ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF COMPUTER SCIENCE
AND ENGINEERING

ARTIFICIAL INTELLIGENCE

MASTER THESIS

in

Natural Language Processing

DEVELOPING AND COMPARING MACHINE REASONING MODELS TO HUMANS IN NLP TASKS

CANDIDATE

Mohammad Reza Ghasemi Madani

SUPERVISOR

Prof. Paolo Torroni

CO-SUPERVISOR

Prof. Pasquale Minervini

Academic year 2022/2023

Session 5th

To my parents, *Mehrdad* and *Mansooreh*, and my brother *Mahdi*, for their enormous love and support.

تقدیم به پدر و مادر مهربانم و برادر عزیزم.

Acknowledgements

I sincerely appreciate my supervisors, Dr. Paolo Torroni and Dr. Pasquale Minervini, for their continuous guidance and unwavering support. Their invaluable advice and insightful perspectives have played a pivotal role in shaping me as a future researcher. My heartfelt thanks go to the University of Bologna, which has become my second home. I am grateful for the wonderful friendships forged at the university and for the exceptional support and care I received from the entire university staff. I also express gratitude to the University of Edinburgh for its encouraging faculty and the numerous inspiring research conversations that have enriched my academic journey. Most importantly, I thank my family, whose unwavering belief, encouragement, and unconditional support have been my greatest assets. Their love has been the cornerstone of my success, and I am truly grateful for their enduring presence in my life.

Abstract

Neural Language Models represent a category of computational systems designed to learn task performance directly from raw textual inputs. Their increasing popularity stems from their versatility and remarkable success across diverse domains, such as their transformative impact on machine translation, surpassing traditional machine learning methods. Despite these achievements, a crucial aspect remains unaddressed: the interpretability of the model's decision-making process. Rationale extraction endeavors to furnish explanations that are both faithful (reflective of the model's behavior) and plausible (convincing to humans) by highlighting influential inputs without compromising task model performance. Prior research has primarily focused on optimizing plausibility using human highlights when training rationale extractors, while jointly training the task model to optimize for predictive accuracy and faithfulness. In this thesis, we delve into the significance of explanations, the associated challenges, and the research landscape in this field. We also introduce REFER, a framework that incorporates a differentiable rationale extractor that facilitates back-propagation through the rationale extraction process. Through joint training of the task model and rationale extractor with human highlights, our analysis demonstrates that REFER achieves significantly improved results in terms of faithfulness, plausibility, and downstream task accuracy on both in-distribution and out-of-distribution data compared to previous baselines.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Neural Language Models (NLMs) are a class of computing systems designed to learn how to perform tasks directly from raw textual inputs. These systems have been gaining in popularity over recent years due to their versatility and revolutionary success in various domains. For example, NLMs have revolutionized machine translation [96], with significant improvements in performance over traditional machine learning methods. Learning without hard-coding task-specific knowledge is a breakthrough in artificial intelligence. However, there is a key missing and that is the interpretation behind a model decision-making process. In the rest of this chapter, we investigate the importance of explanation, challenges we may face, and research outline in this field of study.

1.1 The Importance of Explanations for NLMs

It has been shown that a key factor in the success of NLMs is their capability which allows them to learn features at various levels of abstraction between the raw data and the prediction. However, this comes at the cost of explainability[1], since providing a human-intelligible interpretation of an intricate composition of a large number of non-linear functions is a difficult open

[1] In the following sections, we use explainability and interpretability interchangeably.

question. Thus, in safety-critical applications, such as health diagnosis, credit allowance, or criminal justice, one may still prefer to employ less accurate but human-interpretable models, such as linear regression and decision trees.

The doubts around the decision-making processes of NLMs are justified, as it has been shown that seemingly very accurate systems can easily rely on spurious correlations (or statistical bias) in datasets to provide correct answers [34, 35, 16, 65]. Another source of mistrust in black-box systems comes from the potential subjective bias that these systems might develop, such as racism, sexism, or other kinds of discrimination and subjectivity [9]. For instance, [26] cast doubt on the fairness of the widely used commercial risk assessment software COMPAS for recidivism prediction. Such biases may be learned either from under-representation or irrelevant statistical correlations in the datasets that we train and test our models on. Therefore, an increasing need for explainable models is arising.

Explainability refers to the ability to explain or present the behavior of models in human-understandable terms [25, 28]. Improving the explainability of NLMs is crucial for two key reasons. First, for general end users, explainability builds appropriate trust by elucidating the reasoning mechanism behind model predictions in an understandable manner, without requiring technical expertise. With that, end users are able to understand the capabilities, limitations, and potential flaws of NLMs. Second, for researchers and developers, explaining model behaviors provides insight to identify unintended biases, risks, and areas for performance improvements. In other words, explainability acts as a debugging aid to quickly advance model performance on downstream tasks [92, 2, 103]. It facilitates the ability to track model capabilities over time, make comparisons between different models, and develop reliable, ethical, and safe models for real-world deployment.

In the past few years, several works have been proposed for improving the interpretability of NLMs. In this thesis, we focus on local interpretable methods proposed for Natural Language Processing (NLP) tasks. In chapter 2, we define local methods as those that provide explanations only for specific decisions made by the model - that is, methods that provide explanations for single instances, rather than aiming to provide general descriptions of the model's decision-making process. We explore several recent local interpretation techniques in NLP, which aim to support normal users with no expertise.

1.2 Research Challenges

In this section, we explore key research challenges that warrant further investigation from both the NLP and the explainable AI communities.

1.2.1 Explanation without Ground Truths

Ground truth explanations for NLMs are usually inaccessible. For example, there are currently no benchmark datasets to evaluate the global explanation of individual components captured by NLMs. This presents two main challenges. First, it is difficult to design explanation algorithms that accurately reflect an LM's decision-making process. Second, the lack of ground truth makes evaluating explanation faithfulness and fidelity problematic. It is also challenging to select a suitable explanation among various methods in the absence of ground truth guidance. Potential solutions include involving human evaluations and creating synthetic explanatory datasets.

1.2.2 Sources of Emergent Abilities

NLMs exhibit surprising new capabilities as the model scale and training data increase, even without being explicitly trained to perform these tasks. Elucidating the origins of these emergent abilities remains an open research challenge, especially for proprietary models like ChatGPT and Claude whose architectures and training data are unpublished. Even open-source LMs like

LLaMA currently have limited interpretability into the source of their emergent skills. This can be investigated from both a model and a data perspective.

Model Perspective It is crucial to further investigate the Transformer-based model to shed light on the inner workings of NLMs. Key open questions include: 1) What specific model architectures give rise to the impressive emergent abilities of NLMs? 2) What is the minimum model complexity and scale needed to achieve strong performance across diverse language tasks? Continuing to rigorously analyze and experiment with foundation models remains imperative as NLMs continue to rapidly increase in scale. Advancing knowledge in these areas will enable more controllable and reliable NLMs. This can provide hints as to whether there will be new emergent abilities in the near future.

Data Perspective In addition to the model architecture, training data is another important perspective for understanding the emergent abilities of NLMs. Some representative research questions include: 1) Which specific subsets of the massive training data are responsible for particular model predictions, and is it possible to locate these examples? 2) Are emergent abilities the result of model training or an artifact of data contamination issues [7]? 3) Are training data quality or quantity more important for effective pre-training and fine-tuning of NLMs? Understanding the interplay between training data characteristics and the resulting behavior of the model will provide key insights into the source of emergent abilities in large language models.

1.2.3 Shortcut Learning of NLMs

Recent explainability research indicates that language models often take shortcuts when making predictions. For the downstream fine-tuning paradigm, studies show that language models leverage various dataset artifacts and biases for natural language inference tasks, such as lexical bias, overlap bias, position bias, and style bias [27]. This significantly impacts out-of-distribution generalization performance. For the prompting paradigm, a recent study analyzes how language models use longer contexts [57]. The results show that performance was highest when relevant information was at the beginning or end of the context, and worsened when models had to access relevant information in the middle of long contexts. These analyses demonstrate that both paradigms tend to exploit shortcuts in certain scenarios, highlighting the need for more research to address this problem and improve generalization capabilities.

1.2.4 Safety and Ethics

The lack of interpretability in NLMs poses significant ethical risks as they become more capable. Without explainability, it becomes challenging to analyze or constrain potential harms from issues such as misinformation, bias, and social manipulation. Explainable AI techniques are vital to audit these powerful models and ensure alignment with human values. For example, tools to trace training data attribution or visualize attention patterns can reveal embedded biases, such as gender stereotypes [53]. Additionally, probing classifiers can identify if problematic associations are encoded within the model's learned representations. Researchers, companies, and governments deploying NLMs have an ethical responsibility to prioritize explainable AI. Initiatives such as rigorous model audits, external oversight committees, and transparency regulations can help mitigate risks as NLMs become more prevalent. For example, as alignment systems continue to grow in scale, human feedback is becoming powerless at governing them, posing tremendous challenges to the safety of these systems. Leveraging explainability tools as part of audit processes to supplement human feedback could be a productive approach. Advancing interpretability techniques must remain a priority alongside expanding model scale and performance to ensure the safe and ethical development of increasingly capable NLMs.

1.3 Research Agenda and Outline

This thesis explores various domains comprehensively, with the goal of contributing new insights and advancements to the existing body of knowledge. Following the findings and gaps identified in the preceding section, the subsequent research endeavors are designed to address key research questions and push the boundaries of knowledge.

1.3.1 Effectiveness of Human Supervision

Nowadays, machine learning systems can learn to capture spurious correlations in the data for solving any given task, and often struggle in more challenging cases [65]. When models are allowed to make predictions without considering rationales-related criteria—faithfulness and plausibility—the rationales extracted by the model can be incomprehensible and lack meaningful interpretations [99] and it becomes evident that neural models tend to rely on dataset-specific patterns. As a result, although the model can produce accurate results on in-distribution data, the rationales extracted by the model can be incomprehensible and lack meaningful interpretations [99]. For instance, baseline models often pay attention to stop-words and full-stops, which can be considered as instances of null attention [99]. Conversely, in supervised models, the words that receive the most attention are typically nouns such as ”man”, ”woman” or ”people”, as they frequently serve as the subjects of sentences [35].

