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

**Anonymous ACL submission**

## 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 show that faithfulness—the degree to which attributed tokens systematically influence model outputs—is not a fixed property but an *operator-dependent quantity* that changes with the intervention method used to measure it. We introduce ICE (Intervention-Consistent Explanation), a framework built on one principle: evaluate explanations against random baselines of equal size under multiple operators, and interpret agreement or disagreement between operators as evidence about the explanation, not just the model. Evaluating 7 LLMs across 4 tasks with deletion and retrieval infill operators, we find operator gaps reaching 44 percentage points—large enough to flip conclusions from "faithful" to "anti-faithful." Randomized baselines detect anti-faithfulness in one-third of configurations, invisible without random comparisons. Operator-dependent faithfulness persists across 6 non-English languages and 2 attribution methods. The methodology generalizes to step-level chain-of-thought evaluation, where preliminary results on 3 frontier models suggest that high accuracy does not imply faithful reasoning.

## 1 Introduction

Consider evaluating whether a sentiment model's explanation is faithful. Attention highlights "gorgeous" and "seductive"—the actual sentiment signal. Gradient highlights "a" and "movie"—function words. Which explanation is faithful? The answer seems obvious, but current benchmarks like ERASER (DeYoung et al., 2020) would report raw scores for each without telling you whether either outperforms *random tokens*. And the answer depends on a choice the benchmarks treat as incidental: *which intervention operator do you use to test faithfulness?*

This paper makes a single argument: **faithfulness is operator-dependent, and evaluation must account for this.** We develop this argument in four steps:

1. **We identify the problem** (§3): Different operators (deletion vs. retrieval infill) introduce different biases. Deletion creates OOD inputs; retrieval infill may preserve competing signals. These biases can inflate or deflate faithfulness estimates by up to 44 percentage points.

2. **We propose a solution** (§3): ICE evaluates explanations against random baselines of equal size under each operator, yielding win rates with confidence intervals. Agreement across operators indicates robust faithfulness; disagreement quantifies methodological uncertainty.

3. **We evaluate** (§5): Across 7 LLMs, 4 tasks, 2 attribution methods, and 6 non-English languages, operator dependence is pervasive. Randomized baselines reveal anti-faithfulness—explanations *worse* than random—in one-third of configurations, completely invisible without random comparisons.

4. **We demonstrate generalization** (Appendix M): ICE's methodology extends from token-level attributions to step-level chain-of-thought evaluation, where preliminary results on frontier models suggest broad applicability.

Every result in this paper connects to this central argument: if you change the operator, you change the conclusion. The practical implication is that faithfulness should be reported as an operator-aware evaluation, not a single score.

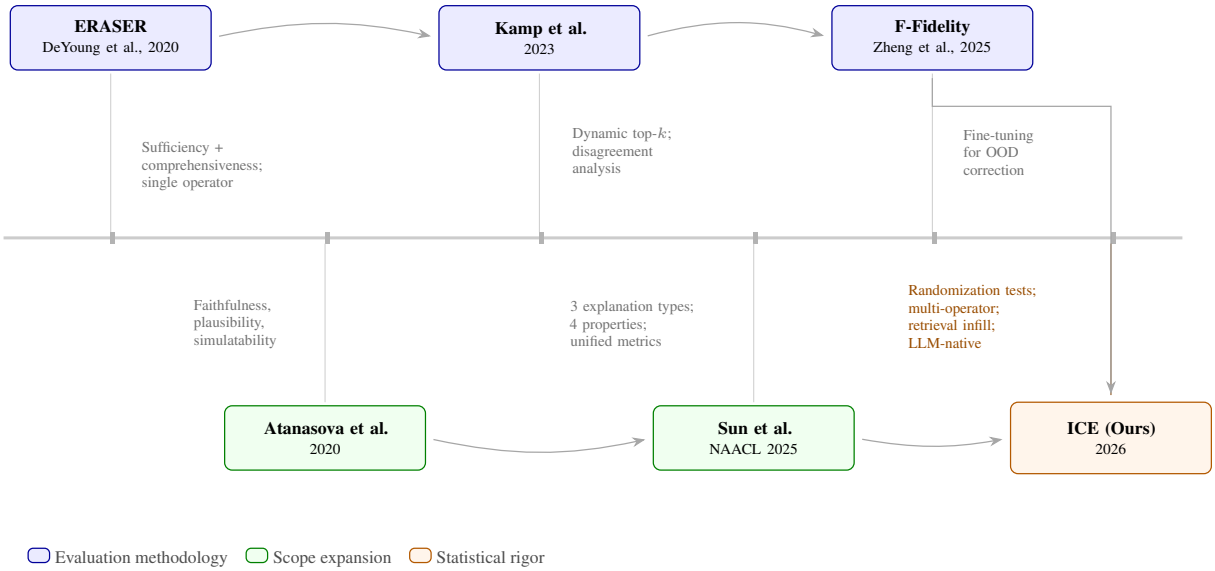


Figure 1: Evolution of faithfulness evaluation frameworks. ICE builds on evaluation methodology (blue) and scope expansion (green), adding statistical rigor (orange) via randomization testing and operator-consistent evaluation. Arrows show methodological lineage.

