

# ICE: Intervention-Consistent Explanation Evaluation with Statistical Grounding for LLMs

Abhinaba Basu<sup>1,2</sup> and Pavan Chakraborty<sup>1</sup>

<sup>1</sup>Indian Institute of Information Technology, Allahabad (IIITA)

<sup>2</sup>National Institute of Electronics and Information Technology (NIELIT)

abhinaba.basu@iiita.ac.in

## Abstract

Evaluating whether explanations faithfully reflect a model’s reasoning remains an open problem. Existing benchmarks use single interventions without statistical testing, making it impossible to distinguish genuine faithfulness from chance-level performance. We introduce ICE (Intervention-Consistent Explanation), a framework that compares explanations against matched random baselines via randomization tests under multiple intervention operators, yielding win rates with confidence intervals. Evaluating 7 LLMs across 4 English tasks, 6 non-English languages, and 2 attribution methods, we find that faithfulness is operator-dependent: operator gaps reach up to 44 percentage points, with deletion typically inflating estimates on short text but the pattern reversing on long text, suggesting that faithfulness should be interpreted comparatively across intervention operators rather than as a single score. Randomized baselines reveal anti-faithfulness in one-third of configurations, and faithfulness shows zero correlation with human plausibility ( $|r| < 0.04$ ). Multilingual evaluation reveals dramatic model-language interactions not explained by tokenization alone. We release the ICE framework and ICEBench benchmark.

## 1 Introduction

Consider a sentiment model that correctly classifies “gorgeous, witty, seductive movie” as positive with 78% confidence. You ask: *which words drove the prediction?* Gradient attribution highlights “a” and “movie”—function words with no sentiment. Attention highlights “gorgeous” and “seductive”—the actual sentiment signal. The gradient explanation is not just unhelpful; it is **anti-faithful**, performing worse than randomly selected tokens in 100% of trials.

This example illustrates a fundamental problem: we lack rigorous tools to tell faithful explanations from misleading ones. ERASER (DeYoung

et al., 2020), the standard benchmark, reports raw sufficiency and comprehensiveness scores without statistical testing—a sufficiency of 0.62 might be noise. It uses a single intervention (deletion), which can inflate faithfulness by creating out-of-distribution (OOD) inputs that degrade *any* prediction, not just those relying on the highlighted tokens (Hase et al., 2021). And it cannot detect anti-faithfulness at all, since it lacks random baselines.

We introduce ICE (Intervention-Consistent Explanation), a framework that addresses these gaps. ICE asks a simple question: *do the tokens identified by an explanation method outperform randomly selected tokens?* By comparing explanations against matched random baselines under identical interventions, ICE cancels out-of-distribution (OOD) artifacts and provides statistically grounded answers with confidence intervals.

Our central claim is that attribution-faithfulness is not a single number that can be read off from one intervention. It is an *operator-dependent quantity*: different operators introduce different biases—deletion creates OOD inputs that can inflate or deflate estimates depending on text length, while retrieval infill preserves surface form but may introduce competing signals. When operators agree, the evidence for genuine faithfulness is strong regardless of these biases; when they disagree, the gap quantifies methodological uncertainty. ICE makes this dependence explicit through randomized baselines and operator-aware evaluation.

We evaluate 7 LLMs across 4 English tasks, 6 non-English languages, and 2 attribution methods (attention, gradient). Our key contributions:

1. **Operator-aware faithfulness:** Intervention choice materially changes conclusions, with operator gaps reaching 8–44 percentage points. The direction of the gap is text-length dependent. Faithfulness should be interpreted

as an operator-dependent quantity, not a single score.

2. **Statistical grounding:** Randomized baselines with win rates, effect sizes, and bootstrap CIs distinguish genuine faithfulness from chance and detect anti-faithfulness—invisible without random baselines—in nearly one-third of configurations.
3. **Cross-lingual and cross-architecture generalization:** These effects generalize across 6 non-English languages (4 scripts, 5 families) and encoder models, exposing model-language interactions and failures invisible to standard English-only benchmarks.

More broadly, our results suggest that faithfulness evaluation should be interpreted comparatively across intervention regimes rather than reduced to a single benchmark score.

**Roadmap.** We position ICE within faithfulness evaluation (§2), present the framework (§3), describe our experimental setup (§4), report results (§5), and analyze key findings including operator effects (§6).

## 2 Related Work

### 2.1 Faithfulness Evaluation Frameworks

The evolution from single-metric evaluation to statistically grounded frameworks (Figure 1) motivates ICE’s design.

**ERASER and its legacy.** ERASER (DeYoung et al., 2020) introduced sufficiency and comprehensiveness as standard faithfulness metrics but has three limitations ICE addresses: no statistical testing, a single deletion operator that conflates faithfulness with OOD degradation, and no random baselines to detect anti-faithfulness. Our retrieval infill results (§6.2) show operator gaps reaching up to 44 percentage points.

**Subsequent frameworks.** Sun et al. (2025) compare explanation types across several evaluation properties under fixed interventions; ICE asks whether the evaluation itself remains stable when the intervention regime changes. This is a deeper methodological question: Sun treats operator choice as implementation; ICE elevates it to a first-class variable. Crucially, Sun et al. still rely on raw scores without random baselines. Kamp et al.

(2023) show that dynamically estimating top- $k$  reduces attribution disagreement; our  $k$ -sensitivity analysis (Appendix D) confirms faithfulness conclusions can reverse depending on  $k$ . Kamp et al. (2025) find that training on sufficient rationales does not consistently improve faithfulness—our “Lucky Tokens” category captures a similar phenomenon. F-Fidelity (Zheng et al., 2025) addresses OOD via fine-tuning; ICE controls for OOD at the evaluation protocol level. Zaman and Srivastava (2025) apply a causal lens and find no single metric works across all tasks, echoing our finding that operator choice changes conclusions.

### 2.2 The Faithfulness-Plausibility Distinction

Jacovi and Goldberg (2020) define faithfulness as a graded property distinct from plausibility, arguing both require independent evaluation. Parcalabescu and Frank (2024) argue many faithfulness tests actually measure self-consistency rather than mechanistic transparency. ICE acknowledges this limitation but argues that behavioral faithfulness, when statistically grounded, remains the most scalable evaluation for billion-parameter models. We contribute quantitative evidence: zero correlation ( $|r| < 0.04$ ) between ICE faithfulness and human rationale alignment across three models.

### 2.3 Attribution Methods for LLMs

Attention remains foundational for transformers; gradient faces memory and instability challenges in large LLMs. Madsen et al. (2024) find faithfulness is explanation-type, model, and task-dependent, reinforcing ICE’s multi-method evaluation. We include Integrated Gradients for encoder baselines but omit it for 7B+ LLMs due to memory constraints.

### 2.4 Multilingual Explainability and Statistical Methods

Surveys (Resck et al., 2025) highlight the need for cross-lingual faithfulness analysis, while studies (Zhao and Aletras, 2024) suggest larger multilingual models may produce less faithful explanations. Prior multilingual work evaluates plausibility rather than faithfulness and lacks statistical rigor. Permutation-based testing (Mandel and Barnett, 2024; Biswas et al., 2025) provides the statistical foundation ICE builds on.

