Ask Twice: Bidirectional Cross-Encoder Scoring for Faithfulness Evaluation

Arnav Mantro[1], **Leslie Barrett**[1]
[1]Bloomberg
Correspondence: lbarrett4@bloomberg.net

Abstract

Current LLM-based question answering systems provide cited evidence for their claims. However, determining whether each claim within the generated responses is fully supported by its cited evidence is still an ongoing research problem. The task of evaluating attribution quality, or faithfulness to the source text, goes beyond the traditional idea of lexical matching used for relevance in early QA tasks, and the use of lexical models based on semantic distance or word overlap has declined in favor of NLI-based models and large fine-tuned models. The larger fine-tuned models tend to perform best on current special-purpose benchmarks such as AttributionBench despite struggling with OOD data. Still, exactly why human judgments of faithfulness should align better with LLM evaluators than those based on lexical entailment has never been fully explained. We address the problem with current entailment models by introducing bidirectionality and codifying that idea into a new evaluation modeling approach. We show that this approach is both lightweight and extensible yet competitive with SOTA LLM-based methods, avoiding the pitfalls of both standard NLI and word-overlap approaches. We further show that our model architecture is motivated by a deeper understanding of the gap between lexical distance and faithful claim attribution, exposing the conceptual mechanism behind attribution quality evaluation.

1 Introduction

Current QA systems leverage the power of LLMs to generate claims based on cited source data. The resulting information allows users to ask increasingly complex questions with grounded answers. However, such systems are also prone to hallucinations and the addition of irrelevant material (Ji et al., 2023; Zhang et al., 2025). Evaluating the fidelity of the claim with respect to the source text has become a recent area of research and remains an open problem.

Early methods for evaluating information retrieval and question answering relied on the concepts of textual distance and relevance to judge quality (Salton et al., 1975). As QA evolved, lexical distance, word overlap and coherence measures were used. For example, ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) use n-gram and word-sequence matching, and BLEU (Papineni et al., 2002) uses n-gram comparison with a brevity penalty. Chen et al. (2019) provide an overview of these metrics, showing their shortcomings for QA as datasets evolve further, becoming more abstractive in nature.

Recent variations of these models such as BLEURT (Sellam et al., 2020) and BERTScore (Zhang et al., 2020) were developed to use semantic distance in place of lexical comparisons, and have been used for evaluating the faithfulness quality of QA systems.

In contrast to these models, there are recent entailment-based models such as SummaC (Laban et al., 2022), AutoAIS (Gao et al., 2023) and AlignScore (Zha et al., 2023), trained on NLI corpora, that treat faithfulness as an entailment function between the claim and the source text.

None of these models match the evaluation quality of LLMs on current benchmarks such as AttributionBench (Li et al., 2024), where the best performing models are Flan-T5 (Chung et al., 2022) and GPT-3.5 (OpenAI, 2023). However, these models, despite being fine-tuned to achieve the best performance on in-domain data, do not perform as well on out-of-distribution (OOD) data and thus are inflexible to new domains.

The transition from traditional QA evaluation to evaluation created to accommodate the idea of faithful attribution has received very little attention in the literature. Why, for example, are even the most sophisticated textual-distance-based met-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 528–536
November 6-10, 2026 ©2026 Association for Computational Linguistics

rics not up to the task of evaluating faithfulness when they excelled at doing so for relevance? Relevance is an integral part of question answering quality, even with LLM-generated answers. Hu et al. (2025) refer to relevance in their study on complex attribution, on the grounds that the source must contain relevant facts to support the answer. Similarly, He et al. (2025), in their study on claim verification, show that models in a dynamic setting must retrieve relevant information to assess a claim's validity.

We investigate this question as we propose a new lightweight and extensible approach to modeling faithfulness evaluation, based on a combination of Relevance and Coverage. We propose that faithful attribution must consist of both elements in a particular relation.

- Relevance: Does the reference address the topic of the claim?
- Coverage: Does the claim stay within what the reference supports?

