

# Why and When to Use Statistical Significance in Explanation Evaluation: A Systematic Survey and Guidelines for Explainable NLP

**Fariz Ikhwantri**  
Simula Research Laboratory  
Oslo, Norway  
fariz@simula.no

**Dusica Marijan**  
Simula Research Laboratory  
Oslo, Norway  
dusica@simula.no

## Abstract

In this paper, we present a systematic survey and methodological guidelines for explanation evaluation and statistical significance testing in explainable NLP (XAI). We systematically reviewed 494 verified papers from the ACL Anthology (2020–2025) to identify how explanations are evaluated and the frequency of statistical validation. Our analysis identifies a persistent **rigour gap**: although the **number of explainable NLP papers increased approximately fivefold** during the study period, there was **no corresponding increase in the use of explanation-specific statistical testing**. Across the full corpus, only 25.30% (125/494) of studies substantiate evaluation claims with statistical significance analysis. This remains a minority practice in every surveyed year. We categorise the primary barriers to rigour into three issues: implicit generalisation, inferential reporting gap, and incomplete rigour. To address these, we propose a bipartite framework that maps research hypotheses to appropriate statistical testing regimes. Our goal is to clarify when statistical testing is necessary and to provide a guideline for more rigorous, reproducible explainability research in the LLM era.

## 1 Introduction

Explainable AI (XAI) has become an important part of NLP research. This importance is driven by the increasing trends and deployment of LLMs and agentic frameworks in high-stakes domains such as legal decision support (Spangher et al., 2024), medical text analysis (Vladika et al., 2025; Chen et al., 2025; Sohn et al., 2025), automated fact-checking (Xing et al., 2025), and financial risk (Ong et al., 2025; Long et al., 2025). In those settings, explanations are no longer auxiliary visualisations; they are often used to justify model behaviour (Lei et al., 2016; De Cao et al., 2020; Sheta et al., 2025), support human decision-making (Lu et al., 2022), detect biases (Costa-Jussà, 2019; Al-

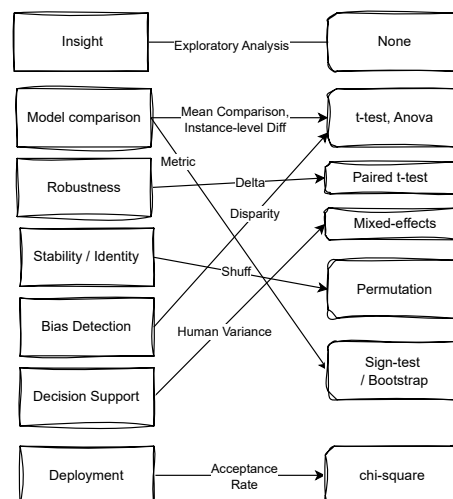


Figure 1: Our simplified mapping between the type of explainability claim and the appropriate statistical testing regime, especially in high-stakes NLP settings.

negheimish et al., 2022), and decide deployment choices (Sohn et al., 2025). As a result, the evaluation and the strength of the evidence used to support explainability claims have become a critical methodological concern.

A growing body of work proposes evaluation metrics for evaluating explanation (Jacovi and Goldberg, 2020), such as faithfulness (DeYoung et al., 2020; Atanasova et al., 2023) and plausibility (Atanasova et al., 2020; Jacovi and Goldberg, 2021; Dai et al., 2023; Wang and Atanasova, 2025). However, despite this progress, a fundamental question remains insufficiently addressed: **why should one care about rigorous quantitative evaluation of an explanation, and when does such evaluation require statistical significance analysis rather than descriptive or illustrative evidence alone?** While performance evaluation in NLP has benefited from extensive methodological scrutiny, including best practices for statistical significance testing (Dror et al., 2018). Our work provides analogous guidance for scientific rigour for

explainability evaluation (Danilevsky et al., 2020; Lyu et al., 2024), which is still fragmented and hard to comprehend.

Explanation evaluations in NLP exhibit substantial methodological heterogeneity. Some studies rely on qualitative case analyses or single-instance demonstrations. Others report aggregate metrics without measures of uncertainty, and only a subset applies statistical tests when comparing explanation methods or drawing conclusions about human impact. Essentially, the choice of evaluation strategy is rarely justified in relation to the **type of explanation claim being made**. In addition, explanations are often used to support claims, such as improved decision support, bias detection, or deployment readiness, **most of the time without appropriate statistical evidence** (Figure 3). These issues are particularly important in high-stakes domains, where weak or inappropriate inference can lead to misplaced trust (Ehsan and Riedl, 2024), fairwashing unfair outcomes (Aivodji et al., 2019; Balagopalan et al., 2022), or regulatory non-compliance (Damen et al., 2025).

This paper presents a **methodological survey of evaluation and statistical significance testing practices in explainable NLP**. We focus on clarifying **why** and **when** statistical testing is necessary. Instead of advocating universal hypothesis testing, we argue that the need for statistical evidence depends on the nature of the research claim (Sad-eqi Azer et al., 2020), the role of explanations in the evaluation pipeline, and the downstream risks associated with model use. Figure 1 summarises this perspective and our conceptual mapping. We present a **simplified bipartite mapping between explainability evaluation claims and statistical testing** regimes. We distilled the concept from our systematic survey. The figure illustrates that explainability claims form a range of the spectrum from exploratory analysis by providing illustrative insights to end-user testing for deployment justification. Each category entails different requirements for statistical rigour.

Our survey has two primary aims:

- First, we systematically review **current evaluation practices and statistical testing methods used to support explainability claims in NLP**. We first identify how often and in what form statistical testing is employed in explainable or interpretable NLP from the ACL

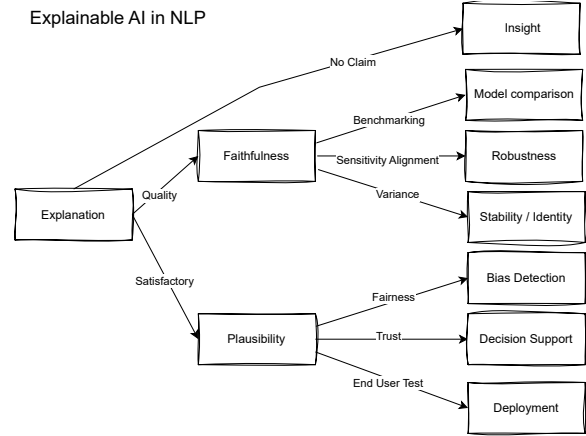


Figure 2: Decision tree concept mappings of explanation evaluation and claims.

anthology from 2020 to 2025<sup>1</sup>.

- Second, we synthesise these findings into a **simplified, decision-oriented guideline** of bipartite mapping which maps research hypotheses and explainability claims to appropriate evaluation strategies and statistical tests. We cover common research claims such as model comparison, robustness analysis, human-subject studies, and bias detection.

This work aims to bridge the gap between descriptive explainability analyses and rigorous empirical validation by providing an overview of concrete hypothesis formulation and principled statistical testing. We hope that this survey provides both a diagnostic view of current practices and a practical framework for researchers seeking to make defensible explainability claims, especially when those claims would carry real-world consequences.

## 2 Background

### 2.1 The Nature of Explanation Evaluation

We categorise XAI’s evolution from its Exploratory (Pre-2019) through the Benchmark and Evaluation Foundation (2020) and Comparative Evaluation (2021–2022) to the rise of Generative/LLMs (2023–Present). Our survey focuses on the latter two periods observed during the shift to LLMs (§A.1). While exploratory work often relied on *illustrative analysis* to demonstrate new techniques, modern explainable NLP has transitioned toward *comparative* and *deployment-oriented* claims. These include asserting improvements over baselines, iden-

<sup>1</sup>Code and Data available at [significancetestXNLP](https://significancetestXNLP.github.io/)

tifying demographic biases, or verifying robustness against adversarial noise (Alvarez-Melis and Jaakkola, 2018).

Explanation evaluation is rarely an assessment of a method in isolation; rather, it is a test of whether an interpretability method provides valid evidence in support of a specific research claim (Jacovi and Goldberg, 2020). As illustrated in the decision-tree abstraction in Figure 2, the appropriate evaluation protocol is dictated by the nature of the claim being supported. For example, a method may be highly *plausible* (aligning with human intuition) while remaining entirely *unfaithful* to the model’s actual decision process (Jain et al., 2020; Wiegrefe and Pinter, 2019). A detailed formalisation of this notation is provided in Appendix A.2.

When researchers move beyond illustration to make these broader generalisations, they implicitly assert that the observed differences are systematic and not artefacts of random variation in data sampling, model initialisation, or human rater variability (Card et al., 2020). In such cases, descriptive point-estimates are insufficient to ensure scientific reliability (Dror et al., 2018). This motivates the two central questions of this survey: *why* does explanation evaluation necessitate statistical significance testing, and *when* should such testing be applied in a principled manner?

## 2.2 Importance of Statistical Testing in Explanation Evaluation

Statistical significance testing determines whether an observed effect reflects a systematic phenomenon rather than random variation. In XAI, this is necessitated by three primary sources of noise: (i) **Data Sampling**, as metrics are computed over finite test sets; (ii) **Model Stochasticity**, including initialisation and optimisation variance; and (iii) **Evaluation Noise**, such as the instability of perturbation-based scores or the variance in human rater judgments (Søgaard, 2013; Dror et al., 2017).

Common testing paradigms in NLP include paired tests to control for instance-level variability, unpaired tests for differing samples, and resampling-based methods (e.g., bootstrapping or permutation) that estimate sampling distributions without strong parametric assumptions (Søgaard et al., 2014; Dror et al., 2018). However, explainability evaluations frequently challenge core assumptions such as independence and exchangeability. For instance, explanations generated for the same input are not independent (Alvarez-

Melis and Jaakkola, 2018; Krishna et al., 2025), perturbation-based evaluations often induce correlated measurements (Hooker et al., 2019), and human-subject studies may involve learning or fatigue effects (de Winter and Dodou, 2017). Consequently, the appropriateness of a test depends on its alignment with the experimental design, and failure to account for multiplicity across various datasets or metrics can lead to inflated false-positive rates.

We argue that significance testing in XAI serves a specific diagnostic purpose: it establishes the **strength of evidence used to support a research claim**. A p-value or confidence interval does not establish the absolute "truth" of an explanation. Still, it indicates whether the observed "faithfulness" or "plausibility" is a stable property of the system or a result of experimental noise. In the following sections, we provide empirical evidence from our survey to demonstrate the extent to which these requirements are currently met in the literature.

## 3 Survey Method and Results

### 3.1 Survey Method

First, we perform a systematic literature review by filtering the title and year (Stage-I), followed by automatic classification with GPT-4o (Stage-II), categorising paper venue (Stage-III), filter-in the paper’s main conference and findings (Stage-IV), and Manual review of statistical testing on explanation evaluation (Stage-V). The details of our systematic review can be read in Appendix C.

**Validation of LLM Classification and Human Annotation.** To validate the LLM (GPT-4o) filtering (Stage-II), we compared LLM labels against author(s)’ manual review ( $N = 593$  (input of Stage-V)), yielding 94.1% agreement and 88.2% Prevalence-Adjusted Bias-Adjusted Kappa (PABAK) (Byrt et al., 1993). Senior author overlap review ( $N = 41$ ) further confirmed alignment (PABAK=75.6%). While class imbalance led to a moderate Cohen’s  $\kappa \approx 54.8\%$ , high PABAK scores support the LLM’s efficacy as a high-recall filter. We report PABAK only as a secondary prevalence-adjusted measure and do not treat it as sufficient evidence of reliability. The Stage-V manual review output is a final verified corpus of  $N = 494$  papers (329 Main, 165 Findings). The details of our Inter-Annotator Agreement can be read in Appendix C.6.