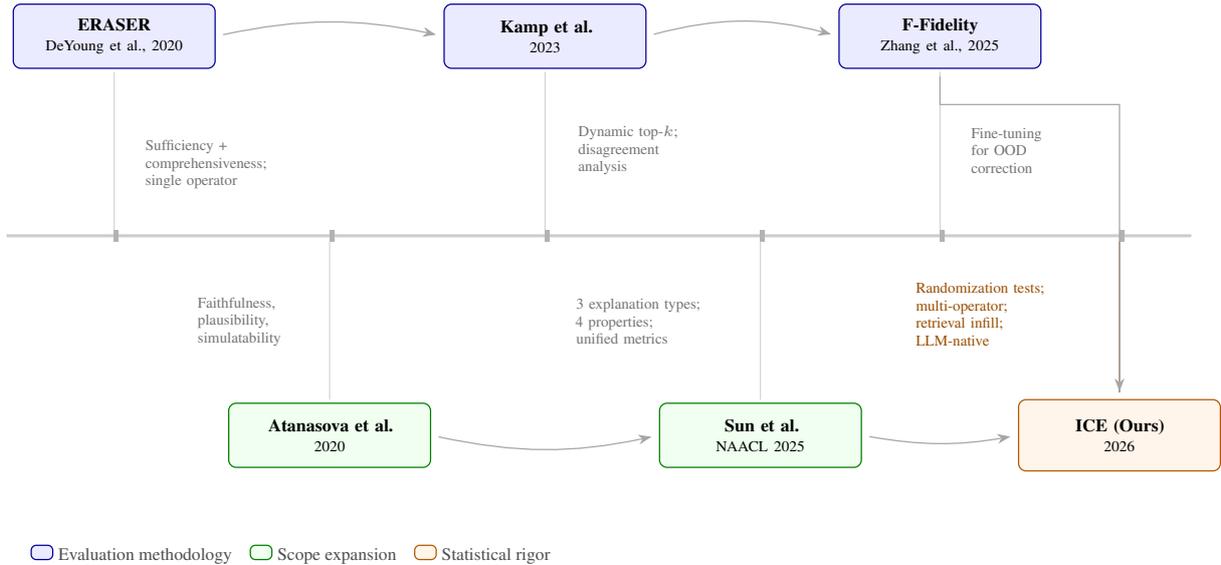


Figure 1: Evolution of faithfulness evaluation frameworks. ICE builds on evaluation methodology (blue) and scope expansion (green), adding statistical rigor (orange) via randomization testing and operator-consistent evaluation. Arrows show methodological lineage.

Aspect	ERA	Sun	Mad	F-Fid	ICE
Stat. testing	✗	✗	✗	✗	✓
Uncertainty	✗	✗	✗	✗	✓
LLM support	✗	✗	Ltd	✗	Native
Multilingual	✗	✗	✗	✗	6 langs
Multi-operator	✗	Part	Part	✗	✓
OOD mitigation	✗	✗	✗	✓	✓

Table 1: Comparison of faithfulness frameworks. ERA=ERASER (DeYoung et al., 2020), Sun=Sun et al. (2025), Mad=Madsen et al. (2024), F-Fid=F-Fidelity (Zheng et al., 2025). ✓=supported, ✗=not, Ltd=Limited, Part=Partial.

## 2.5 Comparison with Prior Frameworks

Table 1 positions ICE relative to prior work. Among current frameworks, ICE is unique in combining statistical significance testing, native LLM support, and multi-operator evaluation.

## 3 The ICE Framework

### 3.1 Problem Formulation

Given a model  $f$ , input  $x$  with tokens  $(t_1, \dots, t_n)$ , and an explanation method  $E$  producing importance scores  $E(x) = (e_1, \dots, e_n)$ , we evaluate whether the top- $k$  fraction of tokens (rationale  $r$ , e.g.,  $k=0.2$  retains the top 20%) identified by  $E$  are genuinely important for  $f$ 's prediction. The central question: *does the explanation method identify tokens that are more important than randomly selected tokens?*

### 3.2 NSR and Randomization Testing

We define Normalized Score Retention (NSR) to measure how much of the original prediction is preserved when only rationale tokens remain:

$$\text{NSR}(r) = \frac{s(x_o^r) - s(\emptyset)}{s(x) - s(\emptyset)} \quad (1)$$

where  $s(x)$  is the prediction score on original input,  $s(x_o^r)$  is the score with only rationale tokens under operator  $o \in \{\text{delete}, \text{retrieval}\}$ , and  $s(\emptyset)$  is the baseline (empty input).  $\text{NSR} \in [0, 1]$ : 1 means perfect retention, 0 means complete loss.

**Concrete Example.** Consider “a gorgeous, witty, seductive movie” classified as positive with  $s(x) = 0.78$ . Attention highlights {gorgeous, seductive} as the rationale  $r$ . Using deletion as the operator, keeping only these tokens yields  $s(x_o^r) = 0.72$ . With empty input  $s(\emptyset) = 0.50$ . Then  $\text{NSR} = (0.72 - 0.50)/(0.78 - 0.50) = 0.79$ —the rationale preserves 79% of the prediction signal. Repeating with 50 random token sets: if the rationale beats 46 of 50, the win rate is 92%, indicating strong faithfulness.

Figure 2 illustrates this pipeline, showing how the choice of intervention operator (deletion vs. retrieval infill) can change the faithfulness verdict.

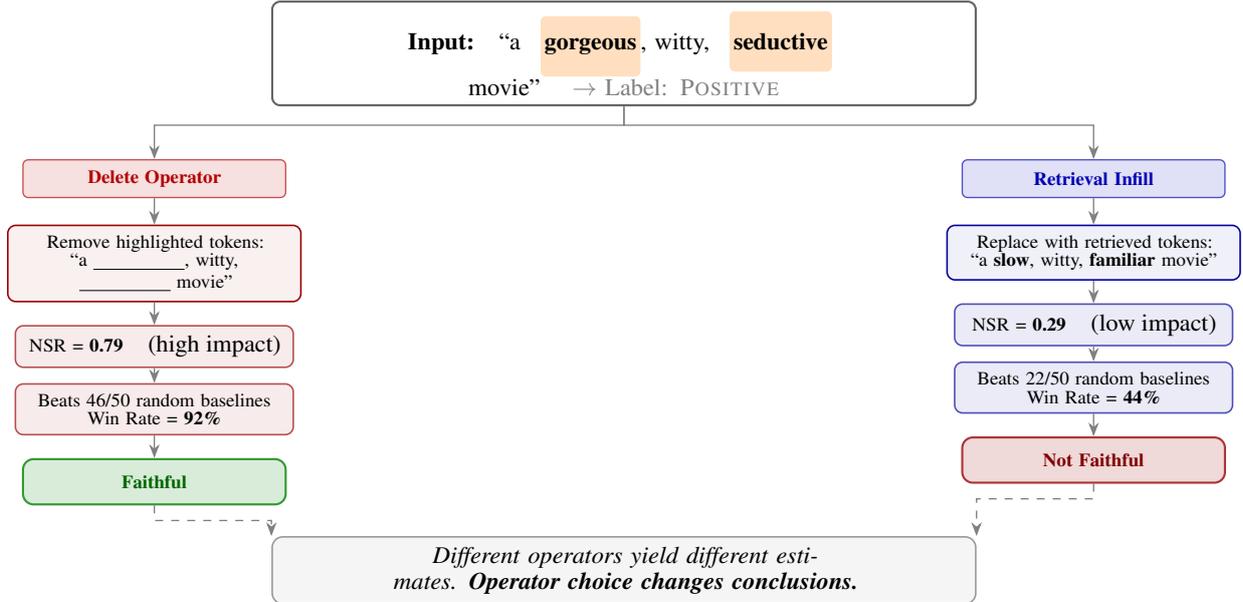


Figure 2: ICE pipeline on a sentiment example. Attention identifies “gorgeous” and “seductive” as the rationale. **Delete** removes all *other* tokens, leaving only the rationale—an unnatural input that preserves the prediction (WR = 92%), but this may reflect OOD artifacts rather than genuine faithfulness. **Retrieval Infill** replaces non-rationale tokens with tokens randomly sampled from other corpus examples (“slow, familiar”), preserving natural surface form. The rationale must now dominate in realistic context rather than in isolation (WR = 44%). Label tokens are blacklisted, but replacement text may carry incidental sentiment—this is by design, as it tests robustness of the attribution signal. Same rationale, same model, same metric—only the operator differs, yet the verdict changes.