In certain contexts, faithful explanations are crucial – for example, they can be used to determine whether a model relies on protected attributes, such as gender or religious group [77]. [65] propose the hypothesis that neural natural language inference (NLI) models might rely on three fallible syntactic heuristics: (i) lexical overlap, (ii) subsequences, and (iii) constituents. These heuristics are used by the models to make predictions in NLI tasks. To evaluate whether the models have indeed adopted these heuristics, we use Heuristic

Analysis for NLI Systems (HANS, [65]), which includes a variety of examples where such heuristics fail, providing a means to assess a model's reliance on these heuristics. table 1.1 shows instances of these heuristics in the HANS dataset.

Table 1.1: The heuristics targeted by the HANS dataset, along with examples of incorrect entailment predictions that these heuristics would lead to.

Heuristic	Definition	Example
Lexical overlap	The premise entails all hypotheses constructed from its own words.	The **judges admired** the **doctors**. $\xrightarrow{\text{Wrong}}$ The **doctors admired** the **judges** .
Subsequence	The premise entails all of its contiguous subsequences.	**The lawyers believed the bankers** resigned. $\xrightarrow{\text{Wrong}}$ The lawyers believed the bankers.
Constituent	The premise entails all complete subtrees in its parse tree.	Probably **the tourists waited**. $\xrightarrow{\text{Wrong}}$ The tourists waited.

Faithfulness refers to the degree to which an explanation provided by a model accurately reflects the information utilized by the model to make a decision [37]. Without understanding the factors and information that influence the predictions of the model, it becomes difficult to trust or explain its outputs. In certain contexts, faithful explanations are crucial – for example, they can be used to determine whether a model is relying on protected attributes, such as gender or religious group [77]. Ensuring faithfulness in model explanations is therefore imperative to address issues of transparency, fairness, and accountability in the deployment of AI systems.

1.3.2 Imitation of Human Reasoning

Human rationales are often derived from their extensive background knowledge and understanding of various concepts. While language models (LMs) possess some degree of this knowledge, they face challenges in balancing between optimizing for task performance and meeting the criteria for extractive explanations. Therefore, balancing plausibility, faithfulness, and task accuracy presents a challenging task. A model can reflect its inner process to make a prediction (faithful), but it may not make sense for humans (implausible). On the other hand, a model that returns convincing rationales (plausible) without using them during decision-making is not very useful (unfaithful). The ideal explanation regularization model would achieve human-level performance on the task output while providing faithful rationales for the task prediction in a manner that is convincing to humans or plausible. However, it is often necessary to reach a trade-off between plausibility, faithfulness, and predictive accuracy. In some cases, optimizing for plausibility may require sacrificing faithfulness or task performance, and vice versa.

Optimizing for plausibility in these models necessitates continuous human-in-the-loop feedback. Acquiring such feedback is often impractical, leading many researchers to resort to human-annotated gold rationales as a more cost-effective form of plausibility supervision [24, 71]. [36] introduce Saliency Guided Training (SGT) which regularizes a task model to produce faithful Attribution Algorithm (AA)-based rationales. Still, AAs can be a bottleneck for plausibility, as producing human-like rationales is a complex objective requiring high-capacity rationale extractors [71, 24]. A more detailed discussion of these baseline models is provided in chapter 5.

1.3.3 Highlights Efficiency

Humans can efficiently learn new tasks with only a few examples by leveraging their prior knowledge. Recent approaches for rationalizing rely on a large number of labeled training examples, including task labels and annotated rationales for each instance. Obtaining such extensive annotations is often infeasible for many tasks. Additionally, fine-tuning LMs, which typically have billions of parameters, can be expensive and prone to overfitting. Ideally, each training instance would be annotated with a gold rationale, allowing direct minimization of the plausibility loss for each instance. To this end, we analyze the impact of having varying amounts of supervision on the extracted rationales during the training process. Given the high cost of human annotations, a more practical approach involves incorporating a limited amount of human supervision. We investigate the characteristics of effective rationales and demonstrate that making the neural model aware of its rationalized predictions can significantly enhance its performance, especially in low-resource scenarios.

1.3.4 Out-of-distribution Generalization

Out-of-distribution (OOD) generalization refers to the ability of a model to accurately handle data samples that deviate from the distribution of its training data. OOD generalization is a critical challenge in NLP tasks and plays a pivotal role in ensuring the reliability and effectiveness of NLP models in real-world applications. Effective OOD generalization in NLP requires models to capture and understand the underlying linguistic properties and generalizable patterns rather than relying on memorization or overfitting specific training instances. However, despite the growing interest in OOD generalization, existing evaluations in the field of explanation robustness have been limited in scope and coverage.

The poor performance of models on OOD datasets can stem from limitations in the model's architecture, insufficient signals in the OOD training set, or a combination of both [65]. An NLI system that correctly labels an example may not do so by understanding the meaning of the sentences but rather by relying on the assumption that any hypothesis with words appearing in the premise is entailed by the premise [21, 70]. [35] raises doubts about whether models trained on the SNLI dataset truly learn language comprehension or primarily rely on spurious correlations, also known as artifacts. For instance, words like ”friends” and ”old” frequently appear in neutral hypotheses, while ”animal” and ”outdoors” are prevalent in entailment hypotheses, and ”nobody” and ”sleeping” are common in contradiction hypotheses. They also show that a premise-agnostic model, which only considers the hypothesis as input and predicts the label, achieves 67% accuracy on the test set [10].

[65] delve into two factors contributing to the adoption of heuristics by SoTA models trained on standard NLI datasets like SNLI [8] or MNLI [101]. Firstly, the MNLI training set predominantly consists of examples that align with these heuristics, offering limited contradicting instances. Even among the 261 contradicting cases in MNLI, their impact on challenging the heuristics is limited. Notably, 133 of these cases involve negation in the premise but not the hypothesis. Rather than employing these cases to override the lexical overlap heuristic, models may learn to associate contradiction labels specifically with negation in the premise, but not in the hypothesis [64]. To analyze this, we evaluate our model on contrast sets [32] as well as unseen data, which are (mostly) label-changing small perturbations on instances to understand the true local boundary of the dataset. Essentially, they help us understand if the rationale extractor has learned any dataset-specific shortcuts.

Considering what we discussed so far, this thesis will focus on answering the following research questions in the following chapters:

- Does training the model on human highlights improve the generalization properties of the model?
- How can we make machines imitate human rationales?
- Does training the model on a small number of human highlights improve its generalization properties?
- Do the learned rationale extractors generalize to OOD data?

Chapter 2

Background on Explanatory Methods for NLMs

In this chapter, we review explanation techniques for NLMs trained with the pre-training and downstream fine-tuning paradigms. First, we introduce approaches to provide local explanations and global explanations. Here, the local explanation aims to provide an understanding of how a language model makes a prediction for a specific input instance, while the global explanation aims to provide a broad understanding of how the NLM works overall. Next, we discuss how explanations can be used to debug and improve models.

2.1 Local Explanation

The first category of explanations refers to explaining the predictions generated by NLM. Consider a scenario where we have a language model and we input a specific text into the model. The model then produces a classification output, such as sentiment classification or a prediction for the next token. In this scenario, the role of explanation is to clarify the process by which the model generated the particular classification or token prediction. Since the goal is to explain how the NLM makes the prediction for a specific input, we call it the local explanation. This category encompasses two main streams of approaches for generating explanations including feature attribution-based explanation (also known as extractive rationales) and natural language explanation.

2.1.1 Feature Attribution-Based Explanation

Feature attribution methods aim to measure the relevance of each input feature (e.g., words, phrases, text spans) to a model's prediction. Given an input text x comprised of n word features $x_1, x_2, ..., x_n$, a finetuned language model f generates an output $f(x)$. Attribution methods assign a relevance score $r(x_i)$ to the input word feature x_i to reflect its contribution to the model prediction $f(x)$. The methods that follow this strategy can be mainly categorized into three types: perturbation-based methods, gradient-based methods, and surrogate models.

Perturbation-Based Explanation Perturbation-based methods work by perturbing input examples such as removing, masking, or altering input features and evaluating model output changes. The most straightforward strategy is leave-one-out, which perturbs inputs by removing features at various levels including embedding vectors, hidden units [52], words [51], tokens and spans [102] to measure feature importance. The basic idea is to remove the minimum set of inputs to change the model prediction. The set of inputs is selected with a variety of metrics such as confidence score or reinforcement learning. However, this removal strategy assumes that input features are independent and ignores correlations among them. Additionally, methods based on the confidence score can fail due to pathological behaviors of overconfident models [31]. For example, models can maintain high-confidence predictions even when the reduced inputs are nonsensical. This overconfidence issue can be mitigated via regularization with regular examples, label smoothing, and finetuning models' confidence [31]. Besides, current perturbation methods tend to generate out-of-distribution data. This can be alleviated by constraining the perturbed data to remain close to the original data distribution [78].

Gradient-Based Explanation Gradient-based attribution techniques determine the importance of each input feature by analyzing the partial derivatives of the output w.r.t. each input dimension. The magnitude of derivatives reflects the sensitivity of the output to changes in the input. The basic formulation of raw gradient methods is described as $s_j = \frac{\partial f(x)}{\partial x_j}$, where $f(x)$ is the prediction function of the network and x_j denotes the input vector. This scheme has also been improved as gradient×input [42] and has been used in various explanation tasks, such as computing the token-level attribution score [68]. However, vanilla gradient-based methods have some major limitations. First, they do not satisfy the input invariance, meaning that input transformations such as constant shift can generate misleading attributions without affecting the model prediction [42]. Second, they fail to deal with zero-valued inputs. Third, they suffer from gradient saturation where large gradients dominate and obscure smaller gradients. The difference-from-reference approaches, such as integrated gradients (IG), are believed to be a good fit to solve these challenges by satisfying more axioms for attributions [95]. The fundamental mechanism of IG and its variants is to accumulate the gradients obtained as the input is interpolated between a reference point and the actual input. The baseline reference point is critical for reliable evaluation, but the criteria for choosing an appropriate baseline remain unclear. Some use noise or synthetic reference with training data, but performance cannot be guaranteed [60]. In addition, IG struggles to capture output changes in saturated regions and should focus on unsaturated regions [66]. Another challenge of IG is the computational overhead to achieve high-quality integrals. Since IG integrates along a straight line path that does not fit well the discrete word embedding space, variants have been developed to adapt it for language models [88, 85, 29].

