Plausible Extractive Rationalization through Semi-Supervised Entailment Signal

Anonymous ACL submission

Abstract

The increasing use of complex and opaque black box models requires the adoption of interpretable measures, one such option is extractive rationalizing models, which serve as a more interpretable alternative. These models, also known as Explain-Then-Predict models, employ an explainer model to extract rationales and subsequently condition the predictor with the extracted information. Their primary objective is to provide precise and faithful explanations, represented by the extracted rationales. In this paper, we take a semi-supervised approach to optimize for the plausibility of extracted rationales. We adopt a pre-trained natural language inference (NLI) model and further fine-tune it on a small set of supervised data (10%). The NLI predictor is leveraged as a source of supervisory signals to the explainer via entailment alignment. We show that, by enforcing the alignment agreement between the explanation and answer in a question-answering task, the performance can be improved without access to ground truth labels. We evaluate our approach on the ERASER dataset and show that our approach achieves comparable results with supervised extractive models and outperforms unsupervised approaches by $> 100\%$.

1 Introduction

Large language models such as Google's BERT (Devlin et al., 2018) and OpenAI's GPT series (Brown et al., 2020) are gaining widespread adoption in natural language processing (NLP) tasks. These models achieved impressive performance in multiple NLP tasks ranging from solving text generation to information extraction (Liu et al., 2023). However, little is known regarding how answers are generated or which portion of the input text the model focuses on. These flaws highlight concerns surrounding trust and fear of undesirable biases in the model's reasoning chain. Explainable AI (XAI) is currently an active field of research aimed at addressing these issues (Adadi and Berrada, 2018; Cambria et al., 2023; Yeo et al., 2023). In this work, we focus on extractive rationalizing models (Lei et al., 2016), which are also known as Explain-Then-Predict (ETP) models, and are designed towards producing highlights serving as **faithful** explanations. Faithfulness is defined as serving an explanation that represents the model's reasoning process for a given decision, while plausibility refers to the level of agreement with humans (Jacovi and Goldberg, 2020). An advantageous characteristic of ETP models is that they concurrently produce the explanation and the task label, eliminating the necessity for an added layer of interpretation.

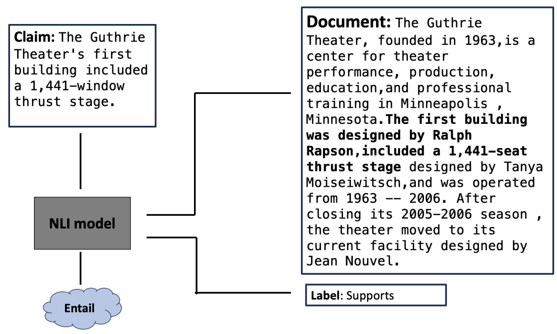

Figure 1: An example from the FEVER dataset, where the bold statement is the annotated rationale. Given the document and claim, the label denotes that the document contains evidence supporting the claim. The NLI predictor interprets this as a form of entailment between the claim and rationale.

This differs from post-hoc techniques such as LIME (Ribeiro et al., 2016) or SHAP (Lundberg and Lee, 2017), specifically tailored to interpret black-box models. Although these techniques are model-agnostic by design, they are computationally expensive and do not guarantee faithfulness nor optimized for plausibility. Chain-of-thought (CoT) (Wei et al., 2022) is another popular approach, aimed at prompting Large Language Models (LLM) such as OpenAI's GPT4 to elucidate its own prediction, in the form of reasoning steps which is said to be a form of explanation. However, we note that though the reasoning steps are seemingly plausible and convincing, there is no guarantee of the reasoning being faithful towards the supported output. This is largely due to the issue of hallucination in LLMs which causes it to generate erroneous answers (Huang et al., 2023), meaning that the explanation can be equally hallucinated, thereby denying its faithfulness. ETP models instead constrain the predictions on a compressed subset of the input, referred to as rationales, thereby guaranteeing the output to be solely conditioned on the subset. This can be seen as a binary form of feature relevance.

In our work, we focus on improving the plausibility of rationales, measured via matching human annotations. Several work has established benchmark datasets that consist of both the task label as well as human-annotated rationales (Bao et al., 2018; DeYoung et al., 2019). Current works in extractive rationalization mostly implement a pipeline procedure of training an explainer and a predictor (DeYoung et al., 2019), trained either jointly or separately. The training approach for these models can be bifurcated into two primary methods: supervised or unsupervised rationale extraction. In our methodology, we strike a balance by leveraging a minimal subset of annotated rationales ($\leq 10\%$) to refine an ETP model. This refinement is applied to a separate NLI predictor, functioning as an auxiliary instructor for the explainer in the event of limited annotated rationales. More importantly, the explainer has no access to the annotations, which are exclusively presented to the NLI predictor.

