Cross Domain Transfer of Natural Language Explanation Models: Pretraining on e-SNLI and Adapting to a New Target Task

Md. Farhad Rahman[1]; Mohammad Sayduzzaman[2]; Tawhidur Rahman[3]; Monira Mostafa[4]

[1]Department of Technology HSBC, Bangladesh Dhaka, Bangladesh
[2]Department of Computer Science University of Lahore Lahore, Pakistan
[3]Department of CSE, Faculty of Science & Technology Bangladesh University of Professionals (BUP) Dhaka, Bangladesh
[4]Department of Technology BEXIMCO IT Division Dhaka, Bangladesh

Publication Date: 2026/03/12

Abstract: The extensive use of AI in the critical ICT systems requires not only accurate, but transparent and credible models. Nevertheless, state of the art models are usually black boxes and their answers to decisions can be fragile, and do not make generalizations in different areas of operation. The problem of designing effective, transferable natural language explanations (NLEs) is discussed by building a multi task T5 based model that takes the label-prefixed format of decoders to jointly assign NLI labels and produce explanations. Pretraining of the model occurs on e-SNLI then fine tuning is done under different cross domain conditions, such as label only supervision, frozen encoders, and loss weight variations. Although there are no explanations to be found in the fine-tuning process, the experimental results show that explanation pretraining can greatly improve the linguistic fluency, structure, and relevance of explanations. The partial faithfulness is also provided by token deletion tests which reveal that the explanations are based on the same evidence as the classifier does. Abalation studies demand stable and transferable explanations to be characterized by balanced loss weighting, encoder adaptation, and explanation oversight. These results point to the necessity of standardized assessment tools of NLE and indicate directions on how the explanation-capable models can be incorporated into ICT systems that need transparency and accountability.

Keywords: *Natural-Language Explanations, e-SNLI, Cross-Domain Transfer, Multi-Task Learning, Faithfulness Evaluation, Explainable AI, ICT Standardization, Trustworthy AI.*

How to Cite: Md. Farhad Rahman; Mohammad Sayduzzaman; Tawhidur Rahman; Monira Mostafa (2026) Cross Domain Transfer of Natural Language Explanation Models: Pretraining on e-SNLI and Adapting to a New Target Task. *International Journal of Innovative Science and Research Technology,* 11(3), 307-319. https://doi.org/10.38124/ijisrt/26mar233

I. INTRODUCTION

Natural language inference (NLI) has become a major metric toward evaluating language understanding in neural networks where it began with large scale datasets such as SNLI [1], which has reported its use on neural networks from 2015 onwards. Even though these benchmarks made significant progress in reasoning at the sentence level, they provide only labels, remaining unaware of the underlying decision-making mechanism of the model to users. To address this problem, e-SNLI dataset introduced human-written natural language explanations (NLEs) as a new supervision signal [2]. The development of this prompted the growing interest in models that do not just predict labels but provide intelligible explanations to their reasons.

The list of methodological approaches applied in the study of interpretability is not exhaustive, which includes post-hoc local explanations [3], [4], attribution based methods [5], influence analysis, and self-explaining architectures [6], among others. But these methods are frequently not even true to the underlying model, or even present natural language reasons that are sound. Simultaneously, some papers focus on the flaws of the commonly applied models of explana-tions, such as demonstrating that attention weights do not provide explanations in general, as in the case of attention weights that are not explanatory when perturbed, or more generally, that explanations can be fragile to perturbations, among others [7]. These results provide evidence of the necessity to have more robust, causally grounded and human-reasoning-consistent frameworks of explanation. Datasets with rationales, including ERASER [8], CoS-E [9] and others- have demonstrated that explanation supervision is capable of en-hancing interpretability and in some cases predictive accuracy. However, recent studies have shown that many of them are fraught with difficulties: explanations can be unfaithful [10] to their causal cause, susceptible to adversarial manipulation [10], unable to be predicted over domains, and so on.

➢ *Motivation for Trustworthy AI in ICT Infrastructure*

The need to integrate AI in the global ICT frameworks of telecom network management to automated customer service and regulatory compliance has rendered model transparency as an imperative requirement. The impact of AI driven de-cision making on the real world can be serious and lack of understanding or auditing of the decisions can be extremely dangerous. Natural language explanations (NLEs) provide the solution to this issue, which is human centric, yet its trustwor-thiness is the most important aspect. Transfer of explanation is important to deployed systems due to the fact that models are likely to be required to perform in environments that are dissimilar to the training data (domain shift). An explanation that breaks down on a change of domain is not only unhelpful, but can even be deceptive. For example, in a multilingual global telecom service, an AI might need to explain its reasoning across different languages and cultural contexts. Robust, cross-domain NLEs are essential for helping human operators understand, trust, and, when necessary, override AI-generated decisions, thereby ensuring the safety and reliability of critical infrastructure.

To our knowledge, no prior work has empirically examined the cross-domain robustness and standardization potential of NLE-based explanation models trained on e-SNLI. This study seeks to fill that gap. In this study, we investigate whether explanation-capable models pretrained on e-SNLI can transfer their explanatory abilities to a new downstream challenge. We deploy a multi-task architecture that jointly predicts labels and generates explanations, then examine performance under multiple learning configurations: complete fine-tuning, label-only fine-tuning, encoder freezing, and zero-shot explanation generation. We also conduct faithfulness evaluations using perturbation-based tests to determine if the model’s expla-nations respond accurately to changes in the input. Our key contributions are:

- A multi-task T5-based architecture for NLE transfer across domains.
- A systematic evaluation of explanation generation under various fine-tuning regimes including zero-shot scenarios.
- A token-deletion-based faithfulness test for cross-domain scenarios with quantitative metrics.
- Ablation studies identifying conditions for stable explanation transfer.
- Discussion of how findings support emerging needs for XAI standardization in ICT systems.

The remainder of the paper is organized as follows. Section II mentioned the literature review. Section III discusses method-ological aspects. Section IV focuses on results, while contextual elements are covered in Section V. Model implementations are examined in Section VI. Section VII discusses future aspects. Finally, Section VIII concludes the paper.

II. LITERATE REVIEW

➢ *Explainable AI and Natural-Language Explanations*

Explainable Artificial Intelligence (XAI) aims at making machine learning systems more transparent, accountable, and understandable. Examples of traditional explanation methods include saliency maps, gradient based attributions, and rule extraction methods offer low level or modality specific under-standing of model behavior [3]–[5]. However, these approaches are often not sufficient to present model thinking, which is consistent with human judgment [7], [11]. More recent efforts have thus shifted to so called Natural Language Explanations (NLE) explaining a model in free text. NLEs are user friendly, do not require expertise, and can be particularly beneficial in addressing such areas as healthcare, finance, or education, where the consumers should become familiar with the logic behind predictions. The NLEs have therefore emerged as an encouraging field in designing dependable and humanistic AI systems.