Surrogate Models Surrogate model methods use simpler, more comprehensible models to explain individual predictions of black-box models. These surrogate models include decision trees, linear models, decision rules, and other white-box models that are inherently more understandable to humans. The explanation models need to satisfy additivity, meaning that the total impact of the prediction should equal the sum of the individual impacts of each explanatory factor. Also, the choice of interpretable representations matters. Unlike raw features, these representations should be powerful enough to generate explanations yet still understandable and meaningful to human beings. An early representative local explanation method called LIME [83] employs this paradigm. To generate explanations for a specific instance, the surrogate model is trained on data sampled locally around that instance to approximate the behavior of the original complex model in the local region. However, it is shown that LIME does not satisfy some properties of additive attribution, such as local accuracy, consistency, and missingness [59]. SHAP is another framework that satisfies the desirable properties of additive attribution methods [59]. It treats features as players in a cooperative prediction game and assigns each subset of features a value reflecting their contribution to the model prediction. Instead of building a local explanation model per instance, SHAP computes Shapley values using the entire dataset. Challenges in applying SHAP include choosing appropriate methods for removing features and efficiently estimating Shapley values. Feature removal can be done by replacing values with baselines like zeros, means, or samples from a distribution, but it is unclear how to pick the right baseline. Estimating Shapley values also faces computational complexity exponential in the number of features. Approximation strategies including weighted linear regression, permutation, and other model-specific methods have been adopted to estimate Shapley values. Despite its complexity, SHAP remains popular and widely used due to its expressiveness for large deep models. To adapt SHAP to Transformer-based language models, methods such as TransSHAP have been proposed [43]. TransSHAP mainly focuses on adapting SHAP to sub-word text input and providing sequential visualization explanations that are well-suited for understanding how NLMs make predictions.

2.1.2 Natural Language Explanation

Natural language explanation refers to explaining a model's decision-making on an input sequence with generated text. The basic approach for generating natural language explanations involves training a language model using both original textual data and human-annotated explanations. The trained language model can then automatically generate explanations in natural language [81]. As explanations provide additional contextual space, they can improve downstream prediction accuracy and perform as a data augmentation technique [61, 103]. Apart from the explain-then-predict approach, predict-then-explain and joint predict-explain methods have also been investigated. The choice of methods depends on the purpose of the task. However, the reliability of applying generated explanations still necessitates further investigation. It is worth noting that both the techniques introduced in this section and the CoT explanations covered later in Section 4 produce natural language explanations. However, the explanations covered here are typically generated by a separate model, while CoT explanations are produced by the NLMs themselves.

2.2 Global Explanation

Unlike local explanations that aim to explain a model's individual predictions, global explanations offer insights into the inner workings of language models. Global explanations aim to understand what the individual components (neurons, hidden layers, and larger modules) have encoded and explain the knowledge properties learned by the individual components. Global explainers are particularly useful, for example, for quick model diagnostics of possible biases or knowledge discovery. Since global explainers aim to explain the behavior of the entire target model, usually via distilling the target model into an interpretable one, they implicitly provide local explanations as well. However, it is difficult, or impossible, for an interpretable model to accurately capture all the irregularities learned by a highly non-linear model. Hence, local explanations derived from global explainers might not always be accurate. The majority of the current works in the literature focus on designing local explainers.

Three main approaches exist for examining global explanations: probing methods that analyze model representations and parameters, neuron activation analysis to determine model responsiveness to input, and concept-based methods. Since the focus of our research is on local explanations, we avoid discussing each method in detail.

2.2.1 Probing-Based Explanations

The self-supervised pre-training process empowers language models, such as BERT [23] and T5 [80], to acquire extensive linguistic knowledge from training data. Probing techniques such as classifier-based probing and parameter-free probing are employed to explore the knowledge embedded in these models.

Classifier-based probing involves training a shallow classifier on top of pre-trained language models with frozen parameters. These classifiers identify linguistic properties or reasoning abilities acquired by the model. Studies categorized under vector representations explore the model's embedded knowledge, revealing that lower layers capture word-level syntax, while higher layers excel in sentence-level syntax and semantic knowledge. Probing syntactic information involves parse trees and structural probes, but debates persist on whether classifiers truly learn syntax or merely task-related information.

Parameter-free probing tailors datasets to linguistic properties without using classifiers. The model's performance on these datasets serves as an indicator of its ability to capture specific properties. Data-driven prompt search, exploring language models' text generation abilities, is another approach. However, concerns arise about regularities in training datasets influencing results and obscuring real factual knowledge captured by language models.

2.2.2 Neuron Activation Explanation

Neuron analysis in language models delves into individual dimensions, examining crucial neurons for model performance or linked to specific linguistic properties. This line of work typically involves two steps: identifying important neurons unsupervisedly and establishing relations between linguistic properties and individual neurons through supervised tasks. Shared neurons across models learning similar properties are ranked using metrics like correlation measurements or learned weights [4]. Conventional supervised classification can also be used for this purpose. The importance of ranked neurons is validated quantitatively through ablation experiments, and methods like greedy Gaussian probing have emerged to identify crucial neurons.

2.2.3 Concept-Based Explanation

Concept-based interpretability algorithms map the inputs to a set of concepts and measure the importance score of each pre-defined concept to model predictions. By introducing abstract concepts, models can be explained in a human-understandable fashion rather than on low-level features. Information in latent space can also be transformed into comprehensible explanations.

2.3 Making Use of Explanations

In the previous subsections, we introduced methods to provide explanations for NLMs. In this subsection, we discuss how explainability can be used as a tool to debug and improve models.

The inherent complexity of neural models has given rise to concerns regarding their opacity [84], particularly about the societal implications of employing neural models in high-stakes decision-making scenarios [5]. Therefore, explainability is of utmost importance for fostering trust, ensuring ethical practices, and maintaining the safety of NLP systems [25, 55].

2.3.1 Differnt Usage of Explanations in Training

Here, we discuss a few applications that arise on how explanations can be used in modeling a task, in a standard supervised learning setup.

Learning to Explain Rationalization offers local explanations by providing a unique explanation for each prediction instead of a global explanation that covers the entire model [1, 82]. These explanations yield valuable insights for various purposes, including debugging, quantifying bias and fairness, understanding model behavior, and ensuring robustness and privacy [69]. However, obtaining direct supervision in the form of human-labeled rationales during training is not always feasible, which has led to the development of datasets that include human justifications for the true labels. These efforts enhance the interpretability of NLP models and address the limitations associated with direct supervision in learning to explain.

Post-hoc Explanations Post-hoc explanations are another branch of interpretability research. These explanations often involve token-level importance scores. In the quest for effective post-hoc explanations, a balance must be struck between the clarity of semantics and the avoidance of counter-intuitive

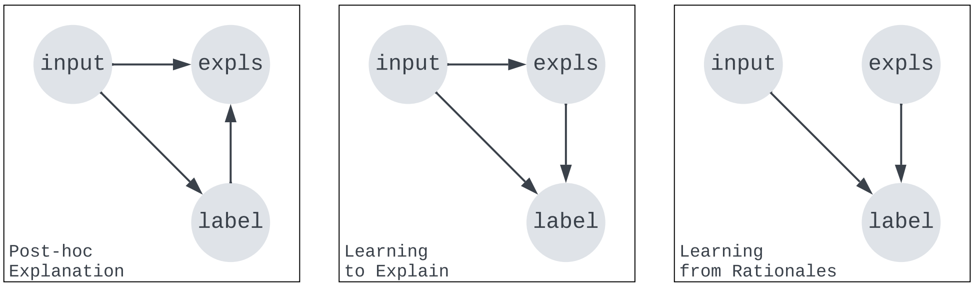


Figure 2.1: Computation graphs describing the relationships between post-hoc explanations, learning to explain, and learning from rationales.

behaviors. Gradient-based explanations [94, 90] provide clear semantics by describing the local impact of input perturbations on the outputs of the model. However, they can sometimes exhibit inconsistent behaviors [31], and their effectiveness relies on the differentiability of the model. Alternatively, there are model-agnostic methods that do not rely on specific model properties. One notable example is Local Interpretable Model-agnostic Explanations [LIME, 82]. These approaches approximate the behavior of the model locally by repeatedly making predictions on perturbed inputs and fitting a simple, explainable model over the resulting outputs.

Learning from Human Rationales Recent research has focused on leveraging rationales to enhance the training of neural text classifiers. [108] introduced a rationale-augmented Convolutional Neural Network that explicitly identifies sentences supporting categorizations. [93] demonstrated that incorporating rationales during training improves the quality of predicted rationales, as preferred by humans compared to models trained without explicit supervision [93]. In addition to integrated models, pipeline approaches have been proposed, where separate models are trained for rationale extraction and classification based on these extracted rationales [48, 18]. These approaches assume the availability of explicit training data for rationale extraction.

2.3.2 Applications of Explanations for Further Works

Debugging Models Post-hoc explainability methods can be used to analyze model feature importance patterns to identify biases or limitations in its behavior [27]. For example, if the model consistently attends to certain tokens in the input sequence regardless of the context, this may indicate that the model relies on heuristics or biases rather than truly understanding the meaning of the input sequence. Recent work has used Integrated Gradients to debug trained language models in natural language understanding tasks, finding that they use shortcuts rather than complex reasoning for prediction [27]. Specifically, these models favor features from the head of long-tailed distributions, picking up these shortcut cues early in training. This shortcut learning harms model robustness and generalization to out-of-distribution samples. Integrated Gradient explanations have also been used to examine the adversarial robustness of language models [17]. The explanations reveal that models robust to adversarial examples rely on similar features, while non-robust models rely on different key features. These insights have motivated the development of more robust adversarial training methods.