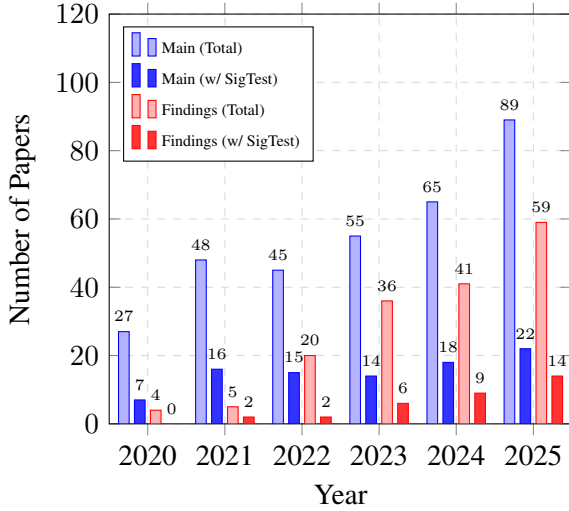


Figure 3: Trends in explainable NLP papers that evaluate explanations and those that additionally report statistical significance testing (2020–2025). In addition, a comparison of research volume and statistical rigour between the main conference and findings (2020–2025). The plot highlights that the main conference track remains the accounts for the larger share of papers in our surveyed corpus.

### 3.2 Survey Result

Across all papers that perform explanation evaluation, 125 of 494 (25.30%) substantiate explanation-specific claims with statistical significance testing. As illustrated in Figure 3, the adoption of statistical significance testing has remained consistently low, fluctuating between 21% and 34% year-over-year. While absolute research volume has increased fivefold since 2020, the proportion of studies with significant test-backed claims has **not increased correspondingly**. The yearly proportions were 22.58% (7/31) in 2020, 33.96% (18/53) in 2021, 26.15% (17/65) in 2022, 21.98% (20/91) in 2023, 25.47% (27/106) in 2024, and 24.32% (36/148) in 2025.

We emphasise that our reported 25.30% testing rate represents an **optimistic upper bound** on genuine rigour. This means our survey tracked the *presence* of significant testing, but we intentionally refrained from judging the appropriateness of the chosen tests. Furthermore, some portion of the study operates in high-stakes domains where raw evaluation data is often withheld due to privacy constraints, making front-end methodological rigour even more critical.

Our corpus includes 329 papers from the main conference tracks and 165 from *findings*. While both venues exhibit a similar overall lack of statis-

tical substantiation, we observed distinct longitudinal trends. As shown in the detailed venue review in Figure 3, the main conference consistently provided a larger and more stable absolute volume of significant testing ( $n=92$ ) compared to *findings* ( $n=33$ ).

Among the 125 papers that conduct statistical significance analysis, we performed a deep dive into the specific dimensions of explanation<sup>2</sup> Within the substantiated subset, significance testing is most prevalent in *Plausibility* (70.40%, 88/125 papers), followed by *Faithfulness* (43.20%, 54/125), while *Utility* is the indirect explanation evaluation dimension (26.40%, 33/125).

**Analysis: Ratio of paper evaluates their explanation claim backed with statistical significant testing** We performed a targeted re-analysis of the 80 samples in the main conference and findings papers that do not use significance testing. We used the notes that we captured during manual annotation. We categorised the gap into three substantive rigour-gap categories, with Illustrative Insights treated as a fourth, non-inferential category.

**Implicit Generalisation.** The largest category (46.25%) (37/80) makes competitive claims, stating that a specific method or refinement is superior to baselines. This category indicates generalisation without appropriate statistical support. Many papers mainly use an aggregate of result tables/plots without measuring if these differences exceed the noise of the experimental setup.

**Inferential Reporting Gap.** Among papers reporting correlations (Spearman/Pearson), associated p-values are absent in over 22.50% (18/80). We observed that the absence of p-values in some of the paper due to a technical hurdle in applying multiplicity corrections (e.g., Bonferroni) to complex XAI evaluation designs.

**The Incomplete Rigour.** In 21.25% of (17/80) papers, we found that papers actually conduct significance testing for performance but do not use it for the explanation evaluation. This category indicates a priority gap in how rigour is applied in the field. Although in these types of gaps of significant tests, it might also be due to the paper’s primary focus on applying new methods and evaluating performance. Only around 10% (8/80) of these papers are **illustrative insights**.

<sup>2</sup>Note that these categories are not mutually exclusive on a single paper. A single paper could perform different evaluations for multiple dimensions.

These findings motivate our bipartite mapping (Section 4). Once the role of the explanation and the evaluation signal are identified, the statistical structure (independent vs paired vs grouped/repeated) could be determined, of why and when statistical testing is necessary, and which test families are appropriate.

## 4 A Bipartite Mapping from Explanation Evaluation to Statistical Testing

Figure 4 summarises the central contribution of this paper: a bipartite mapping that connects how explanations are used in NLP research to the structure of evaluation data and, consequently, to appropriate forms of statistical testing.

Rather than mapping abstract claims directly to statistical tests, we draw a figure that encodes a sequence of decisions that arise naturally in empirical practice. Specifically, statistical significance testing becomes meaningful only after three prior questions are answered: (i) what role the explanation plays, (ii) what evaluation signal is derived from it, and (iii) what statistical structure this signal induces.

### 4.1 Explanation Evaluation Claim

We categorise the common research objectives in explainable NLP into six primary claim types with the expectation of statistical testing, summarised in Table 1. These categories range from *Illustrative Insights*, which are purely qualitative and do not require statistical substantiation, to *Deployment Readiness*, where statistical testing is mandatory to justify real-world use. The levels of statistical rigour defined in our framework: *Mandatory* and *Strongly Required*, are derived from a **risk-based evidence framework** (European Parliament and Council of the European Union, 2024; Raji et al., 2020). Following the principle of **proportional rigour** (Card et al., 2020), we argue that the evidential burden must scale with the potential responsibility of the claim. As shown in Table 1, the transition from descriptive reporting to empirical validation is driven by the **downstream risk**: claims involving human-centric decision support or bias detection carry a direct human impact in medicine (Bluhm, 2016), necessitating strict hypothesis testing. For Model Comparison, Robustness, Stability and Identity, we expect from recommended to required. We align these expectations with established normative standards in NLP (Dror et al., 2018; Sogaard et al.,

2014; Sadeqi Azer et al., 2020), where significance testing is *required* to distinguish systematic model properties from stochastic noise.

**Explainability Claims Categories.** We categorise research objectives into six primary claim categories based on the intended role of the explanation. On the left side of the mapping (Figure 4), **insights** serve qualitative analysis (Lei et al., 2016); these involve instance-level differences where statistical testing is not required. **Model comparison** arises from benchmarking (Camburu et al., 2018; DeYoung et al., 2020; Ma et al., 2024), yielding aggregate signals like mean faithfulness or sensitivity alignment scores across instances. **Robustness** (Alvarez-Melis and Jaakkola, 2018) and **stability/identity** (Kindermans et al., 2022) focus on the faithfulness of the explanation to the model’s decision-making process. In contrast, **fairness** and **bias detection** (Bolukbasi et al., 2016) focus on plausibility, typically involving the measurement of disparities in attribution mass or robustness across demographic groups. Finally, humans evaluate **decision support** through signals like acceptance rates (Yin and Neubig, 2022), task success (Zhao et al., 2024), or preference distributions.

**Statistical Distribution and Structure.** Each evaluation signal induces a specific data structure that dictates the appropriate statistical regime. Instance-level metrics in model comparisons are typically **paired** across models on identical inputs. In contrast, bias detection often induces **independent or grouped** samples. Human-subject evaluations introduce **repeated measurements** and inter-subject variability, requiring mixed-effects modelling to distinguish explanation utility from rater-specific variance. Furthermore, the data type, **continuous** (e.g., attribution scores) versus **categorical** (e.g., task success), determines whether parametric assumptions are viable. Aligning the claim with these structural properties is essential for choosing between simple paired tests and complex generalised linear mixed models (GLMMs) (Stroup et al., 2024).

### 4.2 Explanation Evaluation Claim-Statistical Testing Pairs

We map the common explanation evaluation claims to their corresponding statistical testing. Table 2 provides representative examples of these pairs. We align evaluation signals with appropriate tests and

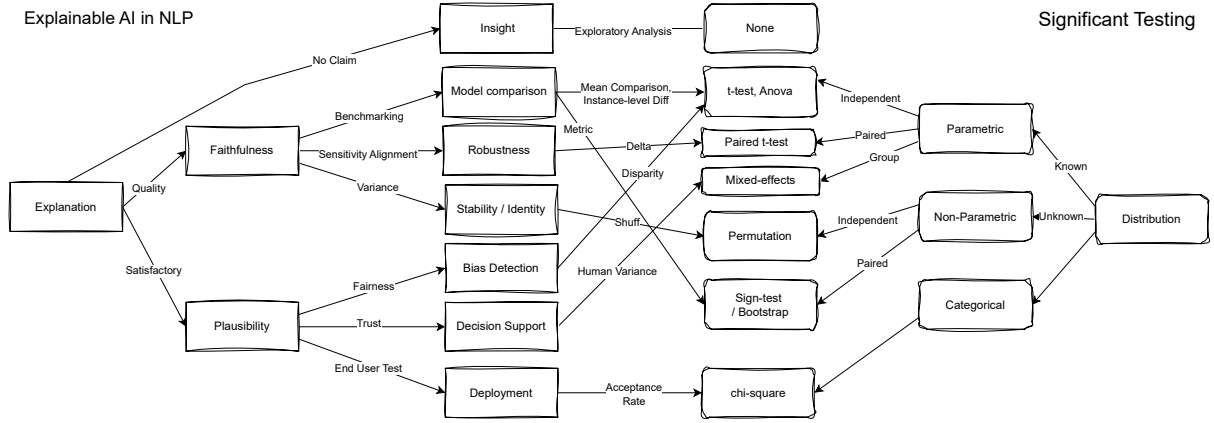


Figure 4: Merging two leaves of evaluation explanation, research claim and statistical significant testing concept from each root of the respective decision trees.

| Explainability Claim                   | Role of Explanation                                             | Evaluation Methods                                            | Stats. Test Expectation      | Foundational & Survey Exemplars                                                                                                                                   |
|----------------------------------------|-----------------------------------------------------------------|---------------------------------------------------------------|------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Illustrative insight                   | Qualitative understanding of individual predictions             | Case studies, visualisation, qualitative analysis             | Not required                 | <b>Foundational:</b> Lei et al. (2016)<br><b>Exemplars:</b> Lepori et al. (2025); Menta et al. (2025)                                                             |
| Models comparison (M)                  | Compare explanation methods or model variants                   | Perturbation-based metrics, axiomatic checks, simulatability  | Recommended                  | <b>Foundational:</b> Nguyen (2018); DeYoung et al. (2020)<br><b>Exemplars:</b> Malladi et al. (2025); Saakyan et al. (2025)                                       |
| Robustness (R), Stability/Identity (S) | Generalisation claim consistency across data, seeds, or domains | Robustness tests, stability metrics, multi-dataset evaluation | Required                     | <b>Foundational:</b> Alvarez-Melis and Jaakkola (2018)<br><b>Exemplars:</b> Bastings et al. (2022); Held and Yang (2023)                                          |
| Bias detection (B)                     | Evidence of fairness reliance on sensitive or spurious features | Perturbation analysis, white-box tests, controlled datasets   | Strongly required            | <b>Foundational:</b> Bolukbasi et al. (2016)<br><b>Exemplars:</b> Howard et al. (2025); Peled-Cohen et al. (2025)                                                 |
| Decision support (H)                   | Improved trust, usability, or task performance                  | Human-subject studies, plausibility evaluation                | Required (human variability) | <b>Foundational:</b> Hase and Bansal (2020); Bućinca et al. (2020)<br><b>Exemplars:</b> Yin and Neubig (2022); Menon et al. (2023)                                |
| Deployment readiness (D) (high-stakes) | Justification for real-world use or compliance                  | Combined faithfulness, plausibility (experts) evaluations     | Mandatory                    | <b>Foundational:</b> Raji et al. (2020); European Parliament and Council of the European Union (2024)<br><b>Exemplars:</b> Patel et al. (2025); Niu et al. (2025) |

Table 1: Taxonomy of explainability claims, evaluation roles, and expectation statistical requirements. We ground each claim type in foundational literature that established these evaluative norms and provide exemplars from our 2020–2025 verified study set.

reporting conventions based on the induced data structure.