As we show in the following sections, our cross-encoder-based approach to modeling these dimensions is competitive with current models and consistent with the intuition that faithful attribution must meet relevance and coverage criteria.

Our contributions are threefold:

1. We show that our bidirectional cross-encoder scoring captures a faithfulness dimension that unidirectional scoring systematically misses.
2. We reveal that NLI-based faithfulness evaluation, the field's dominant paradigm, captures only one of two necessary signals needed for robust attribution evaluation, and demonstrate that applying our method improves NLI-based models.
3. We show that bidirectional scoring composes with fine-tuning, showing competitive results on three AttributionBench out-of-distribution datasets.

2 Background and Prior Work

The dominant approaches for faithfulness fall into three categories, each with fundamental limitations. Text overlap metrics such as ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) measure surface-level token overlap between claim and reference. These were designed for tasks where the output closely matches a reference text, based on n-gram comparisons. Because ROUGE in particular measures the recall of n-gram overlap between texts, a high score can indicate that the model retained more of the key points from the source. However, both metrics are overlap-based and poor measures of faithful attribution, as a claim can be fully supported without sharing specific words, and these approaches do not account for hallucinated facts.

Metrics based on embedding similarity such as BERTScore (Zhang et al., 2020), BARTScore (Yuan et al., 2021) and BLEURT (Sellam et al., 2020) capture semantic relatedness better than n-gram-based approaches but do not measure directional support and thus incorrectly conflate the concepts of "a similar idea" and "a supported claim".

Entailment-based methods have become the more common approach most recently. SummaC (Laban et al., 2022) and AlignScore (Zha et al., 2023) are both RoBERTa-based models. SummaC is fine-tuned on NLI datasets splitting source and generated texts into sentences and computing entailment scores for each pair; AlignScore is fine-tuned with a unified alignment loss function. AutoAIS (Gao et al., 2023) was created primarily for QA systems. Based on the NLI classifier TRUE of Honovich et al. (2022), it indicates whether a question and answer are entailed by the provided attribution, and common implementations use T5 (Raffel et al., 2023) as a base model. The original AIS approach (Attributable to Identified Sources) was proposed in (Rashkin et al., 2022), which considered a binary notion of attribution such that a text passage y is attributable to a set of evidence A just in cases where the statement "according to A, y" is valid.

LLM-based methods show strong results in certain studies on certain datasets. For AttributionBench OOD datasets in zero-shot mode, leaders are Flan-UL2 (Tay et al., 2023) and GPT-4 CoT. For fine-tuned mode, the leaders are T5-XXL-TRUE and AttrScore-Flan-T5. In the AttributionBench study (Li et al., 2024), older embedding and similarity-based methods were not compared. We do so in the next section and find that indeed larger, newer models do better, but as pointed out in Li et al. (2024), size is not directly linked with performance.

3 Data and Methods

3.1 Methodology

Our proposed new method is based on our observation that cross-encoder models, which jointly attend over both claim and reference inputs with order-sensitive positional encoding, produce different scores when inputs are swapped. We show that these two scores - forward and reverse - capture the two complementary dimensions of faithfulness.

Combining both scores with a single learned weight - a zero-parameter technique we call bidirectional scoring - produces consistent improvements across virtually every cross-encoder family we test. As we will show, the gains are modest for models already strong at relevance but transformative for NLI models, which we show capture coverage but systematically miss relevance when used in a single direction.

3.2 Task and Data

We use AttributionBench (Li et al., 2024), which provides human-annotated (claim, reference) pairs with binary attribution labels. It was designed for training and evaluating automatic attribution and faithfulness evaluators. The data is sampled from seven different datasets covering different domains of questions, claims and information sources. Of the seven datasets, three have out-of-distribution (OOD) test sets and the rest have full train, development and test divisions. For all datasets, tasks are evaluated as binary classification with the labels "Attributable" and "Not Attributable". We chose to test our approach on the OOD test datasets to demonstrate our model's extensibility to new domains and highlight that the model design, rather than size or fine-tuning, provides a truly new insight that is domain-agnostic.