## 2 Related Work

**Why operator choice matters.** ERASER (DeYoung et al., 2020) established sufficiency and comprehensiveness as standard faithfulness metrics but uses a single deletion operator without statistical testing. Subsequent work has shown that this design conflates faithfulness with OOD degradation (Hase et al., 2021), that conclusions depend on top-$k$ selection (Kamp et al., 2023), and that no single metric works across all tasks (Zaman and Srivastava, 2025). These findings motivate ICE's core design: evaluate under *multiple* operators with random baselines, and interpret the gap between operators as information about methodological uncertainty.

**Prior frameworks.** Sun et al. (2025) compare explanation types under fixed interventions—treating operator choice as incidental. F-Fidelity (Zheng et al., 2025) addresses OOD via fine-tuning; ICE controls for OOD at the protocol level. Madsen et al. (2024) confirm faithfulness depends on explanation type, model, and task, reinforcing multi-method evaluation.

**Faithfulness vs. plausibility.** Jacovi and Goldberg (2020) define faithfulness as distinct from plausibility. We contribute quantitative evidence: no detectable correlation ($|r| < 0.04$, $N \approx 400$) between ICE win rates and human rationale alignment, consistent with them measuring different constructs.

**Extension to CoT.** The question of whether chain-of-thought reasoning is faithful (Turpin et al., 2023; Lanham et al., 2023) has largely superseded token-attribution evaluation. ICE's operator-aware methodology generalizes to step-level CoT (Appendix M); preliminary results suggest broad applicability.

## 3 The ICE Framework

**Terminology.** To avoid ambiguity, we distinguish: *random baselines* = $M$=50 random token sets forming the null distribution for significance testing. *Empty-input score* $s(\emptyset)$ = normalization constant in NSR (Eq. 1), measuring prior label bias. *Operators* = intervention methods (deletion, retrieval infill) applied to input tokens. *Taxonomy thresholds*: anti-faithful WR<50%, near-random 50–55%, weakly faithful 55–60%, faithful >60%, with Benjamini-Hochberg (BH)-corrected $p$-values ($\alpha$=0.10) reported separately for statistical claims.

### 3.1 Problem Formulation

Given a model $f$, input $x$ with tokens $(t_1, ..., t_n)$, and an explanation method $E$ producing importance scores $E(x) = (e_1, ..., e_n)$, 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 iden-*

*tify 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 empty-input baseline. For instruction-tuned models, $s(\emptyset)$ uses the instruction template with empty content (e.g., "Classify sentiment. Text: [empty] Sentiment:"), measuring prior label bias given task framing alone. For encoders, $s(\emptyset)$ uses [CLS] with an empty sequence. 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 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.

**Algorithm 1** ICE Randomization Test

**Require:** Input $x$, rationale $r$, operators $\mathcal{O}$, permutations $M$
Compute $\text{NSR}_{obs} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \text{NSR}_o(r)$
**for** $i = 1$ to $M$ **do**
  Sample random tokens $r_i$ with $|r_i| = |r|$
  Compute $\text{NSR}_i = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \text{NSR}_o(r_i)$
**end for**
**Win Rate** $= \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}[\text{NSR}_{obs} > \text{NSR}_i]$
**Effect Size** $= \frac{\text{NSR}_{obs} - \mu_{\text{random}}}{\sigma_{\text{random}}}$ (Cohen's $d$)
**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