### Algorithm 1 ICE Randomization Test

**Require:** Input  $x$ , rationale  $r$ , operators  $\mathcal{O}$ , permutations  $M$

- 1: Compute  $\text{NSR}_{obs} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \text{NSR}_o(r)$
- 2: **for**  $i = 1$  to  $M$  **do**
- 3:   Sample random tokens  $r_i$  with  $|r_i| = |r|$
- 4:   Compute  $\text{NSR}_i = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \text{NSR}_o(r_i)$
- 5: **end for**
- 6: **Win Rate**  $= \frac{1}{M} \sum_{i=1}^M \mathbb{1}[\text{NSR}_{obs} > \text{NSR}_i]$
- 7: **Effect Size**  $= \frac{\text{NSR}_{obs} - \mu_{\text{random}}}{\sigma_{\text{random}}}$  (Cohen’s  $d$ )
- 8: **return** Win Rate, Effect Size,  $p$ -value

**The Randomization Test.** The  $p$ -value uses a one-sided test:

$$p = \frac{1 + \sum_{i=1}^M \mathbb{1}[\text{NSR}_i \geq \text{NSR}_{obs}]}{M + 1}$$

with +1 terms for conservative finite-sample correction. We primarily report **win rate**, which remains stable even when NSR denominators are small. Unless noted, we report per-operator results rather than the multi-operator average, to reveal operator-specific effects.

### 3.3 Operators

Any intervention creates inputs the model was not trained on. Different interventions introduce

different OOD artifacts. Using a single operator risks confounding faithfulness with operator-specific degradation (Hase et al., 2021). ICE employs multiple operators:

- **Deletion:** Removes tokens from the sequence. Fast but produces unnatural truncated text.
- **Retrieval Infill:** Replaces tokens with contiguous spans sampled from other examples, excluding the current example (leave-one-out) and filtering label-indicative tokens (label-blacklisted). Preserves surface form while destroying task-relevant content. Replacement spans are drawn from the same distribution, which may introduce label-correlated artifacts; label blacklisting mitigates but does not eliminate this risk.

For encoder models, we additionally evaluate mask-based operators (Appendix A). For autoregressive LLMs, masking creates degenerate outputs; we use deletion and retrieval infill.

### 3.4 Statistical Framework

We compute 95% bootstrap confidence intervals ( $B = 200$  resamples) for uncertainty quantification

and apply Benjamini-Hochberg FDR correction ( $\alpha = 0.10$ ) for multiple testing. Full details appear in Appendix C.

## 4 Experimental Setup

We focus on autoregressive LLMs; encoder and Chain-of-Thought evaluations appear in Appendices H and M.

We evaluate 7 LLMs (1.5B–8B parameters; GPT-2, Llama 3.x, Qwen 2.5, Mistral, DeepSeek, LFM2) on 4 English datasets (SST-2, IMDB, e-SNLI, AG News) and 6 non-English languages using native sentiment data (French, German, Hindi, Chinese, Turkish, Arabic—covering 4 scripts and 5 language families). We compare attention and gradient attribution methods. Full model and dataset details appear in Appendix B.

We use  $k = 0.2$  (top 20% tokens),  $N = 500$  examples per dataset,  $M = 50$  permutations for LLMs ( $M = 100$  for encoders), and 512-token truncation. A  $k$ -sweep in Appendix D justifies  $k = 0.2$  as a middle ground. We release ICEBench with pinned dataset versions for reproducibility (Appendix B).

## 5 Results

Our experiments test three hypotheses implied by the ICE framework: (1) operator choice materially changes faithfulness estimates, (2) operator agreement signals reliable attribution, and (3) these patterns generalize across languages and architectures.

### 5.1 English Benchmark Results

Figure 3 shows win rates under both operators (deletion and retrieval infill) and both attribution methods, revealing four patterns.

**Operator effects.** Comparing left (deletion) and right (retrieval infill) panels reveals systematic differences. On short text, deletion produces more green cells, particularly on SST-2 (gaps of 8–9 pp) and e-SNLI (up to 44 pp for Llama-3.2). On IMDB (long text), the pattern reverses for most models—retrieval infill yields *higher* win rates than deletion (e.g., Llama-3.2: 91.8% vs. 71.3%), likely because deleting most of a long review creates severely degraded input, while retrieval preserves natural text length. This suggests the operator gap direction is text-length sensitive.

**Short vs. Long Text.** Under both operators, attention beats gradient on short text (SST-2). Both converge on long text (IMDB: ~85–95% for capable models under both operators). Attention captures local sentiment signals; gradient benefits from accumulated context.

**Task-Specific Patterns.** NLI favors attention under deletion (Llama 3.2: 86.4%) but shows more nuanced patterns under retrieval infill (Llama 3.2 attention drops to 42.6%, while gradient rises to 83.4%). Topic classification favors gradient under both operators.

**Base vs. Instruct.** Llama 3.1-8B Base shows task-dependent faithfulness under both operators: near-random on SST-2 but exceptional on e-SNLI (97.2% gradient under deletion, 90.1% under retrieval—the highest across all configurations).

### 5.2 Multilingual Results

Figure 4 reveals striking cross-lingual variation under both operators. No single model dominates under deletion: Qwen 2.5-7B leads on German (82.7%) and Turkish (79.3%), Llama 3.1 on French (80.8%), while GPT-2 shows anti-faithfulness on French (15.8%) and Turkish (29.7%). Tokenization does not predict faithfulness: GPT-2 achieves 66% Hindi gradient despite  $8.1\times$  token expansion, while French ( $1.8\times$ ) yields near-random results.

Under retrieval infill (right panel), the operator gap varies by language: GPT-2 Hindi *increases* from 65.4% to 68.8%—retrieval exceeding deletion, as also observed on IMDB for most models—while Llama 3.1 drops to anti-faithful (35.1% Hindi, 37.3% Chinese). Overall, cross-lingual faithfulness is not explained by tokenization alone; model-language interactions, morphological structure, and operator sensitivity jointly determine whether attributions remain faithful. Full language-specific analysis appears in Appendix F.

### 5.3 Effect Sizes and Anti-Faithfulness

Effect sizes quantify faithfulness magnitude beyond win rates. Llama 3.1-8B e-SNLI gradient achieves  $d = 2.50$  (extraordinarily large), while GPT-2 French shows  $d = -2.36$  (severe anti-faithfulness). We find anti-faithfulness (win rate  $< 50\%$ ) in nearly one-third of English deletion configurations (18 of 56), predominantly gradient-based, where gradient assigns highest importance to sentence-initial function words while ignoring task-relevant content. Anti-faithful explanations

