

Are XAI Metrics Meaningful Proxies for Human Judgment? Evaluating Explanation Quality in Claim Matching

Jaroslav Kopčan¹, L'uboš Kriš¹, Qiwei Peng², Miroslav Blšták¹, Vera Schmitt^{3,4,5,6}

¹Kempelen Institute of Intelligent Technologies, ²University of Copenhagen,

³Technische Universität Berlin, ⁴Johannes Gutenberg-University Mainz,

⁵German Research Center for Artificial Intelligence (DFKI),

⁶Centre for European Research in Trusted AI (CERTAIN)

Correspondence: jaroslav.kopcan@kinit.sk

Abstract

AutoXAI frameworks automatically select and optimize explainability methods using quantitative metrics such as fidelity and plausibility. But do these metrics align with human perception of explanation quality? We investigate this question for attribution-based explanation methods, which assign token-level relevance scores to the input, in the context of claim matching, a core fact-checking task in which a bi-encoder model matches social media posts to previously fact-checked claims. Combining a large-scale evaluation of 11 XAI methods with a controlled user study, we find that methods recommended by AutoXAI metrics are perceived by users as most helpful for understanding the claim-post match, while deprioritized methods are judged as misleading. We further show that simple post-processing of explanations can improve human-perceived usefulness without degrading fidelity. Because AutoXAI recommends the top-ranked method in practice, this alignment at the decisive top-versus-bottom boundary positions these metrics as reliable indicators for distinguishing clearly strong from clearly weak methods.

1 Introduction

As neural models are increasingly deployed in settings where reliability, transparency, and user trust are essential, Explainable Artificial Intelligence (XAI) methods have emerged as a key tool for understanding and validating model behavior (Mohammadi et al., 2025; Dhaini et al., 2025b). By providing human-interpretable rationales for model predictions, XAI techniques aim to support error analysis, debugging, and informed human decision-making (Ribeiro et al., 2016; Miller, 2019). However, even a well-generated explanation is only valuable if we can reliably assess its quality, and poorly chosen evaluation metrics may lead practitioners to select high-scoring explanations that fail to aid human understanding. This

RQ1: Do quantitative fidelity and plausibility metrics identify which explanation methods users perceive as clearly more versus clearly less helpful?

RQ2: Can post-processing of attribution-based explanations improve human-perceived usefulness without degrading their fidelity to the underlying model?

RQ3: Given this alignment, can AutoXAI rely on these metrics to flag strong and weak candidate methods without costly human evaluation?

Table 1: The three research questions addressed in this work.

has led to growing interest in systematic evaluation frameworks for XAI methods, resulting in both automated, metric-based approaches and human-centered assessments.

Explanation quality is assessed along two complementary axes: quantitative metrics such as fidelity and plausibility, which are reproducible, scalable, and inexpensive to compute (Jacovi and Goldberg, 2020; DeYoung et al., 2020; Markus et al., 2021), and qualitative human evaluation, which is costlier and harder to scale but is widely regarded as the strongest evidence of real-world usefulness (Herman, 2017; Doshi-Velez and Kim, 2018). Despite broad agreement on the importance of both, whether automated metrics actually reflect human perception remains underexplored, a gap that is particularly consequential for AutoXAI approaches, which use quantitative metrics as optimization targets (Cugny et al., 2022; Kriš et al., 2025): if misaligned with human judgment, they may produce technically optimal yet practically uninformative explanations. In practice, AutoXAI recommends the top-ranked method for deployment, so the decisive question is whether that recommendation can be trusted without costly human evaluation. To investigate this, we combine a quantitative evaluation with a controlled user study, addressing the three research questions in Table 1.

Our study makes three contributions. First,

we provide a large-scale quantitative evaluation of 11 attribution-based XAI methods (which produce token-level relevance scores over the input) for claim matching, spanning fidelity and plausibility on a bi-encoder retrieval model. Second, we pair this with a controlled user study (15 participants) that tests whether these metrics track human-perceived explanation quality, and we provide an initial calibration of token-level overlap (IoU) against perceived-quality categories for this setting. Third, we show that simple post-processing improves the human-perceived usefulness of explanations for understanding the claim-post match without degrading fidelity. Together, these results show that commonly used metrics reliably separate the strongest from the weakest methods (a 1.62-point gap on a 5-point scale between AutoXAI’s top- and bottom-ranked methods), validating metric-driven selection at the tested top-versus-bottom extremes (we do not test whether finer metric differences among mid-ranked methods track human judgment) and offering a replicable methodology for other task-architecture settings.

2 Related Work

XAI and AutoXAI. XAI methods have been developed to enhance the understanding of the decision-making process in machine learning models (Mersha et al., 2024). Such methods vary from attribution-based approaches, such as SHAP (Lundberg and Lee, 2017), LIME (Ribeiro et al., 2016), and Occlusion (Zeiler and Fergus, 2014), to gradient-based approaches, such as LRP (Bach et al., 2015) and Integrated Gradient (Sundararajan et al., 2017), focusing on isolating the specific input features that drive a model’s prediction. However, selecting the optimal XAI method for a specific task remains a non-trivial challenge, as its performance is highly sensitive to hyper-parameter configurations and specific tasks. To address this, *AutoXAI* frameworks automate the selection and configuration of XAI methods by optimizing quantitative evaluation metrics, analogous to how AutoML automates model selection (He et al., 2021). Cugny et al. (2022) introduced the first such framework, automating LIME and SHAP configuration based on user-defined objectives. Shawi and Jamel (2025) proposed a meta-learning approach that recommends explanation techniques based on dataset characteristics. More recently, the o-MEGA framework (Kriš et al., 2025) expanded this scope to

cover a wide range of XAI methods and evaluation metrics, with a focus on claim matching and text classification. Complementarily, Dhaini et al. (2025a) introduce EvalxNLP, a framework that benchmarks post-hoc attribution methods for transformer-based NLP models along fidelity, plausibility, and complexity to help practitioners select an appropriate explainer. Because these frameworks rely on quantitative metrics as optimization targets, validating whether those metrics align with human judgment is critical.

XAI Evaluation. As XAI methods proliferate, evaluating the quality and reliability of generated explanations has become critical. Fidelity (also termed *faithfulness*), the degree to which an explanation accurately reflects the model’s logic, and plausibility, how comprehensible such generated explanations are to humans, are proposed to evaluate the effectiveness of XAI approaches (Nauta et al., 2023; Bello et al., 2025). While these metrics provide insightful assessment, they often fail to capture the actual utility of an explanation for end-users in practical contexts. Addressing this gap, recent research has pivoted toward human-centered evaluation, positing that an explanation’s value is best measured by its impact on user performance and trust (Dawoud et al., 2023; Nauta et al., 2023; Kazmierczak et al., 2026). For instance, Schmitt et al. (2025) focused on human-centered evaluation and evaluate different explainability features in AI systems for fact-checking, focusing on their impact on performance, perceived usefulness, and understandability. Within fact-checking specifically, Tan et al. (2025) improve explainable fact-checking by modeling claim-evidence correlations to produce more informative rationales. A parallel line of work scrutinizes the evaluation metrics themselves: Zaman and Srivastava (2025) adopt a causal lens to test whether widely used fidelity metrics measure what they purport to, cautioning that metric scores can be misleading proxies for explanation quality. In this work, we focus on the task of claim matching and systematically examine a range of XAI methods through both quantitative and qualitative analysis, assessing the alignment between human judgments and automatic evaluation metrics.

3 Measuring Explanation Quality

Quantitative evaluation offers a reproducible, scalable approach to measuring explanation quality along two core dimensions consistently identified

Claim: Barack Obama is running for president in Kenya

(a) Human rationale (majority vote)

Barack Obama will be the first man ever to rule two countries in different continents if the reports are true. The former US President has declared interest to run for the 2021 presidential election in Kenya.

(b) Occlusion (word-level) — AutoXAI top-ranked

Barack Obama will be the first man ever to rule two countries in different continents if the reports are true. The former US President has declared interest to run for the 2021 presidential election in Kenya.

(c) Occlusion + smoothing alg.

Barack Obama will be the first man ever to rule two countries in different continents if the reports are true. The former US President has declared interest to run for the 2021 presidential election in Kenya.

Attribution  high  token marked by human annotators

Figure 1: A claim–post pair from MultiClaim (post 6616): the annotated data and the explanations shown to user-study participants. Panel (a) is the majority-vote human rationale that plausibility is scored against (14 of 36 tokens marked); panels (b) and (c) reproduce the attribution maps at the colour intensities participants saw, on a scale shared across panels.

by prior work (*fidelity* and *plausibility* (Markus et al., 2021; Zhou et al., 2021)), assessed with established metrics for textual explanation evaluation (Hedström et al., 2023; DeYoung et al., 2020; Attanasio et al., 2023). Figure 1 grounds both dimensions in a concrete claim–post pair: the human rationale that plausibility is scored against, and the token attributions whose removal fidelity measures.