Improving Models Regularization techniques can be used to improve the performance and reliability of model explanations. Specifically, explanation regularization (ER) methods aim to improve NLM generalization by aligning the model's machine rationales (which tokens it focuses on) with human rationales [41]. For example, a framework called AMPLIFY is proposed that generates automated rationales using post-hoc explanation methods [44]. These automated rationales are fed as part of prompts to NLM for prediction. Experiments show that AMPLIFY improves the accuracy of NLMs by 10-25% for various tasks, even when human rationale is lacking. Another study proposes ER-TEST [41], a framework that evaluates the out-of-distribution (OOD) generalization of ER models along three dimensions: unseen dataset

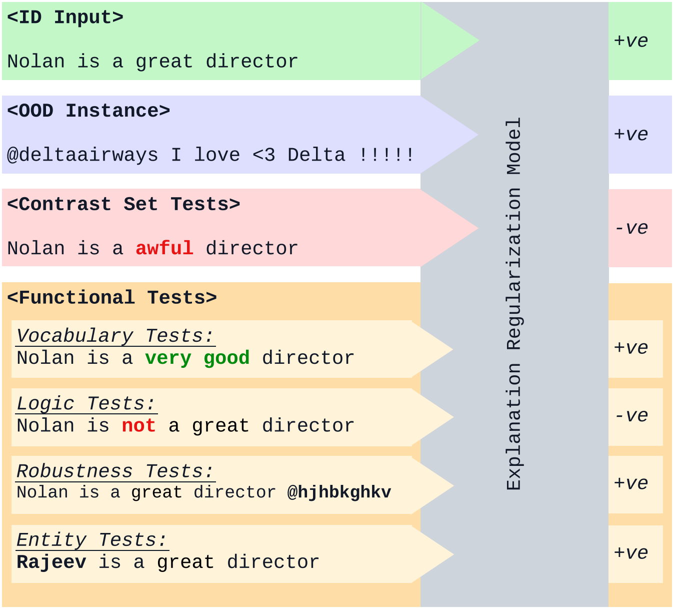

<ID Input>
Nolan is a great director
+ve
<OOD Instance>
@deltaairways I love <3 Delta !!!!!
+ve
<Contrast Set Tests>
Nolan is a awful director
-ve
<Functional Tests>
Vocabulary Tests:
Nolan is a very good director
+ve
Logic Tests:
Nolan is not a great director
-ve
Robustness Tests:
Nolan is a great director @hjhbkghkv
+ve
Entity Tests:
Rajeev is a great director
+ve
Explanation Regularization Model

Figure 2.2: ER-TEST Framework - Apart from existing ID evaluations of ER criteria, ER-TEST evaluates ER's impact on OOD generalization along three dimensions: A. Unseen datasets, B. Contrast set tests, and C. Functional tests.

tests, contrast set tests, and functional tests. This provides a more comprehensive evaluation than just in-distribution performance. They consider three types of explainability methods, including Input*Gradient, attention-based rationale [91], and learned rationale [14]. Across sentiment analysis and natural language inference tasks/datasets, ER-TEST shows that ER has little impact on in-distribution performance but yields large OOD gains. An end-to-end framework called XMD was proposed for explanation-based debugging and improvement [47]. XMD allows users to provide flexible feedback on task- or instance-level explanations via an intuitive interface. It then updates the model in real time by regularizing it to align explanations with user feedback. Using XMD has been shown to improve models' OOD performance on text classification by up to 18%.

Downstream Applications Explainability can also be applied to real-world problems such as education, finance, and healthcare. For example, explainable zero-shot medical diagnosis is an interesting use case. One recent study proposes a framework for explainable zero-shot medical image classification utilizing vision-language models like CLIP along with NLMs like ChatGPT [56]. The key idea is to leverage ChatGPT to automatically generate detailed textual descriptions of disease symptoms and visual features beyond just the disease name. This additional textual information helps to provide more accurate and explainable diagnoses from CLIP [79]. To handle potential inaccuracies from ChatGPT on medical topics, the authors design prompts to obtain high-quality textual descriptions of visually identifiable symptoms for each disease class. Extensive experiments on multiple medical image datasets demonstrate the effectiveness and explainability of this training-free diagnostic pipeline.

Chapter 3

Explanation Evaluation

Following the goals of explainable AI, a model's quality should be evaluated not only by its accuracy and performance but also by how well it provides explanations for its predictions. In the previous chapter, we introduced different explanation techniques and their usages, but evaluating how faithfully they reflect a model's reasoning process remains a challenge. Two key dimensions of evaluations are *plausibility* to humans and *faithfulness* in capturing NLMs' internal logic. Both parts will mainly cover universal properties and metrics that can be applied to compare various explanation approaches. We focus on quantitative evaluation properties and metrics, which are usually more reliable than qualitative evaluations.

Given the young age of the field, unsurprisingly there is little agreement on how explanations should be evaluated. The majority of the works lack a standardized evaluation or include only an informal evaluation, while a smaller number of papers looked at more formal evaluation approaches, including leveraging ground truth data and human evaluation [20]. Human evaluations assess plausibility through the similarity between model rationales and human rationales or subjective judgments. However, these methods usually neglect faithfulness. Subjective judgments may also not align with model reasoning, making such an evaluation unreliable. As argued by [37], faithful evaluation should have a clear goal and avoid human involvement. Automatic evaluations test importance by perturbing model rationales, avoiding human biases. Therefore, developing rigorous automatic metrics is critical for fair faithfulness evaluation, which will be covered under the faithfulness evaluation dimension.

A more direct way to assess the explanation quality is to ask humans to evaluate the effectiveness of the generated explanations. This has the advantage of avoiding the assumption that there is only one good explanation that could serve as ground truth, as well as sidestepping the need to measure the similarity of explanations. Here as well, it is important to have multiple annotators, report inter-annotator agreement, and correctly deal with subjectivity and variance in the responses. The approaches found in this survey vary in several dimensions, including the number of humans involved, as well as the high-level task that they were asked to perform (including rating the explanations of a single approach and comparing explanations of multiple techniques).

On the other hand, several works compare generated explanations to ground truth data in order to quantify the performance of explainability techniques. Employed metrics vary based on task and explainability technique, but commonly encountered metrics include P/R/F1 [11], perplexity, and BLEU [54, 81]. While having a quantitative way to measure explainability is a promising direction, care should be taken during ground truth acquisition to ensure its quality and account for cases where there may be alternative valid explanations. Approaches employed to address this issue involve having multiple annotators and reporting inter-annotator agreement or mean human performance, as well as evaluating the explanations at different granularities (e.g., token-wise vs phrasewise) to account for disagreements on the precise value of the ground truth [11].

While the above overview summarizes how explainability approaches are commonly evaluated, another important aspect is what is being evaluated. Explanations are multi-faceted objects that can be evaluated on multiple aspects, such as fidelity (how much they reflect the actual workings of the underlying model), comprehensibility (how easy they are to understand by humans), and others. Therefore, understanding the target of the evaluation is important for interpreting the evaluation results. [12] provides a comprehensive presentation of aspects of evaluating approaches. In this thesis, we utilized a recently introduced method from ERASER [24], for extractive rationales evaluation which involves evaluating plausibility and faithfulness.

3.1 Evaluating Plausibility

The plausibility of local explanation is typically measured at the input text or token level. Plausibility evaluation can be categorized into five dimensions: grammar, semantics, knowledge, reasoning, and computation [87]. These dimensions describe the relationship between the masked input and human-annotated rationales. Different evaluation dimensions require different kinds of datasets. For example, the sentence “The country [MASK] was established on July 4, 1776.” has the human-annotated rationale “established on July 4, 1776” and the answer to the mask should be “the United States” deriving from fact/knowledge. Although rationales might be in different granularity levels such as token or snippet and dimensions, evaluation procedures are the same except for diversified metrics. Human-annotated rationales are generally from benchmark datasets, which should meet several criteria: 1) sufficiency, meaning rationales are enough for people to make correct predictions; 2) compactness, requiring that if any part in the rationales is removed, the prediction will change [63]. The explanation models are then responsible for predicting important tokens and generating rationales with these tokens. The above two kinds of rationales will be measured with various metrics. Popular metrics can be classified into two classes according to their scope of measurement. Metrics measuring two token-level rationales include Intersection-Over-Union (IOU), precision, and recall. Metrics that measure overall plausibility include the F1 score for discrete cases and the area under the precision-recall curve (AUPRC) for continuous or soft token selection cases [24].

3.2 Evaluating Faithfulness

There are several model-level principles to which explanation methods should adhere to be faithful, which include implementation invariance, input invariance, input sensitivity, completeness, polarity consistency, prediction consistency, and sufficiency. Implementation invariance also known as model sensitivity means that the attribution scores should remain the same regardless of differences in the model architectures, as long as the networks are functionally equal [95]. Even gradient-based approaches usually meet this metric well; the assumption may not be grounded. Input invariance requires attribution methods to reflect the sensitivity of prediction models w.r.t. effective input changes. For example, attribution scores should remain the same over the constant shift of the input [42]. Input sensitivity defines attribution scores should be non-zero for features that solely explain prediction differences [95]. Completeness combines sensitivity and implementation invariance with path integrals from calculus [95], which only applies to differentiable approaches. Polarity consistency points out that some high-ranking features could impose suppression effects on final predictions, which negatively impacts explanations and should be avoided, but mostly not [58]. Prediction consistency confines that instances with the same explanations should have the same prediction. Sufficiency requires that data with the same attributions should have the same related labels even with different explanations [22]. In this class of approaches, researchers aim to prevent certain types of contradictory explanations by formulating axioms and properties. However, each metric can address only one particular facet of faithfulness problems. It is extremely difficult to provide all-in-one solutions within a single framework. Additionally, these approaches focus solely on avoiding inconsistent behaviors of explanation models by designing properties for explanation methods. The overall performance of models' faithfulness is measured with the following metrics by quantitatively verifying the relationship between prediction and model rationales.