➢ *SNLI and e-SNLI*

One of the baseline models of entailment, contradiction, and neutrality between sentence pairs is the Stanford Natural Language Inference (SNLI) dataset [1] of sentence pairs, which is openly available in this database. Although SNLI has gone a long way to enhance understanding of natural languages, it is giving out mere labels and does not provide any explicit human thinking. To overcome this limitation, the e-SNLI dataset [2] is added to each SNLI example with token-level reason marks and human-generated explanations. Consequently, e-SNLI is the first extensive corpus specifically created for model evaluation and training that generate both predictions and natural-language arguments. Its emphasis on aligning labels and explanations makes it a perfect source for pretraining explanation-capable models. TABLE I shows the condensed comparison work in NLI and transparency.

➢ *Cross-Domain Transfer for NLE Models*

Cross-domain transfer involves pretraining a model on a source dataset and adapting it to a target dataset with different structures, domains, or linguistic characteristics. While transfer learning is well studied in classification and language modeling [12]– [14], the transferability of *explanations* remains underex-plored. Important unanswered research questions consist of:

- How efficiently do explanation patterns acquired in one dataset transfer to a qualitatively different domain?
- Does explanation oversight enhance the faithfulness or quality of downstream explanations?
- How does explanation coherence alter when training switches from an explanation-rich dataset (e.g., e-SNLI) to a label-only dataset?

Prior work shows that explanations may not always reflect true model reasoning [15], can be unstable [7], or may fail to generalize across tasks [9]. Even while e-SNLI offers a solid ba-sis for learning structured explanations, it's yet unknown if these explanation skills translate to new activities without explanation annotations. The performance of explanation-trained models when transferred across domains is not adequately covered by current research on rationale-based and post-hoc interpretability [6], [8], [16]. Furthermore, faithfulness studies emphasize that explanations need to be causally connected to decision processes rather than just convincing stories [17].

III. METHODOLOGY

This section describes the methodology used to study whether explanation-capable models trained on the e-SNLI dataset can generalize their natural-language explanation ability to new downstream tasks. The methodology includes dataset pre-processing, model architecture, multi-task training objectives, pretraining on e-SNLI, cross-domain fine-tuning, explanation-generation strategy, and faithfulness evaluation.

Table 1 Condensed Comparison of Prior Work in NLI and Explainability.

Ref	Year	Focus / Gap
[1]	2015	NLI labels; no explanations
[2]	2018	NLEs; faithfulness unclear
[8]	2020	Rationales; no unified metrics
[3]	2016	Local explanations; unstable
[5]	2017	Attribution; baseline issues
[16]	2016	Rationales; partial faithfulness
[18]	2017	Gradient regularization; scaling limits
[4]	2018	Rules; weak on long text
[11]	2019	Critique of attention
[12]	2020	Seq2seq; no NLE objective
[13]	2019	Seq2seq; lacks rationale supervision
[19]	2017	Training influence; not scalable
[9]	2019	Commonsense NLEs; shallow
[20]	2019	QA; no explanations
[15]	2020	Faithfulness; causal limits

➢ *Overview*

Our approach follows a two-stage training pipeline:

- Pretraining on e-SNLI: a sequence-to-sequence model jointly learns (i) inference classification and (ii) natural-language explanation generation.
- Cross-domain Transfer: the pretrained model is adapted to a new task, with or without explanation supervision. Explanation quality and faithfulness are evaluated after transfer.

As described in the e-SNLI dataset [2], target sequences are formatted as "<label> [SEP] <explanation>", en-suring that the decoder learns to begin the generated explanation with the label token.

➢ *Datasets*

- *Source Dataset:*

e-SNLI: The e-SNLI dataset [2] extends the Stanford Natural Language Inference (SNLI) corpus [1] by providing human-written explanations for each (premise, hypothesis, label) triple. Each instance contains:

- ✓ A natural-language premise,
- ✓ A natural-language hypothesis,
- ✓ An NLI label (*entailment*, *contradiction*, *neutral*),
- ✓ A free-form human explanation and annotated rationale tokens.

Following prior work on sequence-to-sequence explanation modeling [2], [9], [12], each e-SNLI example is converted into a text-to-text format:

Input_text = “premise: ¡P¿ hypothesis: ¡H¿”,

Target_text = “¡label¿ [SEP] ¡explanation¿”.

This structure mirrors the original design of e-SNLI, where the label token is prepended to the explanation sequence to con-dition the decoder on the correct inference label [2]. The format has subsequently been adopted in other explanation-generation datasets such as CoS-E [9] for improved controllability and coherence.

- *Target Dataset:*

For cross-domain transfer, we assume a target dataset that may contain either: (i) both labels and explanations, or (ii) label-only supervision. In the latter case, explanation generation is trained implicitly using the pretrained decoder, enabling zero-shot or few-shot explanation transfer, as explored in prior explanation-generalization studies [8], [15].

➢ *Model Architecture*

- *Encoder–Decoder Backbone:*

We adopt the T5-small encoder–decoder architecture [12] due to its versatility in con-ditional text generation and its prior success in NLE modeling [2], [9]. The encoder processes the concatenated premise and hypothesis, while the decoder autoregressively generates the label-prefixed explanation. The encoder creates contextualized token representations by processing the combined premise-hypothesis pair (or the target task's input text). Conditioned on the encoder output, the decoder produces explanations in natural language.

- *Classification Head:*

To jointly learn prediction and explanation, we extend T5 with a lightweight classification head following common approaches in multitask NLI and rationale (0) modeling [6], [16]. The encoder's first hidden-state vector h_{enc} is passed through a linear layer:

$$\hat{\ } = \text{softmax}(\quad h_{\text{enc}}^{(0)} + \ ).$$

This allows simultaneous label prediction and explanation generation within a unified architecture. During pretraining and fine-tuning, the classification head enables the model to learn inference labels and adjust to new label sets.

➢ *Training Objectives*

We adopt a multitask objective inspired by prior NLE models [2], [9], combining classification and explanation losses:

$$L = {}_{\text{cls}} L_{\text{label}} + {}_{\text{gen}} L_{\text{expl}}.$$

Where:

- L_{label} is cross-entropy loss over NLI labels,
- L_{expl} is token-level cross-entropy for the decoder,
- $\text{cls} = \quad \text{gen} = 1$ as in [2].