Fidelity measures whether an explanation accurately reflects the model’s decision-making behavior. We adopt a *functional fidelity* perspective (Jacovi and Goldberg, 2020; DeYoung et al., 2020): features highlighted as important should genuinely influence predictions when perturbed. We quantify it with two *Area Over the Perturbation Curve* (AOPC) metrics, comprehensiveness and sufficiency, which track how the prediction changes as the most important tokens are progressively removed or kept. At each perturbation step the selected tokens are deleted from the input and the shortened text is re-encoded, rather than being masked or replaced with a special placeholder token.

Plausibility measures alignment between explanations and human judgments of relevance, independent of model internals (Jacovi and Goldberg, 2020; DeYoung et al., 2020): do the highlighted features correspond to what humans consider important? We quantify it against human annotations with four metrics: token-level F1 and *Intersection over Union* (IoU), capturing exact-token overlap with human rationales, and *Area Under the Precision-Recall Curve* (AUPRC) and Average Precision Score, which are ranking-based and

more robust to annotation variability. Plausibility thus serves as a quantitative proxy for the human-perceived *quality* of an explanation, the degree to which it is interpretable and useful to humans. Our user study (Section 4) tests whether this proxy matches perceived quality; throughout, *plausibility* denotes the computed overlap metric and *perceived quality* the human judgment it aims to capture (we avoid the term *comprehensibility* to prevent confusion with the fidelity metric *comprehensiveness*).

3.1 Faithful to Models or Friendly to Humans?

Fidelity and plausibility may diverge because models operate over tokenized representations while humans reason over semantic units (Jacovi et al., 2023). Whether this tension is fundamental (Jacovi and Goldberg, 2020) or task-dependent remains empirical. Because of this tension, we explore two families of approaches for balancing the two dimensions (*post-processing* of attribution maps and adjusting the *perturbation unit of analysis*) and evaluate their effect in our experiments (Sections 6 and 6.2). In our claim-matching setting we find the two dimensions to be largely complementary rather than antagonistic: methods with high fidelity also achieve high plausibility, and both approaches can further improve perceived quality without compromising fidelity.

Post-processing. Post-processing refines attribution maps to improve perceived quality while preserving the attribution signal, adapting techniques from Herdt et al. (2025) and Smilkov et al. (2017). All strategies operate on the 1D sequence of token-

level scores $\mathbf{a} = (a_1, \dots, a_n)$ produced by an XAI method, after normalization by one of four schemes (L2, mean, min-max, or z-score with tanh squashing). Figure 1(c) shows the effect on our running example, and Table 4 reports results for every configuration.

Gaussian smoothing replaces each a_i by a Gaussian-weighted average over a local window, following SmoothGrad’s noise-reduction principle (Smilkov et al., 2017):

$$\tilde{a}_i = \frac{\sum_{j \in \mathcal{N}(i)} \exp\left(-\frac{(i-j)^2}{2\sigma^2}\right) a_j}{\sum_{j \in \mathcal{N}(i)} \exp\left(-\frac{(i-j)^2}{2\sigma^2}\right)} \quad (1)$$

where $\mathcal{N}(i)$ is the window centered at position i and σ controls the degree of smoothing: larger σ suppresses high-frequency noise but blurs the boundary between salient and non-salient regions.

Median smoothing instead takes the median within a sliding window of size w , $\tilde{a}_i = \text{median}(\{a_j : j \in \mathcal{N}_w(i)\})$, which is more robust to outlier tokens and preserves sharp transitions while still suppressing isolated noise.

Span convolution search, adapted from the saliency-map verbalization framework of Feldhus et al. (2023), slides a fixed-size window over the sequence, ranks windows by cumulative attribution, and retains the top-ranked contiguous spans. This groups individual token attributions into coherent multi-token spans (e.g., named entities), aligning explanation granularity with human semantic units.

Clustering (K-means, GMM, DBSCAN) discretizes continuous scores into a small number of importance categories. K-means and GMM partition tokens into k clusters by attribution value; DBSCAN identifies dense regions without a predefined k , separating salient groups from background noise. All three reduce cognitive load by turning continuous attribution gradients into categorical levels. Per-method analysis is in Appendix D.

Perturbation unit of analysis. A second approach adjusts the perturbation unit used during attribution computation. For perturbation-based methods, perturbing whole words rather than individual subword tokens can yield explanations more naturally aligned with human semantic units, though at the cost of reduced granularity and applicability to perturbation-based methods only. In our experiments we instantiate this by comparing word-level and token-level variants of Occlusion (Table 3), letting us test directly whether aligning

the perturbation unit with human word boundaries yields more human-aligned explanations.

3.2 Attribution Polarity Alignment with Human Annotations

Attribution methods assign continuous positive and negative scores to tokens, whereas human annotators in our task identify only *supporting evidence* as binary masks. Since our fidelity analysis (Section 6.1) shows that negative attributions exhibit low functional fidelity in bi-encoder models, we focus on positive attributions for plausibility evaluation. Prior to computing plausibility metrics, we (1) filter out negative attributions and (2) binarize remaining scores via top- k selection to match human annotation format.

4 Qualitative Evaluation

While quantitative metrics provide automated, reproducible measurements of explanation quality, they cannot fully capture whether explanations are genuinely useful to end users in real-world tasks. Qualitative evaluation serves as a critical verification step that assesses the alignment between computed metrics and human subjective perception, directly addressing the research questions posed in Table 1. By combining quantitative and qualitative approaches, we can validate that our computational metrics correlate with human-perceived explanation quality and identify gaps between theoretical explanation quality and practical user experience (Herman, 2017; Nauta et al., 2023).

4.1 User Study Design

We employ a mixed experimental design: XAI method is a *within-subjects* factor (each participant rates all five methods, each on a different subset of pairs). The method-to-pair mapping is counter-balanced across five participant groups, ensuring that every method is evaluated on every pair across groups while no participant sees more than one explanation per pair. Each claim-post pair is evaluated with a single attribution method per participant, avoiding within-pair comparison effects (see Appendix A for design advantages). The evaluation followed a two-step procedure: (1) participants independently highlighted text they considered relevant to the claim, then (2) were presented with computed explanations and rated their quality with respect to their own selections and the task context. Attribution method presentation order was randomized.

Part I - Quantitative Evaluation (AutoXAI)

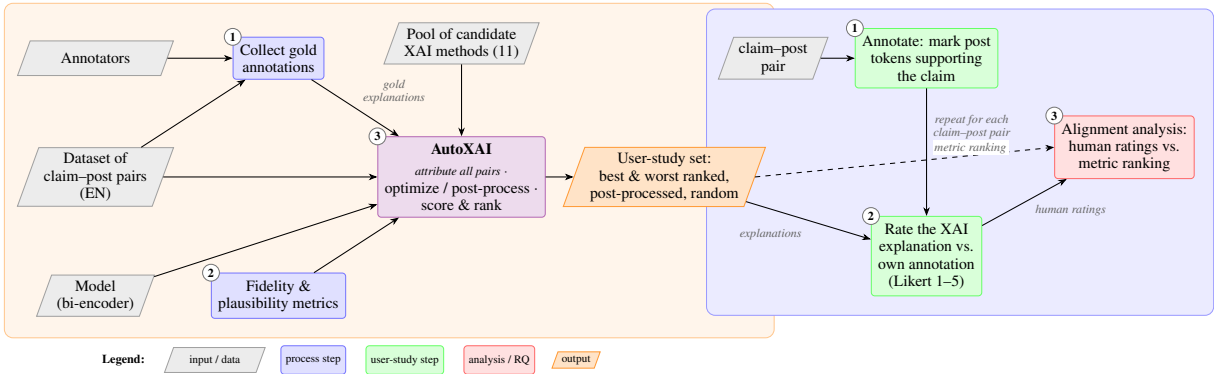


Figure 2: Overview of the two-part pipeline. **Part I (AutoXAI)** attributes, post-processes, and scores explanations from a pool of candidate XAI methods, then ranks them by fidelity and plausibility (using gold annotations as ground truth). **Part II (user study)** shows the best- and worst-ranked methods, their post-processed variants, and a random baseline to 15 participants, who annotate the supporting tokens and rate each explanation on a 5-point Likert scale (mixed design: method within-subjects). The alignment analysis tests whether human ratings match the metric ranking.

Evaluation Metrics. We measure token-level overlap via IOU (Intersection Over Union) between user-selected text passages and computed explanations. Subjective evaluation uses a 5-point Likert scale (Appendix B) at the instance level, answering: “Was the most important information with respect to the claim present in the XAI-highlighted text passages?” Results are aggregated across methods to identify quality patterns.

Quality Control. Attention-check questions were inserted every 15 samples, participants failing multiple checks were excluded. Random explanations (tokens selected randomly, matching real explanation length distributions) were included as a baseline to verify that attribution methods outperform chance.

5 Methodology and Data

Task. We evaluate explanation quality on the *claim matching* task, a core component of automated fact-checking. Given a social media post containing a checkable claim, the system retrieves the most semantically similar previously fact-checked claim from a repository. Explanations aim to highlight which parts of the post support the retrieved match. Figure 2 provides an overview of entire research pipeline, illustrating both the quantitative evaluation via the AutoXAI framework and the qualitative human evaluation protocol.