- *Comprehensiveness* (Comp): the change in probability for the original predicted class before and after top-ranked important tokens are removed, which means how influential the rationale is. It is formulated as Comp $= m(\mathrm{x}_i)_j - m(\mathrm{x}_i \backslash \mathrm{r}_i)_j$. A higher score indicates the importance of rationales [24].
- *Sufficiency* (Suff): the degree to which the parts within the extracted rationales can allow the model to make a prediction, which is defined as Suff $= m(\mathrm{x}_i)_j - m(\mathrm{r}_i)_j$ [24].

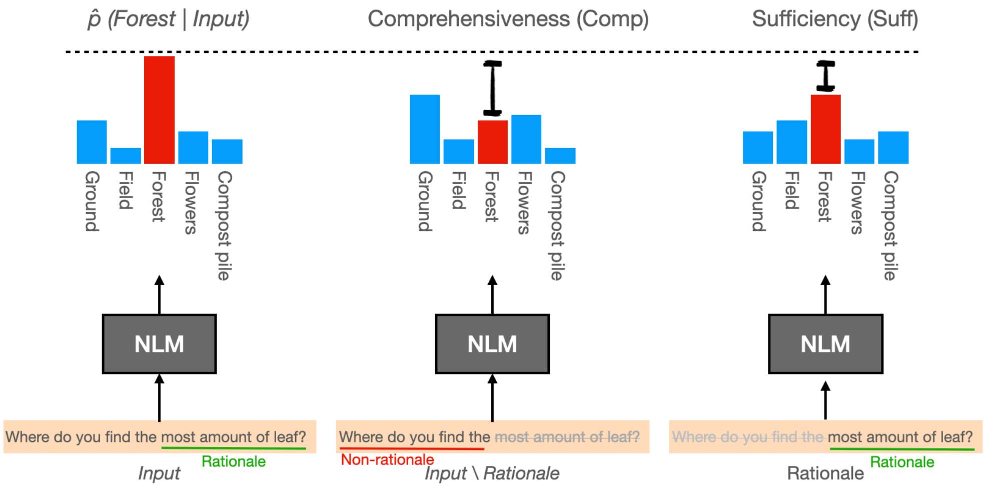


Figure 3.1: Sufficiency and Comprehensiveness computation procedure.

In ERASER [24], related tokens are classified into groups ranked by importance scores so that tokens can be masked in a ranked order and gradually observe output changes. The correlation between output changes and the importance of masked tokens denotes models' ability to correctly attribute feature importance. Two key questions persist when evaluating explanation models, regardless of the specific metrics used: 1) how well does the model quantify important features? 2) can the model effectively and correctly extract as many influential features as possible from the top-ranked features? However, existing evaluation metrics are often inconsistent with the same explanation models. [19] demonstrates that attention-based importance metrics are more robust than non-attention ones whereas regarding attention as an explanation is still debatable [38].

Chapter 4

Datasets

Highlights are subsets of the input elements (words, phrases, or sentences) that explain a prediction. [49] coin them extractive rationales, or subsets of the input tokens of a textual task that satisfy two properties: (i) compactness, they are short and coherent, and (ii) sufficiency, they suffice for prediction as a substitute of the original text. [104] introduce a third criterion, (iii) comprehensiveness, that all the evidence that supports the prediction is selected, not just a sufficient set. Since the term "rationale" implies human-like intent, [37] argue to call this type of explanation highlights to avoid inaccurately attributing human-like social behavior to AI systems. They are also called evidence in fact-checking and multi-document question answering (QA) —a part of the source that refutes/supports the claim. To reiterate, highlights should be sufficient to explain a prediction and compact; if they are also comprehensive, we call them comprehensive highlights. table 4.1 gives examples of sufficient vs. non-sufficient and comprehensive highlights.

The main datasets in literature for this task are the CoS-E [81] and e-SNLI [10] datasets, all of which have gold highlights from ERASER [24]. For the OOD generalization evaluation, we consider MNLI [101] and HANS [65]. Given the discrepancies in the characterization of highlights and the specific instructions provided to annotators, we can conclude that relying solely on

Table 4.1: Examples of highlights differing in comprehensiveness and sufficiency

Instance with Highlight	Type of Highlight
Premise: People are stretching on yoga mats. ***Hypothesis***: They stretched on bikes. ***Label***: contradiction	***Premise***:People are stretching on yoga mats. ***Hypothesis***:They stretched on bikes. (sufficient)
Premise: People on bicycles waiting at an intersection. ***Hypothesis***: There are people on bicycles. ***Label***: entailment	***Premise***: People on bicycles waiting at an intersection. ***Hypothesis***:There are people on bicycles. (comprehensive)
Premise: People on bicycles waiting at an intersection. ***Hypothesis***: Some people on bikes are stopped at a junction. ***Label***: neutral	***Premise***: People on bicycles waiting at an intersection. ***Hypothesis***: Some people on bikes are stopped at a junction. (¬ sufficient)

a general description of data collection for post-hoc assessment of comprehensiveness is prone to errors. In addition, it is important to note that gold rationales are typically annotated based on the gold task label (y_i) rather than the predicted label (y_i^*) generated by the task model, which is unknown beforehand. Therefore, if the predicted label (y_i^*) differs from the gold label (y_i), the supervision provided by the gold rationales may introduce some noise or inconsistencies. However, our fully end-to-end framework addresses this challenge by backpropagating the sufficiency and comprehensiveness loss from the task model to the rationale extractor. This allows the framework to mitigate the impact of noisy supervision and enhance the overall performance and alignment between the rationale extractor and the task model.

4.1 CoS-E

Deep learning models perform poorly on tasks that require commonsense reasoning, which often necessitates some form of world knowledge or reasoning over information not immediately present in the input. CoS-E [81] consists of multiple-choice questions and answers taken from the work of [97]. It includes supporting rationales for each question-answer pair in two forms. Extracted supporting snippets and free-text descriptions that provide a more detailed explanation of the reasoning behind the answer choice.

4.2 e-SNLI

e-SNLI [10] is an augmentation of the SNLI corpus [8] and includes human rationales as well as natural language explanations, although they were not utilized in our work. It is worth noting that the authors of e-SNLI collect non-comprehensive highlights. In their annotation instructions, they specifically instruct annotators to highlight only words in the hypothesis and not in the premise for neutral pairs. Furthermore, they consider explanations involving contradiction or neutrality to be correct as long as at least one piece of evidence in the input is highlighted. Focusing on the hypothesis and allowing partial highlighting of evidence leads to the collection of non-comprehensive highlights in the dataset.

4.3 MNLI

MNLI [101] covers a broader range of written and spoken text, subjects, styles, and levels of formality compared to SNLI. It was introduced to determine the logical relationship between two given sentences. With over 400,000 sentence pairs, MNLI provides a rich and representative dataset that enables researchers to develop and evaluate models for natural language inference across different domains.

To conduct additional OOD generalization evaluation, we utilized two OOD Contrast Sets called **MNLI-Contrast** and **MNLI-Original**. These contrast sets were created by slightly modifying the original MNLI instances [50]. In MNLI-Contrast, the modification changes the original label, while in MNLI-Original, the original label remains the same. Examples of these contrast sets are shown in table 4.3. Besides, to evaluate the plausibility metrics on OOD data, we performed a random sampling of 50 instances from the MNLI validation split and annotated them manually w.r.t. gold labels. We referred to this particular subset of data as **e-MNLI**. table 4.2 shows instances

Table 4.2: e-MNLI instances for different labels. Following e-SNLI for neutral labels only tokens in hypothesis are highlighted.

Instances with Highlights	***Label***
Premise: They drive it around the country in a dilapidated ice-cream truck trying to keep it cool. ***Hypothesis***: They used an ice cream truck to try and keep it from getting warm.	entailment
Premise: Then he turned to Tommy. ***Hypothesis***: He talked to Tommy.	neutral
Premise: but i've lived up here all my life and i'm fifty eight years old so i i could ***Hypothesis***: I have moved somewhere else in my life.	contradiction

Table 4.3: MNLI Contrast Test Set. In the MNLI-Original the original label is unchanged while in the MNLI-Contrast the label is also changed based on changes in premise or hypothesis.

Model	**Contrast Set Instance**
MNLI-Contrast	***Premise***: yeah well that's not really immigration. $\xrightarrow{\text{past simple}}$ Yeah well that wasn't immigration. ***Hypothesis***: That is not immigration. $\xrightarrow{\text{future simple}}$ That won't be immigration. ***Label***: entail→ neutral
MNLI-Original	***Premise***: Clearly, GAO needs assistance to meet its looming human capital challenges. $\xrightarrow{\text{it cleft: ARG1}}$ Clearly it is GAO who needs assistance to meet its human capital challenges looming. ***Hypothesis***: GAO will soon be suffering from a shortage of qualified personnel. $\xrightarrow{\text{it cleft: ARG1}}$ It is GAO who soon will be suffering from a shortage of personnel qualified for. ***Label***: neutral→ neutral

from e-MNLI for different labels.

4.4 HANS

HANS [65] is designed to evaluate the capability of NLI systems to rely on heuristics and patterns instead of genuine understanding. HANS consists of sentence pairs carefully crafted to mislead models using three heuristic categories: Lexical Overlap, Subsequence, and Constituent. Instances for each heuristic are given in table 1.1. By evaluating models on the HANS dataset, researchers can gain insights into the limitations and robustness of NLI systems and foster the development of models that rely on genuine linguistic understanding rather than shallow heuristics.