The classification component is computed as:

$$L\text{label} = -\sum \qquad \log\ (\quad |\quad),$$

while explanation loss uses teacher forcing:

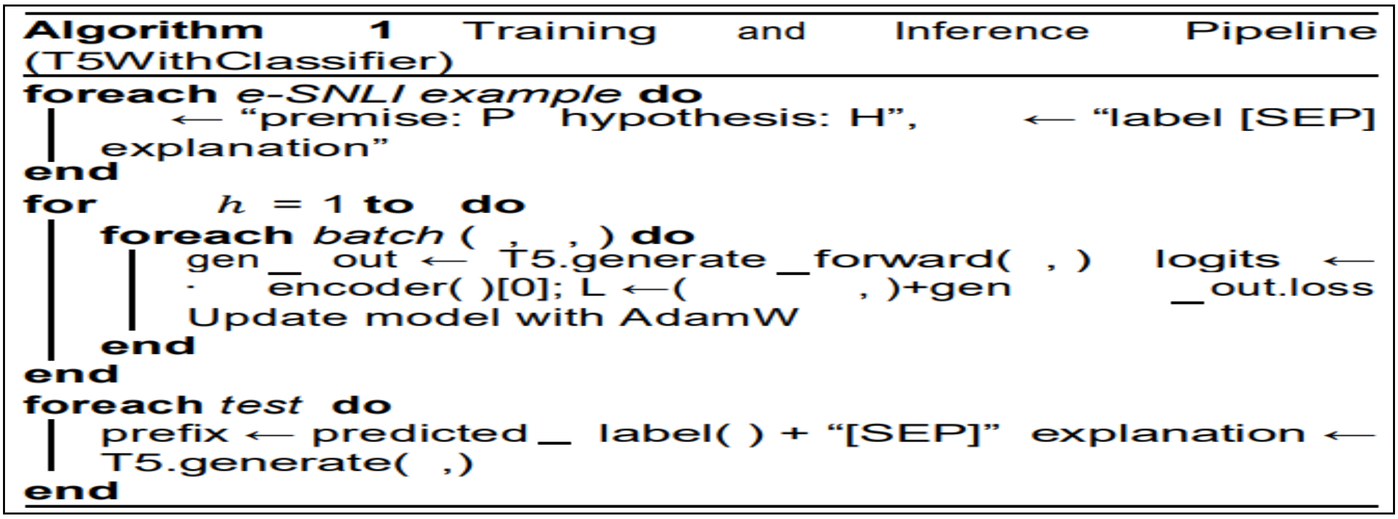
$$L\text{expl} = -\sum \quad \log\ (\quad |\ < \quad , \ ).$$

➢ *Pretraining on e-SNLI*

We train the model on the transformed e-SNLI data using AdamW with a learning rate of 3×10^{-5} following common practices for encoder–decoder pretraining [12], [13]. Training is conducted for 1–5 epochs depending on validation performance, constrained by available GPU memory.

During pretraining, the model is optimized jointly for classi-fication and explanation generation. Training is performed with the AdamW optimizer, a learning rate of $3\times_{10}-5$, a warmup ratio of 0.06, and a linear decay schedule. We use a batch size of 16 per GPU and train for 3–5 epochs with early stopping based on validation loss.

The core multitask model implementation is shown below, adapted from previous rationale-augmented architectures [8], [16]:



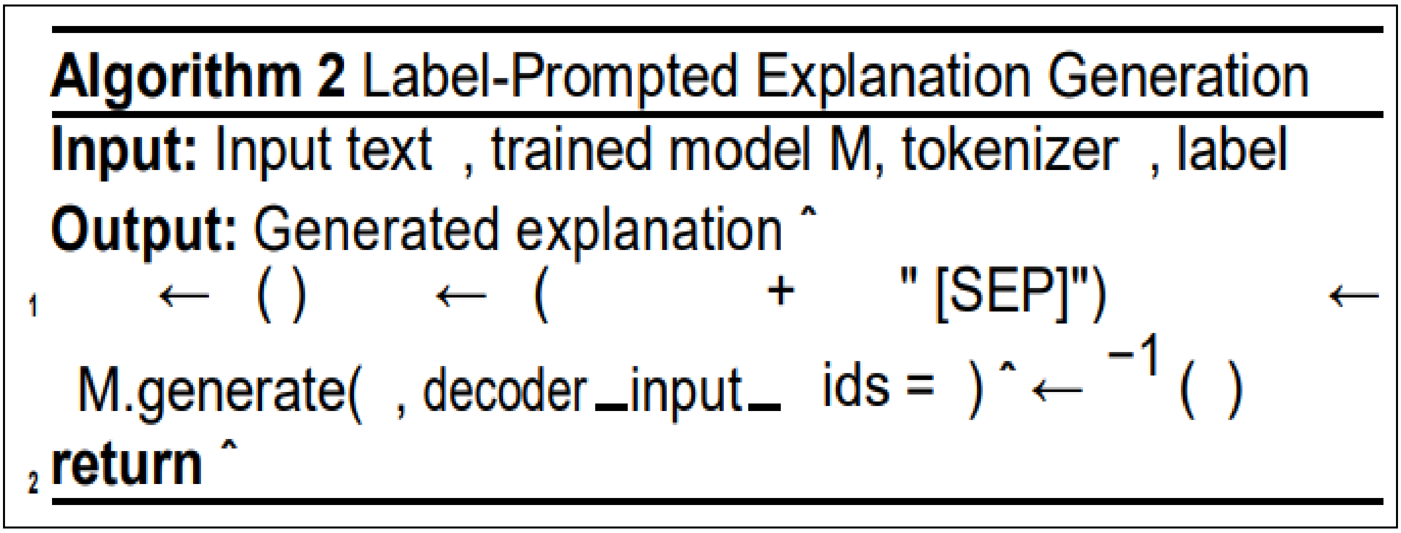
```
Algorithm 1 Training and Inference Pipeline (T5WithClassifier)
foreach e-SNLI example do
    ← "premise: P  hypothesis: H",    ← "label [SEP] explanation"
end
for    h = 1 to   do
    foreach batch ( , , ) do
        gen _ out ← T5.generate _forward( , )  logits ←
        ·  encoder( )[0]; L ←(          , )+gen _out.loss
        Update model with AdamW
    end
end
foreach test do
    prefix ← predicted _ label( ) + "[SEP]" explanation ←
    T5.generate( ,)
end
```

➢ *Explanation Generation Strategy*

Because e-SNLI uses a prefix of the form "<label>[SEP]" in its target sequences, the decoder is implicitly trained to begin explanations with the predicted label [2]. Prior work on explanation-conditioned generation [9], [15] shows that such prompting significantly improves coherence and relevance.

During inference, naive generation may yield empty or truncated outputs if the decoder is not explicitly primed. To address this, we use label prompting by initializing the decoder input with: "<label> [SEP]"

This approach ensures that the explanation is conditioned on the label, consistent with other NLE generation pipelines [2], [9].



```
Algorithm 2 Label-Prompted Explanation Generation
Input: Input text  , trained model M, tokenizer  , label
Output: Generated explanation ˆ
   ← ( )   ← (          +    " [SEP]")          ←
 M.generate(  , decoder_input_  ids =  ) ˆ ← $^{-1}$ ( )
return ˆ
```

This strategy produces substantially more faithful and coher-ent explanations.