Subset	Pairs	Used for
English evaluation pool	6,386	Fidelity
Human-annotated	512	Plausibility
User-study sample	100	User study
LLM-annotated	5,874	HP search

Table 2: How the data is used. The two annotated subsets partition the evaluation pool ($512 + 5,874 = 6,386$); the user-study sample is drawn from the human-annotated one.

Data. We use the MultiClaim dataset¹ (Pikuliak et al., 2023), a multilingual resource for claim retrieval. Our experiments focus on the English subset. Fidelity evaluation is conducted on 6,386 post-claim pairs. For plausibility evaluation, we use a human-annotated subset of 512 post-claim pairs selected stratified by length, in which annotators highlight post spans that explain why a given claim is a match. Majority voting is used to ensure annotation quality. Table 2 summarizes how each subset is used.

Model. All claim matching experiments use the multilingual E5-large bi-encoder (Wang et al., 2024), which embeds posts and claims into a shared vector space.

Methods. We evaluate 11 explanation methods spanning gradient-based, perturbation-based, and decomposition approaches, including Input×Gradient, Saliency, Guided Backpropagation, Occlusion, Feature Ablation, LIME, SHAP

¹<https://zenodo.org/records/7737983>

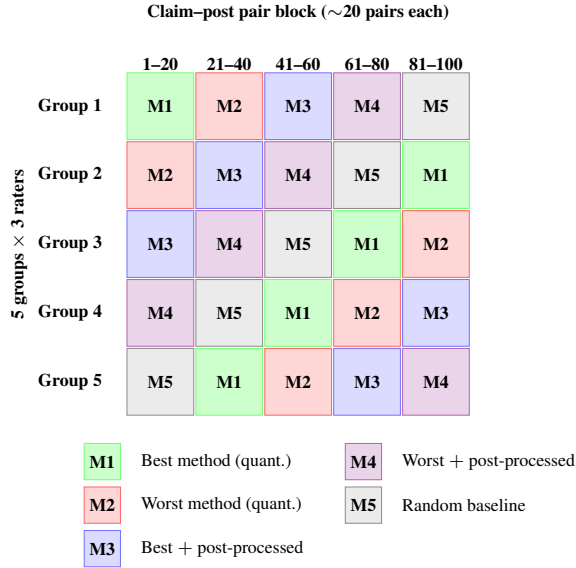


Figure 3: Human evaluation design: 15 participants in 5 groups of 3, each annotating all 100 claim-post pairs. The five methods (M1–M5) are counterbalanced across groups, so each covers ~20 pairs per group and every pair block is seen under a different method across groups. Pair blocks are illustrative.

variants, GAE, and LRP. Each method produces token-level relevance scores. To align with human annotations, which mark only supporting evidence, we retain positive attributions for plausibility evaluation.

Hyperparameter Optimization Protocol. XAI methods for human evaluation are selected based on quantitative performance. Specifically, we include the best and worst-performing methods according to computed metrics, along with post-processed variants designed to improve interpretability (Section 3.1). Hyperparameters are optimized on 5,874 post-claim pairs using LLM-generated explanation annotations (Appendix G), maximizing a single combined objective that weights fidelity and plausibility equally (0.5/0.5), i.e., the aggregate score used to rank methods, while the final evaluation reported in Table 3 is conducted exclusively on the human-annotated subset. Full optimization details and configurations are provided in the Appendix H.

User Study Protocol. We conduct a controlled human evaluation (Figure 3) with 15 annotators in 5 groups of 3. Five XAI methods are evaluated (Section 7), the best and worst methods from quantitative evaluation, their post-processed variants, and a random baseline. Each group evaluates

the same 100 post-claim pairs, with XAI methods randomly assigned to pairs such that all methods are evaluated equally across groups. Method-to-pair mappings differ across groups, controlling for text difficulty while enabling inter-annotator agreement analysis. The annotation tool is described in Appendix I.

6 Quantitative Evaluation Results

Table 3 presents a quantitative evaluation of 11 XAI methods across multiple evaluation metrics. For each method, hyperparameters are tuned using the o-MEGA autoXAI framework (Kriš et al., 2025).

To facilitate a unified comparison across evaluation dimensions, we report a single aggregated score computed via hierarchical averaging. Metrics are first averaged within each dimension (fidelity and plausibility) using equal weights, and the two resulting dimension-level scores are then averaged to obtain the final score. All metrics are bounded in the [0,1] range under their standard definitions and therefore require no normalization prior to aggregation. We use this aggregation consistently for all methods.

Occlusion-based methods demonstrate strong overall performance, with word-level Occlusion achieving the highest combined evaluation, followed closely by GAE Explain and token-level Occlusion. Gradient-based methods such as Gradient SHAP, Guided Backpropagation, and Saliency show comparable performance around 0.709, while ConservativeLRP exhibits the lowest performance, primarily due to weaker plausibility scores despite competitive fidelity metrics.

6.1 Results on Fidelity Evaluations

We evaluated fidelity using AOPC metrics measuring comprehensiveness and sufficiency, with each perturbation step modifying 10% of the text length (Figure 4, detailed interpretation in Appendix C). Both metrics consistently show that negative attributions exhibit low fidelity across all methods, reflecting structural properties of bi-encoder architectures - independent encoding without token-level interaction (Reimers and Gurevych, 2019; Thakur et al., 2021), cosine similarity on mean-pooled embeddings where token removal induces only local perturbations (Zhou et al., 2022), and anisotropic embedding spaces with high baseline similarity (0.885 between empty posts and claims) (Ethayarajh, 2019). Positively attributed tokens show

Method	Fidelity		Plausibility				Final
	AOPC _{suff}	AOPC _{comp}	Avg Prec	AUPRC	Token F1	Token IoU	
Occlusion (word-level)	0.988	0.903	0.548	0.698	0.620	0.496	0.775
GAE Explain	0.991	0.911	0.542	0.693	0.597	0.477	0.770
Occlusion (token-level)	0.999	0.915	0.740	0.688	0.585	0.456	0.767
Lime	0.997	0.915	0.607	0.647	0.557	0.422	0.749
Feature Ablation	1.000	0.916	0.677	0.657	0.529	0.397	0.743
Kernel SHAP	0.990	0.911	0.542	0.640	0.542	0.409	0.740
Gradient SHAP	0.981	0.911	0.510	0.608	0.475	0.335	0.709
Guided Backprop	0.984	0.912	0.535	0.603	0.473	0.332	0.709
Saliency	0.983	0.912	0.526	0.604	0.473	0.332	0.709
Input X Gradient	0.978	0.916	0.526	0.600	0.465	0.326	0.705
ConservativeLRP	0.962	0.923	0.472	0.454	0.303	0.165	0.625

Table 3: Performance for different XAI methods across fidelity and plausibility dimensions. Methods are ranked by the final aggregated score.

higher fidelity as they align directionally with claim embeddings. Consequently, explanations should focus on positive evidence, and we restrict plausibility evaluation to positive attributions accordingly.

6.2 Results on Plausibility Evaluation

We present plausibility results in three steps: (1) quantitative plausibility scores across methods (Table 3), (2) the impact of post-processing on both plausibility and fidelity (Table 4), and (3) how these quantitative differences relate to human-perceived quality categories (Table 7 in Section 7).

Plausibility evaluation in Table 3 reveals substantial variation in alignment with human annotations. Token-level metrics (F1, IoU) are sensitive to annotation variability, while ranking-based metrics (AUPRC, Average Precision) are more robust. Despite these limitations, a clear trend emerges: perturbation-based methods outperform gradient-based approaches, with word-level Occlusion achieving the strongest alignment and ConservativeLRP showing poor performance, motivating our qualitative user study to validate these quantitative patterns.

Overall, plausibility scores are substantially lower than fidelity across all methods. However, our qualitative evaluation (Table 7) reveals that even highly rated explanations achieve approximately 60% IOU, indicating that current plausibility metrics may underestimate actual usefulness.

Table 4 reports both dimensions for the two focal methods on the same human-annotated subset used for plausibility evaluation (Table 2). Post-processing substantially enhances plausibility while simultaneously improving fidelity. This occurs because AOPC metrics select top- k to-

Post-processing	Norm.	Avg	Fid.	Plaus.
<i>ConservativeLRP</i>				
GMM	L2	0.743	0.939	0.547
K-means	L2	0.743	0.939	0.547
DBSCAN	L2	0.726	0.938	0.514
Span convolution	Mean	0.678	0.933	0.423
Smooth median	Mean	0.678	0.941	0.415
Smoothing alg.	Mean	0.671	0.941	0.401
None	L2	0.580	0.877	0.283
<i>Occlusion</i>				
Smoothing alg.	Mean	0.837	0.951	0.724
Smooth median	Mean	0.829	0.954	0.705
K-means	Mean	0.824	0.954	0.693
GMM	Mean	0.823	0.954	0.692
Span convolution	Mean	0.817	0.954	0.679
None	L2	0.776	0.893	0.659
DBSCAN	L2	0.769	0.943	0.596

Table 4: Post-processing results for ConservativeLRP and Occlusion across post-processing strategies and normalization schemes. *None* is the unprocessed baseline. Best scores per XAI method are bolded within each group.

kens by attribution magnitude at each perturbation step, post-processing concentrates attribution mass onto coherent spans, making this selection more precise and thus increasing the measured fidelity signal without altering the underlying attribution mechanism (Appendix D). For ConservativeLRP, clustering-based methods improved performance from 0.580 to 0.743 (28% gain), for Occlusion, smoothing achieved 0.837 (8% gain).