Our approach is inspired by recent work in ensuring factual consistency in abstraction summarization (Roit et al., 2023), which has been found useful in cases of hallucination. The authors use the entailment signal as a reward in reinforcement learning to ensure factuality in summarization tasks. We instead optimize for plausibility and constrain the explanation to be aligned with the given query. Our proposed approach is simple to implement yet effective in providing effective learning signals. Firstly, we create an augmented dataset based on the alignment between the provided rationales and the NLI classes. This is used to provide further fine-tuning to the NLI predictor. The NLI predictor is then used to annotate each sentence such that it can be used to train the explainer.

NLI models are designed to determine whether a hypothesis contradicts, entails, or is neutral to a given premise. As such, they provide useful signals to align a given explanation to the answer produced by the predictor. An example shown in Figure 1, in a fact verification example, the purpose of the rationale is to act as evidence to either support or refute the given claim. This can be interpreted alternatively as an NLI task where the claim acts as the premise while the rationale is the hypothesis, in this case entailing the premise. We further note that this simple principle not only addresses the scenario of scarce supervisory labels but can also act as a counter-checker against the predictor. As seen later on, this can have some desirable effects on the robustness of rationales (Chen et al., 2022) and enhanced predictive performance. In summary, the three key contributions of this work are the following:

- A simple yet effective approach that improves the plausibility and robustness of extracted rationales, while simultaneously improving task performance. The approach achieves competitive results against supervised models while outperforming unsupervised models by a large margin (>100%).
- To the best of our knowledge, this is the first work to utilize an auxiliary NLI predictor in a semi-supervised fashion for extractive rationalization.
- Our approach has low resource requirements, using models of <300M parameters, and a small set of human-annotated rationales.

2 Methodology

2.1 Problem setting

Given an input document consisting of N sentences, $x_i = \{x_{i,1}, x_{i,2}, ..., x_{i,N}\}$. The task objective can be decomposed into two steps, namely rationale extraction, and task prediction. An explainer, f_θ takes in the input document and generates a binary mask over the sentences indicating the rationales, $f_\theta(\hat{z}_i|x_i) \in \{0, 1\}_N$.

The predictor, g_ϕ can only consider the masked inputs during inference, since the initial reason for extractive rationalization is to present the rationales as a faithful explanation towards the task prediction, $\hat{y}_i = g_\phi(\hat{z}_i \odot x_i)$, $\odot$ is the element-wise multiplication. As rationales are designed to be a concise representation of the original text, there naturally exists a trade-off between generating a sparse z and retaining sufficient information to accurately infer the task label. In various studies, optimization strategies are generally consistent, differing mainly in the use of human-annotated labels for training rationale extractors. Our approach, however, employs a semi-supervised method using an auxiliary predictor optimized for NLI, denoted as f_{NLI}.

2.2 Semi-supervised NLI signal

Humans tend to prefer explanations that are aligned with the supported answer, similar to how NLI tasks involve generating the alignment between two sentences. As such, NLI predictors naturally serve as helpful supervision in the absence of annotated rationales. This is especially applicable in a fact-verification scenario where the task is to infer if a given claim is supported by the provided document. For example, given a document containing the following annotated rationale: *"Kung Fu Panda opened in 4,114 theaters, grossing \$20.3 million on its opening day"* along with a claim: *"Kung Fu Panda made more than \$1 million on opening day."*. The rationale acts as supporting evidence if the corresponding label, y_i = *SUPPORT*, indicates that the claim should be supported given the document and vice versa. The NLI predictor is fine-tuned based on this simple heuristic, to match each sentence in the document against the query. It is trained on the augmented dataset created via a label transformation technique shown in Algorithm 1. Note that the transformation operates under the assumption that there are no contradictory sentences against the label, ie in $x_{i,j}$ contradicts the claim when the label is entailment. The transformation takes into account both annotated labels only during training and predictions otherwise.

During training, the NLI predictor acts as the source of supervision in place of the human-annotated rationales. As the explainer is trained to predict a binary mask, Algorithm 1 can be implemented in reverse to transform the NLI outputs back to rationale labels, $\tilde{z}$ for the explainer's training, (see Appendix for more details). We note that the above approach is likewise applicable to binary true/false tasks where the predictor has to indicate if the answer is true or false concerning the question. This extends the applicability towards most NLP tasks since they can always be rephrased as such.

