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

Abhinaba Basu[1,2] **and Pavan Chakraborty**[2]
[1]National Institute of Electronics and Information Technology, New Delhi, India
[2]Indian Institute of Information Technology Allahabad, Prayagraj, India
mail@abhinaba.com, pavan@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 show that faithfulness 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 that evaluates explanations against random baselines of equal size under multiple operators. Evaluating 7 LLMs across 4 tasks with deletion and retrieval infill operators, we find that switching operators crosses the positive-evidence threshold in 18% of configurations (5 of 28 attention comparisons), with gaps reaching 44 percentage points. Randomized baselines detect anti-faithfulness (explanations worse than random) in nearly one-third of English deletion configurations, invisible without random comparisons. These patterns persist 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 argues that faithfulness is operator-dependent, and that evaluation must account for this. While prior work has noted that individual evaluation choices matter (Hase et al., 2021; Kamp et al., 2023), no framework has systematically quantified operator dependence or provided a protocol for interpreting it.

We first show that different operators introduce different biases (§3): deletion creates OOD inputs while retrieval infill may preserve competing signals, inflating or deflating faithfulness estimates by up to 44 percentage points. To address this, we introduce ICE (Intervention-Consistent Explanation), which evaluates explanations against random baselines of equal size under each operator, yielding win rates with confidence intervals. When operators agree, the evidence for faithfulness is strong; when they disagree, the gap quantifies methodological uncertainty.

Evaluating 7 LLMs across 4 tasks, 2 attribution methods, and 6 non-English languages (§5), we find operator dependence is pervasive. Randomized baselines reveal that explanations perform *worse* than random in nearly one-third of English deletion configurations (18/56)—a finding invisible without random comparisons. The methodology also extends to step-level chain-of-thought evaluation (Appendix M), where preliminary results suggest broad applicability.

ICE is not simply "re-running ERASER with a second operator." It adds randomization testing against size-matched baselines (which no prior framework provides), an integrated protocol for interpreting operator agreement, and retrieval infill as an in-distribution alternative to deletion.

2 Related Work

Why operator choice matters. ERASER (DeYoung et al., 2020) established sufficiency and comprehensiveness as standard faithfulness metrics but

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 22349–22365
October 24-29, 2026 ©2026 Association for Computational Linguistics