7 Alignment Analysis by Qualitative Evaluation

Our method selection mirrors how AutoXAI is used in practice: pipelines rank methods by quantitative metrics and practitioners deploy the top-ranked one. We therefore test whether that recommendation is

reliable at the decisive extremes: selecting Occlusion (highest combined fidelity and plausibility, Table 3) as the AutoXAI-recommended method and ConservativeLRP as the worst, each paired with a post-processed variant (Occlusion+smoothing alg. and ConservativeLRP+GMM, Table 4) to test whether post-processing improves user experience, plus a random baseline as a sanity check, directly probing the upper and lower bounds of metric-driven recommendation quality.

Importantly, the quantitative plausibility scores in Table 3 were computed against 512 majority-voted annotation pairs from the MultiClaim dataset (Section 5), whereas the user ratings reported below were collected from 15 independent participants on a *separate* 100-pair subset with no annotator overlap. The alignment we report is therefore a predictive relationship across distinct human judges and data.

Detailed results from the user study are provided in Tables 5 and 6. Occlusion, the top-ranked method, received the highest mean rating from participants (3.28), while ConservativeLRP received a substantially lower mean rating (1.66). The difference of 1.62 points on the 5-point scale represents a large perceptual gap, demonstrating that the method AutoXAI metrics would recommend is indeed perceived as substantially more useful than the method it would deprioritize. A Friedman test confirms a significant effect of explanation method on ratings ($\chi^2(4) = 48.0, p < 10^{-9}$), and Holm-corrected pairwise Wilcoxon signed-rank tests partition the five methods into three significance tiers (superscripts in Table 5; full matrix in Appendix F): the two Occlusion variants are statistically indistinguishable from each other and each significantly outranks all remaining methods, whereas ConservativeLRP forms a significantly lower tier than every other method.

Notably, ConservativeLRP ($M = 1.66$) was rated significantly *below* the random baseline ($M = 1.92$; tiers c vs. b in Table 5): confidently selective but wrong highlights appear to mislead users more than visibly arbitrary ones. This nuances our central claim: metrics reliably separate strong from weak methods, but a low score alone does not distinguish a merely *less useful* method from one *actively worse* than no explanation, which matters when using metrics as a lower-bound filter in AutoXAI pipelines. Post-processing lifts ConservativeLRP+GMM ($M = 2.02$) to statistical parity with random (both tier b), indicating the

Method	Mean	Median	Count
Occlusion+smooth alg.	3.46 ^a	4.0	247
Occlusion	3.28 ^a	3.0	231
ConservLRP+GMM	2.02 ^b	2.0	257
Random	1.92 ^b	2.0	246
ConservativeLRP	1.66 ^c	1.0	256

Table 5: Summary statistics of user ratings across each utilized explanation method (5-point Likert scale: 1=Misleading, 5=Superior). Superscripts denote homogeneous subsets: methods that do not share a letter differ significantly (Holm-corrected pairwise Wilcoxon signed-rank tests, $p < 0.05$), yielding three tiers $a > b > c$. The full pairwise matrix is given in Appendix F.

deficit stems partly from calibration and presentation rather than absent signal.

Impact of Post-processing on Perceived Quality. Post-processing measurably improved human ratings. Smoothed Occlusion gained 0.18 points (5.5%) over the original, with the median rising from 3.0 to 4.0 and 59.1% of ratings at level 4 (“Similar and equally useful”). ConservativeLRP+GMM improved more in relative terms (0.36 points, 21.7%; median 1.0→2.0), shifting the distribution from predominantly “Misleading” toward “Somewhat misleading”: less problematic, though still far from ideal.

Perceived Utility of Explanations. Attribution highlights are the interface through which fact-checkers judge *why* a post was matched to a claim. Our study measures whether users *perceive* these highlights as capturing claim-relevant information (their perceived usefulness) rather than whether they improve downstream verification performance; the Likert anchors are worded accordingly (Appendix B). By this measure, 86.3% of Occlusion+smoothing explanations received ratings ≥ 3 , compared to only 18.4% for ConservativeLRP (Table 6): AutoXAI-recommended explanations are predominantly perceived as useful for understanding the claim-post match, while deprioritized methods are overwhelmingly perceived as misleading. Whether this perceived usefulness translates into faster or more accurate verification is a natural next step.

Token Overlap (IOU) by Rating Category. We analyze the relationship between IOU (computed token-level overlap) and Likert ratings (holistic subjective judgments, Appendix B), computing mean IoU per rating category with top- k binarization (k

Method	Rating 1	Rating 2	Rating 3	Rating 4	Rating 5
Occlusion+smooth alg.	12 (4.9%)	15 (6.1%)	61 (24.7%)	146 (59.1%)	9 (3.6%)
Occlusion	6 (2.6%)	31 (13.4%)	86 (37.2%)	93 (40.3%)	12 (5.2%)
ConservLRP+GMM	62 (24.1%)	123 (47.9%)	62 (24.1%)	5 (1.9%)	1 (0.4%)
Random	84 (34.1%)	104 (42.3%)	53 (21.5%)	3 (1.2%)	2 (0.8%)
ConservativeLRP	140 (54.7%)	69 (27.0%)	41 (16.0%)	6 (2.3%)	0 (0.0%)

Table 6: Distribution of Likert scale ratings (1-5) for each explanation method. Modal ratings are bolded. Rating 1 = Misleading, Rating 5 = Superior.

Rating	N	Mean (SD)	Median
1	324	18.1% (16.8)	15.8%
2	393	23.5% (18.7)	20.0%
3	329	36.4% (21.4)	33.3%
4	308	61.9% (25.5)	61.5%
5	36	37.4% (27.4)	33.3%

Table 7: IOU between computed explanations and human annotations across rating categories. SD shown as percentage points. Rating 4 shows the strongest alignment.

set to the number of tokens each annotator selected, controlling for explanation length).

Table 7 shows a clear progression: Category 1 (Misleading) explanations reach only 18.1% IOU with human selections, rising to 61.9% for Category 4 (Similar and equally useful), and ratings correlate significantly with overlap overall (Spearman $\rho = 0.335$, $p < 1.3 \times 10^{-38}$). Subjective ratings are thus grounded in token-level alignment, though the moderate coefficient indicates overlap is not the sole driver; presentation clarity and cognitive load also shape judgments. This yields an initial, setting-specific calibration of overlap levels to perceived quality for bi-encoder claim matching, not a task-general threshold.

Inter-Annotator Agreement. Agreement on text selection was substantial (mean $\kappa = 0.667$), validating the human annotations used for plausibility. Agreement on quality ratings was moderate (mean $\kappa = 0.389$), consistent with XAI evaluations where quality judgments are inherently preference-laden (Herman, 2017; Kazmierczak et al., 2026). Despite this subjectivity, ratings systematically track computed token overlap (Table 7), and between-method effect sizes (1.6-1.8 points) far exceed individual disagreement. Detailed analysis is in Appendix E.

Statistical Robustness. Consistent with our mixed design (method as a within-subjects factor), a repeated-measures ANOVA on participant-level means ($N=15$) confirms a highly significant

method effect ($F(4, 56) = 92.55$, $p < 10^{-23}$, partial $\eta^2=0.869$), corroborated by a non-parametric Friedman test ($\chi^2(4) = 48.0$, $p < 10^{-9}$). The large effect size (Cohen’s $f=0.919$) reflects our deliberate selection of ranking extremes; we use it for prospective sample planning (G*Power sensitivity analysis, Appendix F) rather than uninformative post-hoc power (Hoenig and Heisey, 2001). Together with the IOU calibration and the 1.62-point between-method gap, this supports the robustness of our comparative findings.

8 Conclusion

This work evaluates whether quantitative XAI metrics align with human perception of explanation quality for attribution-based methods in claim matching using a bi-encoder retrieval architecture. Our findings address the research questions posed in Table 1:

RQ1. The top-ranked method (Occlusion) is consistently rated most useful for understanding the claim-post match, while the lowest-ranked (ConservativeLRP) is judged misleading (a 1.62-point gap), and ratings correlate significantly with token-level overlap ($\rho=0.335$), which the IOU calibration (Table 7) maps to perceived-quality categories. *Both metrics capture meaningful aspects of human-perceived quality in this setting.*

RQ2. Post-processing improves fidelity and plausibility together (Table 4) and lifts human ratings, with smoothed Occlusion rated highest overall. *No prohibitive tradeoff exists: interpretability can be improved at the presentation level without sacrificing fidelity.*

RQ3. Quantitative metrics reliably separate the strongest from the weakest methods, flagging clearly helpful and inadequate ones without costly human evaluation. Since practitioners deploy the top-ranked method, we validate this top-versus-bottom boundary (the decision that matters in practice) and leave finer-grained discrimination among mid-ranked methods to future work.