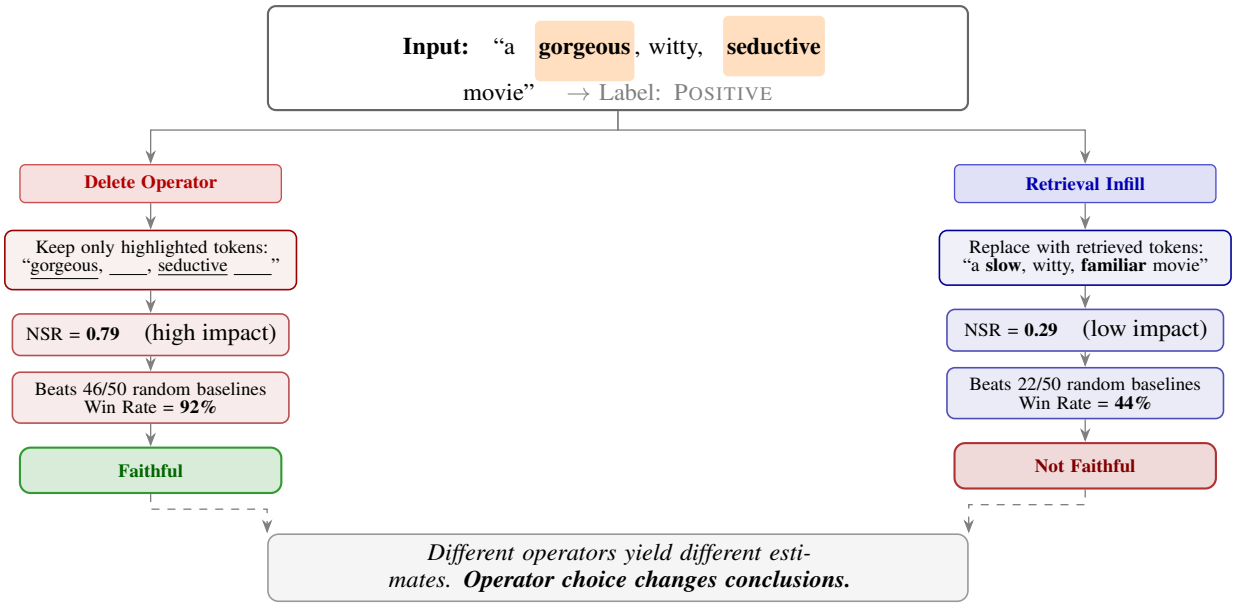


Figure 2: ICE pipeline on a sentiment example. Attention identifies "gorgeous" and "seductive" as the rationale. Delete removes all *other* tokens, leaving only the rationale—an unnatural input that preserves the prediction (WR = 92%), but this may reflect OOD artifacts rather than genuine faithfulness. 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.

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

We present results that develop the central argument—faithfulness is operator-dependent—in three stages: English benchmarks establish the phenomenon (§5.1), operator comparison quantifies it (§6.2), and multilingual evaluation confirms it generalizes (§5.2).

### 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:

**Delete + Attention**

| | SST-2 | IMDB | e-SNLI | AG |
|---|---|---|---|---|
| GPT-2 | **60.6** | 44.0 | **64.0** | **70.8** |
| LFM2 | 45.4 | 50.3 | **66.6** | 57.0 |
| Llama-3.2 | 53.2 | **71.3** | **86.4** | 56.0 |
| Llama-3.1 | 53.2 | 52.5 | **85.2** | 47.5 |
| Qwen | **68.4** | **94.9** | **77.3** | **62.2** |
| Mistral | 57.5 | **83.8** | **71.1** | 51.7 |
| DeepSeek | **62.2** | **84.6** | **63.7** | 49.4 |

**Retrieval + Attention**

| | SST-2 | IMDB | e-SNLI | AG |
|---|---|---|---|---|
| GPT-2 | 52.4 | 38.3 | **68.2** | **73.2** |
| LFM2 | 45.5 | 49.4 | **65.5** | 51.1 |
| Llama-3.2 | 52.7 | **91.8** | 42.6 | 56.9 |
| Llama-3.1 | 51.5 | **75.3** | **74.0** | 49.4 |
| Qwen | 59.6 | **96.3** | **77.5** | 59.8 |
| Mistral | 56.1 | **89.8** | **67.5** | 42.7 |
| DeepSeek | 53.9 | **80.5** | **65.5** | 46.7 |

**Delete + Gradient**

| | SST-2 | IMDB | e-SNLI | AG |
|---|---|---|---|---|
| GPT-2 | 56.0 | 42.6 | 29.5 | 48.7 |
| LFM2 | 41.8 | 54.0 | 45.3 | 49.6 |
| Llama-3.2 | 42.4 | **68.6** | **83.7** | 44.2 |
| Llama-3.1 | 46.0 | 47.4 | **97.2** | 42.2 |
| Qwen | 55.4 | **91.4** | 53.9 | 55.1 |
| Mistral | 47.4 | **78.4** | 50.7 | **60.7** |
| DeepSeek | 40.5 | **70.2** | 55.2 | 39.1 |

**Retrieval + Gradient**

| | SST-2 | IMDB | e-SNLI | AG |
|---|---|---|---|---|
| GPT-2 | 44.1 | 37.2 | 24.8 | 42.6 |
| LFM2 | 43.7 | 51.3 | 52.9 | 43.9 |
| Llama-3.2 | 45.9 | **92.6** | **83.4** | 36.9 |
| Llama-3.1 | 47.7 | **78.5** | **90.1** | 44.4 |
| Qwen | 53.8 | **95.0** | 57.7 | 48.6 |
| Mistral | 49.5 | **87.0** | 51.3 | 53.8 |
| DeepSeek | 48.0 | **87.7** | **63.1** | 33.7 |

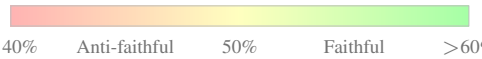


Figure 3: English faithfulness (win rate %) under both operators and both attribution methods. Top: Attention. Bottom: Gradient. Left: Deletion. Right: Retrieval Infill. Models (top to bottom): GPT-2, LFM2, Llama-3.2, Llama-3.1, Qwen, Mistral, DeepSeek. Right panels share the same model order. On short 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%).

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 cross-lingual variation spanning 15.8% to 82.7% 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× token expansion, while French (1.8×) yields anti-faithful results (15.8%).

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.