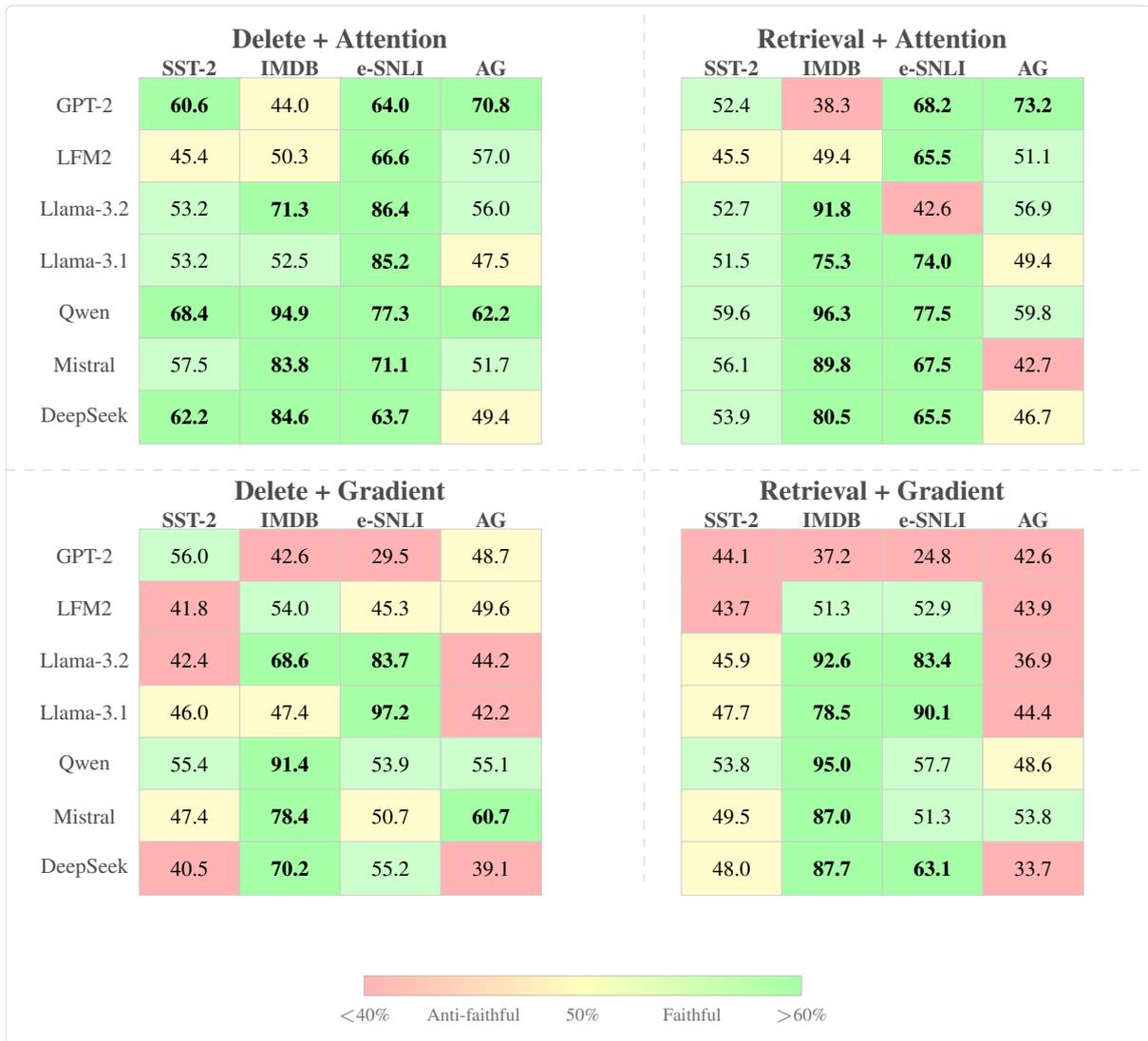


Figure 3: English faithfulness (win rate %) under both operators and both attribution methods. Top: Attention. Bottom: Gradient. Left: Deletion. Right: Retrieval Infill. Models (top to bottom): GPT-2, LFM2, Llama-3.2, Llama-3.1, Qwen, Mistral, DeepSeek. Right panels share the same model order. On short text (SST-2), deletion yields higher estimates; on IMDB (long text), the pattern reverses for most models. Green = faithful (>60%), yellow = random, red = anti-faithful (<40%).

actively mislead users; practitioners must verify faithfulness on their specific configuration. Full effect size tables and bootstrap CIs appear in Appendix I.

## 6 Analysis

### 6.1 Faithfulness vs. Plausibility

All models show near-zero correlation between human rationale alignment (IoU) and ICE win rate ( $|r| < 0.04$ ,  $p > 0.5$ ; Figure 5). This consistency across 1.5B–8B models on e-SNLI provides strong evidence that faithfulness and plausibility are orthogonal evaluation axes. Plausibility benchmarks do not measure faithfulness; both require indepen-

dent assessment (full statistics in Appendix J).

### 6.2 Retrieval Infill: Does Operator Choice Matter?

As shown in the 4-panel comparison (Figure 3), operator choice substantially changes win rates, with deletion typically exceeding retrieval on short text but the pattern reversing on long text. Figure 6 quantifies this effect on representative configurations.

The results confirm that **operator choice fundamentally changes conclusions**. Delete classifies Llama-3.2 e-SNLI as “Truly Faithful” (86.4%), but Retrieval Infill downgrades it to anti-faithful (42.6%)—a 44 percentage point gap. Conversely,

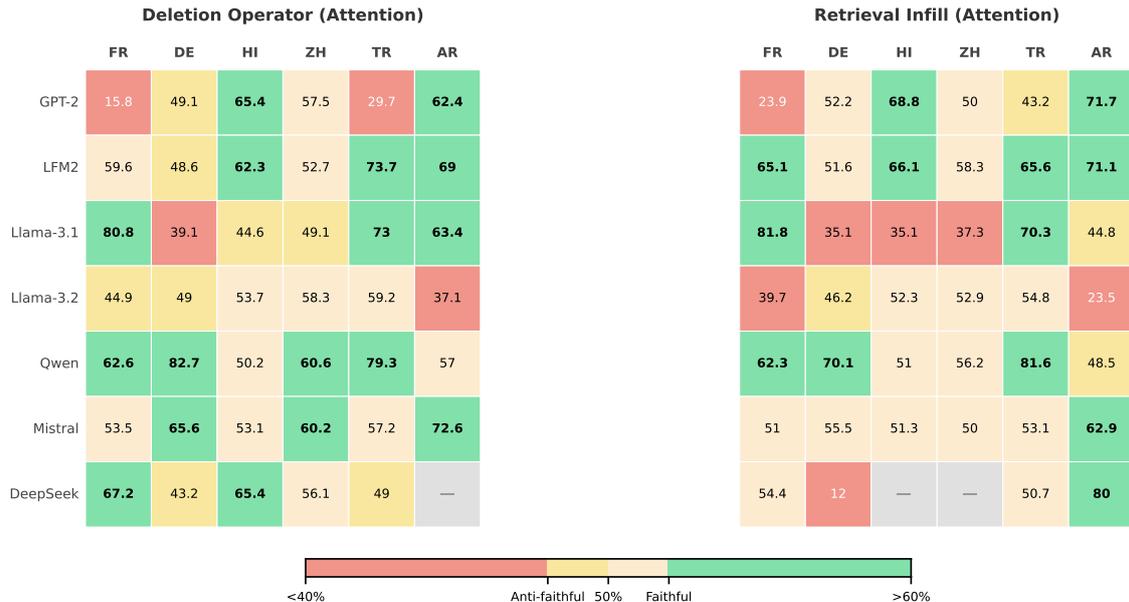


Figure 4: Multilingual faithfulness (attention win rate %) under both operators across 6 languages and 4 scripts. **Left:** Deletion. **Right:** Retrieval Infill. GPT-2 Hindi *increases* under retrieval (65.4%→68.8%), while Llama-3.1 drops to anti-faithful (35.1% Hindi, 37.3% Chinese). Gray = no valid output (DeepSeek Arabic/Hindi/Chinese under retrieval due to tokenizer limitations).

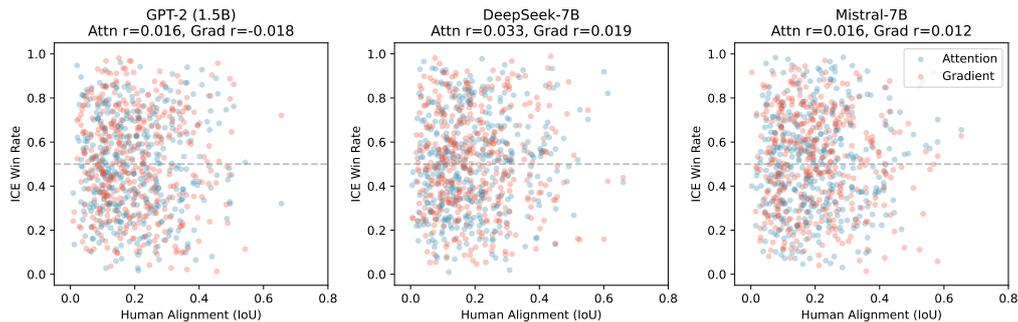


Figure 5: IoU (human alignment, x-axis) vs. ICE Win Rate (faithfulness, y-axis) across GPT-2, DeepSeek, and Mistral on e-SNLI. All  $|r| < 0.04$ : no correlation between human alignment and computational faithfulness.

both operators *agree* on Qwen e-SNLI (~77%), providing stronger evidence of genuine faithfulness there.