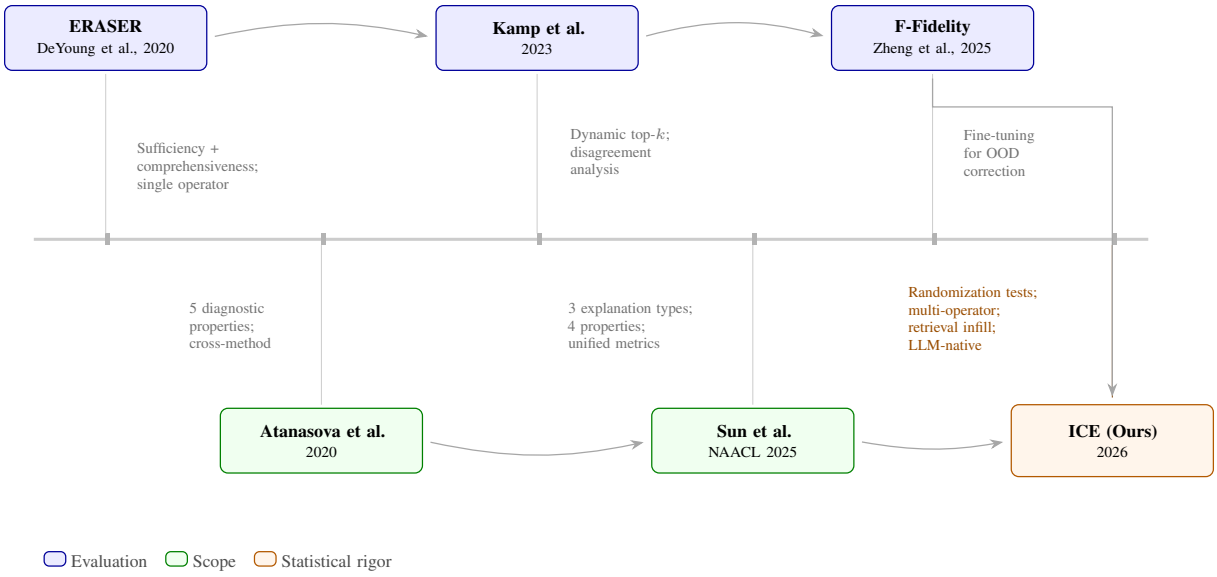


Figure 1: Evolution of faithfulness evaluation frameworks. ICE builds on evaluation methodology (blue) and scope expansion (green), adding statistical rigor (orange) via randomization testing and operator-consistent evaluation. Arrows show methodological lineage.

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. Atanasova et al. (2020) proposed a diagnostic framework evaluating explanations along faithfulness, rationale consistency, and agreement with human annotations. 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. Faithfulness is distinct from plausibility (Jacovi and Goldberg, 2020). We find no detectable correlation between ICE win rates and human rationale alignment ($|r|<0.04$), consistent with them measuring different constructs.

Extension to CoT. ICE's operator-aware methodology generalizes to step-level CoT evaluation (Turpin et al., 2023; Lanham et al., 2023); preliminary results appear in Appendix M.

3 The ICE Framework

Terminology. *Random baselines*: M=50 random token sets forming the null distribution. *Empty-input score* $s(\emptyset)$: normalization constant measuring prior label bias. *Operators*: intervention methods (deletion, retrieval infill). *Taxonomy*: anti-faithful WR<50%, near-random 50–55%, weakly faithful 55–60%, faithful >60%. BH-corrected p-values at α=0.10.

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 is whether the explanation method identifies 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 rationale tokens are retained and non-rationale tokens are intervened on:

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

where $s(x)$ is the prediction score on the original input, $s(x_o^r)$ is the score after retaining rationale r and applying operator o to the non-rationale tokens, 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=1 indicates retention equal to the original prediction signal; NSR=0 indicates the empty-input baseline. Values below 0 or above 1 are possible when the intervened score falls below the baseline or exceeds the original.

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. Table 1 shows how operator choice changes this verdict.

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** $d_{\text{null}} = \frac{\text{NSR}_{obs} - \mu_{\text{random}}}{\sigma_{\text{random}}}$
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.

	Delete	**Retr. Infill**
Input	"a gorgeous witty seductive movie"	
Rationale	{gorgeous, seductive}	
Result	"gorgeous seductive"	"gorgeous *a* seductive *an*"
$s(x_o^r)$	0.72	0.61
NSR	0.79	0.39
WR	92%	68%
Verdict	**Faithful**	**Weakly F.**

Table 1: ICE pipeline on a sentiment example. Non-rationale tokens are removed (delete) or replaced (infill). The same rationale gets different verdicts.

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 non-rationale tokens with tokens sampled from other examples, excluding the current example (leave-one-out) and filtering label-indicative tokens (label-blacklisted). Preserves input length while destroying task-relevant content. Replacement tokens are drawn from the same corpus, 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 4 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 2 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 generally favors attention, with Mistral as the main exception.

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 3 reveals cross-lingual variation spanning 12.0% to 82.7% across 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. The mechanism is that over-segmented tokenizers split semantically coherent units across multiple tokens, diluting per-token attribution scores and creating misalignment between token boundaries and linguistic units (Rust et al., 2021). However, model-language interactions, morphological structure, and operator sensitivity jointly determine whether attributions remain faithful—tokenization is one factor among several.

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 using NLLB-200 (NLLB Team et al., 2022) machine translation to translate English SST-2 into French and German (N=200):
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 gradi-



Figure 2: 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 = near-random, red = strongly anti-faithful (<40%); values below 50% are anti-faithful per the taxonomy in §3.

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 2: 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.

ent achieves $d_{\text{null}} = 2.50$ (2.5 standard deviations above the random baseline), while GPT-2 French gradient shows $d_{\text{null}} = -2.36$ (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.



Figure 3: Multilingual faithfulness (attention win rate %) under both operators across 6 languages and 4 scripts. **Left:** Deletion. **Right:** Retrieval Infill. The color scale highlights strongly anti-faithful values (<40%); values below 50% are anti-faithful under the taxonomy in §3. Gray = no valid output.

