focus, and probing classifiers that analyze what information representations encode. With API access only, perturbation-based methods like LIME remain viable since they only require the ability to query the model with modified inputs, and prompting-based explanations including Chain-of-Thought and self-explanation become attractive options for LLM-based systems. For true black-box scenarios where only inputs and outputs are observable, practitioners must rely exclusively on model-agnostic methods such as LIME, Anchors, and counterfactual generation.

The scope of explanation needed further guides method selection within access constraints. For explaining single predictions, local methods including LIME, attention visualization, and counterfactuals provide focused insights into specific decisions. For characterizing overall model behavior, global methods such as SHAP aggregation across many predictions, probing classifiers that reveal what representations encode, and model distillation that extracts interpretable approximations provide broader insights. Many applications benefit from combining both approaches: local explanations for individual decisions that stakeholders question, alongside global summaries of model behavior that build overall understanding and trust.

Different audiences require fundamentally different explanation strategies, and tailoring explanations to the intended recipients is essential for their effectiveness. End users typically prefer

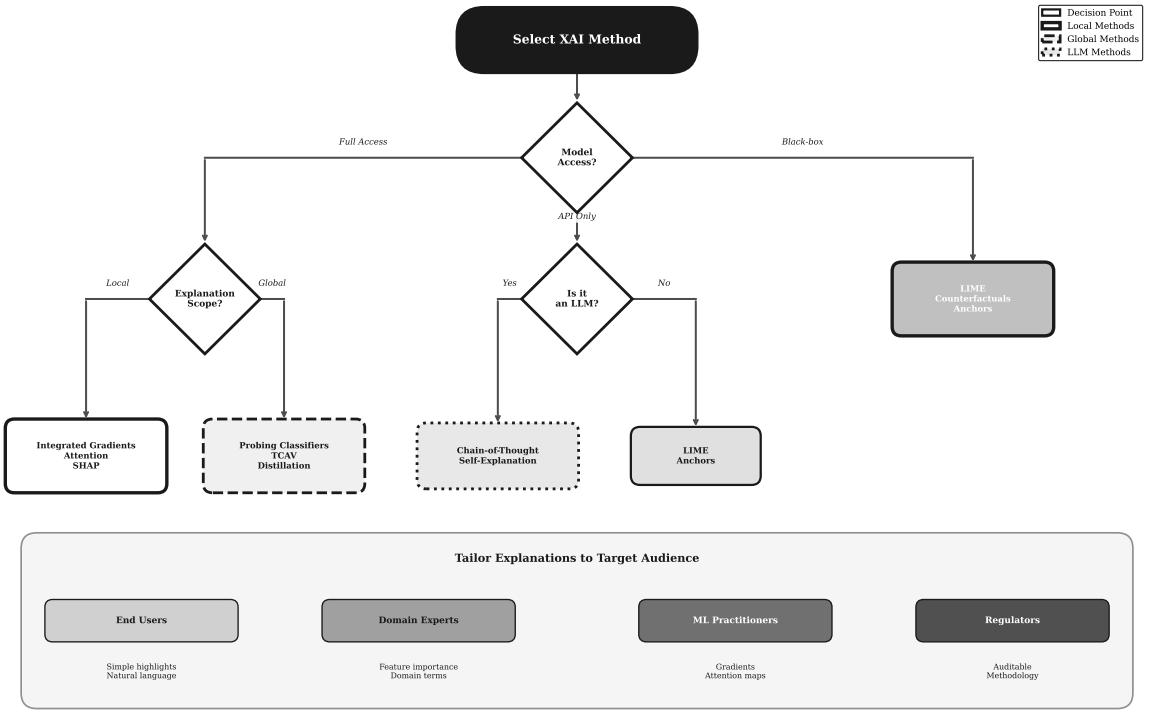

Figure 2: Decision Tree for Selecting Explainability Methods

Figure 5: Decision tree for selecting explanation methods. The framework guides practitioners through key decision points: model access level (full, API, or black-box), explanation scope (local vs. global), and target audience. Recommendations are tailored to each combination of factors, with special consideration for LLM-based systems.

natural language explanations that require no technical background, simple visual highlights that draw attention to relevant input features, and actionable counterfactuals that suggest what could be different. Domain experts such as physicians, lawyers, or financial analysts benefit from feature importance presented using domain-specific terminology and rule-based explanations that map to their professional knowledge and reasoning patterns. ML practitioners working on model development can interpret detailed attributions, attention patterns, and gradient visualizations that provide technical insight into model behavior. Regulators require auditable explanations with documented methodology that can support compliance verification and external review.

5.1.2 Decision Framework for LLM-Era Systems

For systems using Large Language Models, additional considerations apply beyond traditional explainability methods. The first question is whether the model is a black-box API such as GPT-4 or Claude with no access to internal representations. If so, prompting-based explanations through Chain-of-Thought and self-explanation become the primary options. If internal access is available through open-weight models, the degree of access matters: full access enables mechanistic interpretability and attention analysis at scale, while limited access supports gradient-based methods and probing approaches.