We evaluate models on the out-of-distribution dataset consisting of the following combined datasets: BEGIN (Dziri et al., 2021) (dialogue grounding), AttrScore-GenSearch (Yue et al., 2023) (web search), and HAGRID (Kamalloo et al., 2023) (knowledge-grounded QA). We report macro-F1 averaged across the OOD source data, with thresholds tuned on the combined development set from the ID (in-domain) data.

3.3 Base Model

We use ms-marco-MiniLM-L6-v2 (22.7M parameters), a cross-encoder trained for passage re-ranking on MS-MARCO (Bajaj et al., 2018). We selected MS-MARCO(-MiniLM-L6-v2) as the base because passage re-ranking most closely mirrors the attribution task structure (i.e., scoring a (claim, reference) pair for support). Furthermore, MS-MARCO cross-encoders are available across a range of sizes, allowing us to prioritize parameter efficiency and speed during experimentation, especially during fine-tuning. Inference splits each reference into sentences, scores each (claim, sentence) pair, and takes the maximum, a max-over-sentences strategy that localizes the most relevant evidence. We fine-tune on AttributionBench training data using sentence-level pairs, discriminative learning rates (backbone: 2e-6, head: 5e-4), and the preserved MS-MARCO classification head for OOD generalization. We evaluate traditional metrics on the OOD test set to establish the difficulty of the task. We include a basic TF-IDF representation with cosine distance. We use three variants of ROUGE: unigram (ROUGE-1), bigram (ROUGE-2) and longest-common-subsequence (ROUGE-L). We use BERTScore F1, a balanced measure of the similarity between the candidate and reference sentences, and we use BLEU unigram implementation. We also benchmark two NLI implementations from Hugging Face using DeBERTa (Hussain, 2024).

Metric	Type	Params	OOD-F1
ms-marco-MiniLM-L6-v2 (our model - forward + FTed)	Relevance	22.7M	81.9
TF-IDF Cosine	Text Overlap	0	76
ROUGE-2	Text Overlap	0	73.9
BERTScore F1	Embedding	355M	68.2
ROUGE-L	Text Overlap	0	66.5
ROUGE-1	Text Overlap	0	64.9
NLI DeBERTa x-small	NLI (zero shot)	22M	61.1
NLI DeBERTa-large	NLI (zero shot)	304M	59.6
BLEU	Text Overlap	0	55

Table 1: Baseline Models on AttributionBench Data

Reviewing the data in Table 1, text overlap metrics such as BLEU and ROUGE (55-74%) fail because shared words do not guarantee support, and support does not require shared words. BERTScore (68.2%) captures relatedness but not directional support. Zero-shot NLI (DeBERTa x-small and large) performs worst among learned methods (59-62%); the larger model (304M) actually underperforms the smaller (22M), indicating distribution mismatch. Basic TF-IDF Cosine is the best among the traditional metrics.

Our forward-only cross-encoder (81.9%) outperforms all traditional metrics. Yet, error analysis reveals it still falls into a systematic trap that motivates our main contribution.

3.3.1 The Relevance Trap

Error analysis of our model's false positives reveals that the majority are claims topically relevant to the reference but not faithfully supported, falling into the following error classes:

1. Over-generalizations (53%): "horses and other animals" when the reference mentions only "horses".

2. Factual mismatches (29%): "a denturist" vs "prosthodontist."

3. Hallucinated details (12%): "more than 14 million km" when the reference says only "more than...". Despite sharp semantic and contextual differences, these all share the same property, showing high relevance but low fidelity. A single directional score conflates these.

We create a new set of definitions to capture these insights.