Limitations

Task and model specificity. Our analysis targets claim matching with a bi-encoder retrieval architecture. Although claim matching is central to automated fact-checking, the observed metric–human alignment may not generalize to other tasks (e.g., generation, structured prediction) or to architectures with richer token-level interactions. Cross-encoders with joint attention could yield qualitatively different attributions (richer negative contributions and finer-grained importance) but require encoding every query–document pair at inference, which is prohibitive for large-scale retrieval against thousands of fact-checked claims (Reimers and Gurevych, 2019; Thakur et al., 2021). Bi-encoders remain dominant precisely because pre-computed claim embeddings enable sub-linear search, so our evaluation covers the architecture most relevant to deployed systems; extending it to cross-encoder reranking stages is a natural next step.

Scope of human evaluation. Our user study involves 15 participants rating 100 pairs across 5 conditions. Although the between-method effect sizes are large (1.62 points on a 5-point scale) and robust under repeated-measures testing (Section 7), a sensitivity analysis treating the observed effect (Cohen’s $f=0.919$) as a pilot estimate suggests $N\approx 45$ participants for 0.80 power in future replications. Participants were university students and volunteers rather than professional fact-checkers; while identifying claim-relevant text needs no domain expertise, professionals may hold different thresholds for explanation utility. Because we test only the ranking extremes (best vs. worst method, post-processed variants, and a random baseline) to validate the AutoXAI recommendation scenario, our results show that metrics separate these extremes but not whether mid-scoring methods (e.g., Gradient SHAP, LIME, GAE) track human judgment monotonically, a natural extension. For the same reason, the IOU-to-perceived-quality calibration in Section 7 is a preliminary baseline specific to this task and architecture; transferable overlap thresholds would require broader evaluation and larger samples.

Metric coverage. We evaluate alignment for fidelity (AOPC-based) and plausibility (overlap-based) metrics, which are the two dimensions most commonly used in AutoXAI method selection. Other evaluation dimensions such as stability, ro-

bustness, or contrastivity (Nauta et al., 2023) were not included and may exhibit different alignment patterns with human judgment.

Downstream task performance. While our Likert scale captures task-relevant judgments (e.g., whether explanations are “misleading” or “useful” for understanding claim-post matches), we do not measure downstream effects on decision accuracy, speed, or confidence calibration. Once metric-human alignment is established, as our results suggest, the natural next step is a controlled study measuring whether AutoXAI-recommended explanations improve fact-checker performance relative to no-explanation or low-ranked-explanation conditions, building directly on the method selection validated here.

Ethical Considerations

AI Writing Assistance Large language models were used to assist with grammar checking and sentence structure improvements.

Acknowledgments

This work was partially funded by European Union, under the project lorAI - Low Resource Artificial Intelligence, GA No. 101136646, <https://doi.org/10.3030/101136646>, by the Slovak Research and Development Agency under the Contract no. APVV-22-0414, and by the European Union NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I01-03-V04-00007.

References

- Lora Aroyo and Chris Welty. 2015. [Truth is a lie: Crowd truth and the seven myths of human annotation](#). *AI Magazine*, 36(1):15–24.
- Ron Artstein and Massimo Poesio. 2008. [Inter-coder agreement for computational linguistics](#). *Computational Linguistics*, 34(4):555–596.
- Giuseppe Attanasio, Eliana Pastor, Chiara Di Bonaventura, and Debora Nozza. 2023. [ferret: a framework for benchmarking explainers on transformers](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 256–266.
- Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. [On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation](#). *PloS one*, 10(7):e0130140.

- Marilyn Bello, Rosalís Amador, María-Matilde García, Javier Del Ser, Pablo Mesejo, and Óscar Córdón. 2025. [The level of strength of an explanation: A quantitative evaluation technique for post-hoc xai methods](#). *Pattern Recognition*, 161:111221.
- Domenic V Cicchetti. 1994. [Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology](#). *Psychological assessment*, 6(4):284–290.
- Robin Cugny, Julien Aligon, Max Chevalier, Geoffrey Roman Jimenez, and Olivier Teste. 2022. [AutoXAI: A framework to automatically select the most adapted XAI solution](#). In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 315–324.
- Karam Dawoud, Wojciech Samek, Peter Eisert, Sebastian Lapuschkin, and Sebastian Bosse. 2023. [Human-centered evaluation of XAI methods](#). In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 912–921. IEEE.
- 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.
- Mahdi Dhaini, Kafaite Zahra Hussain, Efstratios Zaradoukas, and Gjergji Kasneci. 2025a. [EvalxNLP: A framework for benchmarking post-hoc explainability methods on NLP models](#). In *Joint Proceedings of the xAI 2025 Late-Breaking Work, Demos and Doctoral Consortium*, volume 4017, pages 281–288. CEUR-WS.org.
- Mahdi Dhaini, Tobias Müller, Roksoliana Rabets, and Gjergji Kasneci. 2025b. [Adoption of explainable natural language processing: Perspectives from industry and academia on practices and challenges](#). In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 719–730.
- Finale Doshi-Velez and Been Kim. 2018. [Considerations for evaluation and generalization in interpretable machine learning](#). In *Explainable and interpretable models in computer vision and machine learning*, pages 3–17. Springer.
- Kawin Ethayarajh. 2019. [How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings](#). 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 55–65. Association for Computational Linguistics.
- Nils Feldhus, Leonhard Hennig, Maximilian Dustin Nasert, Christopher Ebert, Robert Schwarzenberg, and Sebastian Möller. 2023. [Saliency map verbalization: Comparing feature importance representations from model-free and instruction-based methods](#). In *Proceedings of the First Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 30–46, Toronto, Canada. Association for Computational Linguistics.
- Xin He, Kaiyong Zhao, and Xiaowen Chu. 2021. [Autotml: A survey of the state-of-the-art](#). *Knowledge-based systems*, 212:106622.
- Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. 2023. [Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond](#). *Journal of Machine Learning Research*, 24(34):1–11.
- Rudolf Herdt, Maximilian Schmidt, Daniel Otero Baquer, and Peter Maaß. 2025. [Smooth deep saliency](#). In *Tenth International Conference on Multimedia and Image Processing (ICMIP 2025)*, page 3.
- Bernease Herman. 2017. [The promise and peril of human evaluation for model interpretability](#). *arXiv preprint arXiv:1711.07414*.
- John M. Hoenig and Dennis M. Heisey. 2001. [The abuse of power: The pervasive fallacy of power calculations for data analysis](#). *The American Statistician*, 55(1):19–24.
- 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. Association for Computational Linguistics.
- Alon Jacovi, Hendrik Schuff, Heike Adel, Ngoc Thang Vu, and Yoav Goldberg. 2023. [Neighboring words affect human interpretation of saliency explanations](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11816–11833, Toronto, Canada. Association for Computational Linguistics.
- Rémi Kazmierczak, Steve Azzolin, Eloïse Berthier, Anna Hedström, Patricia Delhomme, David Filiat, Nicolas Bousquet, Goran Frehse, Massimiliano Mancini, Baptiste Caramiaux, Andrea Passerini, and Gianni Franchi. 2026. [Benchmarking XAI explanations with human-aligned evaluations](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(44):37491–37500.
- Klaus Krippendorff. 2019. [Content Analysis: An Introduction to Its Methodology](#), 4th edition. Sage.
- Ľuboš Kriš, Jaroslav Kopčan, Qiwei Peng, Andrej Ridzik, Marcel Veselý, and Martin Tamajka. 2025. [o-MEGA: Optimized methods for explanation generation and analysis](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 634–642, Suzhou, China. Association for Computational Linguistics.

- J. Richard Landis and Gary G. Koch. 1977. [The measurement of observer agreement for categorical data](#). *Biometrics*, pages 159–174.
- Scott M Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. 2021. [The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies](#). *Journal of biomedical informatics*, 113:103655.
- Melkamu Abay Mersha, Khang Nhut Lâm, Joseph Wood, Ali K. AlShami, and Jugal Kalita. 2024. [Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction](#). *Neurocomputing*, 599:128111.
- Tim Miller. 2019. [Explanation in artificial intelligence: Insights from the social sciences](#). *Artificial intelligence*, 267:1–38.
- Hadi Mohammadi, Robert A. Bagheri, Anastasia Giachanou, and Daniel L. Oberski. 2025. [Explainability in practice: A survey of explainable NLP across various domains](#). *arXiv preprint arXiv:2502.00837*.
- Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. 2023. [From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI](#). *ACM Computing Surveys*, 55(13s):1–42.
- Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023. [Multilingual previously fact-checked claim retrieval](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16477–16500.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). 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 3982–3992. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ["Why Should I Trust You?": Explaining the Predictions of Any Classifier](#). In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.
- Vera Schmitt, Isabel Bezzaoui, Charlott Jakob, Premtim Sahitaj, Qianli Wang, Arthur Hilbert, Max Upravitelev, Jonas Fegert, Sebastian Möller, and Veronika Solopova. 2025. [Beyond transparency: Evaluating explainability in ai-supported fact-checking](#). In *Proceedings of the 4th ACM International Workshop on Multimedia AI against Disinformation*, pages 63–72.
- Radwa El Shawi and Leila Jamel. 2025. [Autoxai: a meta-learning approach for recommendation of explanation techniques](#). *Scientific Reports*, 15(1):41912.
- Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. [Smoothgrad: removing noise by adding noise](#). *Preprint*, arXiv:1706.03825.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. [Axiomatic attribution for deep networks](#). In *International conference on machine learning*, pages 3319–3328. PMLR.
- Xin Tan, Bowei Zou, and Ai Ti Aw. 2025. [Improving explainable fact-checking with claim-evidence correlations](#). In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 1600–1612, Abu Dhabi, UAE. Association for Computational Linguistics.
- Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. [Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310. Association for Computational Linguistics.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. [Multilingual E5 text embeddings: A technical report](#). *arXiv preprint arXiv:2402.05672*.
- Kerem Zaman and Shashank Srivastava. 2025. [A causal lens for evaluating faithfulness metrics](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 29425–29449, Suzhou, China. Association for Computational Linguistics.
- Matthew D Zeiler and Rob Fergus. 2014. [Visualizing and understanding convolutional networks](#). In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13*, pages 818–833. Springer.
- Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022. [Problems with cosine as a measure of embedding similarity for high frequency words](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–423. Association for Computational Linguistics.
- Yao Zhou, Haonan Wang, Jingrui He, and Haixun Wang. 2021. [From intrinsic to counterfactual: On the explainability of contextualized recommender systems](#). *arXiv preprint arXiv:2110.14844*.