A critical consideration for LLM applications is whether faithfulness is essential. In high-stakes applications where explanations must accurately reflect model reasoning, practitioners

should verify Chain-of-Thought faithfulness through intervention tests, use multiple explanation methods for cross-validation, and consider mechanistic analysis where feasible. In applications where plausible explanations suffice for user understanding without strict accuracy requirements, the simpler prompting-based approaches may be adequate despite their faithfulness limitations.

5.2 Task-Specific Recommendations

Table 9 provides recommendations organized by NLP task, summarizing appropriate methods, evaluation focus, and key metrics for each application domain.

Table 9: Task-specific explainability recommendations

Task	**Recommended Methods**	**Evaluation Focus**	**Key Metrics**
Text Classification	LIME, SHAP, attention visualization, Anchors	Faithfulness, human agreement	Sufficiency, comprehensiveness
Sentiment Analysis	LIME, counterfactuals, aspect-based explanations	Plausibility, user trust	Agreement with human rationales
Named Entity Recognition	Token-level attention, CRF feature weights, probing	Span-level accuracy	Token F1, span overlap
Question Answering	Attention rollout, evidence extraction, reading comprehension rationales	Sufficiency, answer faithfulness	Rationale precision/recall
Text Generation	Chain-of-Thought, self-critique, attention to source	Factuality, coherence	Faithfulness tests, human evaluation
Machine Translation	Attention alignment, word-level attributions	Alignment accuracy	Cross-lingual metrics
Summarization	Extractive rationales, attention to key sentences	Coverage, relevance	ROUGE on rationales

5.2.1 Text Classification

For document and sentence classification tasks, a systematic approach yields the best results. Start with attention visualization for quick insights into where the model focuses, then use LIME or SHAP for rigorous feature attribution with theoretical grounding. Apply Anchors for rule-based explanations that practitioners can directly interpret and verify. Finally, validate explanations with human rationale datasets such as e-SNLI and ERASER to ensure alignment with human understanding.

Several pitfalls commonly undermine these efforts and should be actively avoided. Relying solely on attention weights without additional validation is a frequent mistake, as attention does not always correlate with feature importance. Many practitioners ignore out-of-distribution behavior, failing to test whether explanations remain valid under domain shift when models are deployed on data different from training. Most critically, explanations are often deployed without

faithfulness verification, creating a false sense of understanding that can lead to inappropriate trust in model decisions.

5.2.2 Sequence Labeling Tasks

For token-level prediction tasks such as named entity recognition and part-of-speech tagging, explanations must operate at the token level while accounting for context. Token-level gradient attributions identify which context tokens influence predictions for each labeled position, while for structured prediction models that use CRF layers, analyzing transition patterns reveals how the model captures sequential dependencies. Probing classifiers help understand what linguistic information representations encode at different layers. Key considerations include analyzing context windows around entity boundaries, which often prove critical for accurate labeling, and visualizing cross-entity dependencies to understand how recognition of one entity affects predictions for others—the sequential nature of these tasks means explanations should capture both local token features and broader contextual influences.

5.2.3 Question Answering

For extractive question answering, where models select answer spans from context, attention rollout from answer tokens to question and context reveals how the model connects queries to responses, evidence sentence extraction with confidence scores provides interpretable justifications that users can verify, and counterfactual analysis that modifies context to observe answer changes illuminates what information the model actually relies upon. For generative question answering with LLMs, Chain-of-Thought prompting elicits reasoning traces that can be examined and evaluated, self-consistency across multiple reasoning paths indicates reasoning reliability and helps identify confident answers, and source attribution for factual claims helps verify that responses are grounded in provided context rather than hallucinated from the model's parametric knowledge.

5.2.4 Text Generation Tasks

For summarization, translation, and open-ended generation, a multi-faceted approach addresses the complexity of these tasks: attention to source documents reveals which input content influences generated text, Chain-of-Thought prompting provides reasoning transparency for complex generation decisions, self-critique mechanisms allow models to evaluate and improve their own outputs, and tracking factual grounding in source documents helps prevent hallucination. LLM-specific considerations include verifying CoT faithfulness using intervention tests that modify reasoning chains to observe effects on outputs, employing multiple explanation methods for cross-validation to avoid relying on potentially unfaithful self-explanations, and considering mechanistic analysis for applications where explanation accuracy is critical and justifies the additional computational investment.

5.3 Evaluation Protocol Recommendations

5.3.1 Evaluation and Reporting Standards

We recommend the following minimum evaluation protocol for any explainability deployment, along with documentation standards that support reproducibility and enable meaningful comparison across studies. Faithfulness assessment should use at least one rigorous method such as sufficiency and comprehensiveness tests from the ERASER benchmark [15], feature ablation studies that remove attributed features and measure prediction changes, or consistency checks that verify explanations remain stable under minor input perturbations. Plausibility assessment should be conducted when human rationales are available, including token-level F1 against human annotations and Intersection-over-Union (IoU) for span-level explanations that should identify the same text regions humans consider important. Human evaluation is essential for user-facing applications, including comprehensibility ratings, task performance measurements comparing user accuracy with and without explanations, and trust calibration assessment to ensure users appropriately weight AI recommendations.