Across all 21 attention configurations on short-text datasets (7 models  $\times$  3 datasets), deletion exceeds retrieval in 67% of cases, with a median gap of 1.7 pp (mean 4.6 pp) but extreme outliers reaching 44 pp. On long text (IMDB), the relationship reverses for most models. We recommend reporting both operators: **when operators agree, the evidence is strong regardless of which is higher**. When they disagree, the gap quantifies methodological uncertainty and practitioners should treat both estimates as informative bounds. Full numerical breakdowns appear in Appendix K.

### Operator calibration via known-outcome cases.

To validate that operator agreement is a reliable signal, we examine cases with independently verifiable outcomes. For anti-faithful attributions—where gradient selects “a” instead of sentiment adjectives like “gorgeous” (Table 12)—both operators correctly assign  $WR=0\%$  across all 4 examined cases. For genuinely faithful cases (Llama 3.1 e-SNLI gradient, Qwen IMDB attention), both operators converge at  $WR > 77\%$  with gaps  $< 8$  pp. This pattern—agreement on clear positives and clear negatives—validates operator convergence as a calibration signal.

**Comparison with F-Fidelity.** F-Fidelity (Zheng et al., 2025) addresses OOD artifacts by *fine-tuning*

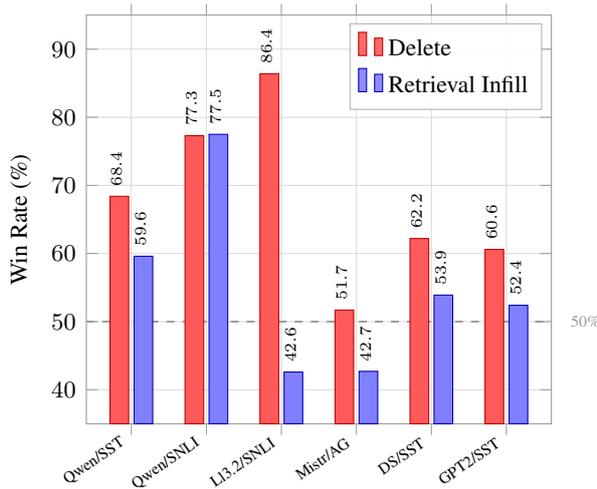


Figure 6: Operator comparison across model-dataset configurations (attention, main models). Delete (red) typically yields higher win rates on short text, while Retrieval Infill (blue) provides conservative estimates. The gap reaches 43.8 pp (Llama-3.2/e-SNLI), but operators agree on Qwen/e-SNLI (77.3% vs. 77.5%). Full results in Appendix Table 15.

the model to adapt to perturbed inputs, then evaluating faithfulness on the adapted model. ICE instead uses retrieval infill to create in-distribution alternatives *without modifying the model*. The approaches are complementary: F-Fidelity requires per-dataset fine-tuning and model weight access, making it inapplicable to black-box or API-only LLMs. ICE is training-free and model-agnostic but relies on the quality of the retrieval pool. Where both are applicable, their agreement would provide the strongest evidence—a direction for future work.

### 6.3 Practical Considerations

Instruction-tuned models show higher sentiment faithfulness than base models, likely from alignment training. LFM2-2.6B shows task-specific behavior—high NLI but near-random sentiment—suggesting efficient architectures trade broad faithfulness for task-specific capability.

Prompt sensitivity analysis (Appendix E) reveals that some prompts induce anti-faithfulness and that confidence does not predict faithfulness. Long text reduces method sensitivity ( $|\Delta| < 2.1\%$  on IMDB vs. 11.8% on SST-2). Our practical guidelines for selecting attribution methods by model, task, and language appear in Appendix Table 17.

## 7 Conclusion

We introduced ICE, a statistically grounded framework for evaluating explanation faithfulness. Evaluating 7 LLMs across 4 English tasks and 6 non-English languages, we draw three main conclusions:

- Operator choice materially changes conclusions:** operator gaps reach up to 44 pp, with deletion typically higher on short text but the pattern reversing on long text and some multilingual configurations. Faithfulness is operator-dependent and best interpreted comparatively across operators, not as a single score.
- Randomized baselines reveal anti-faithfulness:** nearly one-third of configurations perform worse than random, a phenomenon invisible without random baselines.
- Faithfulness and plausibility are orthogonal:**  $|r| < 0.04$  across three models, confirming these require independent evaluation.

Cross-lingual evaluation reveals that these patterns are not explained by tokenization alone; model-language interactions jointly determine faithfulness. Notably, GPT-2 Hindi retrieval infill *exceeds* deletion (68.8% vs. 65.4%), challenging the assumption that retrieval always underestimates faithfulness. We release all code, results, and the ICEBench benchmark.

### Limitations

**Computational cost.** ICE’s randomization tests ( $M = 50$ ) increase computation  $50\times$  over single-point metrics. While this is tractable for the models evaluated here (1.5B–8B), scaling to larger LLMs (70B+) requires either subsampling or adaptive early stopping, which we have not yet validated.

**Attribution method scope.** Our attention extraction averages across all layers and heads—layer-specific or head-specific analysis may yield finer-grained insights (Madsen et al., 2024). We omit Integrated Gradients for 7B+ LLMs due to memory constraints, though this method shows strong results on encoder models (Appendix H).

**Multilingual coverage.** While our 6-language evaluation spans 4 scripts and 5 language families, languages with complex morphology beyond Turkish (e.g., Finnish, Japanese) and tonal languages (e.g., Vietnamese) remain untested. DeepSeek’s tokenizer failure on Hindi and Chinese under retrieval infill highlights that operator applicability is model-dependent.

**Behavioral vs. mechanistic faithfulness.** ICE evaluates *behavioral* faithfulness—whether attributed tokens correlate with prediction changes—rather than *mechanistic* faithfulness, which would require probing internal representations. Recent advances in mechanistic interpretability (Madsen et al., 2024) are complementary; behavioral evaluation remains the most scalable approach for billion-parameter models but does not explain *why* attributions are or are not faithful.

**Chain-of-Thought extension.** Our ICE-CoT extension (Appendix M) is preliminary. Dedicated CoT faithfulness benchmarks have since emerged that provide more comprehensive evaluation of generated reasoning traces.

**Retrieval infill limitations.** While retrieval infill mitigates OOD artifacts from deletion, replacement tokens drawn from the same distribution may preserve some task-relevant signal, potentially underestimating faithfulness. The GPT-2 Hindi case (retrieval exceeding deletion) suggests this concern is language- and model-dependent rather than systematic.

## Ethics Statement

We caution against using faithfulness scores alone for high-stakes domains without domain validation. “Faithful to model” is not “correct reasoning”—a model may faithfully rely on spurious correlations. Our experiments used ~200 GPU-hours on RTX 4090/A100 hardware; we release pre-computed results to reduce replication cost.

## Acknowledgments

**AI Assistance** We used AI assistants (Claude, Gemini) for proofreading, editing, and verification of numerical consistency. All scientific contributions, experimental design, and core writing are the authors’ original work.