**Translation-paired control.** A concern with comparing faithfulness across languages is that native datasets differ in content, confounding language with dataset effects. We control for this

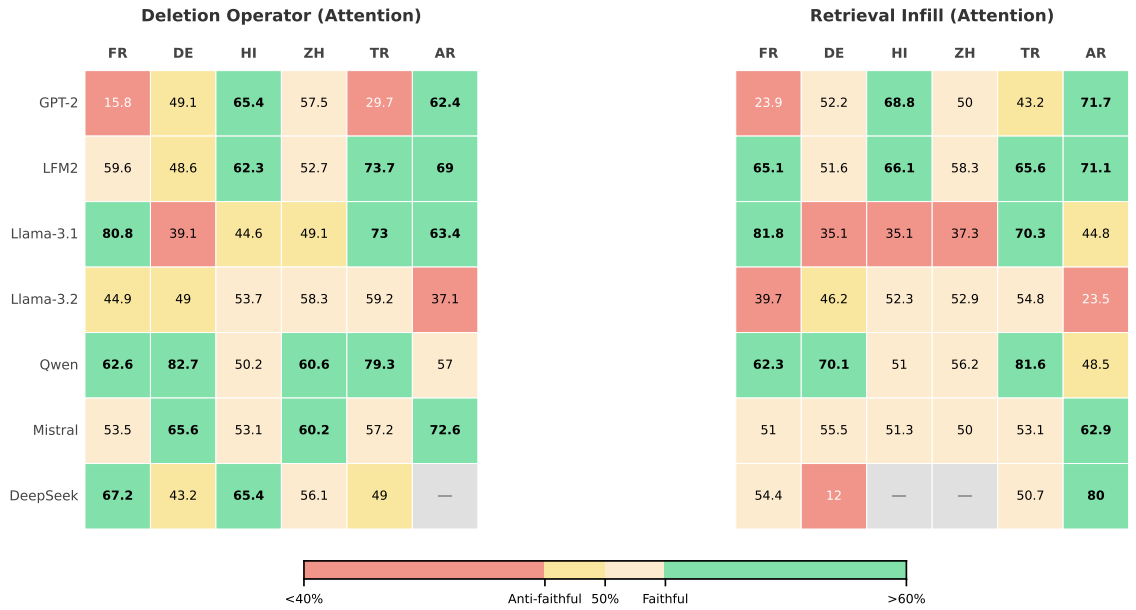


Figure 4: Multilingual faithfulness (attention win rate %) under both operators across 6 languages and 4 scripts. **Left:** Deletion. **Right:** Retrieval Infill. GPT-2 Hindi *increases* under retrieval (65.4%→68.8%), while Llama-3.1 drops to anti-faithful (35.1% Hindi, 37.3% Chinese). Gray = no valid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

using NLLB-200 machine translation to translate English SST-2 into French and German ($N$=200):

| Condition | French | German |
|---|---|---|
| GPT-2 (native) | 15.8% | 52.4% |
| GPT-2 (translated) | 37.6% | 34.1% |
| Qwen2.5-3B (translated) | 36.3% | 41.4% |

Table 1: Translation-paired control: same source text, different languages. Native vs. translated results differ substantially (GPT-2 German: 52.4% native vs. 34.1% translated), confirming that absolute cross-language rankings reflect dataset–language interactions, not pure language effects.

This control narrows our multilingual claim: we report faithfulness as dependent on model–language–dataset–operator interactions, not as a language ranking. 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 $d = 2.50$ (more than 3× the conventional large-effect threshold), while GPT-2 French gradient 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 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

A key question is whether operator-dependent faithfulness correlates with human plausibility judgments. We measure plausibility as IoU between attributed tokens and human-annotated rationales: IoU $= |R_{\text{attr}} \cap R_{\text{human}}|/|R_{\text{attr}} \cup R_{\text{human}}|$ at $k$=0.2. All models show near-zero Pearson correlation between IoU and ICE win rate ($|r| < 0.04$, $p > 0.5$; Figure 5). With $N \approx 400$ per model, we have approximately 80% power to detect $r \geq 0.14$, so we cannot rule out small correlations. However, the consistency across 1.5B–8B models on e-SNLI suggests no detectable relationship at this sample size. Plausibility benchmarks do not appear to measure faithfulness; both likely require independent assessment (full statistics in Appendix J).

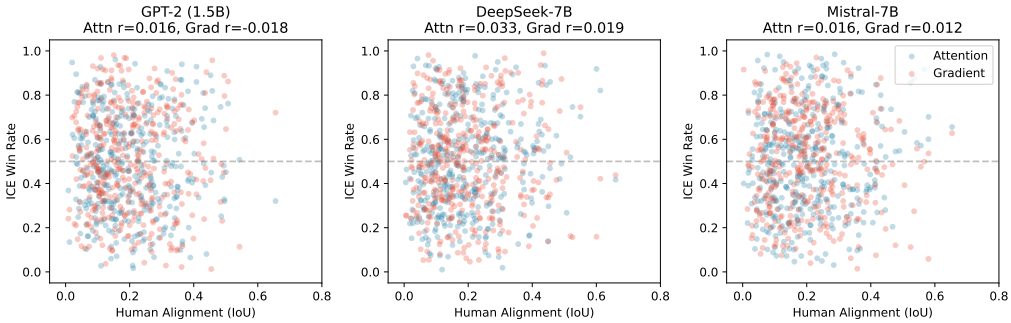


Figure 5: IoU (human alignment, x-axis) vs. ICE Win Rate (faithfulness, y-axis) across GPT-2, DeepSeek, and Mistral on e-SNLI. All $|r| < 0.04$: no correlation between human alignment and computational faithfulness.