A Mixed Design Advantages

Our user study employs a mixed experimental design: XAI method is a within-subjects factor (each participant rates all five methods), while the method-to-pair mapping is a between-subjects factor (varying across the five groups). Presenting only one method per claim-post pair per participant offers several advantages:

- **Eliminates within-pair comparison effects.** Participants evaluate each claim-post pair with a single explanation, without being influenced by alternative explanations for the same content, avoiding anchoring bias or carryover effects at the pair level
- **Naturalistic evaluation.** Mirrors real-world usage where users encounter a single explanation for each piece of content, providing ecologically valid insights into explanation effectiveness
- **Reduces cognitive load.** Participants focus on evaluating one explanation per claim-post pair without needing to compare multiple explanation styles for the same content, reducing fatigue
- **Enables inter-annotator agreement analysis.** Within each group, multiple participants (3 per group) evaluate identical method-pair combinations, enabling reliability assessment of human judgments
- **Controls for text difficulty.** Because method-to-pair mappings vary across groups, each pair is evaluated under different methods, separating method effects from item difficulty

B Annotation Guidelines with Likert Scale for Explanation Quality and Participant Information

B.1 Participant Demographics

The user study was conducted with 15 annotators comprising university students and volunteers who participated without financial compensation. All participants provided informed consent and were briefed on the study objectives and procedures before beginning the annotation task.

B.2 Annotation Guidelines

Participants received the following instructions for the annotation task:

Background The purpose of this annotation is to evaluate and improve the quality of automated text explanations generated by Explainable AI (XAI) methods in the context of fact-checking social media posts. We have created a pipeline that generates visualizations of important parts of social media posts that align with fact-check claims. "Visualization" (Explanation from XAI method) contains two types of colors: white and green, where green indicates the relevance of the text to the fact-check claim, and white indicates that the given part of the post is not relevant to the fact-check claim. Less intense shades of green indicate that the word is less important for the explanation.

Task Your task will be divided into two parts:

1. **Annotate text:** The first part will contain the post-claim pair. As part of our evaluation of the explanation, please highlight which part of the post matches with the fact-check claim. This means that you will search for words and context in social media posts that were the reason for linking the claim to the post. You will mark by clicking on specific words. (You can also hold and drag the correct parts of the words.) We recommend using a PC mouse for this task.
2. **Evaluate the explanation created by AI:** Your evaluation will be on a Likert scale from 1 to 5, where you will rate whether the explanation sufficiently expresses the match between the claim and the post. Try to interpret the post as the user intended.

Participants rated explanation quality using a 5-point Likert scale with the anchors depicted in Table 8:

C AOPC Fidelity Metrics Interpretations

The following section analyzes the AOPC fidelity metrics for XAI methods depicted in the figure 4

For Comprehensiveness The evaluation protocol shows the beginning with the complete post-claim pair and progressively removes tokens according to their attributed importance. It first removes negatively attributed tokens (on the left side of the x-axis) and then positively attributed tokens (on the right side). Removing tokens with negative attributions fails to impact similarity for all methods. Instead, similarity remains approximately

Score	Description
1	Misleading - AI text passages are misleading and do not highlight anything relevant to the claim
2	Somewhat misleading - AI text passages may partially make sense, but they do not highlight relevant information
3	Different but somewhat useful - AI text passages highlight partially different sections from mine, but the information is somehow relevant
4	Similar and equally useful - AI text passages' highlights are almost identical to mine and equally helpful
5	Superior - AI text passages highlight relevant information I missed or present it more clearly

Table 8: Likert scale for evaluating XAI explanation quality

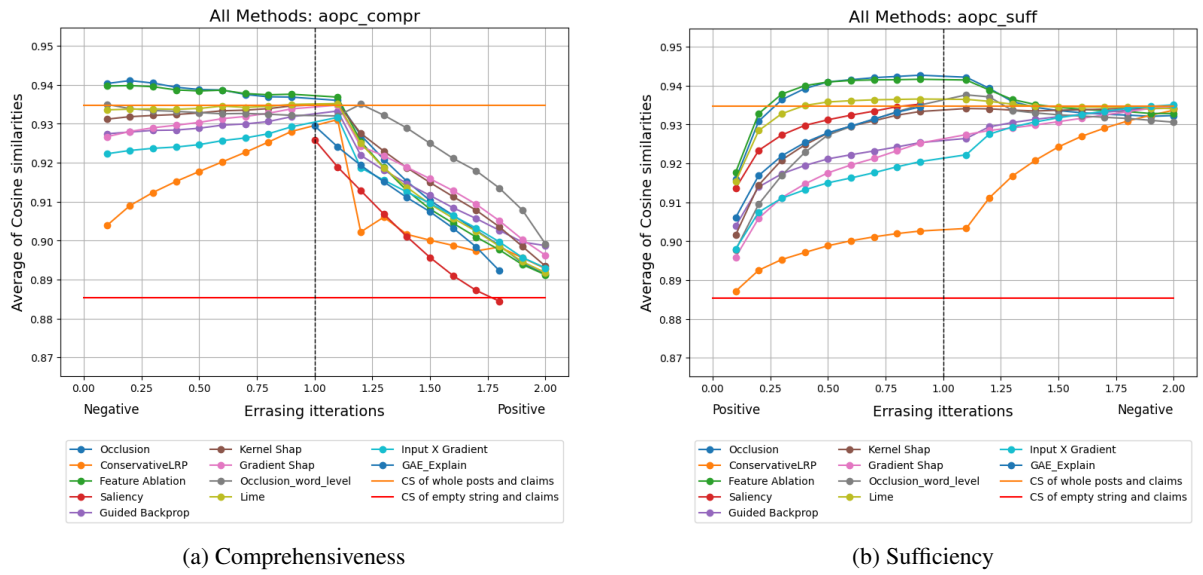


Figure 4: AOPC fidelity metrics for XAI methods. Tokens are removed/added in order of attribution magnitude (negative first, positive second for comprehensiveness, positive first, negative second for sufficiency) in 10% chunks.

constant or shows minimal, inconsistent changes. After the transition, when methods begin removing positively attributed tokens, we observe the degradation in similarity while saliency achieves the steepest decline. demonstrating it successfully identifies tokens that contribute to similarity.

For Sufficiency The evaluation protocol begins with an empty string post and progressively adds tokens according to their attributed importance. Firstly, positively attributed tokens (on the left side of the x-axis) are inserted, then negatively attributed tokens follow (on the right side). The sufficiency results demonstrate similar fidelity compared to comprehensiveness, particularly for identifying positive features. Attribution methods can successfully identify these features because the model’s learned behavior involves building up se-

mantic representations from constituent tokens. Adding tokens with high positive attributions thus reconstructs the semantic content the model relies upon for matching.

C.1 Baseline Similarity and Embedding Space Geometry

The baseline similarity (orange line) serves as a reference point for interpreting meaningful changes in cosine values. The high baseline similarity (0.885) between empty strings as inputs and claims reflects the concentration-of-measure phenomenon in high-dimensional spaces. In the high-dimensional embedding spaces (multilingual-e5-large, T5-large), learned representations cluster within a narrow region, resulting in a narrow operational range of cosine similarity units. This concentration is a

natural consequence of both the curse of dimensionality and the model’s learned semantic space, where texts share common linguistic and thematic features. The XAI methods must therefore identify features that contribute to subtle but meaningful distinctions within this compressed similarity range.

C.2 Understanding Negative Attribution Behavior

The uniform failure of negative attributions across all methods, including both gradient-based and model-agnostic approaches, suggests that this is not method-specific but rather reflects the fundamental characteristics of the semantic similarity task and the bi-encoder architecture.