Building on the granular taxonomy in Table 2, we categorise the most frequent explainability claims and their corresponding statistical signals observed in our study set.

**Faithfulness and Rationale Alignment** The most common evaluation involves comparing models via faithfulness or alignment with human rationales. Since these metrics (e.g., AOPC, Sufficiency) are typically calculated per instance for multiple models on the same data, the samples are inherently **paired**. Within our corpus, Malladi et al. (2025) and Zhao and Daumé III (2025) utilise paired t-tests to validate faithfulness improvements.

To handle non-normal distributions in figurative meaning understanding, Saakyan et al. (2025) employs a paired bootstrap test. When comparing multiple explanation measures, Wojciechowski et al. (2024) demonstrate the use of the non-parametric global Friedman test followed by Nemenyi post-hoc analysis. For rationale alignment, we find that reporting raw correlations is common, but best-in-class studies provide two-tailed p-values to substantiate these coefficients (Xu et al., 2025; Yang et al., 2025; Dehghanighobadi et al., 2025). In cases of fine-grained ranking, Eldifrawi et al. (2025) utilises Pearson and Kendall Tau correlations to measure annotator-evaluator agreement.

| XAI Eval.                                          | Typical signal                                 | Data structure                    | Recommended tests/models                                                                                           | Exemplars from Corpus                                        | Effect size / notes                                                              |
|----------------------------------------------------|------------------------------------------------|-----------------------------------|--------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|----------------------------------------------------------------------------------|
| Faithfulness (deletion/insertion, AOPC) [M, R, S]  | Per-instance curve; AUC or summary             | Paired across instances           | Paired permutation or Wilcoxon signed-rank on per-instance AUC; cluster bootstrap CIs                              | Malladi et al. (2025); Saakyan et al. (2025)                 | Median paired diff; BCa CI; Friedman/Nemenyi for $\geq 3$ methods                |
| Alignment with human rationales [M, H]             | Token/segment F1, IOU, rank corr. per instance | Paired; possibly multi-annotator  | Paired Wilcoxon or permutation on per-instance scores; mixed-effects for multi-annotator designs                   | Xu et al. (2025); Yang et al. (2025); Brandl et al. (2024)   | Rank-biserial; Fisher-z CI for correlations; control for length                  |
| Feature ablation/occlusion [M, R, S]               | $\Delta$ performance after ablating top-k      | Paired per instance/-dataset      | Paired tests on instance-level $\Delta$ ; bootstrap across items; meta-analysis across datasets                    | Fekete and Bjerva (2025); Held and Yang (2023)               | Standardised $\Delta$ ; random-effects meta-analysis                             |
| Ranking/Top-k overlap [M]                          | nDCG@k, Precision@k per instance               | Paired                            | Wilcoxon or sign test on per-instance scores; permutation for complex rank functionals                             | Eldifrawi et al. (2025)                                      | Win-rate with binomial CI; paired diff in nDCG                                   |
| Counterfactual/causal sufficiency-necessity [R, S] | Prediction deltas under interventions          | Paired (same item pre/post)       | McNemar for binary correctness; paired t/Wilcoxon for probabilities; randomization tests for intervention labels   | Dehghanigobadi et al. (2025); Avitan et al. (2025)           | Risk ratio or risk diff; bootstrap CI                                            |
| Simulatability/predict-the-model [R, S]            | Human accuracy with/without explanation        | Paired trials; repeated subjects  | McNemar per subject; or GLMM for binary accuracy with condition fixed effect                                       | Ye et al. (2021); Lamm et al. (2021); Hase and Bansal (2020) | Odds ratio; subject/item random effects                                          |
| Fairness disparity [B]                             | Group-wise attribution mass or metrics         | Independent/groups; covariates    | Welch t or Mann-Whitney; ANOVA/Kruskal-Wallis for $\geq 3$ groups; regression with covariates; cluster-robust SEs  | Peled-Cohen et al. (2025); Islam and Goldwasser (2025)       | Hedges g or Cliff's $\delta$ ; multiple testing correction across groups/metrics |
| Robustness/sensitivity [R, S]                      | Metric under perturbation grid                 | Repeated measures                 | Repeated-measures ANOVA or Friedman; paired tests at each level with FDR control; functional permutation on curves | Ernst et al. (2025); Bhagwatkar et al. (2025)                | Slope/area differences with CI                                                   |
| Human trust/acceptance [H, D]                      | Accept/reject, task success                    | Repeated (subject, item)          | GLMM: logistic with random intercepts for subject and item (and random slopes if suitable)                         | Zhao et al. (2024); Bao et al. (2025)                        | Odds ratio with CI; report ICC and marginal/-conditional $R^2$                   |
| Preference judgments [H, D]                        | Pairwise choices between explanations          | Repeated choices per subject/item | Bradley-Terry-Luce or Thurstone models; likelihood-ratio tests; cluster-robust SEs                                 | Menon et al. (2023)                                          | Worth parameters with CI; win-rate baseline                                      |

Table 2: Representative explanation evaluation–test pairs. Choose tests that match the estimand and dependence structure; report effect sizes and uncertainty. Exemplars denote verified papers from our corpus ( $N = 125$ ) that employ the corresponding testing regimes. The label [M, R, S, B, H, D] indicate the type of explainability claims from Table 1.

### Counterfactual Validity and Feature Ablation

Claims regarding counterfactual validity or robustness through feature ablation often involve measuring performance deltas under targeted interventions. These designs are typically **paired** per item. Avitan et al. (2025) utilise binomial tests to assess human preference for original vs counterfactual text and Chi-square tests to identify significant changes in demographic attributes. In feature-level analysis, Fekete and Bjerva (2025) apply t-tests to evaluate the impact of pruning heads in- and out-of-cluster, while Kim et al. (2025) use circuit ablation to verify the robustness of syllogistic inference.

**Fairness, Bias, and Robustness** Bias detection often involves comparing attribution mass across **independent groups**. Peled-Cohen et al. (2025) utilises logistic regression coefficient analysis to validate feature significance across different eyes, and Islam and Goldwasser (2025) uses binomial tests to determine if fairness-aware predictions exceed a random 50% baseline. For robustness to noise, researchers frequently employ bootstrapping to ensure stability. Ernst et al. (2025) use bootstrap-

ping with a strict p-value threshold ( $p < 0.01$ ) for error-tolerant attribution, while Bhagwatkar et al. (2025) perform bootstrap tests on random samples to assess the significance of different prompting strategies.

### Human Trust, Preference, and Simulatability

Evaluations involving human participants introduce complex variance structures, necessitating repeated-measures designs. To handle subject-level variability in subjective ratings, Zhao et al. (2024) utilise two-related-sample t-tests with **Bonferroni corrections** for multiple comparisons. Menon et al. (2023) illustrates the utility of the paired-sample t-test to demonstrate that users significantly prefer natural language explainers over LIME. For simulatability (the “forward test”), studies often rely on paired t-tests for simulation accuracy (Malladi et al., 2025), though Lamm et al. (2021) demonstrates a Bayesian approach using posterior predictive p-values ( $p = 0.015$ ) to measure the effectiveness of explanations. Finally, in authorship attribution, Bao et al. (2025) utilises post-hoc t-tests on accuracy deltas to quantify the human impact of

incorrect explanations.

## 5 Discussion

Our findings indicate that less than 30% of papers in our surveyed ACL corpus currently substantiate evaluation claims with significance testing. This lack of rigorous quantification has several implications for the reliability of the field and the integrity of the scientific record.

**The Complexity Barrier in the Generative Era.** The recent shift toward generative models has expanded explanation evaluation from model-internal signals such as saliency and attention toward free-text rationales, critiques, and other non-deterministic outputs. These evaluation settings often involve human judgments, correlated measures, multiple metrics, and repeated observations, creating additional statistical challenges. Our survey shows that although the volume of explainable NLP research increased substantially from 2020 to 2025, the proportion of studies using explanation-specific significance testing did not increase correspondingly. We therefore view the generative era as introducing new methodological demands for rigorous explanation evaluation.

### Reliability of Explanation Evaluation Claims

The observed trend toward descriptive-only evaluation may inadvertently facilitate "explanation-washing" or "fairwashing" (Aivodji et al., 2019; Aivodji et al., 2021; Balagopalan et al., 2022), where unverified explanations are used to generate a misplaced sense of trust in a model's behaviour. As noted by Jacovi and Goldberg (2020), claims of faithfulness in interpretability require high standards of evidence. When a study reports an explanation method as "faithful" based on raw metric differences without quantifying uncertainty, it becomes difficult to **distinguish meaningful model properties from stochastic noise**. In high-stakes contexts (Raji et al., 2020; European Parliament and Council of the European Union, 2024), this lack of rigour is not merely a technical oversight but a potential risk to end-users who may rely on these explanations for decision-making. This addresses the **why** of our research focus: **statistical testing serves as the primary safeguard against the scientific and ethical risks of 'explanation-washing,' ensuring that trust in model behaviour is grounded in evidence rather than stochastic artefacts**.

**Guidelines for Risk-Proportionate Rigour** We propose a principle of proportional rigour grounded in our bipartite mapping (§4) to address the question of **when** we should employ statistical testing. Our survey findings suggest that while descriptive evidence may suffice for *Illustrative Insights* in exploratory phases, **the requirement for statistical testing should scale with the potential downstream risk**. Specifically, when explanations move from auxiliary visualisations or text-based rationale to justifications for *Model Comparison*, *Bias Detection*, or *Deployment Readiness*, the burden of proof must shift from illustrative to appropriate statistical testing for evidence.