## References

- Sanad Biswas, Nina Grundlingh, Jonathan Boardman, Joseph White, and Linh Le. 2025. [A target permutation test for statistical significance of feature importance in differentiable models](#). *Electronics*, 14(3).
- Théophile Blard. 2020. French sentiment analysis with bert. <https://github.com/TheophileBlard/french-sentiment-analysis-with-bert>.
- 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, pages 9560–9572.
- DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiusi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, and 69 others. 2024. [Deepseek llm: Scaling open-source language models with longtermism](#). *Preprint*, arXiv:2401.02954.
- Erkin Demirtas and Mykola Pechenizkiy. 2013. [Cross-lingual polarity detection with machine translation](#). In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM ’13)*. ACM.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. [Eraser: A benchmark of datasets for evaluating rationalizable nlp systems](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458.
- Sumanth Doddapaneni, Rahul Aralikkatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. [Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Peter Hase, Harry Xie, and Mohit Bansal. 2021. [The out-of-distribution problem in explainability and search methods for feature importance explanations](#). In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS ’21.

- 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.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Jonathan Kamp, Lisa Beinborn, and Antske Fokkens. 2023. [Dynamic top-k estimation consolidates disagreement between feature attribution methods](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6190–6197. Association for Computational Linguistics.
- Jonathan Kamp, Lisa Beinborn, and Antske Fokkens. 2025. Learning from sufficient rationales: Analysing the relationship between explanation faithfulness and token-level regularisation strategies. *arXiv preprint arXiv:2511.16353*.
- Liquid AI. 2025. [Lfm2 technical report](#). *arXiv preprint arXiv:2511.23404*.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. [Learning word vectors for sentiment analysis](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. Association for Computational Linguistics.
- Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. [Are self-explanations from large language models faithful?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337. Association for Computational Linguistics.
- Francesca Mandel and Ian Barnett. 2024. [Permutation-based hypothesis testing for neural networks](#). In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’24/IAAI’24/EAAI’24*. AAAI Press.
- Mahmoud Nabil, Mohamed Aly, and Amir F. Atiya. 2015. [ASTD: Arabic sentiment tweets dataset](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519. Association for Computational Linguistics.
- Letitia Parcalabescu and Anette Frank. 2024. [On measuring faithfulness or self-consistency of natural language explanations](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089. Association for Computational Linguistics.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). Technical report, OpenAI.
- Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. [Explainability and interpretability of multilingual large language models: A survey](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20454–20486. Association for Computational Linguistics.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. [Recursive deep models for semantic compositionality over a sentiment treebank](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. Association for Computational Linguistics.
- Jingyi Sun, Pepa Atanasova, and Isabelle Augenstein. 2025. [Evaluating input feature explanations through a unified diagnostic evaluation framework](#). 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 10559–10577. Association for Computational Linguistics.
- Songbo Tan and Jin Zhang. 2008. [An empirical study of sentiment analysis for chinese documents](#). *Expert Systems with Applications*, 34(4):2622–2629.
- Michael Wojatzki, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann. 2017. Germeval 2017: Shared task on aspect-based sentiment in social media customer feedback. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 1–12.
- 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*, pages 29413–29437. Association for Computational Linguistics.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. [Character-level convolutional networks for text classification](#). In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1, NIPS’15*, pages 649–657.

Zhixue Zhao and Nikolaos Aletras. 2024. [Comparing explanation faithfulness between multilingual and monolingual fine-tuned language 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 3226–3244. Association for Computational Linguistics.

Xu Zheng, Farhad Shirani, Zhuomin Chen, Chao hao Lin, Wei Cheng, Wenbo Guo, and Dongsheng Luo. 2025. [F-fidelity: A robust framework for faithfulness evaluation of explainable AI](#). In *The Thirteenth International Conference on Learning Representations*.

## A Operator Ablation

Operator	Attn WR	Grad WR
Deletion	59.8%	52.8%
Mask-UNK	12.9%	22.0%
Mask-PAD	13.4%	21.6%

Table 2: Operator ablation on GPT-2/SST-2 (N=100). Masking produces degenerate outputs for autoregressive LLMs, with win rates 4–5× lower than deletion.

## B Models and Datasets

Model	Size	Type	Description
GPT-2	1.5B	Base	Baseline autoregressive (Radford et al., 2019)
LFM2	2.6B	Base	Efficient architecture (Liquid AI, 2025)
Llama 3.2	3B	Inst	Small instruction-tuned (Grattafiori et al., 2024)
Llama 3.1	8B	Base	General reasoning (Grattafiori et al., 2024)
Qwen 2.5	7B	Inst	Multilingual focus (Qwen et al., 2025)
Mistral	7B	Inst	Efficient instruction (Jiang et al., 2023)
DeepSeek	7B	Chat	Multilingual/Chat (DeepSeek-AI et al., 2024)

Table 3: Evaluated LLMs spanning diverse sizes and capabilities.

English datasets: SST-2 (binary sentiment, short text) (Socher et al., 2013), IMDB (binary sentiment, long text) (Maas et al., 2011), e-SNLI (NLI with human rationales) (Camburu et al., 2018), AG News (4-class topic classification) (Zhang et al., 2015).

Multilingual: French (Allocine (Blard, 2020)), German (GermEval 2017 (Wojatzki et al., 2017)), Hindi (IndicSentiment (Doddapaneni et al., 2023)), Chinese (ChnSentiCorp (Tan and Zhang, 2008)),

Turkish (Turkish Sentiment (Demirtas and Pechenizkiy, 2013)), Arabic (Arabic Sentiment (Nabil et al., 2015)).

Dataset	Revision SHA
glue (SST-2)	bcdcba79d07bc86...
imdb	e6281661ce1c48d...
esnli	a160e6a02bbb8d8...
ag_news	eb185aade064a81...

Table 4: Pinned dataset revisions for ICEBench.

## C Statistical Framework Details

Bootstrap 95% confidence intervals are computed over  $B = 200$  resamples:

$$CI_{95\%} = [NSR_{2.5}^*, NSR_{97.5\%}^*] \quad (2)$$

We apply Benjamini-Hochberg FDR correction at  $\alpha = 0.10$  when evaluating multiple examples. Cohen’s  $d$  is interpreted using standard thresholds: 0.2 (small), 0.5 (medium), 0.8 (large). Negative  $d$  indicates anti-faithfulness.

## D K-Sensitivity Analysis

We evaluate faithfulness sensitivity to rationale length  $k \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$  on multilingual data (French/German) with GPT-2 and LFM2-2.6B.

Model	Extr.	0.1	0.2	0.3	0.4	0.5
<i>German Win Rate (%)</i>						
GPT-2	Attn	47.2	52.7	57.0	61.7	<b>66.6</b>
GPT-2	Grad	<b>54.5</b>	46.2	49.7	47.1	42.0
LFM2	Attn	42.2	52.3	46.1	49.3	<b>55.2</b>
LFM2	Grad	50.9	47.0	51.1	52.1	<b>55.9</b>
<i>French Win Rate (%)</i>						
GPT-2	Attn	48.1	49.0	48.2	47.6	47.7
GPT-2	Grad	48.7	49.6	49.8	49.2	<b>50.2</b>
LFM2	Attn	76.6	64.0	66.2	77.1	<b>79.1</b>
LFM2	Grad	60.2	66.9	71.8	75.3	<b>78.1</b>

Table 5: K-sensitivity: win rate (%) by rationale length  $k$ . Bold = best per config. GPT-2 German gradient is non-monotonic (peaks at  $k=0.1$ ), while attention increases with  $k$ .

## E Prompt Sensitivity Analysis

## F Language-Specific Patterns

**German:** High variance across models. Most struggle (34–67%), but Qwen achieves 83% attention and 80% gradient—the highest multilingual result. German’s compound words and flexible word order challenge position-sensitive methods.