Definition:

Let $f_\theta(A, B) \in \mathbb{R}$ denote the score from a cross-encoder for ordered pair (A, B). For a claim c and reference r:

$$s_{\text{fwd}} = f_\theta(c, r) \tag{1}$$

$$s_{\text{rev}} = f_\theta(r, c) \tag{2}$$

where the first term is a relevance function denoting whether the reference addresses the claim, and the second term is a coverage term denoting whether the claim stays within the reference.

For faithful claims, both are high. For relevance-trap errors, forward is high but reverse drops. The bidirectional score:

$$s_{\text{bidir}} = \alpha \cdot s_{\text{rev}} + (1 - \alpha) \cdot s_{\text{fwd}}, \quad \alpha \in [0, 1] \tag{3}$$

expresses a weighted combination of coverage and relevance such that claims with high relevance and low coverage - the relevance-trap errors above - receive lower scores than claims where both are high. α is tuned on a held-out development set, independent of the out-of-distribution evaluation data, and takes a value between 0 and 1.

3.3.2 Theoretical Grounding

Why should combining forward and reverse scores capture faithfulness better than either alone?

Our method is based on the observation that cross-encoder models employ order-sensitive positional encoding while attending over claim and reference. Thus they will produce different scores when inputs are swapped. We show that these two scores - forward and reverse - capture complementary dimensions of faithfulness, including both relevance and coverage.

We motivate bidirectional scoring through an analogy to pointwise mutual information (PMI) (Church and Hanks, 1989). PMI measures the strength of association between two events by comparing their joint probability to what would be expected under independence:

$$\text{PMI}(C; R) = \log \frac{p(C, R)}{p(C) \cdot p(R)} \tag{4}$$

The key structural property of PMI is that it is high *only* when both events predict each other: the joint probability $p(C, R)$ must substantially exceed the product of marginals $p(C) \cdot p(R)$. Equivalently, PMI can be decomposed into either directional condition (via the chain rule), each paired with a marginal correction. For PMI to be large, neither direction can be weak - if C is unlikely given R, or R is unlikely given C, the ratio in Equation 4 stays low.

Faithfulness evaluation has the same conjunction structure. A claim C is faithfully attributed to a reference R only when two conditions hold simultaneously:

- **Relevance Condition:** The reference addresses the topic of the claim - without this, the reference cannot support the claim at all.
- **Coverage Condition:** The claim stays within the scope of what the reference states - without this, the claim may add unsupported detail.

Neither condition alone is sufficient. A claim about "horses and other animals" is relevant to a reference discussing horses (high relevance) but exceeds its scope (low coverage). Conversely, a generic claim like "the sky is blue" has trivially high coverage against any weather reference but fails relevance unless the reference explicitly discusses sky color. Just as PMI requires both directional terms to be strong, faithfulness requires both relevance and coverage to be high.

3.3.3 From Analogy to Implementation

Cross-encoders compute $f_\theta(A, B) \in \mathbb{R}$, a learned scalar reflecting how well the ordered pair (A, B) satisfies the model's training objective. The analogy to PMI is structural, not mathematical: it identifies the conjunction requirement - both directions

must score highly - that bidirectional scoring implements.

Concretely, for a claim c and a reference r, a cross-encoder produces two scores: $s_{\text{fwd}} = f_\theta(c, r)$ and $s_{\text{rev}} = f_\theta(r, c)$. These differ because the model's positional encoding and segment markers create order-dependent representations. As shown in the error analysis above, the forward score (s_{fwd}) is high whenever the reference is topically related to the claim, but remains high even for overgeneralizations and hallucinated details where it captures relevance but not coverage. The reverse score (s_{rev}) drops for exactly these cases: when the claim introduces content absent from the reference, scoring the reference against the claim exposes the mismatch. The bidirectional blend (Equation 3) recovers the conjunction: both scores must be high for the combined score to be high, just as both directional terms must be strong for PMI to be large.

This structural parallel generates three testable predictions that we evaluate empirically:

1. The blending weight α should correlate with the relative informativeness of the reverse signal - models where the reverse direction contributes more discriminative information should assign it higher weight.