**Publication Bias and Preregistration.** The low frequency of significance testing in our survey suggests a systemic challenge regarding the "file drawer" problem (Rosenthal, 1979), where non-significant findings ( $p > 0.05$ ) are less likely to be reported. In XAI research, a non-significant result is a critical scientific contribution; it prevents the community from over-investing in ineffective methods and provides a necessary counter-narrative to "explanation-washing" (Aivodji et al., 2019). To mitigate the risks of *p-hacking* and post-hoc metric selection, the community would benefit from **pre-registration** of research hypotheses and evaluation plans (van Miltenburg et al., 2021). Preregistration is done by specifying metrics and statistical models *a priori*; researchers can ensure that both significant and null results are documented, providing a more transparent and reproducible account of XAI performance.

## 6 Conclusion

In this work, we presented a methodological survey of 494 explainable NLP papers (2020–2025). We identify a persistent rigour gap: despite a fivefold increase in XAI research volume, only 25.30% of the surveyed papers substantiate evaluation claims with statistical significance analysis. This trend addresses the fundamental question of **why** significance testing is necessary for explainability or interpretability: **without quantifying uncertainty, observed metric differences often reflect stochastic noise rather than systematic model properties, facilitating "explanation-washing" and misplaced trust in model behaviour**.

To address answer for **when** researchers should employ statistical significance analysis, we proposed a bipartite mapping grounded in a **down-**

**stream risk** framework. We argue that **the burden of proof must be proportional to the potential impact of the research claim**. While descriptive evidence remains appropriate for exploratory insights, statistical testing becomes essential for comparative benchmarking and mandatory for justifying deployment readiness in high-stakes contexts. By providing a guideline that pairs XAI data structures with validated testing regimes, we aim to transition the field from illustrative reporting toward a reproducible and statistically grounded science of model understanding.

## 7 Ethical Considerations

The ethical implications of our findings concern both the integrity of the scientific process and the safety of deploying AI in sensitive real-world contexts.

**High-Risk Domains and Human Impact** Our survey identified numerous studies operating in high-risk domains, including clinical diagnosis (Wu et al., 2024), legal outcome prediction (Valvoda and Cotterell, 2024), and mental health analysis (Ghosh et al., 2025; Yang et al., 2023). In these settings, a misleading or unverified explanation is not merely a technical error but a potential source of direct human harm. If a practitioner (e.g., a doctor or judge) relies on an explanation that lacks statistical grounding, the consequences can be life-altering. Rigorous, transparent evaluation is therefore not only a technical requirement but an ethical prerequisite for the deployment of AI in society.

### Human Participants and Automated Judges

Finally, many XAI evaluations rely on human subjects to measure trust and acceptance. We advocate for the ethical treatment of annotators, including fair compensation and consideration of the mental load involved in evaluating sensitive content. Furthermore, the increasing reliance on "LLM-as-a-judge" for explanation evaluation poses an ethical risk if these automated metrics mirror the internal biases of the evaluator model without human-in-the-loop verification. We recommend that automated evaluations always be validated against statistically grounded human-subject studies to avoid a "hallucination loop" in interpretability research.

## 8 Limitations

Our study has several limitations stemming from the data collection and analysis pipeline (illustrated in Figure 6).

**Keyword and Title-Based Filtering** Our initial filtering step relied on regex and rule-based methods applied primarily to paper titles. While this is an efficient way to handle the large volume of the ACL Anthology, it may omit relevant interpretability research that does not use explicit "XAI" or "explainability" keywords in the title, for instance, papers focused on "probing," "model analysis," or "diagnostic datasets" that nevertheless evaluate explanation-like signals. This filter may result in a slight under-representation of the total volume of interpretability research.

**Automatic Text Classification for Filtering** Our pipeline employs an LLM (GPT-4o) for the automated classification of the initial 2,009 candidate papers. While we acknowledge that relying on titles and abstracts may omit papers where explanation evaluation is relegated to the appendix, we argue that this approach is a logistically necessary and scientifically validated method for large-scale systematic reviews. Following the methodology of Calderon and Reichart (2025), who successfully used LLMs to characterise trends across a corpus of 2,000 relevant papers, we treat the LLM as a **high-recall librarian** rather than a final evaluator. This classification task is fundamentally one of *bibliographic metadata extraction*. We aim to identify whether authors explicitly claim an XAI focus, a task where LLMs have demonstrated over 90% agreement with human experts (Calderon and Reichart, 2025). Crucially, the "automated bottleneck" was only used to narrow the field for our candidate study set. The most critical analysis is verifying the ties between statistical tests and explanation claims. It was conducted via manual review of the 593 candidate papers (Stage V input). This stage ensures that while the "top of the funnel" benefited from automated scale, the final systematic literature review (Stage V) is grounded in human verification.

**Reliability of Automated Classification.** The Inter-Annotator Agreement (IAA) between our Stage II LLM filter and the Stage V manual review achieved substantial PABAK agreement (0.61 - 0.90) on XAI and explanation evaluation identification (Appendix C.6, Table 3), but moderate

PABAK agreement (0.41-0.60) on auditing of the significant testing (Sig. Test.) on explanation evaluation. This moderate agreement on Sig Test shows that research abstracts are often noisy: many papers mention “interpretability” as a motivation without implementing a reliable evaluation. This justifies our 100% manual full-text review to ensure accuracy.

**Venue Scope and Generalizability** Our analysis focuses on 494 verified main conference and finding papers. We focus our discussion on the main conference papers to reflect established practices. We excluded workshops for the same reason we prioritise main conference papers: Experimental Maturity. Workshops are designed for ‘work-in-progress’ and nascent ideas where the primary goal is often proof-of-concept rather than definitive, statistically substantiated benchmarking. Including workshops would have unfairly penalised the field’s statistics by mixing ‘ideation-stage’ papers with ‘confirmatory-stage’ research.

**Predominantly Single-annotation.** Most papers in the final corpus were reviewed by one human annotator: only 41 of 494 papers received an independent second review, leaving 453 papers (91.7%) single-annotated. Annotation errors or subjective judgements may therefore remain, particularly when determining whether a reported statistical test was tied specifically to explanation evaluation rather than to task performance. The structured extraction protocol and recorded evidential trail improve transparency but do not eliminate this limitation. Accordingly, our findings characterise the surveyed ACL main-conference and Findings corpus and should not be interpreted as a comprehensive estimate for explainable NLP or XAI research generally.

**Subjectivity in Gap Analysis.** The evaluation gap categories (*Implicit Generalisation*, *Inferential Reporting Gap*, and *Incomplete Rigour*) in Section 3.2 are based on a qualitative review of 80 papers. These labels involve human interpretation of author intent and methodological choice. To mitigate researcher bias, we have made our full labelling criteria and verified dataset public for external reviewing.

**Temporal Cutoff** Our survey includes papers from 2020 to 2025. Given the timing of our study and the rolling nature of anthology updates, the Dec 2025 sample represents the available data at

the time of writing. As the year progresses and more venues are archived, the absolute counts after EMNLP 2025 added to bibtex might increase. However, we believe the observed ratios and trends in statistical reporting remain representative of practices within the surveyed ACL main-conference and Findings corpus at the time of collection.

## Acknowledgments

We thank the anonymous reviewers for their constructive feedback, which helped improve the manuscript. The authors acknowledge funding from the CERTIFAI project under the European Union’s Horizon Europe programme, Grant Agreement No. 101120606.

## References

- Ulrich Aivodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. [Fairwashing: the risk of rationalization](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 161–170. PMLR.
- Ulrich Aivodji, Hiromi Arai, Sébastien Gambs, and Satoshi Hara. 2021. [Characterizing the risk of fairwashing](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 14822–14834. Curran Associates, Inc.
- Sarah Alnegheimish, Alicia Guo, and Yi Sun. 2022. [Using natural sentence prompts for understanding biases in language models](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2824–2830, Seattle, United States. Association for Computational Linguistics.
- David Alvarez-Melis and Tommi S. Jaakkola. 2018. [On the robustness of interpretability methods](#). *Preprint*, arXiv:1806.08049.
- Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. [Faithfulness tests for natural language explanations](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.
- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. [A diagnostic study of explainability techniques for text classification](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

- Matan Avitan, Ryan Cotterell, Yoav Goldberg, and Shauli Ravfogel. 2025. [A practical method for generating string counterfactuals](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3267–3286, Albuquerque, New Mexico. Association for Computational Linguistics.
- Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. 2022. [The road to explainability is paved with bias: Measuring the fairness of explanations](#). In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, page 1194–1206, New York, NY, USA. Association for Computing Machinery.
- Calvin Bao, Connor Baumler, Hal Daumé III, and Marine Carpuat. 2025. [Who’s the author? how explanations impact user reliance in AI-assisted authorship attribution](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25312–25330, Suzhou, China. Association for Computational Linguistics.
- Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. 2022. [“will you find these shortcuts?” a protocol for evaluating the faithfulness of input salience methods for text classification](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 976–991, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Rishika Bhagwatkar, Syrielle Montariol, Angelika Romanou, Beatriz Borges, Irina Rish, and Antoine Bosselut. 2025. [CAVE : Detecting and explaining commonsense anomalies in visual environments](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27110–27151, Suzhou, China. Association for Computational Linguistics.
- Robyn Bluhm. 2016. The hierarchy of evidence, meta-analysis, and systematic review. *The Routledge companion to philosophy of medicine*, pages 223–231.
- Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS’16*, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.
- Stephanie Brandl, Oliver Eberle, Tiago Ribeiro, Anders Søgaard, and Nora Hollenstein. 2024. [Evaluating webcam-based gaze data as an alternative for human rationale annotations](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6544–6556, Torino, Italia. ELRA and ICCL.
- Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. [Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems](#). In *Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI ’20*, page 454–464, New York, NY, USA. Association for Computing Machinery.
- Ted Byrt, Janet Bishop, and John B. Carlin. 1993. [Bias, prevalence and kappa](#). *Journal of Clinical Epidemiology*, 46(5):423–429.
- Nitay Calderon and Roi Reichart. 2025. [On behalf of the stakeholders: Trends in NLP model interpretability in the era of LLMs](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 656–693, Albuquerque, New Mexico. Association for Computational Linguistics.
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: natural language inference with natural language explanations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS’18*, page 9560–9572, Red Hook, NY, USA. Curran Associates Inc.
- Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. [With little power comes great responsibility](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.
- Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. [Benchmarking large language models on answering and explaining challenging medical questions](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, Albuquerque, New Mexico. Association for Computational Linguistics.
- Marta R Costa-Jussà. 2019. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, 1(11):495–496.
- Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023. [Plausible may not be faithful: Probing object hallucination in vision-language pre-training](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2136–2148, Dubrovnik, Croatia. Association for Computational Linguistics.
- Vincent Damen, Menno Wiersma, Gokce Aydin, and Rens van Haasteren. 2025. [Explainable ai for eu ai act compliance audits](#). *Maandblad voor Accountancy en Bedrijfseconomie*, 99(4):231–242.
- Marina Danilevsky, Kun Qian, Ranit Aharonov, Yanis Katsis, Ban Kawas, and Prithviraj Sen. 2020. [A](#)

- survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.
- Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. 2020. [How do decisions emerge across layers in neural models? interpretation with differentiable masking](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3243–3255, Online. Association for Computational Linguistics.
- Karin de Langis and Dongyeop Kang. 2023. [A comparative study on textual saliency of styles from eye tracking, annotations, and language models](#). In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 108–121, Singapore. Association for Computational Linguistics.
- Joost de Winter and Dimitra Dodou. 2017. *Human subject research for engineers: A practical guide*. Springer.
- Zahra Dehghanighobadi, Asja Fischer, and Muhammad Bilal Zafar. 2025. [Can LLMs explain themselves counterfactually?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7815, Suzhou, China. Association for Computational Linguistics.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. [ERASER: A benchmark to evaluate rationalized NLP models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.
- Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. [Replicability analysis for natural language processing: Testing significance with multiple datasets](#). *Transactions of the Association for Computational Linguistics*, 5:471–486.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. [The hitchhiker’s guide to testing statistical significance in natural language processing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.
- Upol Ehsan and Mark O. Riedl. 2024. [Explainability pitfalls: Beyond dark patterns in explainable ai](#). *Patterns*, 5(6):100971.
- Islam Eldifrawi, Shengrui Wang, and Amine Trabelsi. 2025. [FinGrAct: A framework for FINE-GRained evaluation of ACTIONability in explainable automatic fact-checking](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9882–9901, Suzhou, China. Association for Computational Linguistics.
- Ori Ernst, Aviv Slobodkin, Meng Cao, Sihui Wei, and Jackie CK Cheung. 2025. [Where did that come from? sentence-level error-tolerant attribution](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6400–6417, Suzhou, China. Association for Computational Linguistics.
- European Parliament and Council of the European Union. 2024. [Regulation \(EU\) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence \(Artificial Intelligence Act\)](#). 12 July 2024.
- Alvan R. Feinstein and Domenic V. Cicchetti. 1990. [High agreement but low kappa: I. the problems of two paradoxes](#). *Journal of Clinical Epidemiology*, 43(6):543–549.
- Marcell Fekete and Johannes Bjerva. 2025. [Linguistically grounded analysis of language models using shapley head values](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 850–865, Albuquerque, New Mexico. Association for Computational Linguistics.
- Alona Fyshe, Partha P. Talukdar, Brian Murphy, and Tom M. Mitchell. 2014. [Interpretable semantic vectors from a joint model of brain- and text- based meaning](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 489–499, Baltimore, Maryland. Association for Computational Linguistics.
- Mohammad Reza Ghasemi Madani and Pasquale Minervini. 2023. [REFER: An end-to-end rationale extraction framework for explanation regularization](#). In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 587–602, Singapore. Association for Computational Linguistics.
- Soumitra Ghosh, Gopendra Vikram Singh, Shambhavi, Sabarna Choudhury, and Asif Ekbal. 2025. [Just a scratch: Enhancing LLM capabilities for self-harm detection through intent differentiation and emoji interpretation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27428–27445, Vienna, Austria. Association for Computational Linguistics.
- Peter Hase and Mohit Bansal. 2020. [Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.
- Larry V. Hedges, James E. Pustejovsky, and William R. Shadish. 2013. [A standardized mean difference effect size for multiple baseline designs across individuals](#). *Research Synthesis Methods*, 4(4):324–341.

- William Held and Diyi Yang. 2023. [Shapley head pruning: Identifying and removing interference in multilingual transformers](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2416–2427, Dubrovnik, Croatia. Association for Computational Linguistics.
- Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. [A benchmark for interpretability methods in deep neural networks](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Phillip Howard, Kathleen C. Fraser, Anahita Bhiwandiwala, and Svetlana Kiritchenko. 2025. [Uncovering bias in large vision-language models at scale with counterfactuals](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5946–5991, Albuquerque, New Mexico. Association for Computational Linguistics.
- Fariz Ikhwantri, Hiroaki Yamada, and Takenobu Tokunaga. 2024. [Analyzing interpretability of summarization model with eye-gaze information](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 939–950, Torino, Italia. ELRA and ICCL.
- Tunazzina Islam and Dan Goldwasser. 2025. [Post-hoc study of climate microtargeting on social media ads with LLMs: Thematic insights and fairness evaluation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15838–15859, Suzhou, China. Association for Computational Linguistics.
- Alon Jacovi and Yoav Goldberg. 2020. [Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.
- Alon Jacovi and Yoav Goldberg. 2021. [Aligning faithful interpretations with their social attribution](#). *Transactions of the Association for Computational Linguistics*, 9:294–310.
- Sarthak Jain and Byron C. Wallace. 2019. [Attention is not Explanation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.
- Sarthak Jain, Sarah Wiegrefe, Yuval Pinter, and Byron C. Wallace. 2020. [Learning to faithfully rationalize by construction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.
- Geonhee Kim, Marco Valentino, and Andre Freitas. 2025. [Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10074–10095, Vienna, Austria. Association for Computational Linguistics.
- Pieter-Jan Kindermans, Sara Hooker, Julius Adembayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2022. [The \(Un\)reliability of Saliency Methods](#), page 267–280. Springer-Verlag, Berlin, Heidelberg.
- Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. 2025. [The disagreement problem in explainable machine learning: A practitioner’s perspective](#). *Preprint*, arXiv:2202.01602.
- Mathis Lamarre, Catherine Chen, and Fatma Deniz. 2022. [Attention weights accurately predict language representations in the brain](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4513–4529, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2021. [QED: A framework and dataset for explanations in question answering](#). *Transactions of the Association for Computational Linguistics*, 9:790–806.
- E.L. Lehmann and H.J.M. D’Abrera. 1975. *Non-parametrics: Statistical Methods Based on Ranks*. Holden-Day series in probability and statistics. Holden-Day.
- Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. [Rationalizing neural predictions](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.
- Michael A. Lepori, Michael Curtis Mozer, and Asma Ghandeharioun. 2025. [Racing thoughts: Explaining contextualization errors in large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3020–3036, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yitao Long, Tiansheng Hu, Yilun Zhao, Arman Cohan, and Chen Zhao. 2025. [FinLFQA: Evaluating attributed text generation of LLMs in financial long-form question answering](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16730–16750, Suzhou, China. Association for Computational Linguistics.
- Jinghui Lu, Linyi Yang, Brian Namee, and Yue Zhang. 2022. [A rationale-centric framework for human-in-the-loop machine learning](#). In *Proceedings of the*

- 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6986–6996, Dublin, Ireland. Association for Computational Linguistics.
- Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. [Towards faithful model explanation in NLP: A survey](#). *Computational Linguistics*, 50(2):657–723.
- Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, Shu Wu, and Liang Wang. 2024. [EX-FEVER: A dataset for multi-hop explainable fact verification](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9340–9353, Bangkok, Thailand. Association for Computational Linguistics.
- Advaith Malladi, Rakesh R Menon, Yuvraj Jain, and Shashank Srivastava. 2025. [Explaining differences between model pairs in natural language through sample learning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28737–28759, Suzhou, China. Association for Computational Linguistics.
- Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. [Few-shot self-rationalization with natural language prompts](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.
- Rakesh Menon, Kerem Zaman, and Shashank Srivastava. 2023. [MaNtLE: Model-agnostic natural language explainer](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13493–13511, Singapore. Association for Computational Linguistics.
- Tarun Ram Menta, Susmit Agrawal, and Chirag Agarwal. 2025. [Analyzing memorization in large language models through the lens of model attribution](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10661–10689, Albuquerque, New Mexico. Association for Computational Linguistics.
- Dong Nguyen. 2018. [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, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.
- Shuai Niu, Jing Ma, Hongzhan Lin, Liang Bai, Zhihua Wang, Yida Xu, Yunya Song, and Xian Yang. 2025. [Knowledge-augmented multimodal clinical rationale generation for disease diagnosis with small language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11011–11024, Vienna, Austria. Association for Computational Linguistics.
- Keane Ong, Rui Mao, Deeksha Varshney, Paul Pu Liang, Erik Cambria, and Gianmarco Mengaldo. 2025. [Deriving strategic market insights with large language models: A benchmark for forward counterfactual generation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11411–11434, Suzhou, China. Association for Computational Linguistics.
- Shruti Palaskar, Akshita Bhagia, Yonatan Bisk, Florian Metzger, Alan W Black, and Ana Marasovic. 2022. [On advances in text generation from images beyond captioning: A case study in self-rationalization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2644–2657, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Ravi Patel, Angus Brayne, Rogier Hintzen, Daniel Jaroslawicz, Georgiana Neculae, and Dane S. Corneil. 2025. [Retrieve to explain: Evidence-driven predictions for explainable drug target identification](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3328–3370, Vienna, Austria. Association for Computational Linguistics.
- Lotem Peled-Cohen, Maya Zadok, Nitay Calderon, Hila Gonen, and Roi Reichart. 2025. [Dementia through different eyes: Explainable modeling of human and LLM perceptions for early awareness](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13834–13860, Suzhou, China. Association for Computational Linguistics.
- Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. [Explain yourself! leveraging language models for commonsense reasoning](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.
- Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. [Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing](#). In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT\* '20*, page 33–44, New York, NY, USA. Association for Computing Machinery.
- Robert Rosenthal. 1979. [The file drawer problem and tolerance for null results](#). *Psychological Bulletin*, 86:638–641.
- Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. [Understanding figurative meaning through explainable visual entailment](#). In *Proceedings of the 2025 Conference of the*

- Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1–23, Albuquerque, New Mexico. Association for Computational Linguistics.
- Erfan Sadeqi Azer, Daniel Khashabi, Ashish Sabharwal, and Dan Roth. 2020. [Not all claims are created equal: Choosing the right statistical approach to assess hypotheses](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5715–5725, Online. Association for Computational Linguistics.
- Hala Sheta, Eric Haoran Huang, Shuyu Wu, Ilia Alenabi, Jiajun Hong, Ryker Lin, Ruoxi Ning, Daniel Wei, Jialin Yang, Jiawei Zhou, Ziqiao Ma, and Freda Shi. 2025. [From behavioral performance to internal competence: Interpreting vision-language models with VLM-lens](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 886–895, Suzhou, China. Association for Computational Linguistics.
- Anders Søgaard. 2013. [Estimating effect size across datasets](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 607–611, Atlanta, Georgia. Association for Computational Linguistics.
- Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Hector Martínez Alonso. 2014. [What’s in a p-value in NLP?](#) In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 1–10, Ann Arbor, Michigan. Association for Computational Linguistics.
- Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2025. [Rationale-guided retrieval augmented generation for medical question answering](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12739–12753, Albuquerque, New Mexico. Association for Computational Linguistics.
- Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. [Interpreting attention models with human visual attention in machine reading comprehension](#). In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 12–25, Online. Association for Computational Linguistics.
- Alexander Spangher, Zihan Xue, Te-Lin Wu, Mark Hansen, and Jonathan May. 2024. [LegalDiscourse: Interpreting when laws apply and to whom](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8536–8559, Mexico City, Mexico. Association for Computational Linguistics.
- Walter Stroup, Marina Ptukhina, and Julie Garai. 2024. [Generalized Linear Mixed Models: Modern Concepts, Methods and Applications](#). Chapman and Hall/CRC.
- Josef Valvoda and Ryan Cotterell. 2024. [Towards explainability in legal outcome prediction models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7269–7289, Mexico City, Mexico. Association for Computational Linguistics.
- Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. 2021. [Preregistering NLP research](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 613–623, Online. Association for Computational Linguistics.
- Juraj Vladika, Ivana Hacajova, and Florian Matthes. 2025. [Step-by-step fact verification system for medical claims with explainable reasoning](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 805–816, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yingming Wang and Pepa Atanasova. 2025. [Self-critique and refinement for faithful natural language explanations](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8481–8507, Suzhou, China. Association for Computational Linguistics.
- Sarah Wiegrefe and Yuval Pinter. 2019. [Attention is not not explanation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.
- Adam Wojciechowski, Mateusz Lango, and Ondrej Dusek. 2024. [Faithful and plausible natural language explanations for image classification: A pipeline approach](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2340–2351, Miami, Florida, USA. Association for Computational Linguistics.
- John Wu, David Wu, and Jimeng Sun. 2024. [Beyond label attention: Transparency in language models for automated medical coding via dictionary learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8848–8871, Miami, Florida, USA. Association for Computational Linguistics.
- Rui Xing, Timothy Baldwin, and Jey Han Lau. 2025. [Evaluating evidence attribution in generated fact](#)