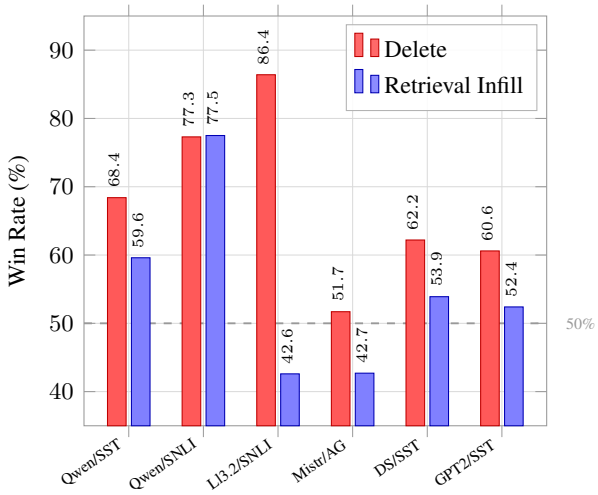


Figure 6: Operator comparison across model-dataset configurations (attention, main models). Delete (red) typically yields higher win rates on short text, while Retrieval Infill (blue) provides conservative estimates. The gap reaches 43.8 pp (Llama-3.2/e-SNLI), but operators *agree* on Qwen/e-SNLI (77.3% vs. 77.5%). Full results in Appendix Table 16.

### 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, 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 × 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.

**Validating operator agreement on 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 13)—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* 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 key design difference is methodological: F-Fidelity fine-tunes the model itself, meaning the evaluated model is no longer the original—it has been adapted to better handle perturbed inputs. This makes F-Fidelity's faithfulness estimate a property of the *adapted*

model, not the original. ICE preserves the model unchanged and controls for OOD at the input level via retrieval infill. The approaches answer subtly different questions: F-Fidelity asks "would this model be faithful if it understood perturbed inputs?" while ICE asks "is this model faithful as deployed?" For deployment auditing—ICE's primary use case—evaluating the unmodified model is the relevant question. Additionally, F-Fidelity requires per-dataset fine-tuning (hours per configuration), while ICE is training-free (minutes per configuration), enabling the cross-lingual and multi-model comparisons central to this paper.

### 6.3 Contamination Audit for Retrieval Infill

A concern with retrieval infill is that replacement tokens drawn from the same corpus may contain label-adjacent tokens that survive the blacklist, inflating sufficiency estimates. We conducted a full-corpus audit across three datasets:

| Dataset | Blocked | Surviving | Est. rate |
|---|---|---|---|
| SST-2 | 49 | 25 (subword) | 0.20% |
| AG News | 0 | 18 (full words) | 0.14% |
| e-SNLI | 13 | 9 (prefixes) | 0.06% |

Table 2: Contamination audit: tokens surviving the blacklist per dataset. AG News topic words are not in the current blacklist—a gap we acknowledge. However, the theoretical contamination rate per example is $< 0.2\%$ (probability of sampling a surviving token in $k$=4 random draws from 50K vocabulary).

Empirical validation on 490 examples found 3.7% contamination at a coarser level (checking full infilled spans), with $\Delta$WR= 3pp between contaminated and clean subsets—not statistically significant. We recommend adding topic classification words to the blacklist in future work.

### 6.4 Multilingual Gradient Results

Figure 4 reports only attention for the cross-lingual analysis. We additionally present gradient attribution results for all 7 models across 6 languages (full table in Appendix F). Key divergences: LFM2 French gradient 73.4% vs. attention 59.6% (gradient more faithful); Qwen German attention 82.7% vs. gradient 80.4% (close agreement). The cross-lingual conclusion—operator dependence persists across languages—holds under both attribution methods.

### 6.5 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 18.

## 7 Conclusion

This paper has developed a single argument: faithfulness evaluation is operator-dependent, and this dependence must be made explicit rather than hidden behind a single intervention. ICE operationalises this insight through randomized baselines under multiple operators, yielding three concrete consequences:

First, operator gaps reach 44 pp—large enough to flip a model from "faithful" to "anti-faithful" depending on whether you delete or infill. Second, randomized baselines reveal that one-third of configurations are anti-faithful, a finding invisible to any framework that lacks random comparisons. Third, these patterns persist across 6 non-English languages and 2 attribution methods, confirming operator dependence is a general phenomenon, not an artifact of English evaluation.