Bi-Encoder Architecture and Token Independence. What attribution methods identify as ‘negative’ features are tokens whose embedding representations are misaligned with the claim embedding in vector space. Because bi-encoders process texts independently, these tokens cannot be semantically oppositional in the way cross-encoders can represent contradiction through joint attention. They simply contribute embeddings pointing in less favorable directions.

Mean Pooling Effects. Under the bi-encoder formulation, removing individual tokens, even those intuitively expected to reduce semantic similarity, has a limited effect on the resulting sentence embedding. Mean pooling aggregates token embeddings into a single centroid, and token removal generally induces small, local perturbations rather than substantial changes in embedding direction. Within the compressed similarity regime (baseline 0.885), removing negatively attributed tokens is unlikely to shift the post-embedding sufficiently far from the claim embedding to be considered a meaningful change.

Positive vs. Negative Attribution Fidelity. At the same time, positively attributed tokens tend to be directionally aligned with the claim embedding and therefore contribute additively to cosine similarity. This makes their removal more likely to induce a measurable decrease in similarity despite the same pooling and geometric constraints. This asymmetry explains why positive attributions demonstrate higher fidelity than negative attributions in bi-encoder architectures.

D Post-processing Analysis Details

This appendix expands on the results in Table 4; the post-processing strategies themselves are defined in Section 3.1.

D.1 ConservativeLRP Post-processing Analysis

For ConservativeLRP, clustering-based methods (GMM and K-means) with L2 normalization achieved the best performance, improving the baseline method’s average score from 0.580 to 0.743. This represents a 28% improvement, driven primarily by dramatic gains in plausibility (from 0.283 to 0.547) alongside modest fidelity improvements (from 0.877 to 0.939).

The enhancement in both dimensions challenges the common assumption that improving human-perceived quality necessarily degrades fidelity to model behavior. The improvements are particularly pronounced for ConservativeLRP, which initially exhibited weaker plausibility, suggesting that post-processing serves as a crucial bridge between raw model attributions and human-interpretable explanations for methods with initially poor perceived quality.

D.2 Occlusion Post-processing Analysis

For Occlusion, smoothing-based methods with mean normalization proved most effective, with the smoothing algorithm achieving the highest average score of 0.837. Post-processing improved the baseline Occlusion method from 0.776 to 0.837, representing an 8% gain. Notably, plausibility increased substantially (from 0.659 to 0.724), while fidelity also improved (from 0.893 to 0.951), further confirming that the fidelity-plausibility tradeoff can be mitigated through appropriate post-processing.

The smaller relative improvement for Occlusion compared to ConservativeLRP reflects that Occlusion already had a strong baseline performance. Nevertheless, the consistent improvements across both dimensions validate that post-processing techniques effectively refine attribution patterns to align with human cognitive units while preserving the underlying attribution mechanism.

D.3 Why Post-processing Improves Fidelity

The improvement in fidelity through post-processing may appear counterintuitive, since post-processing does not alter which tokens the XAI method identified as important. It only refines how

attribution scores are presented. The explanation lies in how AOPC fidelity metrics operate: they evaluate perturbation impact in 10% incremental chunks, selecting the top- k tokens by attribution magnitude at each step. When raw attributions scatter importance across many noisy tokens, the top-10% selected for removal may include low-signal tokens, diluting the measured fidelity signal. Post-processing techniques such as smoothing and clustering concentrate attribution mass onto genuinely important spans by averaging neighboring scores or discretizing into coherent groups. This makes top- k token selection more precise, the tokens selected at each perturbation step are more likely to be truly important, producing a larger measured prediction change. Importantly, post-processing does not “invent” a better XAI method, it refines the presentation of an existing attribution signal, analogous to how SmoothGrad (Smilkov et al., 2017) reduces gradient noise to produce cleaner saliency maps without altering the underlying model.

E Inter-Annotator Agreement Analysis

We report detailed inter-annotator agreement statistics for both text selection and explanation quality ratings.

Text selection agreement. Cohen’s Kappa for text selection (which tokens participants highlighted as relevant) averaged $\kappa = 0.667$ across annotator pairs within groups. This substantial agreement indicates that participants shared a consistent understanding of which post segments are relevant to claims, supporting the reliability of user annotations as a reference for plausibility evaluation.

Rating agreement. Agreement on Likert-scale quality ratings averaged $\kappa = 0.389$. This moderate level is consistent with the inherently subjective nature of explanation quality assessment. Several factors contribute to this variability: (1) participants may apply different internal thresholds for what constitutes “useful” versus “somewhat useful” explanations, (2) individual differences in domain knowledge affect which information is perceived as critical, (3) the granularity of a 5-point scale amplifies disagreement when adjacent categories (e.g., 3 vs. 4) represent subtle distinctions. Importantly, when ratings are collapsed into binary categories (ratings ≤ 2 vs. ≥ 3 , i.e., “misleading” vs. “at least somewhat useful”), agreement increases

substantially, confirming that annotators converge on the broader quality assessment even when they disagree on fine-grained distinctions.

Robustness despite moderate agreement. The systematic relationship between ratings and computed IOU (Table 7), where IOU increases from 18% to 62% across rating categories, demonstrates that individual ratings, despite moderate inter-annotator agreement, meaningfully track genuine quality differences. The large between-method effect sizes (1.6-1.8 Likert points) exceed the noise from individual disagreement by a substantial margin, ensuring that our comparative conclusions are not driven by annotator idiosyncrasies.

F Statistical Power and Robustness Analysis

Repeated-measures analysis. Each of the 15 participants evaluated all five XAI methods across 100 post-claim pairs (~ 20 pairs per method), with method-to-pair assignments randomized within groups (Section 5). We computed participant-level mean ratings per method and applied a repeated-measures ANOVA, which accounts for within-subject correlation by partitioning subject variance from the error term. Results: $F(4, 56) = 92.55$, $p = 5.29 \times 10^{-24}$, partial $\eta^2 = 0.869$. A non-parametric Friedman test yielded $\chi^2(4) = 48.0$, $p = 9.44 \times 10^{-10}$. Both tests confirm that between-method differences are not attributable to individual annotator variation.

Post-hoc pairwise comparison. To identify which specific methods differ, we ran Holm-corrected pairwise Wilcoxon signed-rank tests on the participant-level mean ratings ($N = 15$), the non-parametric repeated-measures counterpart of the Friedman omnibus above. Table 9 reports the adjusted p -values. The tests partition the five methods into three homogeneous tiers (denoted $a > b > c$ in Table 5): the two Occlusion variants are mutually indistinguishable ($p = 0.55$) and each significantly exceeds every other method; ConservativeLRP+GMM and Random do not differ significantly ($p = 0.55$); and ConservativeLRP is rated significantly below every other method, including Random ($p = 0.031$). An item-level analysis treating each of the 100 claim-post pairs as a block (each pair rated under all five methods) reproduces the same tier structure, with the Random vs. ConservativeLRP distinction strengthened to

$p < 0.001$. As a parametric cross-check, a paired t -test between the two focal methods (Occlusion vs. ConservativeLRP) on participant-level means yields $t(14) = 19.1$, $p = 2.0 \times 10^{-11}$ (mean difference 1.62 points), confirming that the gap reported in Section 7 is robust at the participant level.

	Occ+sm	Occl	GMM	Rand
Occlusion	0.55			
ConservLRP+GMM	<0.001	<0.001		
Random	<0.001	<0.001	0.55	
ConservativeLRP	<0.001	0.003	<0.001	0.031

Table 9: Holm-corrected pairwise Wilcoxon signed-rank p -values on participant-level mean ratings ($N = 15$). Occ+sm = Occlusion+smooth alg., Occl = Occlusion, GMM = ConservLRP+GMM, Rand = Random. Non-significant pairs ($p > 0.05$) fall within the same homogeneous subset in Table 5.

Sensitivity analysis for replication planning. Following Hoenig and Heisey (2001), we do not interpret post-hoc achieved power for the significant result above, instead, we use the observed effect size as a pilot estimate for prospective sample planning. Cohen’s f was computed from the rating distributions in Table 6: we derived group-level variances from the observed Likert distributions, computed the pooled within-group standard deviation, and calculated f as the ratio of the weighted between-group standard deviation to the pooled within-group standard deviation, yielding $f=0.919$. Using G*Power (Figure 5), we estimate that approximately $N=45$ participants ($k=5$, $\alpha=0.05$) would provide 0.80 power to detect effects of this magnitude. This estimate uses a between-subjects parameterization and is therefore conservative for a repeated-measures design.

G LLM Annotation Protocol

To enable large-scale hyperparameter optimization, we generated explanation annotations for 5,874 post-claim pairs using GPT-4.1.

G.1 Annotation Quality Assessment Levels

We evaluated LLM-generated annotations on a 4-level ordinal scale:

- **Level 0 (Wrong):** The annotation is missing or completely inaccurate. Does not provide any meaningful information or context. Fails to address the core elements of the post/claim.