- checking explanations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5475–5496, Albuquerque, New Mexico. Association for Computational Linguistics.
- Justin Xu, Yiming Li, Zizheng Zhang, Augustine Yui Hei Luk, Mayank Jobanputra, Samarth Oza, Ashley Murray, Meghana Reddy Kasula, Andrew Parker, and David W Eyre. 2025. [Tree-of-quote prompting improves factuality and attribution in multi-hop and medical reasoning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5605–5622, Suzhou, China. Association for Computational Linguistics.
- Jing Yang, Max Glockner, Anderson Rocha, and Iryna Gurevych. 2025. [Self-rationalization in the wild: A large-scale out-of-distribution evaluation on NLI-related tasks](#). *Transactions of the Association for Computational Linguistics*, 13:314–342.
- Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyang Kuang, and Sophia Ananiadou. 2023. [Towards interpretable mental health analysis with large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.
- Xi Ye, Rohan Nair, and Greg Durrett. 2021. [Connecting attributions and QA model behavior on realistic counterfactuals](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5496–5512, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Kayo Yin and Graham Neubig. 2022. [Interpreting language models with contrastive explanations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Lingjun Zhao and Hal Daumé III. 2025. [A necessary step toward faithfulness: Measuring and improving consistency in free-text explanations](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15799–15813, Suzhou, China. Association for Computational Linguistics.
- Lingjun Zhao, Nguyen X. Khanh, and Hal Daumé III. 2024. [Successfully guiding humans with imperfect instructions by highlighting potential errors and suggesting corrections](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 719–736, Miami, Florida, USA. Association for Computational Linguistics.

## A Explanation Evaluation

### A.1 The Developmental Eras of Explainable NLP

To provide context for the longitudinal trends identified in our survey, we categorise the evolution of explainable NLP into five distinct phases. These eras reflect a progression from qualitative probing to standardised benchmarking, followed by the current challenges of the generative era.

#### Era I: The Exploratory Roots (Pre–2019).

Early interpretability research focused primarily on *probing* the internal representations of RNN and LSTM architectures. This phase was characterised by qualitative analyses of model states (Lei et al., 2016). At this stage, explanations were largely treated as exploratory tools, and evaluation was predominantly anecdotal, focusing on alignment with human linguistic intuition.

#### Era II: The Attention Debate (2019).

The rise of the Transformer architecture triggered a pivotal debate regarding the validity of attention weights as explanations. Jain and Wallace (2019) argued that attention is not an explanation due to its lack of correlation with gradient-based importance, while Wiegrefe and Pinter (2019) countered that attention can provide faithful insights under specific constraints. This debate was foundational for the field, as it helped shift from purely visual interpretability to the requirement for *faithfulness* testing (Jacovi and Goldberg, 2020).

#### Era III: The Benchmark Foundation (2020).

Following the attention debate in 2020, the field increasingly moved toward standardised evaluation. Researchers established formal definitions for XAI properties, such as the distinction between faithfulness and plausibility (Jacovi and Goldberg, 2020), and released shared benchmarks like ERASER (DeYoung et al., 2020). These works introduced the standardised suite of metrics (e.g., sufficiency, comprehensiveness) reviewed in our survey. Because this era focused on establishing measurement protocols, reporting remained predominantly descriptive.

#### Era IV: Comparative Evaluation (2021–2022).

During this phase, the community applied the tools from the previous era to perform comparative reviews of encoder-based models (e.g., BERT, RoBERTa). This was identified in our review,

reaching a 29.66% (35/118) testing ratio in flagship venues. It established a community expectation that comparative claims regarding explanation quality should be supported by statistical significance analysis.

#### Era V: The Rise of Generative Explanation Evaluation (2023–2025).

The increasing use of LLMs has expanded explanation formats toward free-text rationales, critiques, and other generative outputs. Compared with fixed model-internal signals, these evaluations may involve additional sources of variability, including stochastic generation, human judgments, correlated measures, and multiple evaluation dimensions. These developments motivate the need for statistical procedures that account for the structure of generative explanation evaluation.

### A.2 Explanation Evaluation

Let  $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$  denote a trained NLP model with parameters  $\theta$ , mapping an input text  $x \in \mathcal{X}$  to a prediction  $y = f_\theta(x)$ . An *explanation method* is a function

$$E_\phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{Z}, \quad (1)$$

where  $\mathcal{Z}$  denotes the explanation space, e.g., token-level importance scores, free-text rationales, decision rules, or latent representations, and  $\phi$  are the parameters of the explanation method. For simplicity, we could also write  $E(x) \in \mathcal{Z}$  for the explanation produced for input  $x$  and suppress explicit dependence on model parameters and predictions unless needed.

Evaluating explanations assesses whether  $E_\phi(x, f_\theta(x))$  provides valid evidence in support of a specific claim about the behaviour of  $f_\theta$ . Importantly, explanation evaluation is inherently *claim-dependent* and must be interpreted in light of the assumptions encoded in the evaluation protocol.

### A.3 Faithfulness: Evaluation Protocols for Necessary Conditions

Faithfulness characterises the relationship between an explanation and the model’s internal decision-making process, rather than between the explanation and human intuition. As noted in the prior study (Jacovi and Goldberg, 2020), faithfulness does not admit a single formal definition and is instead accomplished through evaluation design.

We therefore define faithfulness relationally, with respect to an evaluation protocol.

#### Faithfulness: Evaluation Protocol

An explanation method  $E_\phi$  is *faithful* with respect to an evaluation protocol  $\mathcal{M}$  if it satisfies the necessary conditions imposed by  $\mathcal{M}$  under its stated assumptions about model behaviour and input perturbations.

Failure under such a protocol constitutes evidence against faithfulness, whereas passing the protocol does not by itself establish faithfulness.

Let  $\mathcal{M} = (\mathcal{A}, \mathcal{T})$  denote an evaluation protocol, where  $\mathcal{A}$  is a set of assumptions and  $\mathcal{T}$  a test procedure. Existing evaluation methods can be understood as testing different necessary conditions of faithfulness. We use the faithfulness evaluation methods from [Lyu et al. \(2024\)](#) as our guideline for structuring explanation evaluation.

**Axiomatic Evaluation** Axiomatic evaluation specifies a set of axioms  $\{a_1, \dots, a_k\}$  that a faithful explanation must satisfy. Each axiom induces a binary test:

$$\mathcal{T}_{a_i}(E_\phi, f_\theta) \in \{0, 1\}. \quad (2)$$

If  $\exists i$  such that  $\mathcal{T}_{a_i} = 0$ , the explanation is deemed unfaithful under this protocol. Passing all tests does not guarantee faithfulness.

**Predictive Power (Simulatability)** Predictive power evaluation measures whether an explanation contains sufficient information to predict the model’s output. Let  $g$  denote a simulator such that

$$\hat{y} = g(x, E_\phi(x, y)). \quad (3)$$

Higher agreement between  $\hat{y}$  and  $y$  is taken as evidence that the explanation encodes decision-relevant information, subject to assumptions about simulator expressivity and information leakage.

**Robustness Evaluation** Robustness evaluation considers small perturbations  $\delta$  applied to the input:

$$x' = x + \delta, \quad f_\theta(x') \approx f_\theta(x). \quad (4)$$

An explanation is considered robust if

$$\mathcal{D}(E_\phi(x, y), E_\phi(x', y')) \leq \epsilon, \quad (5)$$

for a suitable distance measure  $\mathcal{D}$ . This evaluation assumes that similar outputs correspond to similar reasoning processes.

**Perturbation-based Evaluation** Perturbation-based evaluation removes or alters input components according to explanation-derived importance scores. Faithfulness is inferred if the features identified as important by  $E(x)$  are retained.

$$|f_\theta(x) - f_\theta(x^{-k})| \quad (6)$$

It is large when important features  $x^{-k}$  are removed, and small when unimportant features are removed. This evaluation relies on assumptions about feature independence and distributional validity.

**White-box Evaluation** White-box evaluation compares explanations to known ground-truth explanations  $Z^*$ , obtained from transparent models or controlled datasets:

$$\mathcal{D}(E_\phi(x, y), Z^*(x)) \leq \epsilon. \quad (7)$$

#### A.4 Plausibility: Evaluation Protocols for Satisfactory Conditions

Plausibility concerns the alignment between explanations and human expectations. It is typically evaluated through human judgments or agreement with human-annotated rationales. These evaluations do not establish faithfulness unless one assumes that models reason in the same way as humans.

##### Plausibility: Satisfactory Evaluation

An explanation  $E_\phi(x, y)$  is *plausible* if it is perceived by humans as reasonable, intuitive, or consistent with domain knowledge when explaining  $f_\theta(x)$ .

**Human Annotation and Judgment.** The most common approach to plausibility evaluation relies on explicit human judgments. These include rating explanations on dimensions such as helpfulness, coherence, or correctness, as well as measuring agreement between model-generated explanations ([Marasovic et al., 2022](#); [Palaskar et al., 2022](#); [Yang et al., 2025](#)) and human-annotated rationales ([Camburu et al., 2018](#); [Rajani et al., 2019](#); [Ghasemi Madani and Minervini, 2023](#)). For example, prior work has compared explanation highlights or rationales against human-selected evidence spans, treating higher overlap as an indicator of plausibility ([DeYoung et al., 2020](#); [Yin and Neubig, 2022](#)). Other studies directly elicit human preferences between alternative explanations or assess whether explanations improve perceived trust or satisfaction ([Atanasova et al., 2020](#)).

Such evaluations are well-suited for assessing whether explanations meet user expectations, particularly in end-user-facing applications. However, **they inherently reflect human beliefs about how a task *should* be solved, rather than how the model actually reasons. As a result, human annotation primarily provides evidence of plausibility, not faithfulness** (Jacovi and Goldberg, 2020; Lyu et al., 2024).

**Behavioural Proxies: Eye-tracking.** Beyond explicit judgments, recent work has explored behavioural signals as proxies for human reasoning. Eye-tracking data, for instance, captures patterns of visual attention during reading or decision-making, and has been used to approximate which input elements humans consider salient (Sood et al., 2020; de Langis and Kang, 2023; Ikhwantri et al., 2024; Brandl et al., 2024). In plausibility evaluation, **explanations are compared against eye-gaze patterns under the assumption that explanations aligning with human attention are more intuitive or satisfactory to users.**

Eye-tracking offers a less intrusive and more fine-grained signal than explicit annotation, and is particularly attractive for studying real-time interaction with explanations. Nevertheless, gaze patterns reflect human processing strategies rather than model internals, and are influenced by task framing, interface design, and individual differences. Consequently, eye-tracking-based evaluations are plausibility-oriented and should not be interpreted as evidence of faithful model reasoning.