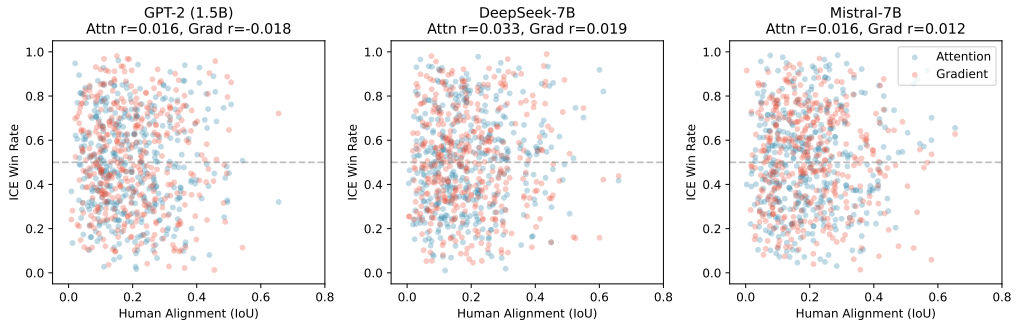


Figure 4: IoU (human alignment, x-axis) vs. ICE Win Rate (faithfulness, y-axis) across GPT-2, DeepSeek, and Mistral on e-SNLI. All $|r| < 0.04$: no correlation between human alignment and computational faithfulness.

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: $\text{IoU} = |R_{\text{attr}} \cap R_{\text{human}}|/|R_{\text{attr}} \cup R_{\text{human}}|$ at k=0.2. All models show near-zero Pearson correlation between IoU and ICE win rate ($|r| < 0.04$, $p > 0.5$; Figure 4). 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).

6.2 Retrieval Infill: Does Operator Choice Matter?

As shown in the 4-panel comparison (Figure 2), operator choice substantially changes win rates, with deletion typically exceeding retrieval on short text but the pattern reversing on long text. Figure 5 quantifies this effect on representative configurations.

The results confirm that operator choice fundamentally changes conclusions. Delete classifies Llama-3.2 e-SNLI as 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

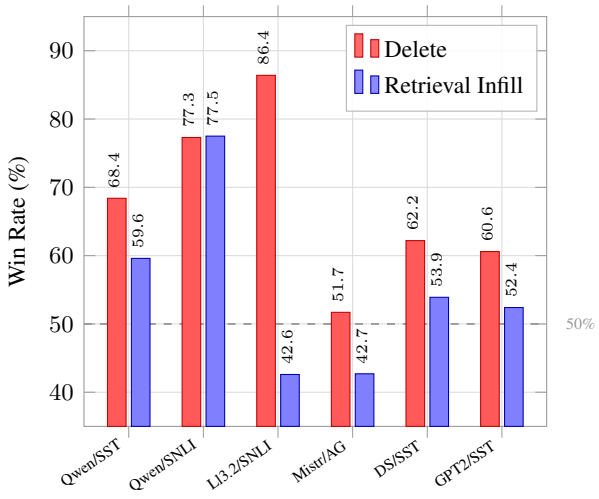


Figure 5: Operator comparison across model-dataset configurations (attention, main models). Delete (red) typically yields higher win rates on short text, while Retrieval Infill (blue) provides conservative estimates. The gap reaches 43.8 pp (Llama-3.2/e-SNLI), but operators *agree* on Qwen/e-SNLI (77.3% vs. 77.5%). Representative comparisons in Appendix Table 18; full attention results in Appendix Table 19.

stronger evidence of genuine faithfulness there.

Across all 28 attention configurations (7 models $\times$ 4 datasets), switching operators crosses the 55% positive-evidence threshold (from weakly-faithful-or-better to near-random-or-worse, or vice versa) in 5 cases (18%), with the most extreme gap reaching 44 pp. Deletion exceeds retrieval in 67% of short-text cases (median gap 1.7 pp, mean 4.6 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 attention results appear in Appendix L; representative paired comparisons 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 15)—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.

Table 3: ERASER-style comprehensiveness vs ICE. In 49/53 disagreements (92%), ERASER yields a positive verdict where ICE does not.