Algorithm 1 Rationale to NLI label transformation

Input: Annotated rationale, z_i, task label, y_i
Output: NLI label, $\tilde{z}_i$

for $z_{i,j}$ in z_i **do**
  if $z_{i,j} = 1 \wedge y_i$ = TRUE **then**
    $\tilde{z_{i,j}}$ = entailment
  else if $z_{i,j} = 1 \wedge y_i$ = FALSE **then**
    $\tilde{z_{i,j}}$ = contradiction
  else
    $\tilde{z_{i,j}}$ = neutral
  end if
end for
return $\tilde{z}_i$

2.3 Sentence-level training

We utilizes a pipeline approach consisting of a shared encoder, along with separate decoders for the explainer and predictor. The input is first encoded into contextualized hidden states, $h_{i,1:L} = enc(x_{i,1:L})$, where L is at the token level. We follow (Paranjape et al., 2020) and transform the token-level hidden states into sentence-level by concatenating the starting and ending tokens and feeding it into an explainer decoder to produce rationales, $\tilde{z}_i = f_\theta(h_i)$, where $h_i = MLP(h_{i,s} \oplus h_{i,e})$, $\oplus$ is the concatenation process.

The predictor is conditioned on the rationales and trained using standard cross entropy.

$$L_{g_\phi} = -E_{z \sim f_\theta(z|x)}[\log(\hat{y}_i|\hat{z})] \tag{1}$$

The explainer loss, L_{f_θ} is similarly computed with (1), but against the augmented targets, $\tilde{z}_i = f_{NLI}(\tilde{z}_i|x_i, y_i) \in \{0,1\}^N$, instead of the annotated targets. The full training and inference approach is depicted in Figure 2, where the NLI predictor is first fine-tuned before training the ETP model. The choice of a shared encoder allows for a form of dependency between e_i and $\hat{y}_i$, as the encoder has to jointly optimize the representation to infer both the task label and rationales accurately. The final loss is thus a combination of both the predictor and explainer cross-entropy loss, $L_{total} = L_{g_\phi} + \lambda L_{f_\theta}$, where λ balances the trade-off between classification and plausibility performance.

The label transformation is only used during training as it requires access to y_i which is not available at test time. However, we will show how f_{NLI} can remain useful during inference by acting as a counter-checker against $\hat{y}_i$.

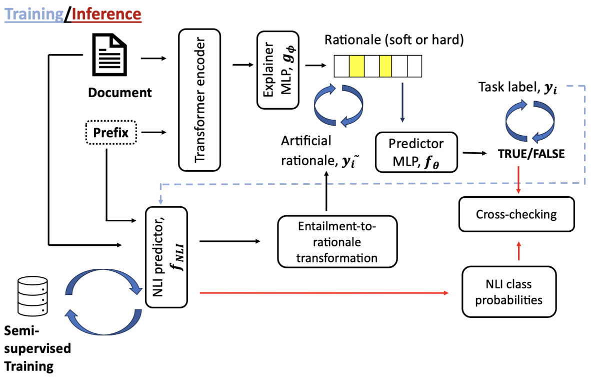

Figure 2: An overview of the proposed approach during training (bold in blue) and inference (bold in red). The NLI predictor only has access to the task label during training. The NLI predictor is initially fine-tuned using a limited set of annotated rationales, before generating artificial targets for the explainer. Cross-checking alignment is conducted during inference against the predictor.

2.4 Inference

During inference, the rationales are extracted solely by the trained explainer, f_θ. However, f_{NLI} can act as a counter-checker against the predictor g_ϕ in the event of a distributional shift in g_ϕ. Given $\hat{z}_i$ and a prefix (claim in fact verification or question-answer pair in Q&A task), f_{NLI} denotes if $\hat{z}_i$ contradicts or entails the prefix. We ignore the neutral probability and re-weight the NLI class probabilities before summing up the n selected sentences in each instance, $\tilde{y}_i^C = \frac{1}{n}\sum_{j=1}^{n}\tilde{y}_{i,j}^C$, where C denotes the NLI class instance. The output task label is then chosen as either generated from the predictor or NLI predictor, whichever is higher, $\hat{y}_i = max(\hat{y}_i, \tilde{y}_i^C)$. This is helpful in the case where g_ϕ is less confident and f_{NLI} steps in by breaking down the task into simpler components such as computing the overall entailment/contradictory score across all sentences or true/false in binary questions.

3 Experiments

3.1 Datasets

We evaluate our approach against unsupervised and supervised baselines across three benchmark tasks from ERASER. ERASER contains a suite of NLP tasks, extended with human-annotated rationales, to assess plausibility.

- **FEVER**: A fact-verification dataset, each instance consists of a claim and a document, where the goal is to determine if the claim is supported or refuted using information from the document.
- **BoolQ**: Question-answering task, containing a context document from Wikipedia and a question, the answer is either true or false. Due to the long sequence, we select the most relevant portion of the context using TF-IDF scoring similar to (Paranjape et al., 2020).
- **MultiRC**: A multi-hop dataset, requiring reasoning over multiple sentences to infer to correct answer. Multiple answer choices can be associated with a single question and the task is to predict if the answer is true or false.