2. α should cluster by training objective, since different objectives induce different biases toward relevance vs. coverage.

3. The blend should outperform the maximum of the two individual scores, confirming the signals are complementary rather than redundant.

Cross-encoders compute classification probabilities, not conditional text probabilities. We do not claim equivalence. But the structure is parallel: just as PMI decomposes association into two directional terms, bidirectional scoring decomposes faithfulness into two directional evaluations. A single direction captures at most one term. Thus we conduct our experiments to test the hypothesis that blended scores will be higher than those of the individual signals.

4 Experimental Setup and Results

4.1 Experimental Models

We conducted our first experiment to learn the effect of applying our bidirectional approach in a variety of training setups with different model architectures. We tested different models associated with six different training paradigms: NLI, STS-B, QNLI, MNLI, FactCheck and MS-MARCO from the Hugging Face (Wolf et al., 2020) library of pretrained models[1]. For NLI, we included all seven cross-encoder checkpoints as well as the RoBERTa-large MNLI model (FacebookAI/roberta-large-mnli), which shares the training paradigm underlying the entailment-based faithfulness evaluators in prior work (SummaC, AlignScore, TRUE). For STS-B we tested the DistilRoBERTa, TinyBERT-L4, and RoBERTa-base cross-encoders. For QNLI, both available models were tested, and for fact-checking, MiniCheck-DeBERTa-v3-Large.

Our experimental design includes 19 cross-encoder models, 7 architecture families and 6 training objectives. Models are all run in zero-shot mode using 4.4M to 355M parameters. Results for F1 score are given for forward, reverse and bidirectional (c, r) pairings. We also show the best α weighting for each model-training objective pairing and the difference (delta) between forward and bidirectional. We add the p-value to indicate the significance value of delta.

In our second experiment we simply compared our bidirectional encoder results with the AttributionBench leaderboard results on OOD data, where each model is fine-tuned on ID data. We used our best forward and bidirectional α weightings from the initial experiment.

4.2 Sufficiency of the Entailment Signal

Most NLI-based faithfulness studies testing factuality (Kryściński et al., 2019) or summarization (Laban et al., 2022) use only the entailment class probability from three-class models, discarding the contradiction and neutral outputs. We tested whether these additional signals provide information for faithfulness beyond what entailment captures. Across the 8 NLI/MNLI models (see Table 2), we extracted all three softmax probabilities in both directions (6 signals per model) and evaluated seven combination strategies on the development set: entailment-only forward, entailment bidirectional, net support (entailment minus contradiction in the reverse direction), net-support bidirectional, and three-way blends incorporating contradiction in either direction. No multi-signal combination improved over entailment-only bidirectional scor-

[1] https://shorturl.at/kL3au

ing on the OOD evaluation. This result is expected: softmax normalization constrains the three class probabilities to sum to one, so contradiction and neutral are linearly determined by entailment and carry no additional information. The discriminative gain from bidirectional scoring comes entirely from the direction swap – evaluating both (c, r) and (r, c) – not from richer use of model outputs within a single direction.

4.3 Results

Table 2 shows the results of our initial experiment. Out of the 19 models tested, all the models that showed significant changes improved in bidirectional mode over either single direction. The 8 NLI/MNLI models improve substantially.