Model / Data	Met.	Raw	WR	Agree?
ERASER-positive, ICE-negative (shaded)				
Llama-70B / CSQA	C	.95	.19	✗
Llama-70B / SST	C	.86	.20	✗
Llama-70B / AG	C	.96	.36	✗
Qwen-32B / GSM	C	.97	.42	✗
Grok4 / ARC	S	.75	.12	✗
DSv3 / MedQA	S	.79	.21	✗
Both frameworks agree				
Gemini3 / GSM	C	.73	.56	✓
Grok4 / MedQA	C	.48	.51	✓
Kimi / MedQA	C	.48	.53	✓

Comparison with F-Fidelity. F-Fidelity (Zheng et al., 2025) addresses OOD artifacts by fine-tuning the model to adapt to perturbed inputs. ICE instead controls for OOD at the input level via retrieval infill, preserving the model unchanged. The approaches answer 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, evaluating the unmodified model is the relevant question. F-Fidelity also requires per-dataset fine-tuning (hours per configuration), while ICE is training-free (minutes), enabling the cross-lingual and multi-model comparisons central to this paper.

6.3 Head-to-Head Comparison with ERASER

To quantify ICE's practical advantage over existing frameworks, we evaluate both ERASER and ICE on 120 step-level CoT configurations (10 frontier API models $\times$ 6 reasoning tasks $\times$ 2 metrics; details in supplementary materials). We threshold ERASER's comprehensiveness at 0.5 and ICE's win rate at 55% for categorical comparison.

Table 3 shows representative configurations. Across all 120, the two frameworks disagree on comprehensiveness in 44% of cases, with ERASER yielding a positive verdict where ICE does not in 92% of disagreements. The core issue is that ERASER's raw scores are typically high (0.65–0.99) in disagreement cases because removing tokens always degrades output, but ICE reveals that random tokens produce similarly high scores—the apparent "faithfulness" is an artifact of input degradation, not explanation quality.

6.4 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 4: 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 token sequences), with ΔWR$=$ 3pp between contaminated and clean subsets—not statistically significant. We recommend adding topic classification words to the blacklist in future work.

6.5 Multilingual Gradient Results

Figure 3 reports only attention for the cross-lingual analysis. We additionally evaluate gradient attribution for all 7 models across 6 languages (complete results in supplementary material; language-specific patterns 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.

6.6 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 20.

7 Conclusion

Faithfulness evaluation depends on the intervention operator, and this dependence has practical consequences. ICE operationalizes this insight through randomized baselines under multiple operators. Across our evaluation, operator gaps reach 44 pp, while randomized baselines reveal that 18 of 56 English deletion configurations are anti-faithful. These patterns hold across 6 non-English languages and 2 attribution methods, and the methodology extends to step-level chain-of-thought evaluation (Appendix M).

For practitioners, the recommendation is straightforward: report faithfulness under at least two operators, compare against random baselines, and treat disagreement between operators as genuine uncertainty about the explanation's quality.

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. We omit Integrated Gradients (Sundararajan et al., 2017) 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 4 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. 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 to step-level CoT evaluation, where the unit of analysis changes from token spans to reasoning sentences. Appendix M presents preliminary evidence that operator sensitivity persists at this level and that high accuracy does not imply faithful reasoning. We scope this paper to the token level, where model internals (attention, gradients) are available for attribution; a comprehensive step-level evaluation across frontier models is treated separately.

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

Code: `https://github.com/abhinaba/ice-faithfulness`
Package: `pip install ice-faithfulness` (`https://pypi.org/project/ice-faithfulness/`)
Data: All result JSON files across 12 experiment categories, each containing per-example scores, win rates, and bootstrap CIs (supplementary archive).
Scripts: `run_ice_llm.py` (English), `run_ice_multilingual.py` (6 languages), `run_ice_llm_retrieval.py` (operator comparison).
Runtime: ∼15 min per model/dataset on RTX 4090 (N=500, M=50). Total: ∼200 GPU-hours. Pre-computed results included.

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, ChatGPT) for proofreading, editing, and verification of numerical consistency. All scientific contributions, experimental design, and writing are the authors' original work.