3.2 Experimental Setup

We use RoBERTa-base (Liu et al., 2019) as the shared encoder between the explainer and predictor. The NLI predictor, f_{NLI} is a DeBERTa-large transformer (He et al., 2021) fine-tuned on multiple NLI datasets, we use the v3 variant. Our approach is agnostic to the choice of the pre-trained transformer for both the backbone encoder and NLI predictor. We selected RoBERTa-base, with its 125M parameters, due to its computational efficiency compared to larger models, while still maintaining high performance. We fine-tune the NLI predictor with only 10% of the training data. We list the full hyperparameter details in A.2. A notable benefit of our approach is that it does not require an expensive search over objective-related hyperparameters.

3.3 Baselines

We evaluate our approach against both supervised and unsupervised models, along with predictors subjected to full context. We refer to **Full-C** as the predictor-only set up to assess the gap in task performance between using the full context as compared to a subset. **Supervised** trains the explainer against human-annotated labels, z_i, instead of $\tilde{z}_i$ in our hard-masking approach, serving as the upper bound for plausibility.

IB is an unsupervised approach from (Paranjape et al., 2020) which optimizes a information-bottleneck objective and selects top $N\%$ according to pre-defined sparse prior. The author additionally introduces a semi-supervised approach of using 25% of the annotated rationales which we refer to as **IB-25%**. Note that this baseline is subjected to higher supervision compared to ours (10%). We did not implement IB with similar supervision since there were minor differences in implementation, though we included the reported results for the sake of fairness (**R**). We choose 10% based on empirical results, serving as a good trade-off between minimal resource requirement and performance, albeit a comparable level of supervision (25%) can be referred from Table 4. All evaluated approach implements an ETP-type setup, consisting of an explainer and predictor except for Full-C.

3.4 Metrics

We report task performance using classification metrics such as accuracy and F1-score, while the plausibility of extracted rationales is assessed using token-F1 (DeYoung et al., 2019) at the sentence level. We leave out any faithfulness metrics such as sufficiency as we assume ETP models to be inherently faithful given that the predictor is only subjected to the extracted explanation. We also assess the robustness by exposing the explainer to adversarial inputs. The adversarial attack is generated by prefixing the context with an adversarially crafted query (Chen et al., 2022), by replacing detected nouns and adjectives with antonyms to distract the explainer. This attack aims to evaluate the explainer's proficiency in disregarding sentences that are subtly incongruent and out of context, while similarly influencing the predictor's context window.

We employ the following equations (Chen et al., 2022) to compute the normalized discrepancy in task performance, Δ_T and plausibility, Δ_P between the original and perturbed inputs as an indicator of robustness. Additionally, we utilize the attack rate, AR to gauge the frequency with which the explainer identifies adversarial sentences.

$$\Delta_T = \frac{1}{N}\sum_{i=1}^{N}\frac{M_t(\hat{y}_i, y_i) - M_t(\hat{y}_i^A, y_i)}{M_t(\hat{y}_i, y_i)} \quad (2)$$

$$\Delta_P = \frac{1}{N}\sum_{i=1}^{N}\frac{M_p(\hat{z}_i, z_i) - M_p(\hat{z}_i^A, z_i)}{M_p(\hat{z}_i, z_i)} \quad (3)$$

$$AR = \frac{1}{N}\sum_{i=1}^{N}\hat{z}_i \cap z^{AS} \quad (4)$$

M_t and M_p is the scoring function for task and plausibility performance, for which we use the F1 and Token-F1 measurement. $\hat{y}_i^A$, and $\hat{z}_i^A$ refer to the generated class label and rationale given the adversarial input. z^{AS} refers to the position of the adversarial prefix.

4 Results

In this section, we will assess our approach against the introduced baselines. All results are averaged over three runs with different seeds. For Full-C, we do not report plausibility performance since there is no explainer module. In the ERASER benchmark, the number of annotated rationale sentences varies between instances as well as tasks. The BoolQ dataset features a greater quantity of annotated sentences and also includes more extended continuous spans of these sentences. The objective of the experiment is to judge the various ETP models' reasoning capabilities over a compressed span of text while having the generated explanation stay as close as possible to human references. Furthermore, we are also interested in studying how an NLI predictor can provide useful learning signals to the explainer in the event of limited annotations.