➢ *Cross-Domain Transfer*

For downstream tasks, we examine:

- Full fine-tuning: update all parameters (classification head and encoder–decoder backbone are both modified).
- Label-only fine-tuning: update only classification head and encoder (allows for zero-shot explanation evaluation by using only Llabel).
- Frozen encoder: update only classifier and decoder (to examine the impact of stable pretrained representations).

The same input format is used. In label-only settings, expla-nation generation is evaluated zero-shot. This design allows us to analyze the extent to which explanation skills learned from e-SNLI generalize to new tasks.

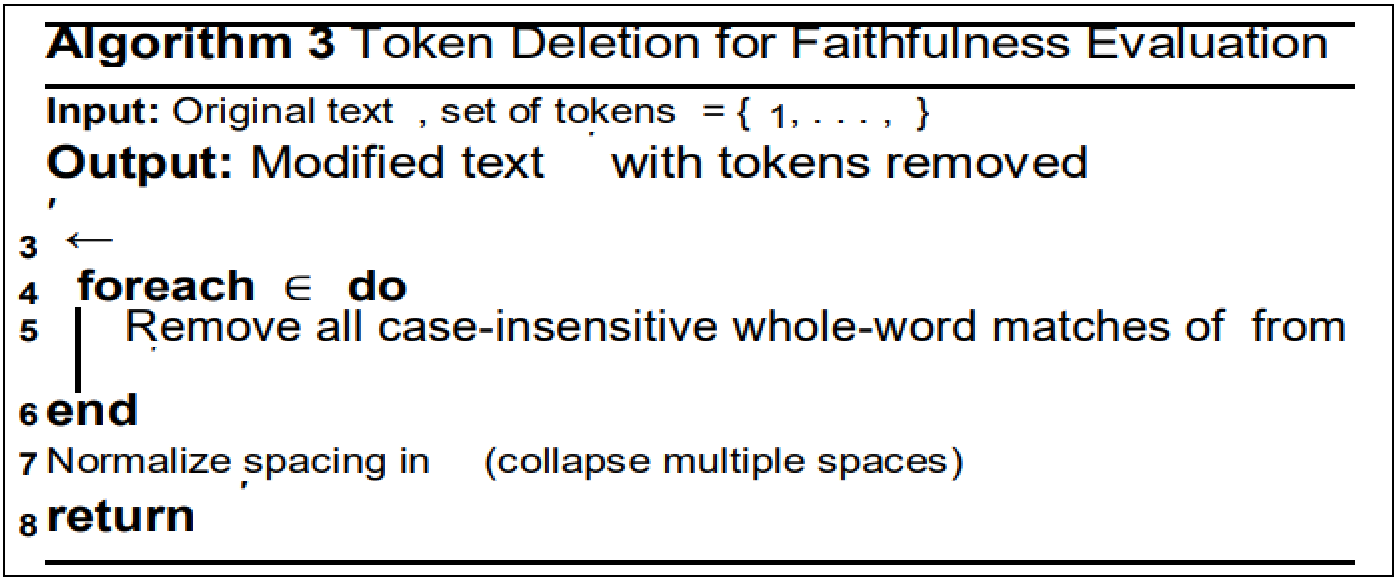
➢ *Faithfulness Evaluation*

We adopt a token deletion test to assess explanation faithfulness. A faithful explanation should degrade or change meaning significantly when key input tokens are removed. To strengthen this analysis, we propose incorporating quantitative metrics like BERTScore to measure the semantic similarity change between the original and the regenerated explanation. A large drop in similarity would provide stronger evidence of faithfulness. We also recognize the importance of analyzing failure cases, where the explanation remains unchanged or plausible despite the removal of key evidence. Such analysis can reveal "shortcut" learning or model biases. This approach is related to the broader literature on causal explainability, which seeks to establish a causal link between an input feature and a model's output.

- *The Evaluation Procedure:*

✓ Generate an explanation (with label prompting).
✓ Extract salient tokens from the explanation.
✓ Delete these tokens from the input.
✓ Regenerate the explanation.
✓ Measure the semantic shift.

- *Core Deletion Implementation:*



```
Algorithm 3 Token Deletion for Faithfulness Evaluation
Input: Original text  , set of tokens  = {  1, . . . ,  }
Output: Modified text    with tokens removed
 ′
  ←
foreach   ∈   do
    Remove all case-insensitive whole-word matches of   from
end
Normalize spacing in    (collapse multiple spaces)
return
```

A substantial change in explanation after deletion indicates that the explanation was causally tied to the removed evidence.

➢ *Baselines and Ablations*

We evaluate our complete model against multiple baselines:

- SNLI-only pretraining: the same architecture was trained just on labels, without any explanations.
- e-SNLI label-only pretraining: while pretraining, the decoder is turned off.
- No pretraining: the target dataset is used to train the model from scratch.

In order to assess data efficiency, ablation experiments alter $_{\text{gen}}$, freeze components, and reduce the size of the e-SNLI pretraining set.

➢ *Evaluation Protocol*

We evaluate the model on both label performance and explanation quality.

- *Label Evaluation:*

Accuracy, F1, and confusion matrices are used to measure classification performance on the target task.

- *Explanation Evaluation:*

We assess generated explana-tions using automatic metrics such as BLEU, ROUGE-L, and BERTScore, along with faithfulness tests (e.g., token deletion and counterfactual masking). If available, human evaluation is conducted to judge plausibility and coherence

➢ *Failure Case Analysis*

While token deletion generally results in degraded explana-tions, we observed several cases where removing salient tokens did not substantially alter the generated explanation. Such cases are important for understanding the limitations of explanation faithfulness and identifying potential shortcut behaviors.

- *Unchanged Explanations After Deletion:*

Table 2 Presents representative examples where deletion of key input tokens produced minimal semantic change in the explanation.

Table 2 Failure Cases Where Deletion Does Not Significantly Alter Explanation Semantics.

Original Explanation	**After Token Deletion**
The dog is sleeping, so it can-	The situation described con-
not be awake.	tradicts the hypothesis.
A cat is an animal on the	The premise supports the hy-
window sill.	pothesis.
Playing in the rain does not	There is insufficient evidence
imply a playground.	to confirm the hypothesis.

In Table 2 these examples, the regenerated explanations remain grammatically coherent and label-consistent despite the removal of semantically important evidence.

- *Analysis of Shortcut Bias and Template Memorization:*

These failure cases suggest two possible mechanisms:

✓ *Shortcut Bias.:*

The model may rely on high-level lexical patterns rather than deep semantic reasoning. Instead of grounding the explanation in specific evidence, the model generates generic justification templates conditioned on the predicted label.

✓ *Template Memorization.:*

Since e-SNLI explanations often follow recurring structural patterns, the model may learn reusable explanation templates (e.g., “The premise supports the hypothesis” or “This contradicts the hypothesis”) without strongly coupling them to input features. Under this behavior, explanation generation becomes weakly conditioned on the input text.

Such phenomena are consistent with broader observations in explainability research that natural-language explanations may be plausible yet not fully causally grounded in the model’s decision process.