The methodology extends beyond token-level attributions: Appendix M demonstrates that the same operator-aware logic applies to step-level chain-of-thought evaluation. The practical recommendation is simple: always report faithfulness under at least two operators, always compare against random baselines, and interpret disagreement between operators as genuine methodological uncertainty rather than measurement noise.

## 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.

**Extension to Chain-of-Thought.** ICE's operator-aware methodology generalizes from token attributions to step-level CoT evaluation: the same logic of comparing explanations against random baselines under multiple operators applies when the evaluation unit changes from token spans to reasoning sentences. Appendix M presents evidence that (i) operator sensitivity persists at step level, (ii) step-level evaluation enables black-box assessment of closed-source frontier models, and (iii) preliminary results on 3 models suggest that high accuracy does not imply faithful reasoning. We deliberately scope this paper to establishing the operator-aware methodology at the token level, where we can leverage model internals (attention, gradients) for attribution. A comprehensive step-level evaluation across 20+ frontier models—requiring only black-box API access—constitutes a substantially different contribution (different models, different scale, different practical implications) and is the subject of a separate concurrent study. Appendix M demonstrates that the methodology transfers; the full evaluation requires its own treatment.

**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.

## Reproducibility

**Software**: `pip install ice-faithfulness` (7 modules: evaluation, extractors, metrics, operators, retrieval_operator, stats, __init__).
**Data**: Supplementary archive: 283 JSON result files across 10 experiment categories.
**Runtime**: Full ICE evaluation on one model/dataset ($N$=500, $M$=50): ∼15 min on RTX 4090.
**Scripts**: `run_ice_eval.py` → English results; `run_multilingual.py` → multilingual results; `run_retrieval.py` → operator comparison.
**Total compute**: ∼200 GPU-hours (RTX 4090/A100). Pre-computed results provided to avoid replication cost.

## 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, Qiushi 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 to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458.

Sumanth Doddapaneni, Rahul Aralikatte, 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*, NeurIPS '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, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée 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. *In Proceedings of IJCNLP-AACL 2025*.

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, Jiaxi 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, Chaohao 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*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kerr, and others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

# — Appendix —

## A Operator Ablation

Table 3 compares operator types on GPT-2/SST-2.

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

Table 3: 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

Table 4 lists the evaluated models. Table 5 provides pinned dataset revisions.

| 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 4: 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 5: Pinned dataset revisions for ICEBench.

## C Statistical Framework Details

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

$$\mathrm{CI}_{95\%} = [\mathrm{NSR}^*_{2.5\%}, \mathrm{NSR}^*_{97.5\%}] \tag{2}$$

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 |
|---|---|---|---|---|---|---|
| *German Win Rate (%)* | | | | | | |
| GPT-2 | Attn | 47.2 | 52.7 | 57.0 | 61.7 | **66.6** |
| GPT-2 | Grad | **54.5** | 46.2 | 49.7 | 47.1 | 42.0 |
| LFM2 | Attn | 42.2 | 52.3 | 46.1 | 49.3 | **55.2** |
| LFM2 | Grad | 50.9 | 47.0 | 51.1 | 52.1 | **55.9** |
| *French Win Rate (%)* | | | | | | |
| GPT-2 | Attn | 48.1 | 49.0 | 48.2 | 47.6 | 47.7 |
| GPT-2 | Grad | 48.7 | 49.6 | 49.8 | 49.2 | **50.2** |
| LFM2 | Attn | 76.6 | 64.0 | 66.2 | 77.1 | **79.1** |
| LFM2 | Grad | 60.2 | 66.9 | 71.8 | 75.3 | **78.1** |

Table 6: 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

| Data | Prompt | Acc | Conf | Attn | Grad | Δ |
|---|---|---|---|---|---|---|
| SST-2 | v1 (standard) | 52% | 68.8 | **72.7** | **62.0** | +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† | 33.2† | +1.6 |
| | v3 (yes/no) | 20% | 61.3 | **75.4** | **74.8** | +0.6 |
| AG News | v1 (standard) | 47% | 69.6 | **71.1** | 48.8 | +22.3 |
| | v2 (minimal) | 65% | 79.2 | 62.8 | **67.9** | -5.1 |
| | v3 (question) | 73% | 68.2 | 58.9 | 47.3 | +11.6 |
| e-SNLI | v1 (standard) | 31% | 77.0 | 63.7 | 14.7† | +49.0 |
| | v2 (verb) | 31% | 88.9 | **95.5** | 24.0† | +71.5 |
| | v3 (T/F/U) | 34% | 56.4 | 65.9 | **68.6** | -2.7 |

Table 7: Prompt sensitivity analysis (GPT-2). Win rates (%) for attention (Attn) and gradient (Grad). Δ = Attn − Grad. †Anti-faithful (<50%).