4.1 Plausibility and Task Analysis

The task and plausibility performance is shown in Table 1. We are unable to replicate the exact results for IB, but for the sake of fairness, we report the performances gathered from the original work (Paranjape et al., 2020). Judging from the results, our approach achieves highly competitive performance against the gold standard for both task (Full-C) and plausibility (Supervised). In FEVER, it even surpasses the full context approach (94.2 vs 93). It goes to show that ETP-like models can benefit from ignoring spurious noise by conditioning the predictor to only text considered essential for inferring the target class. The additional usage of f_{NLI} as a cross-checker during inference also provided considerable improvements across all three benchmarks, at little to no cost in computational resources.

In terms of plausibility, our method delivers a token-f1 score that is on par with the fully supervised approach across all datasets except BoolQ. We note that a likely reason is that the target rationales are largely inconsistent in length, with instances stretching across as many as six contiguous sentences. Since the NLI predictor is optimized toward matching each sentence with the given query. It may fare worse when individual sentences appear to be unrelated to the query but are nonetheless annotated as rationales. Table 2 shows the percentage proportion of sentences annotated as rationales over the target. It's noteworthy that the explainer marks fewer sentences due to the NLI predictor's

	FEVER			MultiRC			BoolQ		
	Task		Plausibility	Task		Plausibility	Task		Plausibility
Approach	Acc	F1	Token-F1	Acc	F1	Token-F1	Acc	F1	Token-F1
Full-C	93	91.8	-	**76**	**72**	-	65.8	53	-
Supervised	90.1	88.4	**83.4**	74.3	70.5	**64.1**	**72.4**	**65.9**	**76**
IB	85.9	85.9	38.9	64.1	63	23.1	64	63.5	10.3
IB w 25%	85.1	85.1	38.4	67.6	67.5	52.7	58.6	52.1	11.4
IB w 25% (R)	-	88.8	63.9	-	66.4	54	-	63.4	19.2
Ours (10%)	$\mathbf{93.7}_{+0.5}$	$\mathbf{92.6}_{+0.5}$	**80.1**	$72.5_{+0.0}$	$68.6_{+0.4}$	**56.4**	$67.4_{+2.1}$	$51.4_{+8.6}$	**29.6**

Table 1: Classification and plausibility performance comparison across the three ERASER tasks. Test results are averaged across 3 seeds. The subscript refers to the case where the NLI predictor is used as a counter checker, in 2.4. Results highlighted in bold refer to the best-performing approach. The supervised approach acts as the upper bound on plausibility performance. **R** is the reported results of the IB approach (Paranjape et al., 2020).

FEVER	MultiRC	BoolQ
100	56.7	20

Table 2: Percentage of extracted over target rationales. BoolQ has the lowest percentage out of all three datasets.

tendency to classify the majority of sentences as neutral, deeming them non-essential for task prediction.

Nonetheless, optimizing the explainer with NLI supervision is proven to be superior compared to the unsupervised information bottleneck objective by (Paranjape et al., 2020). Our approach outperforms the former by large margins in terms of plausibility on FEVER ($> 25\%$) and BoolQ ($> 50\%$), even when provided with a lower amount of supervision (10% vs 25%). The performance gap is even larger when compared to fully unsupervised (IB), with more than twice the scores. The IB method learns a sparse mask over the input document, x_i by maximizing the mutual information between rationale z_i and task label y_i while limiting the extraction budget to a pre-defined prior. However, estimating the prior is difficult and can be detrimental in instances with varying rationale lengths such as in BoolQ. Our approach sidesteps the complicated training yet achieves a better-tuned explainer in extracting plausible rationales.

4.2 Robustness

In this section, we evaluate the robustness of ETP models when faced with inputs prefixed with an adversarial query. The query is unrelated to the document and carries a contrastive meaning with respect to the original. For example, given a claim in FEVER, *"Earl Scruggs was a musician who played banjo."*, the noun, *"Earl Scruggs"* and *"banjo"* is replaced to form the adversarial sentence *"manchester archer was a songwriter who played mandolin ."*. The attack is minimally changed from the query to distract the explainer. A model with limited robustness might interpret the attack as pertinent due to its analogous semantics, thereby influencing the predictor and undermining task performance. The robustness results are reported in Table 3. We found similar findings as compared to (Chen et al., 2022) who note that ETP models exhibit greater robustness compared to predictors subjected to the full context.

In the FEVER dataset, our approach suffers the lowest drop in task and plausibility performance, while having the lowest AR in both datasets. IB has the highest AR, even extracting every adversarial sentence in FEVER. A contributing factor to our approach's low AR rate is that the NLI signal is derived by verifying if a sentence aligns with the query based on the provided task label. This strengthens the explainer's proficiency in dismissing instances that don't satisfy this criterion. On the other hand, the explainer trained with IB is emphasized to maximize the task objective, which can lead to situations where a minimally perturbed sentence is mistakenly perceived as useful. This further proves that training with NLI feedback produces more robust and plausible models.