Data	Prompt	Acc	Conf	Attn	Grad	$\Delta$
SST-2	v1 (standard)	52%	68.8	<b>72.7</b>	<b>62.0</b>	+10.7
	v2 (minimal)	59%	65.2	43.4	55.2	-11.8
	v3 (question)	72%	56.3	47.3	46.3	+1.0
	v4 (completion)	64%	63.9	67.5	64.6	+2.9
	v5 (quoted)	74%	57.0	60.3	50.3	+10.0
IMDB	v1 (standard)	42%	57.9	60.2	58.1	+2.1
	v2 (rating)	2%	63.1	34.8 <sup>†</sup>	33.2 <sup>†</sup>	+1.6
	v3 (yes/no)	20%	61.3	<b>75.4</b>	<b>74.8</b>	+0.6
AG News	v1 (standard)	47%	69.6	<b>71.1</b>	48.8	+22.3
	v2 (minimal)	65%	79.2	62.8	<b>67.9</b>	-5.1
	v3 (question)	73%	68.2	58.9	47.3	+11.6
e-SNLI	v1 (standard)	31%	77.0	63.7	14.7 <sup>†</sup>	+49.0
	v2 (verb)	31%	88.9	<b>95.5</b>	24.0 <sup>†</sup>	+71.5
	v3 (T/F/U)	34%	56.4	65.9	<b>68.6</b>	-2.7

Table 6: Prompt sensitivity analysis (GPT-2). Win rates (%) for attention (Attn) and gradient (Grad).  $\Delta$  = Attn – Grad. <sup>†</sup>Anti-faithful (<50%).

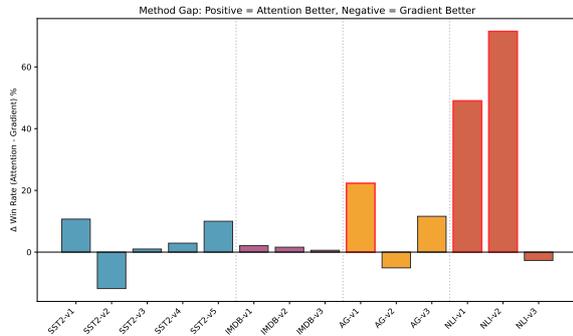


Figure 7: Method gap ( $\Delta$  = Attention – Gradient win rate) by prompt variant.

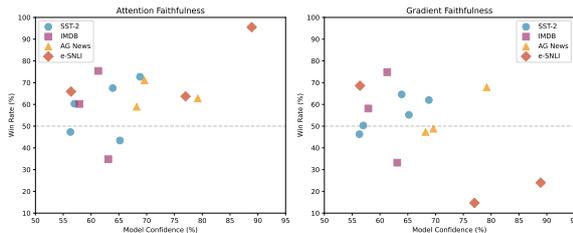


Figure 8: Faithfulness vs. model confidence across prompt variants.

**Chinese:** Moderate deletion results (50–69%). LFM2 gradient (69%) and Llama 3.1 gradient (66%) lead. Under retrieval infill, LFM2 retains 58.3% while Llama 3.1 drops to 37.3%. Character-based tokenization may help align token and semantic boundaries.

**French:** Highly polarized. GPT-2 shows anti-faithfulness (15–16%), while Llama 3.1 (81% attention) and LFM2 (73% gradient) excel.

**Hindi:** Consistent moderate performance un-

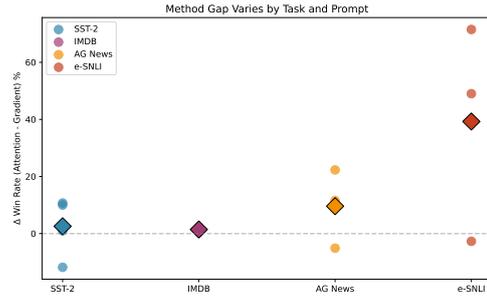


Figure 9: Method gap distribution by task.

der deletion (45–66%). GPT-2 surprisingly strong (65–66%) despite poor tokenization ( $8.1\times$  expansion). Under retrieval infill, GPT-2 reaches 68.8% (exceeding deletion), while Llama 3.1 drops to anti-faithful 35.1%.

**Turkish:** Qwen leads attention (79.3%,  $d = +2.45$ ) while Mistral leads gradient (71.5%,  $d = +0.93$ )—a striking method divergence. GPT-2 shows anti-faithfulness (29.7%). Turkish’s agglutinative morphology may favor gradient’s position-independent scoring.

**Arabic:** Mistral attention achieves 72.6% ( $d = +0.78$ ) while its gradient drops to 39.9%—the largest attention-gradient gap for any language. Arabic’s right-to-left script and root-pattern morphology create unique challenges for positional attribution.

## G Multilingual Detailed Results

Model	FR	DE	HI	ZH	TR	AR
<i>Deletion — Attention Win Rate (%)</i>						
GPT-2	15.8	49.1	65.4	57.5	29.7	<b>62.4</b>
Llama 3.1-8B	80.8	39.1	44.6	49.1	73.0	63.4
Llama 3.2-3B	44.9	49.0	53.7	58.3	59.2	37.1
Qwen 2.5-7B	62.6	<b>82.7</b>	50.2	60.6	<b>79.3</b>	57.0
Mistral 7B	53.5	65.6	53.1	60.2	57.2	<b>72.6</b>
DeepSeek 7B	<b>67.2</b>	43.2	<b>65.4</b>	56.1	49.0	†
LFM2-2.6B	59.6	48.6	62.3	52.7	73.7	69.0
<i>Retrieval Infill — Attention Win Rate (%)</i>						
GPT-2	23.9	52.2	<b>68.8</b>	50.0	43.2	71.7
Llama 3.1-8B	<b>81.8</b>	35.1	35.1	37.3	70.3	44.8
Llama 3.2-3B	39.7	46.2	52.3	52.9	54.8	23.5
Qwen 2.5-7B	62.3	70.1	51.0	56.2	<b>81.6</b>	48.5
Mistral 7B	51.0	55.5	51.3	50.0	53.1	<b>62.9</b>
DeepSeek 7B	54.4	12.0	†	†	50.7	80.0
LFM2-2.6B	65.1	51.6	66.1	<b>58.3</b>	65.6	71.1

Table 7: Multilingual attention win rates (%) under both operators. † = no valid output (Arabic: model limitation; HI/ZH retrieval: tokenizer failure on DeepSeek).

Model	FR	DE	HI	ZH
GPT-2	1.8×	2.0×	8.1×	10.7×
Llama 3.x	1.3×	1.4×	2.5×	4.2×
Mistral	1.4×	1.5×	5.0×	5.7×

Table 8: Token expansion ratios by language ( $\times$  = tokens per character vs. English).

## H Encoder Validation Results

Table 9 reports ICE evaluation on BERT-base-uncased across five ERASER datasets.

Extractor	Suf.	Sig. Rate	AUC-Suf
<i>SST-2 (500 examples)</i>			
LIME	<b>0.617</b>	<b>11.4%</b>	<b>0.302</b>
Integrated Gradients	0.492	0%	0.256
Attention	0.398	0%	0.250
Gradient	0.394	0%	0.253
<i>IMDB (500 examples)</i>			
Gradient	0.519	<b>57.4%</b>	0.345
Attention	0.385	33.2%	0.346
LIME	0.182	0%	0.284
Integrated Gradients	0.149	0%	0.326
<i>e-SNLI (417 examples)<sup>†</sup></i>			
LIME	<b>0.450</b>	0%	0.195
Integrated Gradients	0.406	0%	0.190
Gradient	0.383	0%	0.194
Attention	0.352	0%	0.152
<i>BoolQ (500 examples)</i>			
Gradient	0.071	0%	0.062
Integrated Gradients	0.070	0%	0.056
Attention	0.066	0%	0.055
LIME	0.058	0%	0.046
<i>MultiRC (500 examples)</i>			
Attention	<b>0.103</b>	0%	<b>0.098</b>
Gradient	0.079	0%	0.123
Integrated Gradients	0.071	0%	0.074
LIME	0.068	0%	0.049

Table 9: Encoder results on BERT-base-uncased. <sup>†</sup>417/500 after filtering.

## I Effect Sizes and Anti-Faithfulness Details

Tables 10–13 provide effect sizes, anti-faithful configurations, concrete anti-faithful examples, and bootstrap confidence intervals.

## J Faithfulness-Plausibility Correlation

Table 14 reports IoU-faithfulness correlations across three models on e-SNLI.