**Neuro-cognitive Proxies: fMRI.** A more indirect line of work draws on neuro-cognitive measurements, such as functional magnetic resonance imaging (fMRI), to study correspondences between model representations and human brain activity (Fyshe et al., 2014; Lamarre et al., 2022). In these studies, **similarity between neural activation patterns and model-derived representations or explanations is interpreted as evidence that the model encodes information in a human-aligned manner.**

While such approaches provide intriguing insights into shared representational structure, their relevance to explanation evaluation is limited by strong assumptions. Neural similarity does not imply causal equivalence of reasoning processes, nor does it guarantee that explanations are actionable or reliable for decision support. As with eye-

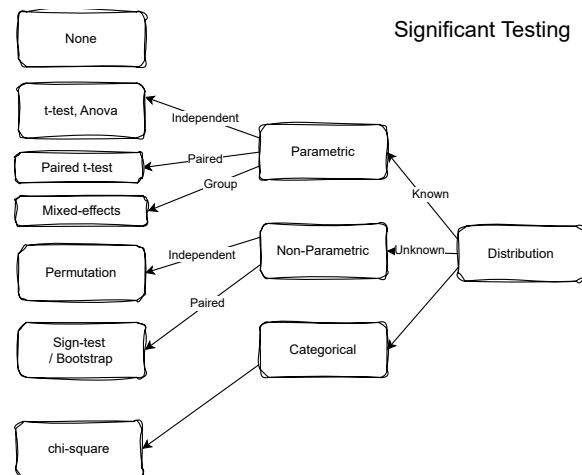


Figure 5: Decision tree concept mappings of significant testing adapted from Dror et al. (2018).

tracking, neuro-cognitive alignment should therefore be viewed as a form of plausibility evidence, reflecting alignment with human cognition rather than faithful model introspection.

**Interpretation and Limitations.** Across these approaches, a common theme is that plausibility evaluations assess how explanations resonate with humans, either explicitly through judgments or implicitly through behavioural and neuro-cognitive proxies. These methods are valuable when explanations are intended to support human understanding, trust, or interaction. However, they do not establish faithfulness unless one assumes that models reason in the same way as humans—an assumption that is generally unwarranted and explicitly discouraged in prior work.

Accordingly, plausibility should be interpreted as a *satisfactory* criterion for explanation evaluation, addressing user-centred concerns, but insufficient on its own to support claims about the internal validity or reliability of model reasoning.

## B Statistical Significant Test

Statistical significance testing provides a principled framework for determining whether an observed effect is likely to reflect a systematic phenomenon rather than random variation. In NLP, such testing is commonly used to support claims about model performance differences, robustness across datasets, or improvements over baselines. However, its role in explainability evaluation remains less clearly articulated, and is often applied inconsistently or omitted altogether.

In this subsection, we review statistical significance testing from a methodological perspective, focusing on the types of variability that arise in NLP experiments and the assumptions underlying common testing procedures. Our goal is not to advocate for universal hypothesis testing, but to clarify the conceptual foundations needed to reason about when and how statistical testing may be appropriate.

### B.1 Sources of Variability in Explainable NLP

Statistical testing is motivated by the presence of variability. **In common NLP benchmarking, variability arises from multiple sources, including random data sampling, stochastic optimisation, model initialisation, and human annotation or judgment.** Let  $m(\cdot)$  denote an evaluation metric computed over a finite sample. Even when evaluating a fixed model, repeated measurement of  $m$  may yield different values due to these sources of randomness (Søgaard, 2013; Dror et al., 2017).

In explanation evaluation, **explanation generation procedures, such as perturbation strategies and human subjective evaluation protocols, introduce additional sources of variability.** As a result, observed differences between explanation methods or evaluation conditions may reflect noise rather than systematic effects.

### B.2 Alternative Hypotheses and Null Hypothesis

Statistical significance testing formalises claims through hypotheses. Typically, a null hypothesis  $H_0$  asserts that there is no difference between two conditions, while an alternative hypothesis  $H_1$  asserts the presence of a difference or effect. For example, a null hypothesis may state that two explanation methods yield equal scores under a given evaluation metric.

Importantly, the formulation of **a hypothesis determines what conclusions can be drawn. A hypothesis test can only provide evidence against the null hypothesis under its assumptions; it cannot establish the truth of an explanation or evaluation method in an absolute sense.** The limitation in explainability research is that evaluation metrics often test necessary rather than sufficient conditions.

### B.3 Common Testing Paradigms in NLP

Several statistical testing paradigms are commonly used in NLP research (Søgaard et al., 2014; Dror

et al., 2017). *Paired tests* compare two methods evaluated on the same data instances, controlling for input-level variability. *Unpaired tests* compare methods evaluated on different samples. *Resampling-based tests*, such as bootstrap or permutation tests, estimate the sampling distribution of a statistic without strong parametric assumptions (Dror et al., 2018).

In addition, multiple-comparison corrections are often required when testing across many datasets, metrics, or explanation methods. Failure to account for such multiplicity can lead to inflated false positive rates, undermining the reliability of reported findings.

### B.4 Statistical Structure Assumptions

All statistical tests rely on assumptions, such as independence of observations, distributional properties of the test statistic, or exchangeability under the null hypothesis. Violations of these assumptions can invalidate test results or lead to misleading conclusions.

Explainability evaluations frequently challenge these assumptions. For instance, **explanations generated for the same input are not independent (Alvarez-Melis and Jaakkola, 2018; Krishna et al., 2025), perturbation-based evaluations may induce correlated measurements (Hooker et al., 2019), and human-subject studies may involve learning or fatigue effects (de Winter and Dodou, 2017).** Consequently, the appropriateness of a statistical test must be evaluated according to the experimental design.

## C Collecting Explainable NLP papers and High-Recall Filtering Strategy

### C.1 Stage I: Title-Based Identification of Explainability Literature

The first stage of our methodological survey aims to construct a high-recall candidate set of NLP papers concerned with explainability or interpretability. To this end, we perform a rule-based filtering of the ACL Anthology BibTeX dump<sup>3</sup>, relying exclusively on title-level lexical cues. This design choice reflects two considerations: (i) titles provide a standardised, curator-independent signal available across venues and years, and (ii) high-recall filtering is preferable at this stage, as subsequent stages further refine and validate inclusion.

<sup>3</sup><https://aclanthology.org/anthology.bib.gz> cutoff date 18 December 2025.

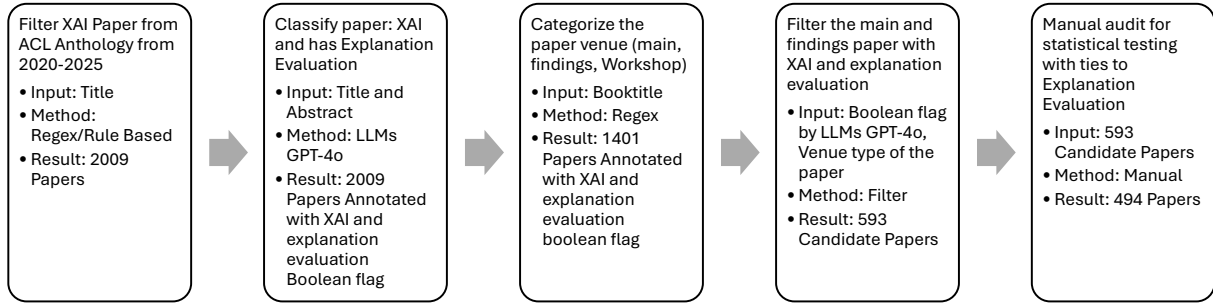


Figure 6: Pipeline of our methodological survey for statistical significant evaluation in explanation evaluation.

We extract all title fields from the anthology BibTeX file and apply a case-insensitive regular expression that targets commonly used terminology associated with explainable AI. The expression includes general terms (e.g., explanation, explainable, interpretable, transparency), evaluation-related concepts (e.g., faithful, plausibility, rationale), and families of explanation methods (e.g., saliency, feature attribution, counterfactual, example-based explanation). We additionally include the names of widely adopted explanation techniques (e.g., LIME, SHAP, Integrated Gradients, Grad-CAM) and benchmark datasets commonly associated with explanation research (e.g., ERASER, e-SNLI, CoS-E, ECQA).

A paper is included in the candidate set if its title matches at least one pattern in the regular expression. The resulting set is intentionally over-inclusive and is not assumed to consist exclusively of explainability papers. Rather, it serves as a high-recall superset for subsequent year filtering and classification. **We emphasise again that this stage is designed for recall rather than precision; inclusion in the candidate set does not imply that a paper substantively contributes to explainability methodology.**

## C.2 Stage II: Abstract-Level Classification via LLMs

In the second stage, we transition from keyword-based filtering to a semantic classification of the candidate set. To manage the volume of 2,009 papers identified in Stage I while ensuring high-quality filtering, we employ Large Language Models (LLMs) to perform text classification for semi-automatic literature review. We provide GPT-4o with the title and abstract of each paper, prompting the model to identify whether a paper (i) fo-

### Prompt

You are an NLP meta-reviewer. Decide if a paper is about explainable/interpretable AI and evaluates explanations. Use only title and abstract.

Consider evaluation signals like perturbation/deletion/insertion tests, sensitivity analysis, faithfulness, comprehensiveness, sufficiency, human agreement/evaluation, utility or simulatability.

Input:

Title: {{title}}

Abstract: {{abstract}}

Return JSON only:

```
{
  "is_xai_interpretable":
    true|false,
  "has_explanation_evaluation":
    true|false,
  "evidence": [verbatim phrases
    from abstract],
  "confidence": 0.0-1.0
}
```

Figure 7: Prompt template for automatic classification of paper. This is the input in stage-II, and the annotation is fixed manually in stage-V

**Prompt**

You are an NLP meta-reviewer.  
For a paper that evaluates explanations,  
determine if it reports  
statistical significance testing.

Look for mentions like  
"statistically significant",  
p-values, t-test, Wilcoxon,  
Mann-Whitney, ANOVA,  
Kruskal-Wallis, Bonferroni/Holm/FDR,  
bootstrap, permutation tests,  
confidence intervals.

Input:  
Title: {{title}}  
Abstract: {{abstract}}

Return JSON only:

```
{
  "has_significance_testing":
    true|false,
  "tests":
    ["e.g., paired t-test,
    Wilcoxon", "..."],
  "p_values":
    ["e.g., p<0.05", "..."],
  "evidence":
    ["verbatim phrases from abstract"],
  "confidence": 0.0-1.0
}
```

Figure 8: Prompt template for significant test presence classification in the paper. This is the input in stage-II, and the annotation is fixed manually in stage-V

cuses on Explainable AI (XAI) and (ii) performs an explicit *explanation evaluation* (e.g., measuring faithfulness, plausibility, or human utility of explanations).

This step is critical because many papers in the candidate set may mention "interpretability" as a motivation without actually evaluating the explanations themselves. The output of this stage is a subset of the 2,009 papers annotated with Boolean flags for both XAI focus and explanation evaluation.

### C.3 Stage III: Venue Categorisation

Simultaneously, we categorise the candidate papers based on their publication venue to allow for cross-venue trend analysis. Using the booktitle field from the BibTeX metadata, we apply regular expressions to categorise papers into four categories: (i) Main Conferences: ACL, EMNLP, NAACL, EACL, AACL/IJCNLP, COLING, LREC, (ii) journals: Computational Linguistics and TACL, (iii) Findings, and (iv) Workshops. This process identified 1,401 papers within these categories that also possessed the XAI and explanation evaluation flags from the previous stage. This categorisation enables us to narrow down the number of papers.