Training	Model	Params	Fwd	Rev	Bidir	Delta	p	α
MS-MARCO+FT	MiniLM-L6 (ours)	22.7M	**81.9**	74.1	**82**	0.1	<0.001	0.28
MS-MARCO	ELECTRA	109M	72.4	71.2	74.1	1.7	<0.001	0.34
MS-MARCO	MiniLM-L12	33.4M	79.2	71	79.2	0	ns	0.01
MS-MARCO	TinyBERT-L2	4.4M	79.1	75.3	77.8	-1.3	ns	0.06
MS-MARCO	MiniLM-L6	22.7M	79.9	74.1	78	-1.8	ns	0.26
MNLI	RoBERTa-L	355M	60.8	73.1	78.4	17.6	<0.001	0.6
NLI	DeBERTa-v3-XS	22M	60.6	70.6	78	17.4	<0.001	0.79
NLI	RoBERTa	125M	60.4	71.1	77.6	17.2	<0.001	0.63
NLI	MiniLM2	82M	59.6	71.2	76.4	16.8	<0.001	0.52
NLI	DistilRoBERTa	82M	59.1	72.3	75.8	16.8	<0.001	0.88
NLI	DeBERTa-v3-B	86M	59.7	67.8	75.5	15.8	<0.001	0.62
NLI	DeBERTa-v3-S	44M	61.4	71.5	75.9	14.5	<0.001	0.82
NLI	DeBERTa-v3-L	304M	59.8	66.5	73.1	13.3	<0.001	0.88
STS-B	DistilRoBERTa	82M	77.3	81.2	79.8	2.5	<0.001	0.9
STS-B	TinyBERT-L4	14.5M	78.3	77.7	78.6	0.3	ns	0.51
STS-B	RoBERTa	125M	80.7	81.5	80.9	0.2	ns	0.17
QNLI	DistilRoBERTa	82M	69	62.3	69.7	0.7	<0.001	0.16
QNLI	ELECTRA	109M	64.9	60	64.3	-0.5	ns	0.07
FactCheck	MiniCheck	304M	65.5	73.1	74.7	9.2	<0.001	0.91

Table 2: Cross-Encoder Results

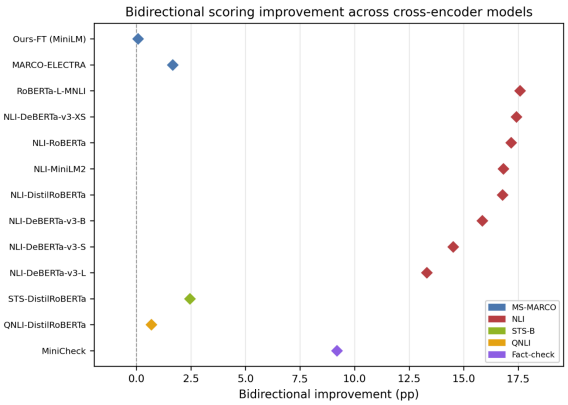


Figure 1: Bidirectional Improvement

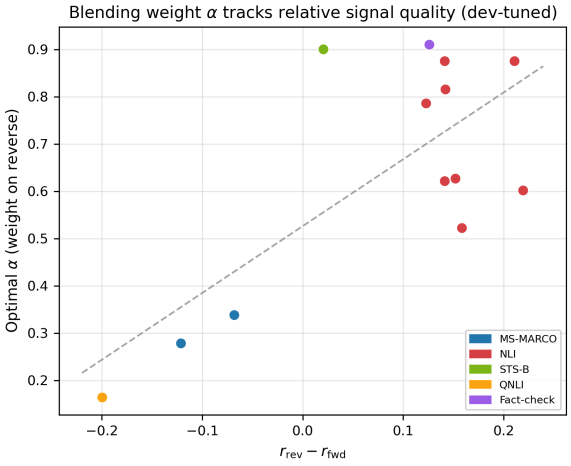


Figure 2: Blended Weight

The performance deltas and their magnitude in percentage improvement are illustrated in Figure 1. We show only results here for the combinations that are significant.

Table 3 shows the Pearson correlation and the effect size with different weightings for the α parameter. The equal-weight sum ($\alpha = 0.5$) achieves r = 0.623 with labels, which is lower than forward only (r = 0.636) because equal weighting dilutes the stronger signal. The optimal blend ($\alpha = 0.28$) achieves r = 0.640. Cohen's d shows a large effect size and AUC mirrors the Pearson correlation with the optimal blend performing best.