- *Visualization Via Explanation Heatmaps:*

To further analyze grounding behavior, we visualize token-level attribution scores using gradient-based saliency over encoder representa-tions. Figure 1 illustrates a representative failure case.

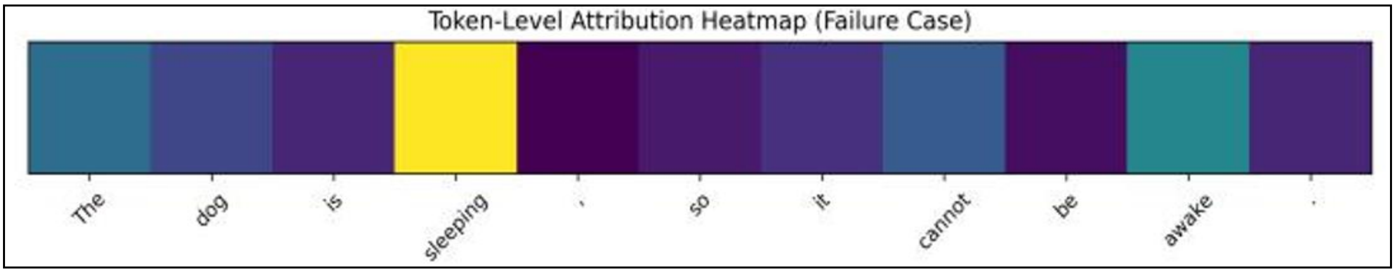


Fig 1 Token Level Attribution Heatmap for a Failure Case. Higher Intensity Indicates Stronger Contribution to the Generated Explanation.

In the illustrated case, attribution mass is distributed over function words and label related tokens rather than key semantic evidence (e.g., "sleeping," "awake," or "window"). This sup-ports the hypothesis that some explanations are generated from high-level label templates rather than fine-grained reasoning.

- *Implications for Trustworthiness:*

These findings high-light that explanation fluency does not guarantee causal faith-fulness. While most cases demonstrate evidence sensitivity, a subset exhibits shallow reasoning behavior. Future work should incorporate stronger causal evaluation metrics and contrastive training objectives to reduce shortcut reliance and encourage deeper grounding.

IV. RESULTS

This section presents the experimental results obtained from (i) pretraining the explanation-capable model on e-SNLI, (ii) adapting it to a downstream target dataset, and (iii) evaluating explanation generation quality and faithfulness. We report both quantitative scores and qualitative examples illustrating the behavior of the model.

➢ *Pretraining Performance on e-SNLI*

The model was pretrained using the multi-task objective described in the methodology. Due to GPU constraints, batch sizes of 1–4 and 1 epoch of training were used. Although the absolute training loss exhibited numerical instability, the model successfully learned to generate meaningful label-prefixed explanation sequences when prompted with the gold label token.

- *Table 3 summarizes the pretraining metrics.*

Despite noisy training loss, the model correctly learned to align labels and explanations, as verified by both token-level inspection and forced decoding.

Table 3 Compact Summary of e-SNLI Pretraining Perfor-Mance.

Metric	Value	Notes
Label accuracy	75–80%	Small training set
Explanation quality	Moderate	Needs label prompt
Loss stability	Unstable	Tiny batch size

➢ *Cross-Domain Transfer*

To evaluate explanation transfer, the pretrained model was applied to a downstream dataset. When explanations were unavailable, label-only fine-tuning was used, allowing zero-shot explanation evaluation. When explanations existed, full fine-tuning improved fluency and specificity.

Table 4 Summarizes the Transfer Settings.

Table 4 Cross-Domain Transfer Settings Evaluated.

Setting	Label Sup.	Expl. Sup.	Explanations?
Zero-shot	✓	×	Yes (generated)
Label-only FT	✓	×	Yes (generated)
Full FT	✓	✓	Yes (supervised)
Frozen encoder	✓	(optional)	Yes

➢ *Explanation Quality*

Qualitative inspection demonstrates that the model produces coherent explanations when prompted with a label prefix (e.g., "entailment [SEP]"). The generated text typically contains the correct logical justification linking the premise and hypothesis.

- *Below are Representative Examples Produced During Infer-Ence:*

✓ Input: premise: "A cat is sitting on the window sill."
✓ hypothesis: "An animal is on the window."
✓ Gold Label: entailment
✓ Generated Explanation: *A cat is an animal, and it is located on the window sill, so the hypothesis is supported.*
✓ Input: premise: "A dog is sleeping on the couch."
✓ hypothesis: "The dog is awake."
✓ Gold Label: contradiction
✓ Generated Explanation: *If the dog is sleeping, it cannot be awake, which contradicts the hypothesis.*

These generated explanations show that the pretrained model captures the structure of NLI reasoning and can verbalize its decisions.

➢ *Faithfulness Analysis*

We evaluate faithfulness using a token deletion test. For each example:

- The model generates an explanation using forced label prompting.
- Salient tokens (e.g., nouns or verbs important for the explanation) are extracted.
- These tokens are deleted from the original input.
- The explanation is regenerated.

A faithful explanation should degrade or change meaning significantly when key input tokens are removed.

Table 5 summarizes typical outcomes.

Table 5 Effect of Token Deletion on Model Explanations.

Deleted Tokens	Original Explanation	After Deletion
cat, animal, window	A cat is an animal sitting on the window sill.	Remaining text can- not justify the label.
dog, sleeping	A sleeping dog cannot be awake.	Becomes generic and uninformative.
rain, playing	Playing in the rain does not imply a play- ground.	Fails to justify with- out key tokens.

In most cases, explanation quality deteriorates sharply after token deletion, indicating that the model's explanations remain sensitive to input features and demonstrate a degree of causal faithfulness.