When reporting results, essential documentation elements include dataset characteristics and size to contextualize results, model architecture and training details that may affect explanation quality, and explanation method hyperparameters that influence output. Reports should present multiple metrics rather than single-point comparisons, along with confidence intervals or significance tests that quantify uncertainty. Computational cost in time and memory helps practitioners assess feasibility, and honest discussion of limitations and failure cases prevents overconfident adoption.

5.4 Tools and Libraries

Table 10 summarizes available implementation resources that practitioners can use to apply the methods discussed in this survey.

Recent developments have produced specialized tools for LLM interpretability that address the unique challenges of large-scale models. For open-weight models with internal access, TransformerLens provides mechanistic interpretability capabilities enabling circuit-level analysis of how computations are implemented, Neuronpedia offers neuron-level analysis and visualization at scales previously impossible, Anthropic's Circuit Analysis tools support identification of computational circuits, and OpenAI Neuron Viewer enables automated interpretation of what individual neurons detect. For API-based LLMs where internal access is unavailable, several frameworks support prompting-based explanation generation: LangChain provides Chain-of-Thought integration and structured outputs for building explanation pipelines, Guidance enables constrained generation with explanation templates that ensure consistent output formats, and DSPy supports programmatic prompting with explanation optimization that can improve explanation quality through automated prompt tuning.

5.5 Implementation Best Practices

A systematic development workflow supports successful explainability implementation. Start with simple methods like attention visualization or LIME to establish baseline understanding before

Table 10: Explainability tools and libraries

Tool	Methods	Notes
Captum	Gradient-based, SHAP, LIME	PyTorch-native, comprehensive attribution methods
SHAP	SHAP, TreeSHAP	Unified interface, fast for tree models
InterpretML	LIME, SHAP, EBM	Microsoft library, glass-box models
Transformers Interpret	Attention, gradients	HuggingFace integration, transformer-specific
LIT (Language Interpretability Tool)	Multiple	Google tool, interactive visualization
Ecco	Attention, gradients	Transformer visualization, neuron analysis
BertViz	Attention visualization	Multi-head attention patterns
AllenNLP Interpret	Gradients, attention	NLP-focused, AllenNLP integration

investing in more complex approaches. Validate faithfulness before deployment by verifying that explanations reflect actual model behavior rather than plausible-sounding rationalizations. Conduct user testing to evaluate whether explanations actually help users in realistic task scenarios, as technically sound explanations may fail to improve user performance. Monitor explanation quality in production as data distributions shift, since explanations valid for training data may become misleading under distribution drift. Finally, iterate based on user feedback and observed limitations, treating explainability as an ongoing process rather than a one-time implementation.

Several pitfalls frequently undermine explainability efforts and should be consciously avoided. Assuming attention equals importance without validation often leads to misleading conclusions since attention serves multiple functions beyond importance weighting. Ignoring faithfulness and accepting plausible explanations can actively mislead users who trust explanations that don't reflect actual model reasoning. Adopting a one-size-fits-all approach neglects the reality that different users need different explanations tailored to their expertise and needs. Underestimating computational cost can make some methods impractical for real-time use in production systems. Over-trusting LLM self-explanations without verification risks propagating unfaithful rationalizations that sound convincing but don't accurately describe model behavior.

For production systems operating at scale, several strategies support practical deployment. Caching explanations for common inputs reduces repeated computation when the same queries recur. Using approximate methods such as KernelSHAP rather than exact Shapley values enables real-time applications with acceptable accuracy trade-offs. Pre-computing global explanations offline amortizes computational cost by investing in comprehensive analysis once rather than repeatedly. Considering explanation-by-design approaches, where models are built with interpretability in mind, can dramatically reduce explanation latency for time-critical systems compared to post-hoc explanation methods.

6 Applications

Explainable NLP has been applied across diverse domains where transparency and accountability are essential. This section surveys key application areas and their specific requirements, examining how XAI techniques are deployed across healthcare, legal, financial, educational, and research contexts, each with distinct requirements and stakeholders.

6.1 Healthcare NLP

Healthcare represents one of the most demanding domains for explainable AI, where decisions directly impact patient outcomes and regulatory compliance is mandatory.

6.1.1 Clinical Decision Support

NLP systems assist clinicians in diagnosis, treatment planning, and risk assessment. These systems support clinical note analysis and coding, drug-drug interaction detection, diagnostic suggestion, and patient risk stratification. Each application demands specific explainability properties to ensure safe and effective clinical use.

Explainability requirements in clinical settings are particularly stringent. Systems must highlight relevant clinical findings within patient records, enabling clinicians to quickly verify the basis for AI recommendations. Evidence-based justifications that reference medical literature are essential, as clinicians require authoritative sources to validate diagnostic suggestions. Furthermore, explanations must support differential diagnosis reasoning, helping clinicians consider alternative interpretations. Critically, systems should enable clinician override with documented rationale, ensuring that human judgment remains paramount.

Several explanation methods have proven effective in clinical contexts. Attention-based highlighting can identify relevant text spans within lengthy patient records. Rule extraction techniques can formalize clinical pathways in interpretable formats. Case-based reasoning presents similar patient examples to support diagnostic conclusions. More recently, Chain-of-Thought prompting has enabled LLMs to articulate diagnostic reasoning in a manner that clinicians can evaluate and critique.