Score	r	Cohen's d	AUC
Forward	0.636	1.65	0.873
Reverse	0.515	1.2	0.802
Sum ($\alpha = 0.5$)	0.623	1.59	0.859
Optimal Blend	0.640	1.75	0.883

Table 3: Signal Contribution

The underperformance of the sum is predicted given the premise that the two signals contribute unequally. Faithfulness is asymmetric: relevance is a necessary condition, coverage a discriminative refinement. We note that α should correlate with which direction is more informative, and α should cluster by training objective. The blend should outperform $\max(s_{\text{fwd}}, s_{\text{rev}})$, and it does so as expected.

Table 4 shows the α clusters by training objective. We note that MS-MARCO and NLI ranges are non-overlapping (0.34 < 0.52). This would not occur with noise smoothing.

Looking at the graph in Figure 2 we see the NLI and FactCheck models tend to cluster their optimal α in the direction of greatest reverse-forward difference. This supports the intuition that optimizing the balance will provide the most gain in cases with the greatest gap between forward and reverse performance.

NLI cross-encoders take (premise, hypothesis). Reverse $f_\theta(r, c)$ = reference-as-premise, claim-as-hypothesis. This is the standard setup

Training	α range	Interpretation
MS-MARCO	0.01 - 0.34	Relevance Dominates
NLI/MNLI	0.52 - 0.88	Coverage Dominates
STS-B	0.17 - 0.90	Variable
QNLI	0.07 - 0.16	Relevance Dominates
FactCheck	0.91	Coverage Dominates

Table 4: Model Alpha Ranges

in NLI-for-attribution models (e.g. SummaC, AlignScore, TRUE). Forward $f_\theta(c, r)$ = claim-as-premise, reference-as-hypothesis is non-standard.

The NLI numbers (59.1-61.4%) shown in Table 5 are the non-standard direction. The standard direction achieves 66.5-73.1%. Neither is sufficient, and both fall below bidirectional results. Figure 3 below brings out this effect.

Direction	F1 Range
Forward (Non-Standard)	59.1 - 61.4%
Reverse (Standard NLI)	66.5 - 73.1%
Bidirectional	73.1 - 78.4%

Table 5: Ranges for All Models

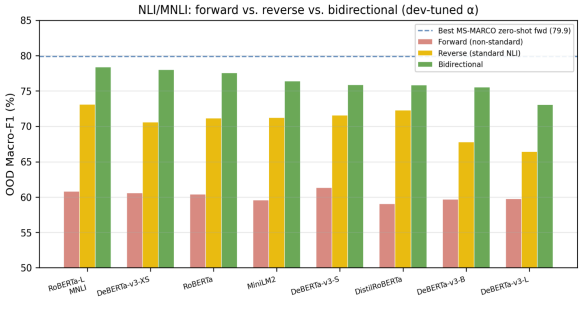


Figure 3: Forward, Reverse and Bidirectional Results

Finally, we show results from our bidirectional cross-encoder approach against all published fine-tuned models on AttributionBench OOD data.

As Table 6 shows, our cross-encoder using the bidirectional approach is competitive with much larger models. In fact, size does not correlate particularly well with performance on the task of recognizing faithful claims.

4.4 Discussion

Results show that our bidirectional approach has a strong positive effect on the models tested in our initial experiment.

To evaluate whether α clustered by training objective, we performed a grid search over $\alpha \in [0, 1]$ in increments of 0.01 for each model on the combined in-domain development set, jointly tuning the classification threshold to maximize macro-F1. We then grouped the resulting optimal α values by training objective, and fixed both α and the threshold for OOD evaluation, with the non-overlapping MS-MARCO and NLI/MNLI ranges. This brought out the systemic differences in the balance of relevance and coverage favored by their training objectives.