Chapter 5

Experiments and Setup

Rationale extraction aims to provide *faithful* (*i.e.,* reflective of the behavior of the model) and *plausible* (*i.e.,* convincing to humans) explanations by highlighting the inputs that had the largest impact on the prediction without compromising the performance of the task model. In recent works, the focus of training rationale extractors was primarily on optimizing for plausibility using human highlights, while the task model was trained on jointly optimizing for task predictive accuracy and faithfulness. In this thesis, we aim to implement various frameworks for rationale extraction and analyze their associated advantages and disadvantages. Besides, we propose **REFER** [33], a framework that employs a differentiable rationale extractor that allows to back-propagate through the rationale extraction process. We analyze the impact of using human highlights during training by jointly training the task model and the rationale extractor. In our experiments, REFER yields significantly better results in terms of faithfulness, plausibility, and downstream task accuracy on both in-distribution and out-of-distribution data. To obtain a single score w.r.t all three criteria, we utilized Normalized Relative Gain (NRG) that maps obtained scores to range 0 to 1 (the higher the better). The final score is composite NRG (CNRG) which is the average over faithfulness NRG (FNRG), plausibility NRG (PNRG), and task NRG (TNRG). On both e-SNLI and CoS-E, our

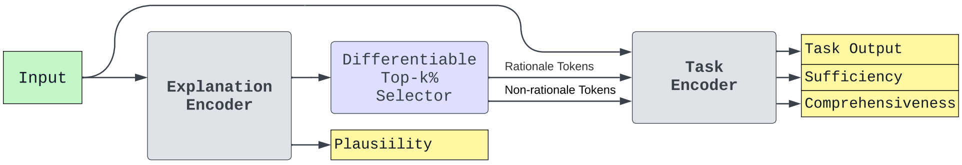


Figure 5.1: The pipeline for explanation regularization is a fully end-to-end approach where the task model's output loss is back-propagated through all components, resulting in a compromised performance that considers all training criteria.

best setting produces better results in terms of CNRG than the previous baselines by 11% and 3%, respectively. Besides, REFER generalizes faithfulness and plausibility over out-of-distribution data.

5.1 REFER Architecture

Task Model Consider $\mathcal{F}_{\text{task}}$ as the task model for text classification, where it consists of an encoder [98] and a head. In modern NLP systems, $\mathcal{F}_{\text{task}}$ usually has a BERT-style architecture [23]. Let $\mathrm{x}_i = [\mathrm{x}_i^t]_{t=1}^n$ be i^{th} input sequence with length n, and $\mathcal{F}_{\text{task}}(\mathrm{x}_i) \in \mathbb{R}^M$ be the logit vector for the output of the task model. We use $y_i = \arg\max_j [\mathcal{F}_{\text{task}}(\mathrm{x}_i)]_j$ to denote the class predicted by task model. Given that cross-entropy loss is used to train $\mathcal{F}_{\text{task}}$ to predict y_i^*, the task loss is defined as follow:

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{CE}}(\mathcal{F}_{\text{task}}(\mathrm{x}_i), y_i^*) \tag{5.1}$$

Rationale Extractor Let $\mathcal{F}_{\text{ext}}$ denote a rationale extractor, such that $\mathrm{s}_i = \mathcal{F}_{\text{ext}}(\mathrm{x}_i)$. Given $\mathcal{F}_{\text{task}}$, x_i, and y_i, the goal of rationale extraction is to output vector $\mathrm{s}_i = [s_i^t]_{t=1}^n \in \mathbb{R}^n$, such that each s_i^t is an importance score indicating how strongly token x_i^t influenced $\mathcal{F}_{\text{task}}$ to predict class y_i. The final rationales are typically obtained by binarizing s_i as $\mathrm{r}_i^{(k)} \in \{0,1\}^n$, via the top-$k\%$ strategy [24, 39, 76, 15]. Other binarization strategies, such as score thresholding and highest-scoring contiguous k-token span, can also be used.

To capture the degree to which the snippets within the extracted rationales are sufficient for a model to make a prediction, we measure the disparity in model confidence when considering the complete input versus only the extracted rationales. A small difference suggests the high importance of extracted rationales.

$$\mathcal{L}_{\text{suff-diff}} = \mathcal{L}_{\text{CE}}(\mathcal{F}_{\text{task}}(\mathrm{r}_i^{(k)}), y_i^*) - \mathcal{L}_{\text{CE}}(\mathcal{F}_{\text{task}}(\mathrm{x}_i), y_i^*) \tag{5.2}$$

Following [13], to avoid negative losses, we can use margin m_s to impose a lower bound on $\mathcal{L}_{\text{suff-diff}}$, yielding the following margin criterion:

$$\mathcal{L}_{\text{suff}} = \max(-m_s, \mathcal{L}_{\text{suff-diff}}) + m_s \tag{5.3}$$

To compute comprehensiveness we create contrast examples for x_i, $\tilde{\mathrm{x}}_i = \mathrm{x}_i \backslash \mathrm{r}_i^{(k)}$, which is x_i with the predicted rationales r_i removed [107]. Similar to eq. (5.2), we measure the difference in model confidence between considering the complete input and the contrast set $\tilde{\mathrm{x}}_i$. A high score here implies that the rationales were influential in the prediction. A negative value here means that the model became more confident in its prediction after the rationales were removed; this would seem counter-intuitive if the rationales were indeed the reason for its prediction.

$$\mathcal{L}_{\text{comp-diff}} = \mathcal{L}_{\text{CE}}(\mathcal{F}_{\text{task}}(\mathrm{x}_i), y_i^*) - \mathcal{L}_{\text{CE}}(\mathcal{F}_{\text{task}}(\tilde{\mathrm{x}}_i), y_i^*) \tag{5.4}$$

Repeatedly, we enforce $\mathcal{L}_{\text{comp-diff}}$ to be positive as follows:

$$\mathcal{L}_{\text{comp}} = \max(-m_c, \mathcal{L}_{\text{comp-diff}}) + m_c \tag{5.5}$$

Due to the change of predicted label during the training, we use y_i^* to compute the comprehensiveness and sufficiency loss instead of using predicted label y_i. This helps the model to be more stable during the training. Finally, the selection of the tokens for matching the human highlights can be cast as a binary classification problem, and the plausibility loss is computed using the binary cross-entropy (BCE) loss function:

$$\mathcal{L}_{\text{plaus}} = -\sum_{t} \mathrm{r}_i^{*,t} \log(\mathcal{F}_{\text{ext}}(\mathrm{x}_i^t)) \tag{5.6}$$

where r_i^* is the gold rationale for input x_i of length t. This leads to the following multi-task learning objective:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{task}} + \alpha_{\text{f}}\mathcal{L}_{\text{faith}} + \alpha_{\text{p}}\mathcal{L}_{\text{plaus}} \\ &= \mathcal{L}_{\text{task}} + \alpha_{\text{c}}\mathcal{L}_{\text{comp, K}} + \alpha_{\text{s}}\mathcal{L}_{\text{suff, K}} + \alpha_{\text{p}}\mathcal{L}_{\text{plaus}} \end{aligned}$$

Back-Propagating Through Rationale Extraction End-to-end models that integrate discrete algorithms into their framework enable various capabilities, such as sampling from discrete latent distributions [40, 75] and solving combinatorial optimization problems [100, 62, 72]. Relying on discrete outputs solely during testing can lead to unexpected behaviors. Conversely, certain situations necessitate the use of discrete outputs during training. For discrete distributions, exact gradient computation for the expected loss becomes intractable. Similarly, in the case of combinatorial optimization problems, the loss function is discontinuous, resulting in gradients that are nearly zero throughout most of the problem space. To address these challenges, the Score Function Estimator (SFE) presents a viable option. However, the SFE suffers from high variance, particularly when dealing with intractable distributions such as $p(z;\theta)$. In this sense, a recent approach called Implicit Maximum Likelihood Estimation (IMLE) [72] offers a solution by combining implicit differentiation through perturbation with the path-wise gradient estimator. This approach alleviates the variance issue associated with the SFE and overcomes the intractability of the distribution $p(z;\theta)$.

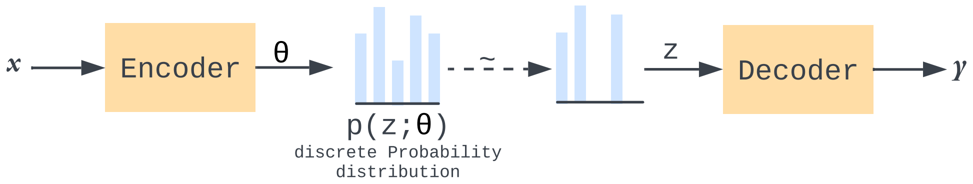


Figure 5.2: Illustration of the learning problem. z is the discrete latent structure, x and y are feature inputs and target outputs, Encoder maps $\mathcal{X} \mapsto \theta$, Decoder maps $\mathcal{Z} \mapsto \mathcal{Y}$, and $p(z;\theta)$ represents the discrete probability distribution. The dashed path indicates non-differentiability.

To back-propagate through the rationale extraction process, we use Adaptive Implicit Maximum Likelihood Estimation (AIMLE) [67], a recently proposed low-variance and low-bias gradient estimation method for discrete distribution that does not require significant hyper-parameter tuning. AIMLE is an extension of Implicit Maximum Likelihood Estimation (IMLE) [72] a perturbation-based gradient estimator where the gradient of the loss w.r.t. the token scores $\nabla_{\mathbf{s}}\mathcal{L}$ is estimated as $\nabla_{\mathbf{s}}\mathcal{L} \approx \mathbf{r}(\mathbf{s}+\epsilon) - \mathbf{r}(\mathbf{s}+\lambda\nabla_{\mathbf{r}}\mathcal{L}+\epsilon)$, where ϵ denotes Gumbel noise, $\mathbf{r}$ denotes the top-k% function, and λ is a hyper-parameter selected by the user. AIMLE removes the need for the user to select λ by automatically identifying the optimal λ for a given learning task.

The proposed pipeline aims to overcome the limitations of previous architectures by taking the advantageous features of the AIMLE to combine discrete exponential family distributions with differentiable neural components. By leveraging the benefits of implicit differentiation and the path-wise gradient estimator, we aim to enhance the stability and efficiency of our framework. In fact, AIMLE enables the backpropagation of the faithfulness loss obtained from the task model through the rationale extractor making it adept at ensuring faithfulness, as it can take into account the performance of the task model when adapting itself.