References

Alexander Amini, Anna Banaszak, Harold Benoit, and others. 2025. LFM2 technical report. *arXiv preprint arXiv:2511.23404*.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3256–3274.

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 9539–9549.

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)*, pages 9:1–9:8. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

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, pages 3650–3666.

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.

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

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.

Meta. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. Meta AI Blog, September 2024.

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.

NLLB Team, Marta R. Costa-jussà, and others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

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

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? On the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the ACL*, pages 3118–3135.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. 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.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, PMLR 70, pages 3319–3328.

Songbo Tan and Jin Zhang. 2008. An empirical study of sentiment analysis for chinese documents. *Expert Systems with Applications*, 34(4):2622–2629.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. 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, pages 74952–74965.

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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, and others. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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 29425–29449. 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.

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*.

— Appendix —

A Operator Ablation

Table 5 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 5: 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 6 lists the evaluated models. Table 7 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 (Amini et al., 2025)
Llama 3.2	3B	Inst	Small instruction-tuned (Meta, 2024)
Llama 3.1	8B	Base	General reasoning (Grattafiori et al., 2024)
Qwen 2.5	7B	Inst	Multilingual focus (Yang et al., 2024)
Mistral	7B	Inst	Efficient instruction (Jiang et al., 2023)
DeepSeek	7B	Chat	Multilingual/Chat (DeepSeek-AI et al., 2024)

Table 6: 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 7: Pinned dataset revisions for ICEBench.

C Statistical Framework Details

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

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

We apply Benjamini-Hochberg FDR correction at $\alpha = 0.10$ when evaluating multiple examples. We use magnitude thresholds 0.2 (small), 0.5 (medium), 0.8 (large) for descriptive interpretation. Negative d_{null} 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 8: 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 9: Prompt sensitivity analysis (GPT-2). Win rates (%) for attention (Attn) and gradient (Grad). Δ = Attn − Grad. †Anti-faithful (<50%).

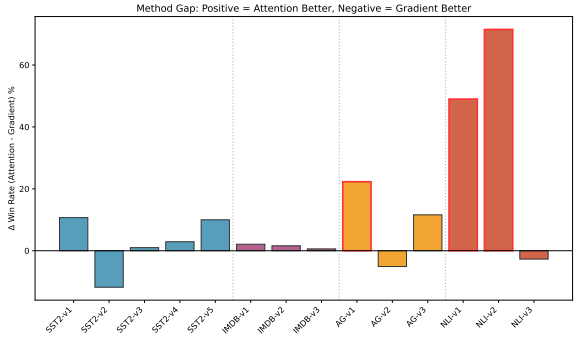


Figure 6: Method gap (Δ = Attention − Gradient win rate) by prompt variant.

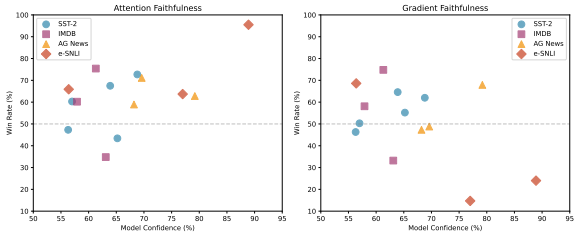


Figure 7: Faithfulness vs. model confidence across prompt variants.

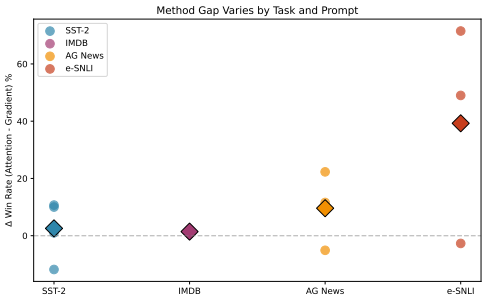


Figure 8: Method gap distribution by task.

F Language-Specific Patterns

German: High variance across models (34–83%). Qwen is the clear outlier at 83% attention, the highest multilingual result. We speculate that compound-word handling in Qwen's tokenizer may play a role, though we cannot isolate this factor.

Chinese: Moderate results (50–69% deletion). LFM2 gradient leads (69%), but drops under retrieval infill, suggesting operator sensitivity extends to character-level tokenization.