5 Ablation

5.1 Importance of NLI training

In this study, we seek to question the usefulness of introducing further fine-tuning using the limited set of annotations. While the NLI predictor is previously fine-tuned on various NLI tasks, the sentence lengths in its training distribution differ from

Approach	FEVER Δ_T	FEVER Δ_P	FEVER AR	MultiRC Δ_T	MultiRC Δ_P	MultiRC AR
Full-C	11.2	-	-	29.6	-	-
Supervised	10.8	37	54.6	14.3	26.7	68.3
IB (25%)	12	35.8	100	**4.9**	**19.1**	93
Ours (10%)	**7.8**	**8.4**	**32.3**	10.2	27.1	**67.2**

Table 3: In both FEVER and MultiRC, we measure robustness with a preference for lower values. Models considering the full context are evaluated solely based on the difference in task performance as they don't engage in rationale extraction. All values are normalized percentages drop computed via (2),(3) and (4)

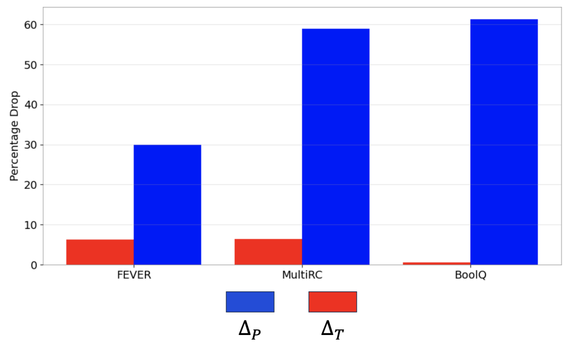

Figure 3: Task and Plausibility performance drop when there is no further fine-tuning on the NLI predictor (10% data). The metrics are computed similarly to robustness using (2) and (3) and are presented in normalized percentages.

those in our experimental datasets. Additionally, domain-specific semantics differences can introduce variations in the NLI predictor's inference process. Consequently, the NLI predictor might not always accurately discern the NLI class, leading to the generation of misleading signals for the explainer. To quantify the effectiveness of further fine-tuning, we compute the drop-in performance on both task and plausibility between an NLI predictor that is fine-tuned, referred to as **FI** and one that is not, **NFI**. The gap in task and plausibility performance is reported in Figure 3.

These results substantiate our initial hypothesis. Without fine-tuning, NFI struggles to provide meaningful feedback to the explainer, primarily because of its limited capability to accurately determine whether a specific sentence should support or contradict the query based on the given task label. Taking a closer look at sentence classifications in Figure 4 reveals that the NFI tends to mistakenly identify neutral sentences as entailments. In the FEVER example, although the initial sentence shares a noun with the claim, it does not address

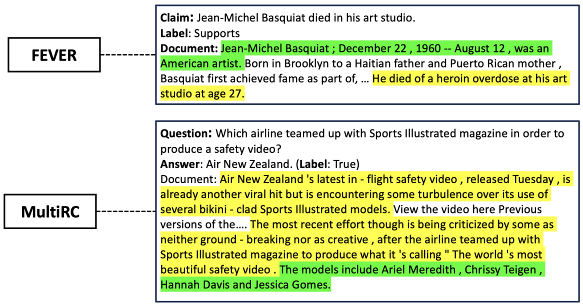

Figure 4: Example of query and input document where the sentences highlighted in green refer to the NLI predicator without fine-tuning. Yellow refers to the annotated rationale as well as extracted by the fine-tuned predictor.

the death of the noun's subject yet the NFI incorrectly recognizes it as entailment. Similarly, in the MultiRC instance, the concluding sentence lacks any significant connection to the given question or answer. This may be the reason why despite a significant drop in accurately extracting the correct rationale, 58.9% in MultiRC and 61.3% in BoolQ, the task performance surprisingly does not incur a huge loss ($< 10\%$). A neutral sentence would not drastically change the class probabilities of the predictor as compared to a contradicting sentence. Nevertheless, incorporating additional fine-tuning on the NLI predictor is still essential in filtering out false positives such as sentences with neutral relationships in inferring the output.

5.2 Model sizes and NLI supervision

Ablation type	Acc	F1	Token-F1
Original (base w 10%)	72.5	68.6	56.4
Large	74.3	71.2	58.1
Base w 25%	73.3	71.1	57.9
Base w 50%	73.8	70.9	58
Without NLI Pre-FT	69.2	64.3	52.6

Table 4: Ablation on model size, % NLI supervision and effects of not doing pre-finetuning of f_{NLI} on general NLI tasks (SNLI, MultiNLI). Implemented on MultiRC.