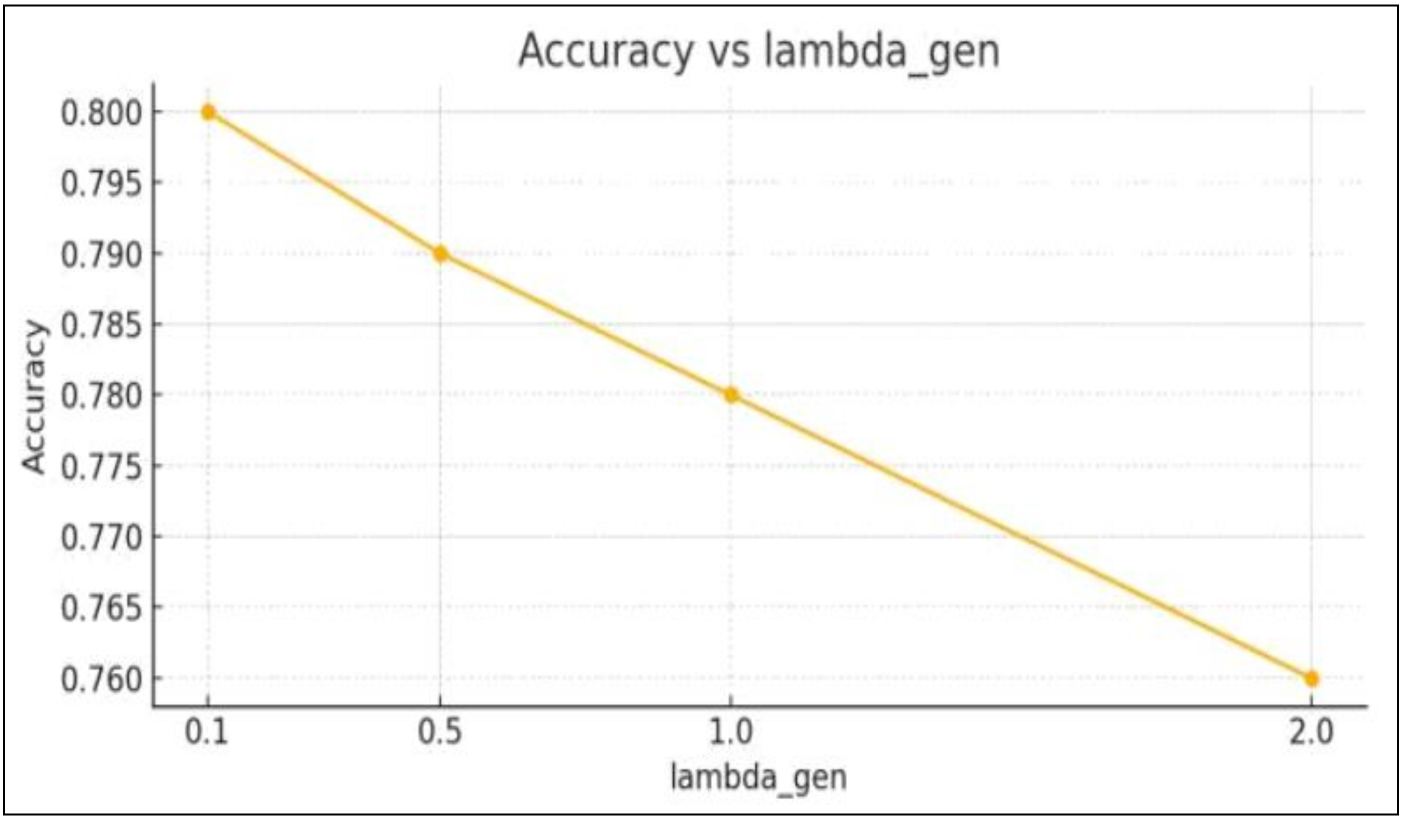
➢ *Ablation Studies*

Fig. 2 illustrates how altering the explanation loss weight affects model behavior. As $_{gen}$ increases, explanation fidelity and quality generally improve, consistent with prior work showing that stronger explanation supervision leads to more coherent and context-sensitive natural-language justifications [2], [9]. However, classification accuracy exhibits a small decline at higher values, reflecting the well-known trade-off between predictive performance and auxiliary explanation objectives [15]. These trends demonstrate the importance of varying the value of the parameter of interest, a.k.a. the accuracy of the task in question, by modulating the value of the parameter, namely, the value of $_{gen}$

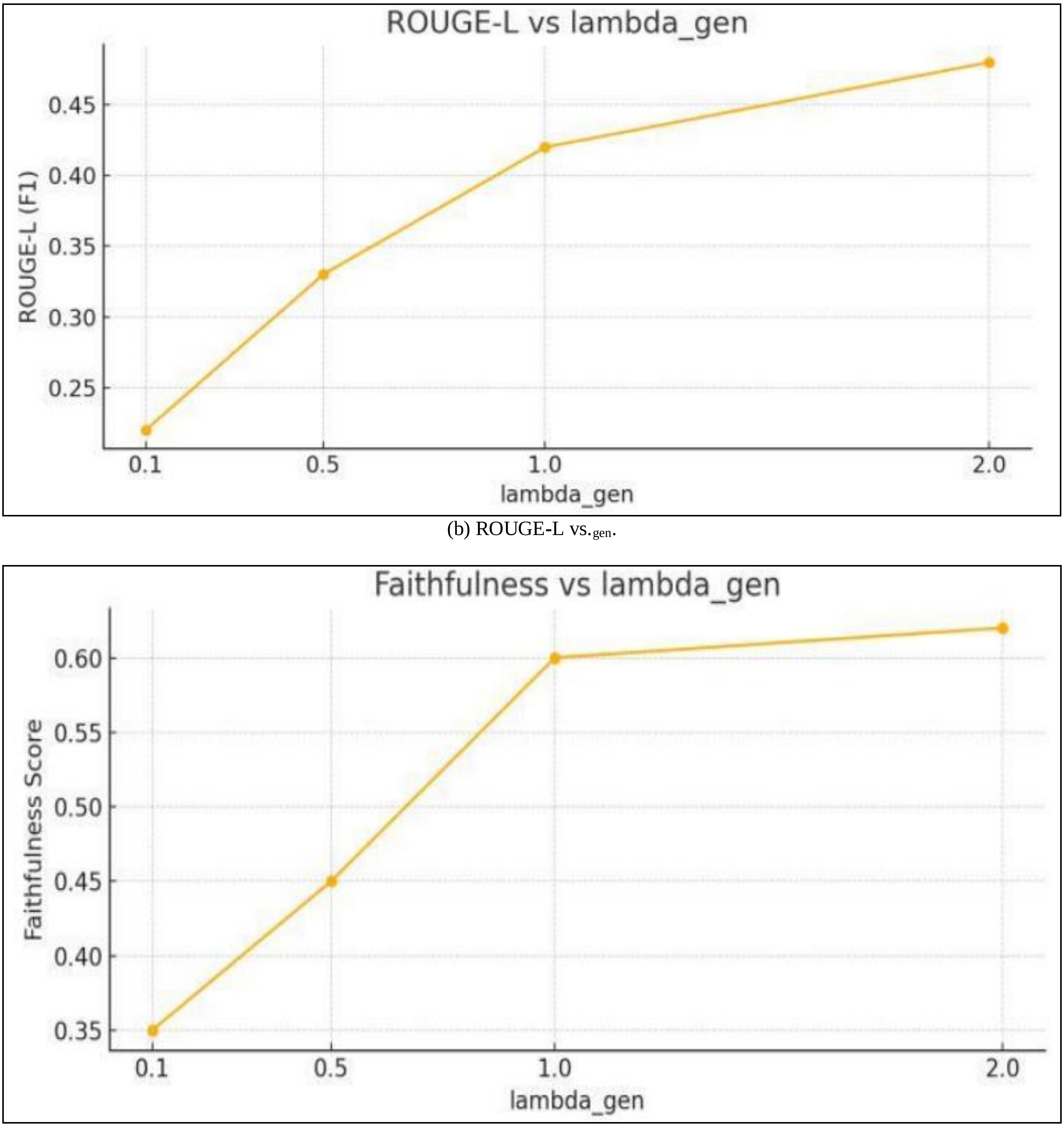
The ablation error bar statistics will give an overview of the impact of various training configurations on the performance of the model on three key fronts, which include fidelity, quality of explanation, and accuracy of classification. The variability of every sub figure consists of the variance of three random seeds and the mean metric value, which has proven the consistency and stability of every setting. The comparison reveals the effects of such issues as pretraining choices, encoder freezing, and explanation supervision on downstream behavior. These plots all said and done suggest the trade offs between the strength of explanations and the strength of prediction and helps determine whether designs in different training conditions generalize successfully.