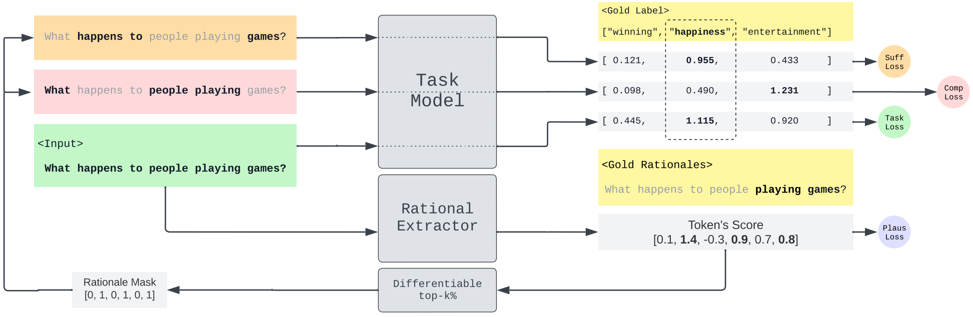

What happens to people playing games?
What happens to people playing games?
<Input>
What happens to people playing games?
Task Model
Rational Extractor
Differentiable top-k%
Rationale Mask [0, 1, 0, 1, 0, 1]
<Gold Label>
["winning", "happiness", "entertainment"]
[0.121, 0.955, 0.433]
[0.098, 0.490, 1.231]
[0.445, 1.115, 0.920]
Suff Loss
Comp Loss
Task Loss
<Gold Rationales>
What happens to people playing games?
Token's Score [0.1, 1.4, -0.3, 0.9, 0.7, 0.8]
Plaus Loss

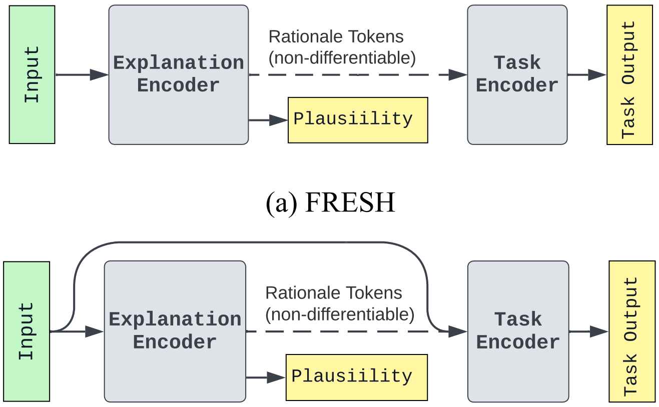
Figure 5.3: **REFER Pipeline**. The Task Model is trained using (i) Task Loss, (ii) Sufficiency Loss, and (iii) Comprehensiveness Loss, while the Rationale Extractor is trained through backpropagation using (i) Plausibility Loss, (ii) Sufficiency Loss, and (iii) Comprehensiveness Loss. This approach ensures a high level of consistency across each criterion, as all components are aware of each other's status and can adapt to strike a balance among the three criteria.

5.2 Baselines Models

To assess the effectiveness of back-propagation through discrete latent space within the designated architecture for REFER, we implemented a set of previous baseline models that don't take advantage of this process. Then we compared their performance on different aspects of extractive rationales and discussed their advantages and disadvantages in chapter 6.

Several prior works have aimed to enhance the faithfulness of extractive rationales using Attribution Algorithms (AAs), which extract rationales via handcrafted functions [94, 36, 89]. AAs do not involve training $\mathcal{F}_{\text{ext}}$ and are applied post hoc (i.e., they do not impact $\mathcal{F}_{\text{task}}$'s training). This class of architectures is not easily optimized and often requires significant computational resources. [89, 86] tackle the computational cost by training a model to mimic the behavior of an AA. Integrated Gradient baseline [AA (IG), 94] is utilized as a baseline for this class. Saliency Guided Training [SGT, 36] is another baseline that uses a sufficiency-based criterion to regularize $\mathcal{F}_{\text{task}}$, such that the AA yields faithful rationales for $\mathcal{F}_{\text{task}}$.

[39, 105, 73, 3, 104, 49] use Select-Predict Pipelines (SPPs) to generate faithful rationales. In SPPs, $\mathcal{F}_{\text{task}}$ is trained to solve a given task using only the tokens chosen by $\mathcal{F}_{\text{ext}}$ [39, 104, 73]; therefore, SPPs aim for ”faithfulness by construction.” However, SPPs only guarantee sufficiency but not comprehensiveness [24], and generally produce less accurate results, since they can only observe a portion of the input, and due to the challenges associated with gradient-based optimization and discrete distributions. FRESH [39] and A2R [104] have been proposed to produce faithful rationales: FRESH relies on training $\mathcal{F}_{\text{task}}$ and $\mathcal{F}_{\text{ext}}$ separately, while A2R aims to improve $\mathcal{F}_{\text{task}}$'s task performance by regularizing it with an attention-based predictor that utilizes the full input [39, 104].



(a) FRESH

(b) A2R

Figure 5.4: Select-Predict-Pipeline Variants.

Regarding the plausibility of the rationales, existing approaches typically involve supervising neural rationale extractors [6] and SPPs [39, 73, 24] using gold rationales. However, LM-based extractors lack training for faithfulness, and SPPs sacrifice task performance to achieve faithfulness by construction. Other works mainly focus on improving the plausibility of rationales [71, 46, 10], often employing task-specific pipelines [81, 45]. In contrast, REFER jointly optimizes both the task model and rationale extractor for faithfulness, plausibility, and task performance and reaches a better trade-off w.r.t. these desiderata without suffering from heuristic-based approaches (e.g., AAs) disadvantages.

UNIREX [13] is the most recent pipeline from Meta research, which considers two main architecture variants: (i) Dual LM (DLM), where $\mathcal{F}_{\text{task}}$ and $\mathcal{F}_{\text{ext}}$ are two separate Transformer-based LMs with the same encoder architecture (ii) Shared LM (SLM), where $\mathcal{F}_{\text{task}}$ and $\mathcal{F}_{\text{ext}}$ share encoder, while $\mathcal{F}_{\text{ext}}$ has its own output head. fig. 5.5 shows the architecture for DLM and SLM in UNIREX. DLM provides more capacity for $\mathcal{F}_{\text{ext}}$, which can help $\mathcal{F}_{\text{ext}}$ provide plausible rationales. While SLM leverages multitask learning and improve faithfulness since $\mathcal{F}_{\text{ext}}$ has greater access to information about $\mathcal{F}_{\text{task}}$'s reasoning process [13]. REFER benefits from both SLM and DLM architectures by establishing communication between separate $\mathcal{F}_{\text{task}}$ and $\mathcal{F}_{\text{ext}}$ using back-propagation.

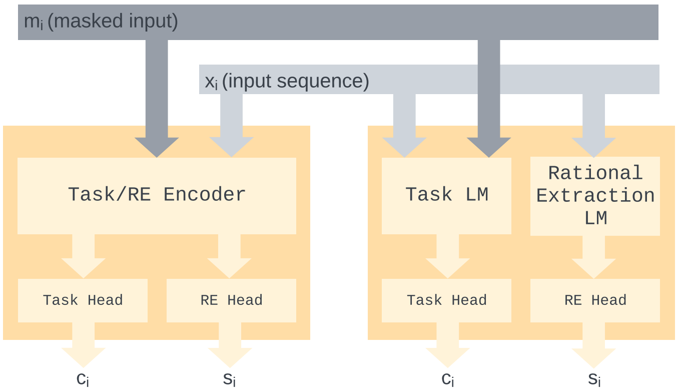


Figure 5.5: Shared LM (left) and Dual LM (right) architecture. Using shared LM, the task model and rational extractor share the same encoder. While in the Dual LM model, they are completely separate

5.3 Models Detail and Hyperparameters

Transformers-based models, such as BERT, have been one of the most successful deep learning models for NLP. Unfortunately, one of their core limitations is the quadratic dependency (mainly in terms of memory) on the sequence length due to their full attention mechanism. To remedy this, [106] proposed BigBird, a sparse attention mechanism that reduces this quadratic dependency to linear. They show that BigBird is a universal approximator of sequence functions and is Turing complete, thereby preserving these properties of the quadratic, full attention model. Along the way, their theoretical analysis reveals some of the benefits of having *O(1)* global tokens (such as CLS) that attend to the entire sequence as part of the sparse attention mechanism. The proposed sparse attention can handle sequences of length up to eight times what was previously possible using similar hardware. Due to the capability to handle longer contexts, BigBird drastically improves performance on various NLP tasks such as question answering and summarization.

In our implementations, we utilize BigBird-Base [106] as the backbone for both $\mathcal{F}_{\text{task}}$ and $\mathcal{F}_{\text{ext}}$. This choice enables us to effectively handle input sequences of considerable length, accommodating up to 4096 tokens. We used AIMLE, which uses adaptive target distribution with alpha and beta initialized to 1 and 0, respectively. Throughout all experiments, we maintain a consistent learning rate of 2×10^{-5} and employ an effective batch size of 32. Our training process spans a maximum of 10 epochs, with early stopping applied after 5 epochs of no significant improvement.

To ensure optimal performance, we focus our hyperparameter tuning efforts on the weights associated with faithfulness and plausibility losses, specifically $\alpha_{\text{c}} = \alpha_s = \alpha_{\text{f}}$, and α_{p} as well as top-$k\%$. We applied a grid search across various configurations and evaluated their impact on comprehensiveness, sufficiency, plausibility scores, and task performance. The entire implementation is carried out using the PyTorch-Lightning framework [74, 30], which provides a streamlined and user-friendly environment for deep learning experiments.

Chapter 6

Discussion and Results

In this chapter, we undertake a systematic analysis of the results to address research questions introduced in chapter 1 and show that our proposed model, REFER, outperforms the SoTA architectures in all the following areas. The rest of this chapter is organized as follows: Initially, we examine various facets of the process of extracting rationales, delving into the influence of human rationales on model behavior, the model's capacity to replicate human-generated rationales and the degree to which the trained models can exhibit generalization on OOD datasets. Subsequently, in addressing these inquiries, we assess the performance of the REFER model in comparison to baseline models and architectures, demonstrating that our proposed model excels over previous models in the majority of instances.