6.1.2 Medical Literature Analysis and Regulatory Compliance

LLMs are increasingly used for literature review and evidence synthesis, including systematic review automation, clinical trial matching, and drug discovery literature mining. These applications require robust source attribution for all claims, clearly communicated confidence levels for extracted information, and sophisticated contradiction detection and flagging mechanisms. Across all healthcare AI applications, systems must comply with multiple regulatory frameworks including FDA guidance on AI/ML-based medical devices [69], the EU Medical Device Regulation (MDR), and HIPAA privacy requirements. Explainability supports regulatory compliance by enabling audit trails for AI-assisted decisions, comprehensive documentation of model behavior, and evidence of bias detection and mitigation efforts.

6.2 Legal and Financial NLP

Legal and financial domains require high-stakes decision-making with significant accountability requirements.

6.2.1 Legal Document Analysis

Legal NLP applications span contract review and analysis, legal research and case law retrieval, regulatory compliance checking, and litigation outcome prediction. Each application involves complex, domain-specific language and often requires processing long document contexts. Explanations in this domain must meet particularly high standards, as they may be subject to legal scrutiny.

Explainability requirements in legal contexts include citation to relevant legal precedents that support the system's analysis, clause-level highlighting in contracts to identify critical provisions, reasoning chains that articulate legal arguments in a defensible manner, and transparent handling of ambiguity when legal language admits multiple interpretations. The need for precise, defensible explanations presents ongoing challenges, as legal reasoning often involves nuanced judgments that resist simple formalization.

6.2.2 Financial Text Analysis

Financial NLP encompasses sentiment analysis for market prediction, credit risk assessment from text, fraud detection in communications, and regulatory filing analysis. These applications face growing regulatory scrutiny and require robust accountability mechanisms.

Explainability requirements in financial contexts include feature attribution for risk factors that influence lending or investment decisions, comprehensive audit trails for lending decisions as required by fair lending regulations, demonstrated compliance with fair lending laws, and transparent documentation of model updates. Failure to provide adequate explanations can result in regulatory penalties and legal liability.

6.2.3 Regulatory Landscape

Financial AI faces an evolving regulatory landscape. The EU AI Act [19] establishes requirements for high-risk AI systems, while GDPR [18] provides rights related to automated decision-making, though the extent of the "right to explanation" remains debated [57, 71]. In the United States, fair lending laws including the Equal Credit Opportunity Act and Fair Housing Act require explainable credit decisions. The SEC has also issued guidance on AI in trading, emphasizing the need for transparency and human oversight.

6.3 Social Science Research

NLP tools increasingly support social science research, where methodological transparency is essential for scientific validity. Social scientists employ NLP for automated coding of qualitative data, discourse analysis at scale, public opinion mining, historical document analysis, survey response coding, theme identification, and sentiment analysis. These diverse applications share common explainability requirements that support scientific validity claims.

Explainability for validity in social science contexts demands transparent coding decisions with clear justification, enabling other researchers to evaluate classification choices. Inter-coder reliability assessment remains important even when one "coder" is an AI system, and documentation of model limitations helps researchers understand the boundaries of automated analysis. Perhaps most importantly, replicability of classification decisions allows other scholars to verify findings. Explanations must align with theoretical frameworks that guide social science inquiry, supporting researcher interpretation rather than replacing scholarly judgment while maintaining transparency about the AI's role in the analysis process. Large-scale social media and text analysis requires reproducible classification pipelines with well-documented methodological choices that support the peer review process. Researchers must acknowledge bias and limitations inherent in automated analysis, and human-in-the-loop validation remains essential for ensuring analytical quality.

6.4 Content Moderation

Content moderation at scale presents unique challenges for explainable AI, balancing efficiency with fairness and transparency. Applications include automated content flagging, priority queue management for human review, user warning systems, and platform policy enforcement, all processing massive volumes of content under significant time pressure while facing distinctive challenges including context sensitivity where sarcasm, quotes, and cultural context can fundamentally alter meaning, adversarial manipulation by users attempting to evade detection, the need for multilingual coverage with consistent explanations across languages, and constantly evolving language including new slurs and coded expressions.

6.4.1 Hate Speech and Toxicity Detection

Explainability requirements for content moderation include identifying specific policy violations that triggered moderation decisions, highlighting problematic content spans within flagged material, providing appeals pathways with clear explanations of the original decision, and distinguishing severity levels to enable proportionate responses. Several datasets support research in this area, including HatEval for hate speech detection with rationales, Jigsaw Toxic Comments for toxicity classification, and various Implicit Hate Speech datasets for more subtle violations.

6.4.2 Misinformation Detection

Misinformation detection encompasses fake news classification, claim verification, and source credibility assessment. Explainability approaches in this domain include evidence retrieval for fact-checking that connects claims to authoritative sources, source attribution and provenance tracking, claim-evidence alignment visualization, and confidence scoring with uncertainty quantification. Research datasets such as FEVER (Fact Extraction and VERification), LIAR-PLUS (fake news with justifications), and ClaimBuster (claim detection) support development and evaluation of these systems. The sheer scale of content moderation—billions of pieces of content requiring rapid decisions—constrains the computational complexity of viable explanation methods, making

efficiency a critical consideration alongside explanation quality.