Configuration	WR	$d$	Interp.
Llama 3.1 e-SNLI Grad	97.2%	2.50	Ext. large
Qwen IMDB Attn	94.9%	1.96	V. large
Qwen IMDB Grad	91.4%	1.84	Large
Llama 3.2 e-SNLI Attn	86.4%	3.77	V. large
Qwen DE Attn	82.7%	1.40	Large
Llama 3.1 FR Attn	80.8%	1.26	Large
GPT-2 FR Attn	15.8%	-2.08	Anti
GPT-2 FR Grad	14.8%	-2.36	Anti
GPT-2 e-SNLI Grad	29.5%	-0.72	Anti
DeepSeek AG Grad	39.1%	-0.53	Anti

Table 10: Effect sizes ( $d$ ). WR=Win Rate.  $d > 0.8$  = large. Anti = anti-faithful.

Model	Config	WR	$d$
GPT-2	FR Attn/Grad	15–16%	-2.1
GPT-2	e-SNLI Grad	29.5%	-0.72
GPT-2	DE Grad	34.2%	-0.39
DeepSeek	SST-2 Grad	40.5%	-0.29
DeepSeek	AG Grad	39.1%	-0.53
Llama 3.2	SST-2 Grad	42.4%	-0.44
Llama 3.1	SST-2 Grad	46.0%	-0.15
Llama 3.1	AG Grad	42.2%	-0.40

Table 11: Anti-faithful configurations (WR < 50%). Negative  $d$  = worse than random.

Text	WR	$d$	Pattern
“gorgeous, witty, seduc- tive”	0%	-2.3	Selects $a$ ; ignores ad- jectives
“tender, heartfelt drama”	0%	-1.1	Selects $a$
“fast, funny, enjoyable”	0%	-2.3	Selects $a$
“high comedy, poignance”	0%	-2.8	Selects $uses$

Table 12: Anti-faithful examples (Llama 3.1-8B/SST-2). Gradient selects initial function words, ignoring sentiment.

Configuration	Win Rate	95% CI
Llama 3.1 e-SNLI Grad	97.2%	[95.4, 99.0]
Qwen IMDB Attn	94.9%	[92.1, 97.7]
Llama 3.2 e-SNLI Attn	86.4%	[82.1, 90.7]
GPT-2 DE Grad	34.2%	[28.7, 39.7]
GPT-2 FR Attn	15.8%	[11.2, 20.4]

Table 13: Bootstrap 95% CIs. Non-overlapping with 50% indicates significant departure from random.

## K Retrieval Infill Detailed Results

Table 15 and Figure 10 provide the full operator comparison data.

## L Relocated English Results Table

Table 16 provides the full numerical English win rates underlying Figure 3. Table 17 summarizes practical recommendations.

Model	Method	r	p	N
GPT-2 (1.5B)	Attention	0.016	0.73	462
	Gradient	-0.018	0.69	493
DeepSeek-7B	Attention	0.033	0.53	370
	Gradient	0.019	0.68	487
Mistral-7B	Attention	0.016	0.77	351
	Gradient	0.012	0.79	485

Table 14: IoU-Faithfulness correlation across three models. No model shows significant correlation ( $|r| < 0.04$ , all  $p > 0.5$ ).

Configuration	Delete		Retrieval	
	WR	Tax	WR	Tax
Qwen / SST-2	<b>68.4</b>	TF	59.6	LT
Qwen / e-SNLI	<b>77.3</b>	TF	77.5	TF
Llama-3.2 / e-SNLI	<b>86.4</b>	TF	42.6	RG
Mistral / AG News	51.7	LT	42.7	RG
DeepSeek / SST-2	<b>62.2</b>	TF	53.9	LT
GPT-2 / SST-2	<b>60.6</b>	TF	52.4	LT

Table 15: Attention win rates for Delete vs. Retrieval Infill ( $k = 0.2$ ). Tax: TF=Truly Faithful, LT=Lucky Tokens, CD=Context-Dependent, RG=Random Guess. Same models as main results.

Model	SST-2	IMDB	e-SNLI	AG News
<i>Deletion — Attention Win Rate (%)</i>				
GPT-2	60.6	44.0	64.0	<b>70.8</b>
Llama 3.2-3B	53.2	71.3	<b>86.4</b>	56.0
Llama 3.1-8B	53.2	52.5	85.2	47.5
Qwen 2.5-7B	<b>68.4</b>	<b>94.9</b>	77.3	62.2
Mistral 7B	57.5	83.8	71.1	51.7
DeepSeek 7B	62.2	84.6	63.7	49.4
LFM2-2.6B	45.4	50.3	66.6	57.0
<i>Retrieval — Attention Win Rate (%)</i>				
GPT-2	52.4	38.3	68.2	<b>73.2</b>
Llama 3.2-3B	52.7	<b>91.8</b>	42.6	56.9
Llama 3.1-8B	51.5	75.3	74.0	49.4
Qwen 2.5-7B	59.6	<b>96.3</b>	<b>77.5</b>	59.8
Mistral 7B	56.1	89.8	67.5	42.7
DeepSeek 7B	53.9	80.5	65.5	46.7
LFM2-2.6B	45.5	49.4	65.5	51.1

Table 16: English attention win rates under both operators. Bold = best per column/operator.

## M Extension to Chain-of-Thought

ICE’s methodology generalizes from feature attributions to generated reasoning. For a model generating reasoning  $R$  followed by answer  $a$ , ICE-CoT applies two tests: the *necessity test* corrupts CoT tokens and checks whether the answer changes, and the *sufficiency test* presents only the reasoning and checks whether the model can recover the answer. Evaluating 6 models on SST-2, e-SNLI,

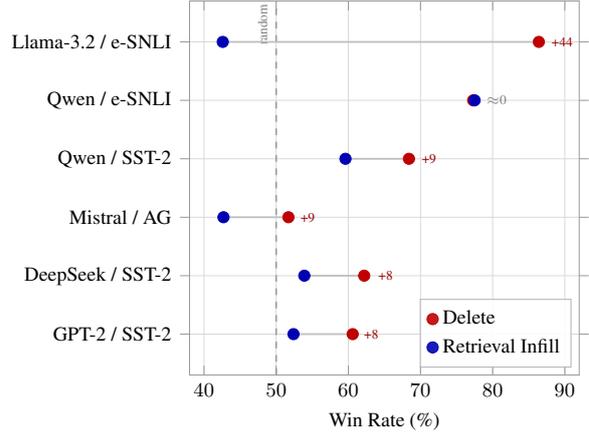


Figure 10: Paired comparison of attention win rates under Delete vs. Retrieval Infill (main models). Delete (red) exceeds Retrieval Infill on most short-text configurations by 8–44 pp, but operators agree on Qwen/e-SNLI ( $\approx 77\%$ ), suggesting genuine faithfulness there.

Scenario	Recommendation
Short text, Sentiment	Use Attention
Long text	Either method works
NLI	Use Attention
Topic Classification	Use Gradient
French	Llama 3.1 Attn (81%) or LFM2 Grad (73%)
German	Qwen Attn (83%) – others struggle
Chinese	LFM2 Grad (69%) or Llama 3.1 Grad (66%)
Hindi	GPT-2 (65–66%) works surprisingly well
Turkish	Qwen Attn (82%) or Mistral Grad (67%)
Arabic	Mistral Attn (73%) – avoid Llama-3.2 (37%)

Table 17: Practical attribution guidelines based on ICE evaluation.

AG News, and GSM8K, most CoT on classification falls into “Lucky Tokens” (44%) or “Random Guess” (43%). Mathematical reasoning (GSM8K) shows an inverted pattern: high necessity but low sufficiency. Retrieval Infill gives more conservative estimates than deletion, consistent with our attribution results.

**Reproducibility.** Code, results, pre-trained extractors, cached win rates, and reproduction scripts are provided in the supplementary material.