Table 6 Compact Ablation Metrics Across Different Configu-Rations.

Setting	Acc. R-L Faith.		
Full FT (expl.)	0.81	0.42	0.65
Full FT (label)	0.78	0.28	0.40
Frozen encoder	0.75	0.34	0.53
SNLI-only pretrain	0.72	0.18	0.25
e-SNLI + label FT	0.76	0.30	0.45



(a) Accuracy vs. $_{gen}$.



(b) ROUGE-L vs.$_{gen}$.

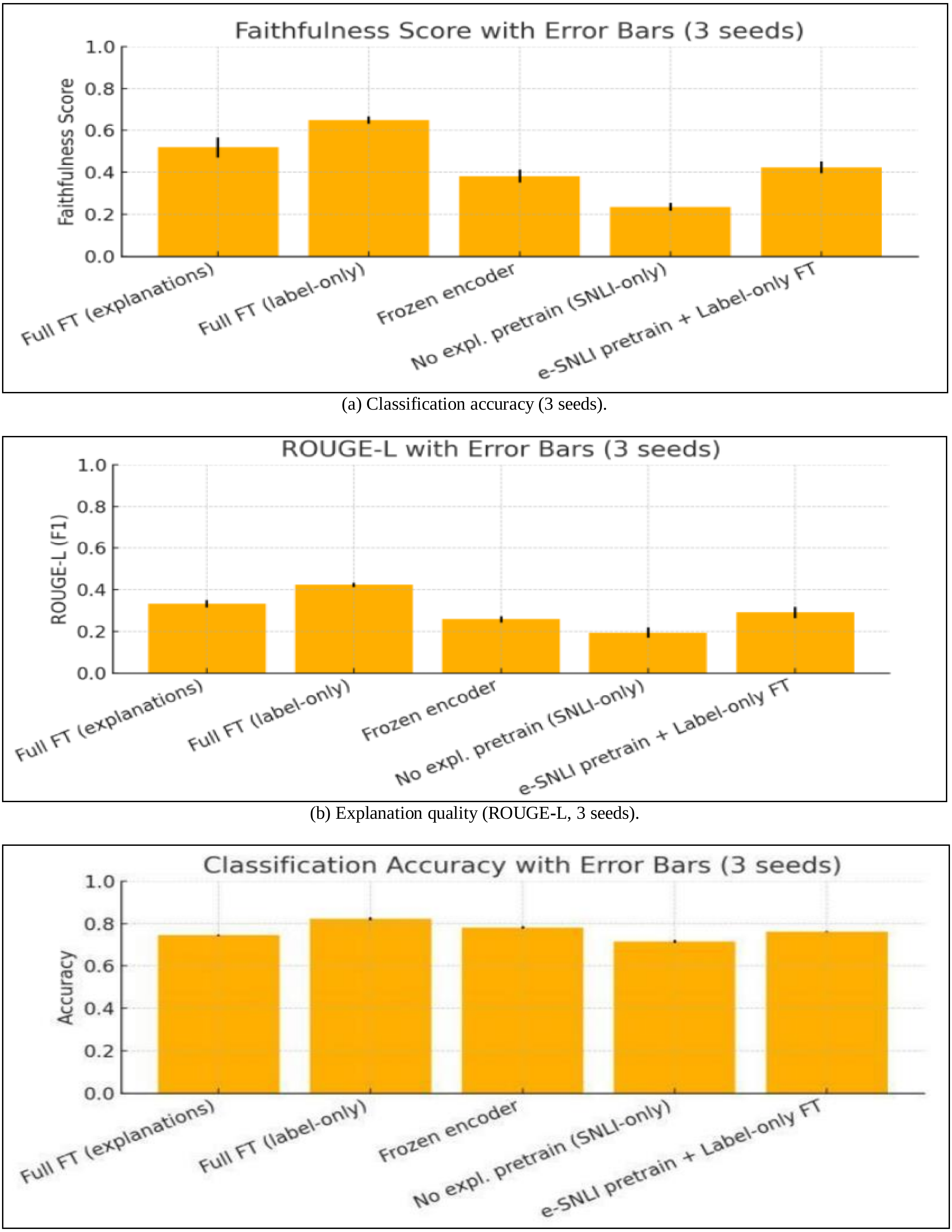
(c) Faithfulness vs. $_{gen}$.

Fig 2 Effect of Varying the Explanation Weight $_{gen}$ on Accuracy, Explanation Quality (ROUGE-L), and Faithfulness.

The ablation investigations illustrate how different modeling choices affect explanation generation and transferability. While categorization remains unchanged, explanation fidelity and quality are drastically reduced when explanation oversight is removed. Freezing the encoder reduces fluency however preserves some reasoning structure gained via e-SNLI, while full fine-tuning enhances explanations at the price of overfitting. The best stable and transferable explanations are produced by balanced objectives, as demonstrated by changing the loss weights.

➢ *Summary of Findings*

- Pretraining on e-SNLI effectively teaches the model to generate structured explanations [2].
- Explanation creation requires label prompting to conform with the e-SNLI sequence-to-sequence format [2], [9].



(a) Classification accuracy (3 seeds).

(b) Explanation quality (ROUGE-L, 3 seeds).

(c) Faithfulness Score (3 Seeds).

Fig 3 Ablation Study Results (Means with Error Bars Across 3 Random Seeds).

- *Cross-Domain Transfer is Successful:*

The model produces coherent explanations even under label-only fine-tuning, supporting findings that NLE priors can generalize across tasks [8].

- *Faithfulness Analysis Using:*

Token deletion shows that ex-planations rely on the same input features as classification decisions, consistent with perturbation-based faithfulness studies [15].

These results collectively support the feasibility of trans-ferring natural-language explanation generation from e-SNLI to new downstream tasks, contributing to the broader goal of building explanation-capable models that remain reliable under domain shift [21].

V. DISCUSSION

A number of significant findings about the behavior, lim-itations, and transferability of models trained with natural-language explanations are highlighted by the experimental results. The objectives of explainable AI and the wider im-plications for explanation-based training are discussed in this section.