6.5 Education and Learning

Educational NLP applications require explanations that support learning objectives rather than merely justifying system outputs. Automated essay scoring systems must provide formative feedback identifying both strengths and areas for improvement, with explanations aligned to rubric criteria that students can understand and act upon, offering specific suggestions for improvement rather than simple scores while maintaining transparency about the criteria underlying scoring decisions. Intelligent tutoring systems support adaptive learning path recommendation, error diagnosis and remediation, and ongoing progress assessment. Explanation needs in educational contexts differ from other domains: explanations must be pedagogically appropriate for the learner's level, offering scaffolded hints rather than direct answers, with the focus on learning and growth rather than mere performance metrics, supporting the educational mission of these systems.

6.6 Customer Service and Business

Commercial NLP applications require explanations that build user trust and support quality assurance. Chatbots and virtual assistants benefit from explainability features including confidence indication for uncertain responses, source attribution for information provided, clear explanation when escalating to human agents, and transparency about conversation flow and system capabilities, all helping users calibrate their trust appropriately. Business intelligence applications encompass customer feedback analysis, market sentiment tracking, and competitive intelligence, requiring trend attribution to specific sources that enables verification, clearly communicated confidence levels for predictions, and actionable insights supported by concrete evidence.

6.7 Cross-Domain Considerations

6.7.1 Common Requirements

Across application domains, we identify common explainability needs:

1. **Accountability**: Ability to audit and justify decisions
2. **Transparency**: Clear communication of system capabilities and limitations
3. **Fairness**: Detection and mitigation of bias
4. **Trust calibration**: Helping users understand when to rely on AI
5. **Human oversight**: Supporting rather than replacing human judgment

6.7.2 Domain-Specific Adaptations

Each domain requires tailored explanations that respect its unique characteristics and stakeholder needs. Healthcare explanations must employ appropriate medical terminology and evidence-based

reasoning that clinicians can evaluate. Legal explanations require precedent citation and formal argumentation structures that meet professional standards. Social science applications demand methodological transparency and theoretical grounding that satisfy scientific norms. Content moderation requires policy-aligned explanations that can operate at massive scale. Educational applications need pedagogically appropriate, growth-oriented feedback that supports learning.

7 Challenges and Future Directions

Despite significant progress, explainable NLP faces fundamental challenges. This section discusses open problems and emerging research directions.

7.1 Open Challenges

7.1.1 Faithfulness Verification

The fundamental challenge of verifying whether explanations accurately reflect model behavior remains unsolved [26]. Several core issues complicate this verification. The field lacks ground truth for what constitutes a "correct" explanation, making objective evaluation inherently difficult. Available faithfulness metrics may not capture true model reasoning, as they often rely on proxy measures rather than direct verification. Post-hoc explanations risk rationalizing observed outputs rather than explaining actual computational processes. Furthermore, different explanation methods frequently disagree with one another, suggesting they may capture different aspects of model behavior or reveal fundamental incompleteness in our understanding.

Large Language Models introduce additional faithfulness challenges. Chain-of-Thought reasoning may not reflect actual computation [68], as the verbalized reasoning process can diverge from the model's internal computations. Self-explanations may be plausible to human readers while remaining unfaithful to the model's actual decision process. Emergent capabilities resist mechanistic explanation, appearing suddenly at scale without clear underlying mechanisms. The sheer scale of these models makes comprehensive analysis computationally infeasible.

Addressing these challenges requires new research directions including causal intervention methods for faithfulness testing, consistency metrics that compare explanations across multiple methods, mechanistic interpretability techniques that can scale to large models, and formal verification approaches that provide mathematical guarantees.

7.1.2 The Attention-Explanation Debate

Whether attention weights constitute valid explanations remains contested (Section 3.4.4), with significant implications for the most widely-used explanation method in NLP.

Current understanding suggests that attention alone is insufficient for explanation, though context matters considerably—attention can be explanatory under specific conditions. Research indicates that gradient-weighted attention may be more faithful than raw attention weights, while multi-head attention complicates interpretation by distributing information across many parallel attention mechanisms.

The field needs research that precisely characterizes when attention serves as valid explanation, methods that integrate attention with other explanation approaches for more robust insights,

and techniques for applying attention-based explanations to the increasingly complex attention patterns in LLMs.

7.1.3 Scaling to Large Language Models

Modern LLMs present unprecedented interpretability challenges across multiple dimensions. Scale challenges include billions of parameters that resist comprehensive analysis, emergent capabilities that appear unpredictably as models grow, training data opacity that prevents tracing connections between inputs and outputs, and the substantial computational costs of mechanistic analysis at this scale.

Capability challenges compound these difficulties. In-context learning mechanisms remain poorly understood despite their practical importance. Instruction-following behavior lacks satisfying explanation, as the connection between training and behavior remains opaque. Reasoning capabilities may be approximate or spurious, producing correct answers through unreliable processes. Safety-relevant behaviors particularly need explanation, yet often prove most resistant to analysis.