Model	F1
AttrScore-Flan-T5 (3B)	84.2
T5-XXL-TRUE (11B)	83.4
ms-marco-MiniLM-L6-v2 (bi)	**82**
ms-marco-MiniLM-L6-v2 (for) (22.7M)	81.9
GPT-3.5 (w/o CoT)	81.9
Flan-T5 (11B)	81.1
AttrScore-Alpaca (7B)	81.1
Flan-UL2 (20B)	81
Flan-T5 (3B)	80.4
Llama-2 (7B)	80.1
Flan-T5 (770M)	78.8
RoBERTa-large-mnli (330M)	62.4

Table 6: AttributionBench All Models

The three-way comparison shown in Table 5 focuses on the F1 scores of the eight NLI models tested, as they represent our largest training objective group that achieved significant scores. We see that bidirectional F1 range is highest.

Also, as shown in the results of Figure 3, we find that Bidirectional exceeds the better single direction by +3.5 to +7.8pp in every case.

An important detail regarding the results in Table 6 is that the T5-XXL-TRUE model is fine-tuned on the TRUE benchmark data, not on AttributionBench data. This is important as these results are on the OOD Attribution test data, which has overlapping datasets with the TRUE benchmark (BEGIN). Thus, it is hard to tell how much this model's performance is due to data leakage.

Finally, despite being third overall, our model actually does achieve SOTA on HAGRID, the most challenging OOD dataset for all models (all scores are lower), where claims often diverge from references in content rather than just phrasing. Here our bidirectional model (77.9%) outperforms the 3B SOTA (76.6%) by 1.3pp. This is exactly the type of error bidirectional scoring is designed to catch: claims that paraphrase and extend knowledge-

grounded answers, where coverage checking is most valuable.

5 Conclusion

We have provided a new insight into faithfulness detection that recognizes its two components: relevance and coverage. Using this insight, we have developed an approach using a cross-encoder base model to create a novel bidirectional scoring system that can be broadly applied. We have applied our system to the claim and source texts in AttributionBench, and our findings indicate that it is able to substantially improve the performance of current NLI models and yield competitive results with LLMs. While we feel that this approach may enhance similar models on other similar tasks such as fact grounding, we focus on faithfulness for this study. We also leave for future work addressing questions of how this approach might be adapted to improve the performance of large language models. As we suggest in the beginning, LLMs perform well on faithfulness evaluation and somehow ingest the balance of components that we have exposed. Still, their OOD performance may be improved by adding each component as an explicit instruction, and we hope that this insight leads to improvements in attribution and faithfulness evaluation metrics generally.

Limitations

This study is limited to a particular task and as such may not be applicable to conceptually similar tasks such as Hallucination Detection. Furthermore, our approach is not designed to detect general mistakes or bias in large language models.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021. Evaluating groundedness in dialogue systems: The begin benchmark. *arXiv preprint arXiv:2105.00071*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Qi He, Cheng Qian, Xiusi Chen, Bingxiang He, Yi R. Fung, and Heng Ji. 2025. Veri-r1: Toward precise and faithful claim verification via online reinforcement learning. *Preprint*, arXiv:2510.01932.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. *Preprint*, arXiv:2204.04991.

Nan Hu, Jiaoyan Chen, Yike Wu, Guilin Qi, Hongru Wang, Sheng Bi, Yongrui Chen, Tongtong Wu, and Jeff Z. Pan. 2025. Can LLMs evaluate complex attribution in QA? automatic benchmarking using knowledge graphs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17096–17118, Vienna, Austria. Association for Computational Linguistics.

Syed Khalid Hussain. 2024. Deberta zero-shot classification.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *Preprint*, arXiv:2307.16883.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *Preprint*, arXiv:1910.12840.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024. Attributionbench: How hard is automatic attribution evaluation? *Preprint*, arXiv:2402.15089.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

OpenAI. 2023. Openai gpt-3.5 [text-davinci-003]. Accessed: 2026-04-15.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2022. Measuring attribution in natural language generation models. *Preprint*, arXiv:2112.12870.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. *Preprint*, arXiv:2004.04696.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. Ul2: Unifying language learning paradigms. *Preprint*, arXiv:2205.05131.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Chen Xu, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.