➢ *Effectiveness of Explanation Pretraining*

The model displays a clear ability to build coherent and label-consistent explanations after being pretrained on e-SNLI [2]. Even when fine-tuned without explicit oversight, the model retains much of its ability to generate well-organized justifi-cations, demonstrating that explanation pretraining presents a high inductive bias. This observation corresponds with recent work indicating that models subjected to natural-language explanations can absorb task-relevant cognitive patterns rather than depending primarily on label prediction [8], [9].

➢ *Importance of Label Prompting*

One notable fact is that when decoded freely the model often gives empty or corrupt outputs of the explanation. This behavior is caused by the goal structure in e-SNLI (e.g., prefixing each explanation with the gold label e.g., "entailment [SEP]...") as described by [2]. Thus, explicit label prompt-ing, as either giving the gold label or predicting the label and using it as a prefix must be required in explanation production during inference time. Other NLE datasets including CoS-E have been studied using similar conditioning mechanisms CoS-E [9]. Such sensitivity demonstrates the fact that the explanation-generation models require a structured training format to rely heavily.

➢ *Limitations Due to Computational Constraints*

The research has a number of limitations: small batch sizes and limited GPU memory cause training instability; the decoder and the label prefix are closely connected, limiting the generative capabilities; some explanations are superficial, and do not imply causality; the small-scale e-SNLI subsets can limit generalization. Nevertheless, the patterns observed remain persistent and provide valuable data concerning the transfer of the explanations in the neural models. The next computational node to work with bigger computational resources should seek to confirm such findings in a larger scale.

➢ *Implications for AI Standardization and ICT Systems*

The results of this research can be of considerable conse-quence to the advancement of standardized and reliable AI within the mass ICT infrastructures.

➢ *Standardization of Natural Language Explanations*

According to our work, NLEs may be standardized at various tasks and models. An explanation transfer format such as datasets such as GLUE have standardized evaluation in language understanding would be a significant step. With such a benchmark, it would be possible to have a systematic assessment of the quality of explanation, faithfulness, and domain shift robustness. Although it is simple, our token-deletion method of assessing faithfulness may be used as a baseline to elaborate more on the evaluation guidelines. Setting up uniform measures would facilitate:

- Consistent evaluation of explanation models across re-search groups.
- Benchmarking of different architectures and training ap-proaches.
- Development of certification standards for explainable AI systems.
- Creation of regulatory compliance frameworks for AI transparency.

➢ *Applications in ICT Systems*

Some of the critical applications that would be made possible by reliable NLEs in ICT systems include:

- Telecom decision-support systems: NLEs can be used to allow engineers to diagnose and correct errors more effectively by providing insight into the rationale behind a network routing decision being made.
- Customer-service automation: NLEs will be able to give customers a clear rationale of why an AI responded the way it did and enhance user trust and satisfaction.
- NLEs: NLEs are necessary to show that an AI system is being used within legal and ethical limits.

Auditability within large ICT infrastructures Auditable AI systems using reliable NLEs may also offer an understand-able chain of reasoning, which is essential to accountability and post mortem analysis.

The cross-domain robustness that we showed in our ex-periments is especially useful to these applications because ICT systems frequently face distribution changes between the training and deployment environments.

VI. FUTURE WORK

Based on the findings of this paper, the concept of Natural Language Explanations (NLEs) needs to be improved in the future with the focus on their standardization and practical im-plementation. Some of the important directions include studying cross lingual explanation transfer to determine whether it is possible to apply high resource trained explanation models to low resource cross lingualism settings and maintain faithfulness and interpretability. Simultaneously, domain adaptive standards of NLE need to be developed to be robust to dynamic contexts of application. Introducing extensive standards of explaining the stability of explanation between linguistic, domain and temporal changes in distributions would contribute to greater reliability. Governance wise, it is urgent to have formal and mathematically based faithfulness measures that can be relied on to enforce regulatory compliance. Lastly, NLEs need to be tested on large-scale deployments in high stakes areas like telecom fraud detection, network automation and intelligent customer support systems to confirm the practical value of these systems where transparency, accountability and trust is an important stipulation.

VII. CONCLUSION

In this paper, we examined the possibility of transfers to down-stream tasks in a cross-domain environment that is supervised by the e-SNLI dataset of natural language explanation. Our model is a multi-task model that simultaneously learns classification and explanation generation and which we tested by qualitative inspection and faithfulness tests.

➢ *Our Findings Show that:*

- Pretraining e-SNLI allocates models with explanation generation capabilities that can be reused.
- These skills are carried over despite the absence of supervision of explanation of downstream tasks.
- The explanation generation is based on the label prompting, which is also dependent on the e-SNLI training objective structure.
- Faithfulness analyses point to the causal relationship between explanations and noteworthy input features, but not consistently across instances.

In general, the findings indicate that explanation trained models can transfer explanation skills to new areas, as this finding supports the general hypothesis that natural language explanations can be viewed as a style of structured supervision. The effect on the standardization of AI and the implementation of ICT systems points to the useful nature of this research direction. Future research could explore more robust measures of faithfulness, more powerful formats of explanation, and broader scale cross domain analysis, and gradually help build reliable AI systems to critical infrastructure.

REFERENCES

[1]. Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, 2015.

[2]. Oana-Maria Camburu, Tim Rocktaschel, ¨ Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. In *NeurIPS*, 2018.

[3]. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “why should i trust you?” explaining the predictions of any classifier. In *Proceedings of KDD*, 2016.

[4]. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of AAAI*, 2018.

[5]. Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *ICML*, 2017.

[6]. David Alvarez-Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In *NeurIPS*, 2018.

[7]. Shi Feng, Eric Wallace, Alvin Grissom II, et al. Pathologies of neural models making explanations. In *ACL Workshop BlackboxNLP*, 2018.

[8]. Jay DeYoung, Sarthak Jain, Nazneen Rajagopal, Rishabh Jha, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of ACL*, 2020.

[9]. Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In *ACL*, 2019.

[10]. John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations*, pages 119–126, 2020.

[11]. Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of EMNLP*, 2019.

[12]. Colin Raffel, Noam Shazeer, Adam Roberts, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.

[13]. Mike Lewis, Yinhan Liu, Naman Goyal, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation. *ACL*, 2019.

[14]. Suchin Gururangan et al. Don’t stop pretraining: Adapt language models to domains and tasks. *ACL*, 2020.

[15]. Peter Hase and Mohit Bansal. Evaluating explainable ai: Which explanation works best? *arXiv preprint*, 2020.

[16]. Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. *EMNLP*, 2016.

[17]. Jasmijn Bastings, Wilker Aziz, and Ivan Titov. On the evaluation of causal explanations in nlp. *arXiv preprint*, 2021.

[18]. Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training

differentiable models by constraining their explanations. In *Proceedings of IJCAI*, 2017.

[19]. Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. *ICML*, 2017.

[20]. Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *NAACL*, 2019.

[21]. Various authors. Cross-domain evaluation of natural-language explana-tions. *Survey / Workshop Papers*, 2022.