Access challenges further limit progress. Proprietary models such as GPT-4 and Claude lack internal access for researchers outside their developing organizations. API-only interaction severely limits applicable explanation methods. Model updates change behavior without transparency, potentially invalidating previous interpretability work.

7.1.4 Evaluation and Human Factors

The field lacks consensus on evaluation methodology, impeding progress and comparison across approaches. Current problems include inconsistent metrics across papers that make comparison difficult, limited benchmarks specifically designed for evaluating LLM explanations, expensive and variable human evaluation protocols, and task-specific evaluation needs that resist standardization. Progress requires standardized evaluation protocols that enable meaningful comparison, LLM-era benchmarks designed specifically for explanation quality assessment, automated evaluation methods that can scale with the field's growth, and cross-task evaluation frameworks that capture general principles of explanation quality.

Explanations must serve human needs, but human-centered evaluation remains challenging. Individual differences in explanation preferences mean that no single explanation format suits all users, cognitive biases affect how humans interpret explanations potentially leading to systematic misunderstanding, users may over-trust plausible but unfaithful explanations creating false confidence, and expertise substantially affects explanation utility with experts and novices requiring different strategies. Research needs include personalized explanation generation that adapts to individual users, cognitive load optimization that balances informativeness with comprehensibility, trust calibration mechanisms that help users appropriately weigh AI recommendations, and studies of long-term effects of explanation use on human decision-making.

7.2 Emerging Research Directions

7.2.1 Mechanistic Interpretability

Understanding models at the level of individual components represents a promising frontier. Current advances include circuit identification in transformers [13, 16] that reveals how specific computations are implemented, automated neuron interpretation using LLMs [9] that scales human-like understanding to many neurons, superposition analysis for understanding how models represent more features than they have dimensions, and discoveries of specific circuits such as induction heads that perform identifiable functions. Nanda *et al.* [49] demonstrated that mechanistic interpretability can track learning dynamics, showing how models develop algorithmic capabilities through the emergence of specific circuit structures.

Future directions include scaling mechanistic analysis to larger models while maintaining interpretability, connecting identified circuits to high-level behaviors that matter for applications, developing automated circuit discovery methods that can map model computation comprehensively, and achieving mechanistic understanding of safety-relevant behaviors that inform alignment efforts.

7.2.2 Faithful Chain-of-Thought

Making LLM reasoning traces more faithful to actual computation represents an active research area. Current approaches include intervention-based faithfulness testing that perturbs reasoning to observe effects, using self-consistency across multiple reasoning paths as a proxy for faithfulness, training methods designed to encourage faithful reasoning, and process supervision that provides feedback on individual reasoning steps.

Key open questions include whether Chain-of-Thought can be made provably faithful rather than merely more faithful, what training methods most effectively improve faithfulness, and how to reliably detect unfaithful reasoning before deployment.

7.2.3 Interactive and Conversational Explanations

Moving beyond static explanations to dialogue-based understanding opens new opportunities. LLMs enable natural language explanation dialogue that can adapt in real time. Users can ask follow-up questions to clarify understanding. Explanations can adapt to user understanding as revealed through the conversation. Collaborative exploration of model behavior becomes possible through iterative inquiry.

However, challenges remain substantial. Maintaining consistency across dialogue turns requires careful attention to coherence. Systems must avoid confabulation in explanations, generating plausible but unfounded claims. All explanations must remain grounded in actual model behavior rather than the system's general knowledge.

7.2.4 Multimodal Explainability

As NLP models incorporate multiple modalities, new explanation needs emerge. Cross-modal attribution must explain text-image relationships and how information flows between modalities.

Vision-language models require explanation methods that span both visual and textual understanding. Audio-text explanation methods must address the temporal nature of audio alongside text. Ultimately, unified multimodal explanation frameworks will be needed to provide coherent explanations across modality boundaries.

7.2.5 Causal and Counterfactual Methods

Strengthening the causal foundations of explanations represents a promising direction. This includes developing causal models of language model behavior that go beyond correlation, counterfactual generation for text that produces meaningful alternatives, causal tracing through model components to identify where decisions are made, and intervention-based explanation methods that test causal hypotheses directly.

7.3 Responsible AI Considerations

7.3.1 Safety and Fairness

Explanations serve critical safety functions in AI systems, including detecting deceptive behavior before deployment, understanding failure modes through interpretability analysis, monitoring for capability jumps or signs of misalignment, and supporting effective human oversight of increasingly capable systems. Hendrycks *et al.* [24] contextualized these concerns within a broader taxonomy of catastrophic AI risks, emphasizing that interpretability serves as a defense against deceptive alignment and unexpected harmful behaviors. Research priorities include developing explanation methods specifically designed for safety-critical behaviors, incorporating explanation analysis into red-teaming methodologies, and using interpretability to verify alignment properties.