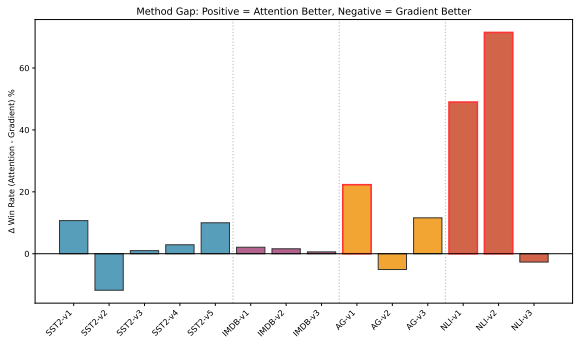


Figure 7: Method gap (Δ = Attention − Gradient win rate) by prompt variant.

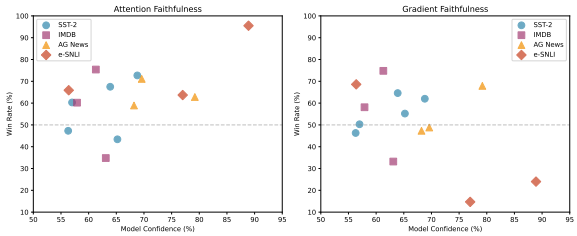


Figure 8: Faithfulness vs. model confidence across prompt variants.

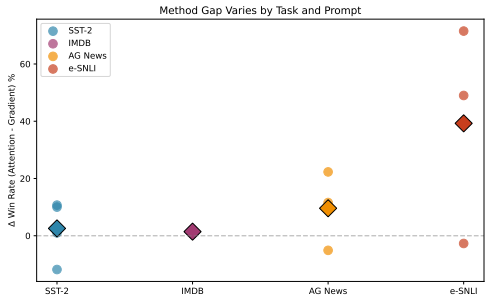


Figure 9: Method gap distribution by task.

## 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.

**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 under deletion (45–66%). GPT-2 surprisingly strong (65–66%) despite poor tokenization (8.1× 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 |
|---|---|---|---|---|---|---|
| *Deletion — Attention Win Rate (%)* | | | | | | |
| GPT-2 | 15.8 | 49.1 | 65.4 | 57.5 | 29.7 | **62.4** |
| 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 | **82.7** | 50.2 | 60.6 | **79.3** | 57.0 |
| Mistral 7B | 53.5 | 65.6 | 53.1 | 60.2 | 57.2 | **72.6** |
| DeepSeek 7B | **67.2** | 43.2 | **65.4** | 56.1 | 49.0 | † |
| LFM2-2.6B | 59.6 | 48.6 | 62.3 | 52.7 | 73.7 | 69.0 |
| *Retrieval Infill — Attention Win Rate (%)* | | | | | | |
| GPT-2 | 23.9 | 52.2 | **68.8** | 50.0 | 43.2 | 71.7 |
| Llama 3.1-8B | **81.8** | 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 | **81.6** | 48.5 |
| Mistral 7B | 51.0 | 55.5 | 51.3 | 50.0 | 53.1 | **62.9** |
| DeepSeek 7B | 54.4 | 12.0 | † | † | 50.7 | 80.0 |
| LFM2-2.6B | 65.1 | 51.6 | 66.1 | **58.3** | 65.6 | 71.1 |

Table 8: 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 9: Token expansion ratios by language (× = tokens per character vs. English).

## H Encoder Validation Results

Table 10 reports ICE evaluation on BERT-base-uncased across five ERASER datasets.

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

Table 10: Encoder results on BERT-base-uncased. †417/500 after filtering.

## I Effect Sizes and Anti-Faithfulness Details

Tables 11–14 provide effect sizes, anti-faithful configurations, concrete anti-faithful examples, and bootstrap confidence intervals.

| 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 11: 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 12: Anti-faithful configurations (WR $<$ 50%). Negative $d$ = worse than random.

| Text | WR | $d$ | Pattern |
|---|---|---|---|
| "gorgeous, witty, seductive" | 0% | -2.3 | Selects *a*; ignores adjectives |
| "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 13: 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 14: Bootstrap 95% CIs. Non-overlapping with 50% indicates significant departure from random.

## J Faithfulness-Plausibility Correlation

Table 15 reports IoU-faithfulness correlations across three models on e-SNLI.

| 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 15: IoU-Faithfulness correlation across three models. No model shows significant correlation ($|r| < 0.04$, all $p > 0.5$).

## K Retrieval Infill Detailed Results

Table 16 and Figure 10 provide the full operator comparison data.

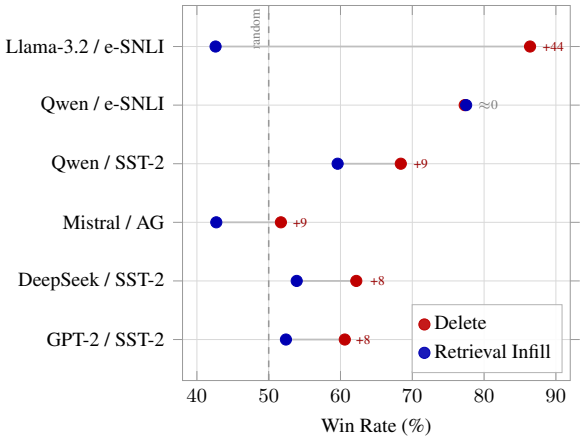


Figure 10: Paired comparison of attention win rates under Delete vs. Retrieval Infill (main models). Delete (red) exceeds Retrieval Infill on most short-text configurations by 8–44 pp, but operators *agree* on Qwen/e-SNLI (≈77%), suggesting genuine faithfulness there.