We carry out further analysis on the effect of both model size and amount of NLI supervision given to f_{NLI}. We compare RoBERTA-large (330M) with the original 10% supervision and the base model with increased level of supervision $\in [25, 50]$. We additionally compare a DeBERTa encoder without prior fine-tuning on NLI datasets, while similarly fine-tuning on 10% of annotated rationales. The benefits of using a larger encoder and increased NLI supervision for f_{NLI}

can be observed from Table 4. Notably, there is little difference in both accuracy and Token-F1 with higher supervision. Furthermore, our approach remains effective using off-the-shelf encoders without prior fine-tuning. This highlights the strength of our approach which remains effective even in low-resource conditions.

6 Related Works

Linguistic Interpretability: Among various interpretability approaches, extractive rationalization has seen plenty of works produced in recent years (Gurrapu et al., 2023). Other similar practices include the use of attention (Mohankumar et al., 2020; Serrano and Smith, 2019) as a form of interpretation to rank text tokens in terms of importance. The topic has caught interested researchers' attention, leading to a division of the field into two groups, either disagreeing or agreeing on the usage of attention as a faithful interpretation. The former states that even when attention scores are randomly scrambled, it has little effect on the predictor's output (Jain and Wallace, 2019), along with the difficulty of generating counterfactuals. Other stances reason that attention weights are biased as they encode information on neighboring tokens (Bai et al., 2021; Tutek and Šnajder, 2022). On the other hand, (Wiegreffe and Pinter, 2019) argues that there exist multiple combinations of weights that could lead to a single output and that the effectiveness of attention weights depends on the definition of faithfulness. This has driven several works to improve the faithfulness of attention-based networks by constraining the attention with the task (Chrysostomou and Aletras, 2021b) or directly penalizing the attention scores corresponding to important words (Chrysostomou and Aletras, 2021a).

Extractive Rationalization: The approach was first introduced by (Lei et al., 2016), who proposed REINFORCE (Williams, 1992) with sparsity regularization to train the explainer through the predictor's learning objective in an end-to-end fashion. (DeYoung et al., 2019) introduces the ERASER benchmark, containing seven NLP tasks. The author utilizes a BERT-to-BERT pipeline for the explainer and predictor and performs sequential training. (Atanasova et al., 2022) constrains the explainer to be consistent and confident while (Lakhotia et al., 2020) adopts the Fusion-In-Decoder (Izacard and Grave, 2020) to process long sequence documents and extract rationales at the sentence level.

However, obtaining supervised rationales is often a privilege that's not readily accessible. This has prompted a large number of research in unsupervised techniques for extracting faithful rationales. (Paranjape et al., 2020) strives for conciseness in rationales via optimizing an information bottleneck objective. (Ghoshal et al., 2022) addressed the issue of spurious correlation in QA tasks by incorporating an additional question generation objective and (Jain et al., 2020) decomposes the joint objective into modular components. (Glockner et al., 2020) encode each sentence separately and aggregate the final task loss using the normalized weights of each sentence. During inference, the sentence with the lowest loss is selected as the rationale. In our approach, we do not make such an assumption, and optimize for every rational sentence.

7 Conclusion

In this paper, we have introduced a simple yet unique way of generating artificial learning signals from an alternative source, to cope with scenarios where human-annotated rationales are scarce. The method harnesses a transformer pre-trained on the NLI task. Through additional fine-tuning, the NLI predictor can produce less biased labels, enhancing the learning process for the explainer.

Through the extensive experiments conducted, we have shown that our work can alleviate the plausibility and robustness of ETP models in a low-resource environment. Notably, with just 10% of the annotated rationale, our method delivers performance on par with fully supervised models and significantly outperforms both semi-supervised approaches that utilize more annotated data as well as unsupervised ones. We have also demonstrated the relevance between an NLI label and the plausibility of the generated explanation. In future directions, we plan to extend this work toward models that generate abstractive explanations, where the NLI signal can act as verification feedback to ensure the mitigation of biased explanations. Another interesting direction is to study how can we extend the NLI predictor's coverage beyond a single sentence, to capture the correspondence between longer documents.

8 Limitations

We only evaluate a singular trait of interpretability: plausibility. We note that multiple other traits of interpretability are equally important and we leave that to further work. The sizes of the encoder models implemented in this work are relatively small, with the biggest consisting of 300M parameters. Though model scaling is the primary objective, we note the importance of extending our work towards larger models given the popularity of NLP research surrounding LLMs.

References

Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Diagnostics-guided explanation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10445–10453.

Bing Bai, Jian Liang, Guanhua Zhang, Hao Li, Kun Bai, and Fei Wang. 2021. Why attentions may not be interpretable? In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 25–34.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. *arXiv preprint arXiv:1808.09367*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. 2023. A survey on XAI and natural language explanations. *Information Processing and Management*, 60(103111).