Explanations can also reveal and help address model biases. Current capabilities include feature attribution that reveals when models rely on demographic proxies, counterfactual analysis that exposes differential treatment across groups, and probing classifiers that detect biases encoded in model representations. Needed advances include systematic bias auditing frameworks that comprehensively evaluate models, explanation-guided debiasing methods that use interpretability to target interventions, intersectional bias detection that considers multiple demographic dimensions simultaneously, and cultural and linguistic bias analysis that extends beyond English-centric approaches.

7.3.2 Regulatory Compliance

Growing regulation demands explainability across jurisdictions. The EU AI Act establishes transparency requirements for high-risk AI systems. GDPR enshrines a right to explanation for automated decisions affecting individuals. The US Executive Order on AI introduces documentation and testing requirements. Sector-specific regulations in healthcare, finance, and other domains add additional requirements.

Compliance challenges include translating technical explanations into formats that satisfy legal requirements, balancing transparency demands with legitimate proprietary concerns, documenting model behavior comprehensively enough for regulatory review, and maintaining valid explanations as models are updated over time.

7.3.3 Democratization of Explainability

Making explanation tools accessible to broader audiences remains an important goal. Current gaps include the fact that most tools require substantial ML expertise to use effectively, computational requirements that limit access to well-resourced organizations, and costly domain-specific adaptation that prevents widespread deployment.

Needed developments include user-friendly explanation interfaces accessible to non-experts, low-resource explanation methods that can run on modest hardware, pre-built explanation pipelines that require minimal customization, and education and training resources that build community capacity.

7.4 Vision for the Future

We envision a future where:

1. **Explanations are standard**: All deployed NLP systems provide appropriate explanations by default
2. **Faithfulness is verifiable**: Robust methods exist to verify explanation accuracy
3. **Human-AI collaboration is seamless**: Explanations support effective human oversight without impeding efficiency
4. **Safety is ensured**: Interpretability tools can detect and prevent harmful AI behaviors
5. **Regulation is satisfied**: Clear standards exist for compliant AI explanations

Achieving this vision requires sustained research effort across technical, human-centered, and policy dimensions.

8 Conclusion

This survey has provided a comprehensive overview of explainability methods for Natural Language Processing, spanning from classical approaches to the latest developments in Large Language Model interpretability.

8.1 Summary of Contributions

We have made three main contributions:

1. Unified Taxonomy: We presented a comprehensive framework categorizing explanation methods along multiple dimensions—locality (local vs. global), directionality (forward vs. backward), model access (agnostic vs. specific), and output form (importance scores, rules, examples, counterfactuals, natural language). This taxonomy encompasses classical methods (LIME, SHAP, gradient-based attribution) through attention-based approaches to cutting-edge LLM techniques (Chain-of-Thought, mechanistic interpretability).

2. Practical Guidelines: We provided actionable decision frameworks for practitioners, including method selection decision trees, task-specific recommendations for classification,

sequence labeling, question answering, and generation tasks, along with evaluation protocol guidance and tool overviews. These guidelines bridge the gap between research advances and practical deployment.

3. 2024 State-of-the-Art: We surveyed recent developments including Chain-of-Thought prompting as explanation, self-explanation capabilities in LLMs, mechanistic interpretability research, and the ongoing attention-explanation debate. This coverage ensures the survey reflects current practice and emerging directions.

8.2 Key Takeaways

For researchers and practitioners, we highlight several key insights:

1. **No single method suffices**: Different explanation methods capture different aspects of model behavior. Combining multiple approaches provides more complete understanding.

2. **Faithfulness is paramount**: Plausible explanations can mislead if they don't reflect actual model reasoning. Faithfulness verification should be standard practice.

3. **Context determines appropriateness**: The right explanation method depends on the model, task, audience, and purpose. One-size-fits-all approaches fail.

4. **LLMs present new challenges**: Scale, emergent capabilities, and API-only access require new explanation paradigms. Chain-of-Thought and mechanistic interpretability represent promising but incomplete solutions.

5. **Human factors matter**: Technical explanation quality is necessary but not sufficient. Explanations must be comprehensible and useful to their intended audience.

6. **Evaluation needs standardization**: The field would benefit from consensus on evaluation protocols, benchmarks, and reporting standards.

8.3 Looking Forward

The field of explainable NLP stands at a critical juncture. As language models become more capable and widely deployed, the need for understanding their behavior grows correspondingly urgent. We identify several priorities for different stakeholders.

For researchers, the path forward requires developing scalable mechanistic interpretability methods that can handle the complexity of modern LLMs. Establishing rigorous faithfulness verification approaches remains essential, as does creating standardized evaluation benchmarks specifically designed for LLM explanations. Additionally, investigating human-centered explanation design will ensure that technical advances translate to practical utility.

For practitioners, success depends on integrating explainability into the development lifecycle from the outset rather than treating it as an afterthought. Validating explanation faithfulness before deployment prevents misleading users with plausible but unfaithful explanations. Matching explanation methods to user needs and task requirements ensures appropriate explanations reach the right audiences. Finally, monitoring explanation quality in production guards against degradation over time.

For the community at large, developing shared benchmarks and evaluation protocols will accelerate progress and enable meaningful comparisons across approaches. Creating accessible tools and educational resources will democratize explainability capabilities. Engaging with policy discussions on AI transparency ensures that technical capabilities inform regulatory frameworks. Fostering interdisciplinary collaboration with cognitive science, HCI, and law will ground technical advances in human needs and societal requirements.