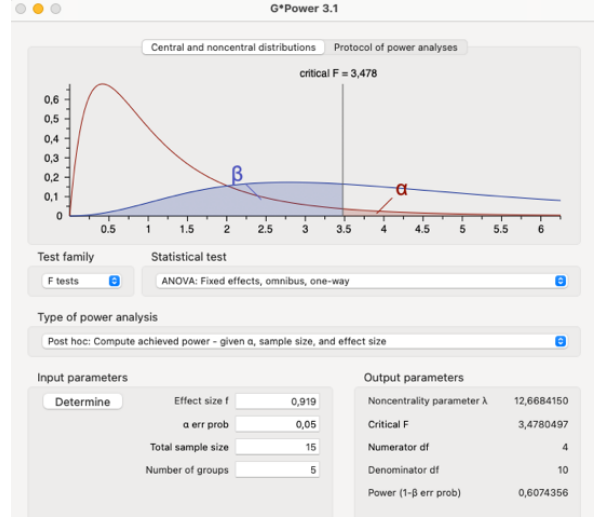


Figure 5: G*Power output for retrospective power analysis (one-way ANOVA, $\alpha=0.05$, $k=5$, $N=15$, $f=0.919$). Achieved power = 0.607.

- **Level 1 (Basic):** The annotation is present but lacks detail and clarity. Provides minimal information and may partially address the post/claim. Requires significant improvement to be fully useful.
- **Level 2 (Adequate):** The annotation is clear and provides sufficient information. Addresses the main elements of the post/claim reasonably well. Could benefit from additional details or minor adjustments.
- **Level 3 (Comprehensive):** The annotation is thorough and highly accurate. Provides complete and detailed information. Fully addresses the post/claim and exceeds expectations in clarity and usefulness.

G.2 Annotation Procedure

Sample and Annotators We randomly sampled 100 post-claim pairs from our annotated dataset for evaluation. Three independent annotators assessed each pair’s LLM-generated annotation quality. All annotators received the same annotation guidelines and examples of each quality level.

G.3 Inter-Annotator Agreement Metrics

Primary Metrics We report Krippendorff’s $\alpha = 0.291$ as our primary agreement metric, as it appropriately accounts for the ordinal nature of our scale (Krippendorff, 2019). This fair agreement ($0.20 \leq \alpha < 0.40$) is consistent with published benchmarks for subjective quality assess-

Metric	Value	Interpret.
Krippendorff's α (ordinal)	0.291	Fair
Fleiss' κ	0.282	Fair
Mean Cohen's κ_w (quadratic)	0.289 ± 0.117	Fair
ICC(2,1)	0.293	Poor
ICC(2,k)	0.555	Fair
Exact agreement (%)	48.5	—
Adjacent agreement (%)	78.2	—
Within ± 2 levels (%)	96.0	—

Table 10: Inter-annotator agreement metrics for LLM annotation evaluation ($n = 101$ samples, 3 annotators). Interpretations follow Landis and Koch (1977) for κ and α , and Cicchetti (1994) for ICC.

ment tasks (Artstein and Poesio, 2008; Aroyo and Welty, 2015).

Agreement Rates Exact agreement (all 3 annotators assigned the same level) was achieved in 48.5% of cases. Adjacent agreement (ratings within ± 1 level) reached 78.2%, demonstrating that annotators generally agreed on quality ranges even when exact ratings differed. Nearly all cases (96.0%) were within ± 2 levels.

G.4 Rating Distribution

Annotator	Level 0	Level 1	Level 2	Level 3
No.1	1 (1.0%)	9 (8.9%)	22 (21.8%)	69 (68.3%)
No.2	3 (3.0%)	13 (12.9%)	23 (22.8%)	62 (61.4%)
No.3	1 (1.0%)	13 (12.9%)	22 (21.8%)	65 (64.4%)
Overall	5 (1.6%)	35 (11.6%)	67 (22.2%)	196 (64.6%)

Table 11: Distribution of quality ratings across annotators. All annotators showed similar distributions with strong positive skew toward Level 3 (Comprehensive).

All three annotators showed similar rating distributions, with 61-68% of annotations rated as Level 3 (Comprehensive), suggesting generally high-quality LLM-generated annotations. Level 0 (Wrong) was rare (1-3% of ratings), indicating that LLM annotations typically provided at least basic useful information.

G.5 Implications for Annotation Quality

The fair agreement ($\alpha = 0.291$) and high adjacent agreement (78.2%) indicate that:

1. Annotators showed conceptual alignment on annotation quality despite some boundary uncertainty between levels.
2. The 4-level scale captures meaningful quality distinctions, particularly for identifying

clearly inadequate (Levels 0-1) versus adequate or excellent (Levels 2-3) annotations.

3. Aggregating judgments from multiple annotators (ICC(2,k) = 0.555, fair) provides more reliable quality estimates than individual judgments.
4. Most disagreements occur at adjacent boundaries (e.g., Level 2 vs. 3), suggesting that exact level assignment is challenging but broad quality categories are more reliable.

- The prompting strategy used - How you instructed the LLM to highlight relevant tokens

H Hyperparameter Optimization Configuration

Figures 6, 7, and 8 present the complete hyperoptimization configuration used in the o-MEGA framework, covering general pipeline settings, model-specific layer and function mappings, and method-specific hyperparameter search spaces, respectively.

```
# Part 1: General Configuration
Hyperoptimization_parameters:
  rationale_path: "path/to/plausibility_annotations"
  task: "post_claim_matching"
  dataset: "path/to/dataset"
  model_path: "intfloat/multilingual-e5-large"
  embeddings_module_name: "embeddings.
    word_embeddings"
  methods: ["GAE_Explain", "Occlusion", "Input X
    Gradient",
    "Guided Backprop", "Feature Ablation", "
    Kernel Shap",
    "Gradient Shap", "Lime", "Saliency", "
    ConservativeLRP"]
  normalizations: ["without_normalize"]
  explanation_maps_token: True
  plausability_weight: 0.5
  faithfulness_weight: 0.5
  multiple_object: False
```

Figure 6: General hyperoptimization configuration parameters for XAI methods evaluation.

I User Study Annotation Tool Interface

The annotation tool used in the user study presents participants with a claim-post pair and allows them to mark relevant portions of the post text. The interface consists of three main sections:

1. **Claim:** Displays the fact-checked claim that needs to be matched with the post content.
2. **Post:** Shows the social media post text. Participants can select text by clicking on individual

```

# Part 2: Model-Specific Parameters
model_param:
  Lime:
    similarity_func:
      function_name:
        - captum.attr._core.lime.
          get_exp_kernel_similarity_function
      parameters:
        distance_mode: ["cosine", "euclidean"]
        kernel_width: [450, 750]
    interpretable_model:
      function_name:
        - captum._utils.models.linear_model.
          SkLearnLasso
      parameters:
        alpha: [1e-19, 1e-25]

  LayerGradientXActivation:
    layer:
      - [model.get_embedding_layer]
    multiply_by_inputs: true

  GAE_Explain:
    implemented_method: true
    layers:
      module_path_expressions:
        - "hf_transformer.encoder.layer.*.attention.
          self.dropout"

  ConservativeLRP:
    implemented_method: true
    layers:
      store_A_path_expressions:
        - "hf_transformer.embeddings"
      attent_path_expressions:
        - "hf_transformer.encoder.layer.*.attention.
          self.dropout"
      norm_layer_path_expressions:
        - "hf_transformer.embeddings.LayerNorm"
        - "hf_transformer.encoder.layer.*.attention.
          output.LayerNorm"
        - "hf_transformer.encoder.layer.*.output.
          LayerNorm"

  Occlusion_word_level:
    implemented_method: true

```

Figure 7: Model-specific parameters for XAI methods hyperoptimization.

words or phrases. Selected text is highlighted in gray, and token indices are displayed below for reference.

3. **Explanation:** This section is revealed only after participants submit their text selection. It displays the XAI-generated explanation with highlighted tokens in green, allowing participants to compare their manual annotations with the computed attributions.

The two-step evaluation process ensures that participants first provide their independent judgment before being exposed to the AI-generated explanations. This design prevents anchoring bias and allows for genuine assessment of explanation quality. Participants can then rate whether the XAI-highlighted text passages sufficiently express the match between the claim and post using the Likert scale questionnaire.

```

# Part 3: Method-Specific Parameters
method_param:
  Lime:
    parameters:
      n_samples: [80, 90]
      token_groups_for_feature_mask: true

  Saliency:
    parameters:
      abs: [true, false]

  Occlusion:
    parameters:
      sliding_window_shapes:
        - [3, 1024]
        - [5, 1024]
      strides:
        - [1, 1024]
        - [1, 512]
      compute_baseline: true

  Gradient Shap:
    parameters:
      stdevs: [0.1, 0.9]
      n_samples: [10, 15]
      compute_baseline: true

  Kernel Shap:
    parameters:
      n_samples: [80, 90]
      compute_baseline: true

  Feature Ablation:
    token_groups_for_feature_mask: true
    compute_baseline: true

  Occlusion_word_level:
    parameters:
      regex_condition:
        - ""
        - ".!?!;:... "

  Integrated Gradients:
    parameters:
      n_steps: [60, 40]

```

Figure 8: Method-specific parameters for XAI hyper-optimization including sampling, baseline computation, and technique-specific settings.