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can rationalization improve robustness? *arXiv preprint arXiv:2204.11790*.

George Chrysostomou and Nikolaos Aletras. 2021a. Enjoy the salience: Towards better transformer-based faithful explanations with word salience. *arXiv preprint arXiv:2108.13759*.

George Chrysostomou and Nikolaos Aletras. 2021b. Improving the faithfulness of attention-based explanations with task-specific information for text classification. *arXiv preprint arXiv:2105.02657*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

Asish Ghoshal, Srinivasan Iyer, Bhargavi Paranjape, Kushal Lakhotia, Scott Wen-tau Yih, and Yashar Mehdad. 2022. Quaser: Question answering with scalable extractive rationalization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1208–1218.

Max Glockner, Ivan Habernal, and Iryna Gurevych. 2020. Why do you think that? exploring faithful sentence-level rationales without supervision. *arXiv preprint arXiv:2010.03384*.

Sai Gurrapu, Ajay Kulkarni, Lifu Huang, Ismini Lourentzou, Laura Freeman, and Feras A Batarseh. 2023. Rationalization for explainable nlp: A survey. *arXiv preprint arXiv:2301.08912*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C Wallace. 2020. Learning to faithfully rationalize by construction. *arXiv preprint arXiv:2005.00115*.

Kushal Lakhotia, Bhargavi Paranjape, Asish Ghoshal, Wen-tau Yih, Yashar Mehdad, and Srinivasan Iyer. 2020. Fid-ex: Improving sequence-to-sequence models for extractive rationale generation. *arXiv preprint arXiv:2012.15482*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. 2020. Towards transparent and explainable attention models. *arXiv preprint arXiv:2004.14243*.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. *arXiv preprint arXiv:2005.00652*.

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.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, et al. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. *arXiv preprint arXiv:2306.00186*.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Martin Tutek and Jan Šnajder. 2022. Toward practical usage of the attention mechanism as a tool for interpretability. *IEEE Access*, 10:47011–47030.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.

Wei Jie Yeo, Wihan van der Heever, Rui Mao, Erik Cambria, Ranjan Satapathy, and Gianmarco Mengaldo. 2023. A comprehensive review on financial explainable ai. *arXiv preprint arXiv:2309.11960*.

A Appendix

In the main paper, we showed how the label transformation technique is used to transform an annotated rationale into an NLI-associated label, for the purpose of fine-tuning the NLI predictor. We will now show how the reverse is applied to facilitate the training of the explainer.

A.1 Reverse label transformation

Given a query, q and each sentence, x_i, we concatenate the query and sentence as input to the NLI predictor, where the NLI class label is generated as $\tilde{y}_i = f_{NLI}(q \oplus x_i)$. This applies to both queries with a single sentence such as the claim in FEVER and BoolQ or double sentences in MultiRC, comprising of both the question and answer. The NLI class, $\tilde{y}_i$ is used together with the task label, y_i to generate $\tilde{z}_i$, used in place of z_i for the semi-supervised explainer. The transformation is detailed in Algorithm 2. T refers to False, F to FALSE, C to Contradiction, and E to Entailment. Note that if the f_{NLI} indicates that the sentence is neutral to the query, the sentence is automatically labeled as a non-rationale. This is similar in the case where if a document is annotated as false with respect to the query, all rationales should be a contradiction and vice versa.

Algorithm 2 Reverse label transformation

Input: query, q_i, input document, x_i, task label, y_i and NLI predictor, f_{NLI}
Output: NLI label,$\tilde{z}_i$

```
1: for each x_{i,j} ∈ x_i do
2:     ỹ_i ← f_NLI(q_i ⊕ x_{i,j})
3:     if (ỹ_i = E and y_i = T) or (ỹ_i = C and y_i = F) then
4:         z̃_{i,j} ← 1
5:     else
6:         z̃_{i,j} ← 0
7:     end if
8: end for
9: return z̃_i
```

A.2 Hyperparameters

We use the AdamW optimizer from (Loshchilov and Hutter, 2017) with ϵ set at 1e-8 and fix the batch size at 8. We use a learning rate warm-up scheduler with the final rate capped at 2e-5 and clip all gradient norms at a value of 1.0 while applying a dropout of 0.2 for the explainer decoder. The explainer decoder is a two-layer MLP with ReLU activation. Early stopping is implemented where the training is stopped if the validation loss does not improve after 3 epochs. We run all our experiments for a maximum of 10 epochs, on NVIDIA A6000s, implemented with PyTorch. We do not find much difference in changing the values of λ and set it to 1.