French: The most polarized language. GPT-2 is anti-faithful (15–16%) while Llama 3.1 excels (81% attention).

Hindi: GPT-2 performs surprisingly well (65–66%) despite 8.1× token expansion. Under retrieval infill, GPT-2 reaches 68.8% (exceeding deletion), while Llama 3.1 drops to 35.1%.

Turkish: Qwen leads attention (79.3%) while Mistral leads gradient (71.5%), the largest method divergence for any language. GPT-2 is anti-faithful (29.7%).

Arabic: Mistral attention achieves 72.6% but its gradient drops to 39.9%, the largest attention-gradient gap across languages.

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 10: 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 11: Token expansion ratios by language (× = tokens per character vs. English).

H Encoder Validation Results

Table 12 reports ICE evaluation on BERT-base-uncased (Devlin et al., 2019) across five ERASER datasets, comparing LIME (Ribeiro et al., 2016), Integrated Gradients (Sundararajan et al., 2017), attention, and gradient extractors.

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

I Effect Sizes and Anti-Faithfulness Details

Tables 13–16 provide effect sizes, anti-faithful configurations, concrete anti-faithful examples, and bootstrap confidence intervals.

Configuration	WR	d_n	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 13: Standardized effect sizes (d_n). $d_n > 0.8$ = large. Anti = anti-faithful.

Model	Config	WR	d_n
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 14: Anti-faithful configurations (WR $<$ 50%). Negative d_n = worse than random.

Text	WR	d_n	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 15: 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 16: Bootstrap 95% CIs for uncertainty quantification. Formal significance is assessed via BH-corrected randomization-test p-values.

J Faithfulness-Plausibility Correlation

Table 17 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 17: IoU-Faithfulness correlation across three models. No model shows significant correlation ($|r| < 0.04$, all $p > 0.5$).

K Retrieval Infill Detailed Results

Table 18 and Figure 9 provide representative operator comparisons.

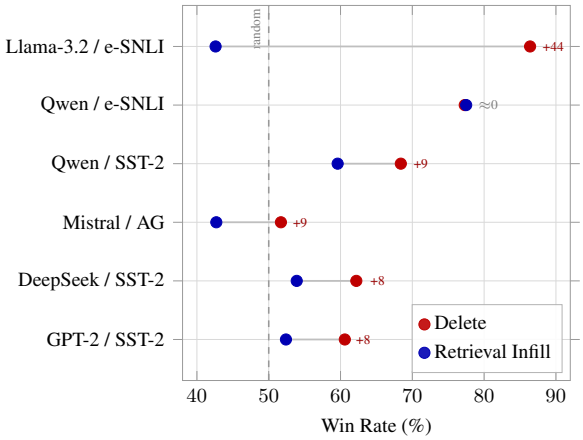


Figure 9: Paired comparison of attention win rates under Delete vs. Retrieval Infill (main models). Delete (red) exceeds Retrieval Infill on most short-text configurations by 8–44 pp, but operators *agree* on Qwen/e-SNLI (≈77%), suggesting genuine faithfulness there.

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

Table 18: Attention win rates for Delete vs. Retrieval Infill ($k = 0.2$). F=Faithful (>60%), WF=Weakly faithful (55–60%), NR=Near-random (50–55%), AF=Anti-faithful (<50%).

L Relocated English Results Table

Table 19 provides the full English attention win rates underlying the top panels of Figure 2. Table 20 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 19: 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 Attention (Gradient for Mistral)
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 (79%) or Mistral Grad (72%)
Arabic	Mistral Attn (73%) – avoid Llama-3.2 (37%)

Table 20: Practical attribution guidelines based on ICE evaluation.

M Preliminary: Extension to Chain-of-Thought

This appendix presents **preliminary** evidence that ICE's operator-aware methodology generalizes from token-level attributions to step-level chain-of-thought (CoT) evaluation. The core insight—comparing against random baselines under multiple operators—transfers when the unit of analysis changes from tokens to reasoning steps. A comprehensive evaluation across frontier models is the subject of separate work.

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 NSR 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 (randomization tests, 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 21: 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 22: 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$\approx$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.