### C.4 Stage IV: Main Conference and Findings Refinement

To ensure the feasibility of manual verification and to focus our analysis on venues that reflect established evaluation norms in the NLP community, we filter the corpus to include only papers published in the *main conference and findings* tracks. By intersecting the Boolean flags from Stage II (XAI and Evaluation) with the venue tier from Stage III (main conference and findings), we identified a candidate set of 593 papers. This filtered subset represents the core corpus of peer-reviewed, high-impact research specifically focused on the evaluation of explanations. **We perform this filter by using excel sheet filter function on the output file of Stage III to create an initial file of Stage V.**

### C.5 Stage V: Manual Analysis for Statistical Testing

The last and most rigorous stage of our pipeline involves a manual, full-text analysis of the 593 papers identified in Stage IV. The inputs are 593 candidate papers for the manual review. Unlike the automated classification in Stage II, this stage is performed manually by the authors to ensure a

100% verification rate. For each paper, we manually check for:

- The presence of statistical significance testing (e.g.,  $p$ -values, confidence intervals, or bootstrap tests).
- Whether these tests are *explicitly tied* to the explanation evaluation metrics rather than just the model’s task performance (e.g., accuracy).
- The specific type of statistical test used and the evaluation claim it supports.

We perform this checking in the main body and appendices of the paper. The output of this stage is the verified dataset of 494 papers (329 main conference and 165 findings) that forms the basis for the trends and guidelines reported in the main body of this paper (see Figure 6 for a visual summary of the full process).

## C.6 Detailed Inter-Annotator Agreement

| Task/Hum.           | $N$ | Obs. Agr. | Cohen’s $\kappa$ | PABAK | Krip. $\alpha$ |
|---------------------|-----|-----------|------------------|-------|----------------|
| <b>XAI</b>          |     |           |                  |       |                |
| First Author        | 593 | 0.941     | 0.548            | 0.882 | 0.546          |
| Senior Author       | 41  | 0.878     | 0.664            | 0.756 | 0.657          |
| <b>Eval. Ident.</b> |     |           |                  |       |                |
| First Author        | 593 | 0.887     | 0.508            | 0.774 | 0.502          |
| Senior Author       | 41  | 0.829     | 0.611            | 0.659 | 0.597          |
| <b>Sig. Test.</b>   |     |           |                  |       |                |
| First Author        | 593 | 0.791     | 0.053            | 0.582 | -0.547         |
| Senior Author       | 41  | 0.780     | -0.082           | 0.561 | -0.123         |

Table 3: Stage-II and Stage-V (LLMs vs Humans) Agreement results. Explanation evaluation identification (Eval. Ident.). Prevalence Bias Adjusted Kappa (PABAK). **This table evaluates automated screening validity and does not represent human inter-annotator agreement.**

| Stage V      | $N$ | Obs. Agr. | Cohen’s $\kappa$ | PABAK        | Krip. $\alpha$ |
|--------------|-----|-----------|------------------|--------------|----------------|
| XAI          | 41  | 0.854     | 0.586            | 0.707        | 0.573          |
| Eval. Ident. | 41  | 0.805     | 0.549            | 0.610        | 0.529          |
| Sig. Test.   | 41  | 0.780     | -0.082           | <b>0.561</b> | -0.123         |

Table 4: Inter-annotator agreement between First and Senior Author. Explanation evaluation identification (Eval. Ident.). Prevalence Bias Adjusted Kappa (PABAK). For Significance Testing (Sig. Test.), the negative  $\kappa$  despite high observed agreement (78%) is an artefact of extreme class imbalance (prevalence of absent testing). PABAK provides additional context under severe prevalence imbalance, but it does not resolve the low negative chance-corrected agreement.

Table 3 provides the comprehensive validation statistics for our automated filtering pipeline. The

first author represents the review of the full candidate pool ( $N = 593$ ), and the senior author represents the validation subset for the annotation guideline and prompting to the GPT-4o ( $N = 41$ ). We report **Prevalence-Adjusted Bias-Adjusted Kappa (PABAK)** (Byrt et al., 1993) alongside Cohen’s  $\kappa$  to account for the extreme class imbalance (low prevalence of substantiated XAI papers) due to the **starting point of the review from filtered XAI Focus and Explanation Evaluation Identification (Eval. Ident.)**.

To validate our consensus underlying the manual review, we report the inter-annotator agreement (IAA) between the first and senior authors across the three core Stage V tasks ( $N=41$ ). As shown in Table 4, we achieved moderate to substantial agreement for identifying **XAI Focus** and **Evaluation Identification**, with PABAK scores of 0.707 and 0.610, respectively.

The review for **Significance Testing** presents a notable statistical case. We observed a high raw agreement of 78.0%, yet a negative Cohen’s  $\kappa$  (-0.082). This result is a classic instance of the “Kappa Paradox” (Feinstein and Cicchetti, 1990), which occurs when a category has an extremely high prevalence in the sample. Because our review was performed on a pre-filtered set of papers where the absence of significance testing is the dominant condition, the probability of agreement by chance is mathematically higher than the observed agreement, resulting in a negative  $\kappa$ .

To provide a more representative measure of consistency that accounts for this prevalence bias, we report the **PABAK score of 0.561**. PABAK provides additional context under severe prevalence imbalance, but it does not resolve the low negative chance-corrected agreement. Qualitatively, the instances of disagreement were highly informative: they almost exclusively occurred in papers categorised as *Incomplete Rigour*. In these cases, the ambiguity of the authorial reporting made it difficult to determine if a reported  $p$ -value was intended to substantiate the explanation metric or merely the model’s overall task performance. This difficulty in reaching a consensus even among experts during a manual full-text review further underscores our core finding: that current XAI reporting standards are too ambiguous to support consistent scientific meta-analysis without the standardised guidelines we propose.

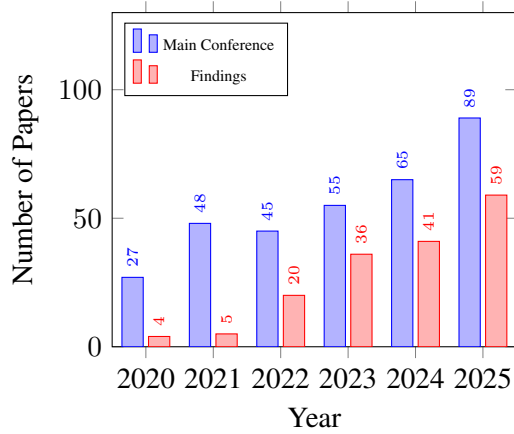


Figure 9: The plot contrasts the absolute volume of papers performing explanation evaluation (XAI+Eval) of main conference tracks versus findings (2020–2025).

## D Labelling Workflow and Criteria

This section details the workflow and criteria used for the systematic review of explainable NLP (XAI) literature. The objective of this task was to identify papers focusing on interpretability and to assess the presence and nature of their explanation evaluations rigorously.

### D.1 Labelling Workflow

We followed a structured five-step process to ensure consistency across the verified corpus ( $N = 494$ ):

1. **Search and Retrieval:** Titles were retrieved from the ACL Anthology. Full-text versions were accessed for every candidate paper identified in Stage IV.
2. **XAI Scoping:** The Abstract, Introduction, and Methods were reviewed to confirm if the paper’s primary contribution was explainability or model analysis.
3. **Evaluation Identification:** The Results and Appendix were searched for explicit evaluations of explanations (e.g., faithfulness, plausibility, or human utility).
4. **Has Significance Test:** Results, tables, figures, and appendix were checked specifically for  $p$ -values, confidence intervals, or bootstrap tests **tied to the explanation metrics**, distinguishing them from model performance tests (e.g., Accuracy/F1).
5. **Gap Categorisation :** If significance testing was absent for the explanation evaluation, the

gap was categorised into one of the four categories (see §D.3).

### D.2 Labeling Criteria

#### 1. XAI/Interpretability Focus Label for (is\_xai\_interpretable)

- **TRUE:** The paper proposes a new explanation method, investigates model internals, or uses interpretability as a core tool to diagnose model behaviour.
- **FALSE:** The paper mentions interpretability as a vague motivation in the introduction, but focuses exclusively on task performance (e.g., F1/Accuracy) without further analysis.

#### 2. Presence of Explanation Evaluation

When labelling (has\_explanation\_evaluation), a paper was marked **TRUE** if it explicitly measured the quality or truthfulness of explanations using signals such as:

**Faithfulness:** Deletion/insertion tests, perturbation analysis, AOPC, sufficiency, or monotonicity.

**Plausibility/Human-Centric:** Plausibility ratings, human evaluation of rationales, or trust/fairness user studies.

**Agreement Metrics:** Alignment with human-provided gold rationales (Token F1, IOU) or inter-annotator agreement ( $\kappa$ ).

**Simulatability:** Forward prediction or leakage-free simulatability tests.

Papers providing only a qualitative “case study” (visual inspection) without a structured evaluation protocol were marked **FALSE**.

#### 3. Statistical Significance

When labelling (has\_significance\_testing), this field identifies whether statistical significance analysis was applied specifically to the evaluation signals listed above.

### D.3 Categorising the Rigour Gap

Where significance testing was absent for an explicit explanation evaluation, annotators recorded the gap using one of the following four categories:

**Category 1: Implicit Generalisation** The paper makes competitive claims (e.g., “Method X is more faithful than Baseline Y”) based purely

on point estimates without measuring if these differences exceed experimental noise.

**Category 2: Inferential Reporting Gap** The paper reports rank correlations (Spearman/Pearson) between human and automated scores, but associated  $p$ -values are absent.

**Category 3: Incomplete Rigour** The paper conducts significance testing for the model’s task performance (e.g., Accuracy) but omits it for the explanation evaluation metrics.

**Category 4: Illustrative Insights** The evaluation is intended for purely qualitative insight, or the evaluation set is too small ( $N$ ) to support meaningful statistical inference.

#### D.4 Evidence and Metadata

For every paper manually reviewed, annotators recorded the specific XAI methods used (e.g., *Integrated Gradients*, *LIME*, or *Attention Weights*) and mechanistic interpretability, such as *logits*, *Sparse Autoencoders* and the specific evaluation signal (e.g., *Faithfulness/AOPC*) to provide an evidential trail for the meta-analysis.

### E Setup and principles of Rigorous Explanation Evaluation

**Define the estimand before the test.** For example, “paired difference in AUC of deletion curves between methods A and B,” or “odds ratio of user acceptance with vs without explanation.”

**Comparison dependence.** Most explanation metrics yield paired observations per instance and method; human and fairness evaluations introduce grouping (subjects, items, protected groups) and repeated measures.

**Prefer robust or non-parametric.** When distributional assumptions are dubious, use permutation/randomisation tests to match complex dependency structures; use mixed-effects models for repeated measures or multi-annotator settings.

**Report effect sizes suited to the scale.** Median paired difference or Hodges–Lehmann estimate (Lehmann and D’Abrera, 1975) for paired designs; odds ratios for binary outcomes; standardised mean differences (Hedges et al., 2013) or non-parametric effects (Cliff’s delta) for group contrasts.