8.4 Final Remarks

Explainability is not merely a technical requirement but a foundation for trustworthy AI. As NLP systems increasingly influence decisions in healthcare, law, education, and daily life, our ability to understand and verify their behavior becomes essential for both safety and accountability.

This survey has aimed to provide both a comprehensive reference for the current state of the field and practical guidance for those seeking to make NLP systems more transparent. We hope it serves researchers developing new methods, practitioners deploying explanations in real systems, and anyone seeking to understand how modern language models work.

The journey toward truly interpretable AI is ongoing. We encourage continued research, collaboration, and dialogue as the field works toward systems that are not only capable but comprehensible.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, 2020.

[2] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[3] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168, 2017.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[5] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140, 2015.

[6] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpretability via model extraction. 2017.

[7] Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. *arXiv preprint arXiv:1801.07772*, 2018.

[8] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 648–657, 2020.

[9] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. *OpenAI Blog*, 2023.

[10] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-SNLI: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[11] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. pages 1721–1730, 2015.

[12] Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. 2024.

[13] Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[14] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable AI for natural language processing. *arXiv preprint arXiv:2010.00711*, 2020.

[15] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, 2020.

[16] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

[17] Jeffrey L Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[18] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data (GDPR). Official Journal of the European Union, 2016. L 119/1.

[19] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act). Official Journal of the European Union, 2024. L 2024/1689.

[20] Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, 2023.

[21] Prashant Gohel, Priyanka Singh, and Manoranjan Mohanty. Explainable AI: current status and future directions. *arXiv preprint arXiv:2107.07045*, 2021.

[22] Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, 2020.

[23] Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, 2020.

[24] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic AI risks. *arXiv preprint arXiv:2306.12001*, 2023.

[25] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

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

[27] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. pages 624–635, 2021.

[28] Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. Aligning faithful interpretations with their social attribution. volume 9, pages 294–310, 2021.

[29] Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3543–3556, 2019.

[30] Gabriel Jarosi. Increasing tree classifier interpretability with SHAP. Medium, 2023. URL https://gabriel-jarosi-ar.medium.com/increasing-tree-classifier-interpretability-with-shap-d058e6b95076. Accessed: 2024.

[31] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, pages 2668–2677. PMLR, 2018.

[32] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR, 2017.

[33] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213, 2022.

[34] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, pages 3–10. IEEE, 2013.

[35] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 29–38, 2019.

[36] Vivian Lai, Han Liu, and Chenhao Tan. "Why is 'Chicago' deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[37] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1675–1684, 2016.

[38] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Faithful and customizable explanations of black box models. pages 131–138, 2019.

[39] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

[40] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. 2016.

[41] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[42] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1): 56–67, 2020.

[43] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing*, pages 305–329, 2023.

[44] Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural NLP: A survey. *ACM Computing Surveys*, 55(5):1–42, 2022.

[45] Sherin Mary Mathews. Explainable artificial intelligence applications in NLP, biomedical, and malware classification: A literature review. *Intelligent Computing-Proceedings of the Computing Conference*, pages 1269–1292, 2019.

[46] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. 35:17359–17372, 2022.

[47] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. 26, 2013.

[48] Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. Towards transparent and explainable attention models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4206–4216, 2020.

[49] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

[50] Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1069–1078, 2018.

[51] Nina Poerner, Benjamin Roth, and Hinrich Schütze. Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 340–350, 2018.

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

[53] Sam Ransbotham, Shervin Khodabandeh, David Kiron, François Candelon, Michael Chu, and Burt LaFountain. Expanding AI's impact with organizational learning. *MIT Sloan Management Review*, 2020.

[54] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[55] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[56] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[57] Andrew D Selbst and Julia Powles. Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242, 2017.

[58] Sofia Serrano and Noah A Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, 2019.

[59] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.

[60] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[61] Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*, 2024.

[62] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[63] Kacper Sokol and Peter Flach. Explainability fact sheets: A framework for systematic assessment of explainable approaches. pages 56–67, 2020.

[64] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[65] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 303–310, 2018.

[66] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Tristan Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Anthropic Research Blog*, 2024.

[67] Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 239–245, 2019.

[68] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. 36, 2023.

[69] U.S. Food and Drug Administration. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. FDA, 2021. URL https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[71] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2):76–99, 2017.

[72] Duyu Wang, Meng Liu, Bing Qin, and Ting Liu. Learning semantic representations of users and products for document level sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1014–1023, 2015.

[73] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

[74] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.

[75] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 11–20, 2019.

[76] Zhengxuan Wu, Atticus Arora, Zifan Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Interpretability at scale: Identifying causal mechanisms in Alpaca. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[77] Rong Yan, Alex G Hauptmann, and Rong Jin. Linear methods for text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 851–860, 2009.

[78] Linyi Yang, Yixuan Feng, Zhuoren Chen, Yuanzhi Wu, and Xirong Li. Generating plausible counterfactual explanations for deep transformers in financial text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6150–6160, 2020.

[79] Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. 35:30378–30392, 2022.

[80] Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[81] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.