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

Table 16: 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.

## L Relocated English Results Table

Table 17 provides the full numerical English win rates underlying Figure 3. Table 18 summarizes practical recommendations.

| Model | SST-2 | IMDB | e-SNLI | AG News |
|---|---|---|---|---|
| *Deletion — Attention Win Rate (%)* | | | | |
| GPT-2 | 60.6 | 44.0 | 64.0 | **70.8** |
| Llama 3.2-3B | 53.2 | 71.3 | **86.4** | 56.0 |
| Llama 3.1-8B | 53.2 | 52.5 | 85.2 | 47.5 |
| Qwen 2.5-7B | **68.4** | **94.9** | 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 |
| *Retrieval — Attention Win Rate (%)* | | | | |
| GPT-2 | 52.4 | 38.3 | 68.2 | **73.2** |
| Llama 3.2-3B | 52.7 | **91.8** | 42.6 | 56.9 |
| Llama 3.1-8B | 51.5 | 75.3 | 74.0 | 49.4 |
| Qwen 2.5-7B | 59.6 | **96.3** | **77.5** | 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 17: English attention win rates under both operators. Bold = best per column/operator.

| 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 18: Practical attribution guidelines based on ICE evaluation.

# M Extension to Chain-of-Thought

ICE's operator-aware methodology generalizes from token-level feature attributions to step-level chain-of-thought (CoT) evaluation. This appendix demonstrates that the core insight—comparing against random baselines under multiple operators—transfers when the unit of analysis changes from tokens to reasoning steps. We present three pieces of evidence: (i) operator sensitivity persists at step level, (ii) the approach enables black-box frontier model evaluation, and (iii) high accuracy does not imply faithful reasoning. A comprehensive evaluation across 20+ frontier models is a substantially different contribution (black-box only, different model set, deployment-focused) and is treated separately.

## M.1 Step-Level Operators

Given a CoT response $C = [s_1, s_2, \ldots, s_n]$ with final answer $a$, we define step-level analogues of ICE's operators:

- **Step-Delete**: Remove step $s_i$, concatenate remaining steps, query the model for a new answer.
- **Step-Retain**: Present only a subset of $k$ steps, query for an answer.

Aggregating over random subsets of size $\lfloor k \cdot n \rfloor$ with $P$ permutations yields step-level necessity and sufficiency win rates, paralleling ICE's token-level NSD decomposition. We combine these into step-level reasoning capacity $= \text{Nec}_{k=0.2} \times (1 - \text{Suf}_{k=0.2})$. ICE's statistical testing (Wilcoxon, bootstrap CIs, BH correction) applies identically.

## M.2 Operator Sensitivity at Step Level

A key finding of our token-level analysis is that faithfulness estimates are operator-dependent. We verify this persists at the step level using 3 open-weight models on GSM8K ($N = 50$):

| Model | Token-level | | Step-level | |
|---|---|---|---|---|
| | Nec | Suf | Nec | Suf |
| Qwen3-0.6B | 0.48 | 0.61 | 0.08 | 0.04 |
| Qwen3-8B | 0.77 | 0.50 | 0.85 | 0.12 |
| DeepSeek-R1-7B | 0.13 | 0.57 | 0.80 | 0.17 |

Table 19: Token-level vs. step-level faithfulness on GSM8K. DeepSeek-R1-7B appears unfaithful at the token level (Nec= $0.13$) but faithful at the step level (Nec= $0.80$), demonstrating that evaluation granularity—tokens vs. steps—is itself an operator choice.

## M.3 Black-Box Frontier Model Evaluation

Unlike token-level ICE, step-level evaluation operates through standard chat completions: remove a sentence from the prompt and re-query. This enables evaluation of closed-source frontier models. We demonstrate on 3 models ($N = 500$):

| | GSM8K | | SST-2 | |
|---|---|---|---|---|
| Model | SRC | Acc | SRC | Acc |
| Model A (RL-trained) | 0.86 | 93% | 0.87 | 97% |
| Model B (RL-trained) | 0.86 | 84% | 0.22 | 87% |
| Model C (instruction) | 0.01 | 95% | 0.00 | 97% |

Table 20: Step-level reasoning capacity (SRC $= \text{Nec}_{0.2} \times (1 - \text{Suf}_{0.2})$) for frontier models. High accuracy does not imply faithful reasoning: Model C matches Model A's accuracy but has SRC≈0 (decorative reasoning).

The results suggest a three-mode taxonomy: genuine (SRC $> 0.4$), scaffolding ($0.15$–$0.4$), and decorative ($< 0.15$) reasoning. The cost is $\sim$\$1–2 per model per task—enabling routine deployment auditing.

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