We label the models with `+P` and `+FP` that are trained by optimizing for plausibility and jointly faithfulness and plausibility, respectively. fig. 6.1 displays the main results for e-SNLI in terms of NRG. Overall, `REFER+FP` achieved the highest composite NRG, improving over the strongest baseline (`UNIREX SLM+FP`) by 12%. Regarding plausibility, models explicitly trained for plausibility (`+P`) or both faithfulness and plausibility (`+FP`) achieved similar results, with `REFER+FP` outperforming the second-best model by 3%. Regarding faithfulness, REFER achieved the highest score in all three configurations. An interesting finding is that even when training REFER and A2R solely for

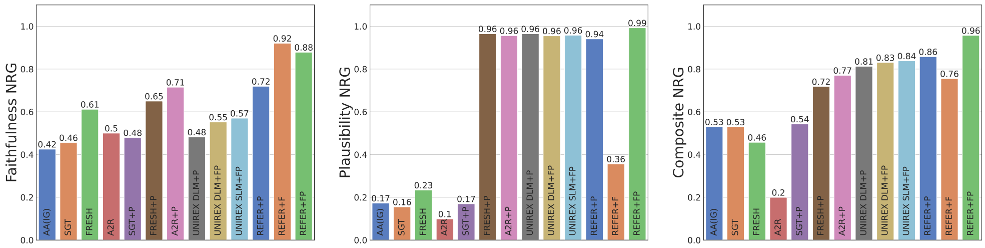


Figure 6.1: Comparison of models w.r.t. faithfulness NRG (FNRG), plausibility NRG (PNRG), and composite NRG (CNRG). +P, +F, +FP indicate whether the model was regularized for plausibility, faithfulness, or both.

plausibility (REFER+P and A2R+P), their faithfulness NRG scores remain considerably higher than all other methods. Detailed results are shown in table 6.2 and table 6.3. Additionally, we analyzed the model's predictions on correctly labeled instances compared to falsely labeled ones, as presented in Table 6.1. Surprisingly, although the model achieves relatively high plausibility scores, the sufficiency and comprehensiveness metrics are low when the model predicts the wrong label. This suggests that even when human rationales are extracted from the inputs, the model does not strongly rely on them in falsely labeled input.

Table 6.1: Comparison of ER metrics for truly predicted labels and falsely predicted labels from REFER. (↑) indicates the higher value is better and (↓) the lower is better.

Metrics	True Predictions	Wrong Predictions
Sufficiency AOPC (↓)	0.0488	0.1566
Comprehensiveness AOPC (↑)	0.3311	0.3057
Plausibility TF1 (↑)	0.8016	0.7012
Plausibility AUPRC (↑)	0.8834	0.7350

The extracted rationales by the model, shown in table 6.4, demonstrate the impact of regularization on explanation regularization. Without ER regularization, the model's reasoning tends to rely on specific data patterns and heuristics rather than meaningful explanations. In contrast, when the model is regularized on ER, the quality of the rationales improves significantly in terms of faithfulness and plausibility. For instance, the example highlights the selection of ”man pushing cart” and ”woman smoking cigarette” as rationales

Table 6.2: Benchmark on CoS-E dataset. Results of the baselines are obtained from the work done by [13].

Configuration		Faithfulness			Plausibility			Task		Composite
Model	End-to-End	Comp (↑)	Suff (↓)	FNRG	TF1 (↑)	AUPRC (↑)	PNRG	Accuracy (↑)	TNRG	CNRG
AA(IG)	FALSE	0.2160	0.3780	0.3306	0.4834	0.4007	0.2935	63.56	0.9772	0.5337
SGT	FALSE	0.1970	0.3240	0.3699	0.5100	0.4368	0.3702	64.35	0.9950	0.5783
FRESH	FALSE	0.0370	0.0000	0.5463	0.3937	0.3235	0.0849	24.81	0.1007	0.2439
A2R	FALSE	0.0140	0.0000	0.5167	0.3312	0.4161	0.1041	21.77	0.0319	0.2176
SGT+P	FALSE	0.2010	0.3280	0.3703	0.4795	0.413	0.3020	**64.57**	**1.0000**	0.5574
FRESH+P	FALSE	0.0130	0.0130	0.5001	0.6976	0.7607	0.9890	20.36	0.0000	0.4964
A2R+P	FALSE	0.0010	0.0000	0.5000	0.6763	0.7359	0.9322	20.91	0.0124	0.4816
UNIREX (DLM+P)	FALSE	0.1800	0.3900	0.2702	0.6976	0.7607	0.9890	64.13	0.9900	0.7497
UNIREX (DLM+FP)	FALSE	0.2930	0.3210	0.4968	0.6952	0.7638	0.9892	62.5	0.9532	**0.8131**
UNIREX (SLM+FP)	FALSE	0.3900	0.4240	0.5000	0.6925	0.7512	0.9714	62.09	0.9439	0.8051
REFER+P	TRUE	0.1831	0.2098	0.4867	**0.6994**	**0.7683**	**1.0000**	61.35	0.9272	0.8046
REFER+F	TRUE	**0.2798**	**0.0000**	**0.8584**	0.3835	0.6691	0.4595	63.21	0.9692	0.7624
REFER+FP	TRUE	0.1206	0.1489	0.4781	0.6881	0.7393	0.9521	64.23	0.9923	0.8075

Table 6.3: Benchmark on e-SNLI dataset. Results of the baselines are obtained from the work done by [13].

Configuration		Faithfulness			Plausibility			Task		Composite
Model	End-to-End	Comp (↑)	Suff (↓)	FNRG	TF1 (↑)	AUPRC (↑)	PNRG	Macro F1 (↑)	TNRG	CNRG
AA(IG)	FALSE	0.3080	0.4140	0.4250	0.3787	0.4783	0.1728	90.78	0.9909	0.5296
SGT	FALSE	0.2880	0.3610	0.4557	0.4170	0.4246	0.1551	90.23	0.9766	0.5291
FRESH	FALSE	0.1200	0.0000	0.6117	0.5371	0.3877	0.2337	72.92	0.5259	0.4571
A2R	FALSE	0.0530	0.0000	0.5000	0.2954	0.4848	0.0989	52.72	0.0000	0.1996
SGT+P	FALSE	0.2860	0.3390	0.4789	0.4259	0.4303	0.1696	90.36	0.9800	0.5428
FRESH+P	FALSE	0.1430	0.0000	0.6500	0.7763	0.8785	0.9649	73.44	0.5394	0.7181
A2R+P	FALSE	0.1820	0.0000	0.7150	0.7731	0.873	0.9562	77.31	0.6402	0.7705
UNIREX (DLM+P)	FALSE	0.3110	0.3710	0.4819	0.7763	0.8785	0.9649	90.8	0.9914	0.8127
UNIREX (DLM+FP)	FALSE	0.3350	0.3460	0.5521	0.7753	0.8699	0.9552	90.51	0.9839	0.8304
UNIREX (SLM+FP)	FALSE	0.3530	0.3560	0.5700	0.7722	0.8758	0.9582	90.59	0.9859	0.8381
REFER+P	TRUE	0.3127	0.1768	0.7193	0.7909	0.8411	0.9409	87.81	0.9136	0.8579
REFER+F	TRUE	**0.3054**	**0.0000**	**0.9207**	0.4443	0.5958	0.3559	90.69	0.9885	0.7551
REFER+FP	TRUE	0.3091	0.0399	0.8786	**0.8126**	**0.8713**	**0.9927**	**91.13**	**1.0000**	**0.9571**

to predict the label contradiction. The evaluation metrics for faithfulness on e-SNLI in table 6.6 further support the notion that the model genuinely relies on these rationales for its predictions.

fig. 6.2 shows the distribution of the results for different combinations of faithfulness and plausibility loss weights on the CoS-E validation set. We trained the model for $(\alpha_f, \alpha_p) \in \{0.0, 0.5, 1.0\}^2$. Based on the results, there is a slight reverse correlation between plausibility and faithfulness. However, the task shows relatively stable behavior over faithfulness and plausibility variation. This means that, with our pipeline, we cannot reach a higher plausibility and faithfulness trade-off from a certain level on CoS-E.

We conducted experiments to investigate how our model behaves when different percentages of human-annotated data are included in the training set.

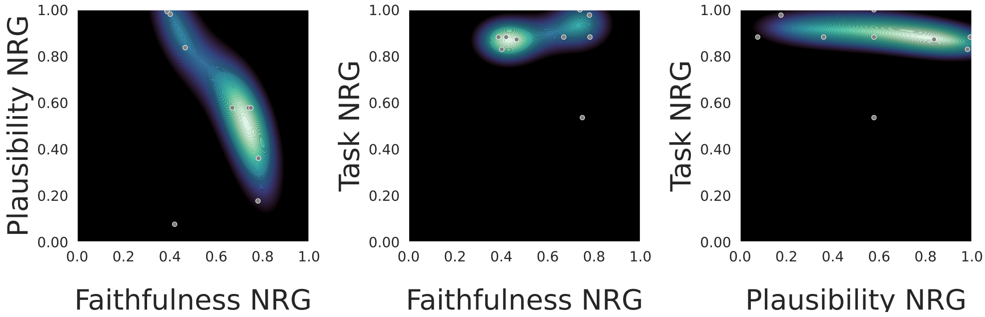


Figure 6.2: REFER results distribution of CoS-E dev split for different faithfulness and plausibility weights and k=50%. Kernel Density Estimation is used to have smoothed distribution over discrete data points for visualization purposes.

Table 6.4: REFER highlights on e-SNLI. Instead of visualizing hard tokens selected by the model, we highlighted all the words w.r.t. their score.

Model	Highlights
Original Instance	***Premise***: A man in green pants and blue shirt pushing a cart. ***Hypothesis***: A woman is smoking a cigarette. ***Label***: contradiction
REFER without ER regularization	***Premise***: A man in green pants and blue shirt pushing a cart . ***Hypothesis***: A woman is smoking a cigarette . ***Predict***: contradiction
REFER with ER regularization	***Premise***: A man in green pants and blue shirt pushing a cart . ***Hypothesis***: A woman is smoking a cigarette . ***Predict***: contradiction

fig. 6.3 showcases the outcomes obtained for all training criteria when varying percentages of human annotation were used: 0.1%, 1%, 10%, 20%, 50%, and 100%. The results indicate that until 10% of the data is annotated by humans, the plausibility remains consistent. On the other hand, REFER achieves comparable plausibility to 100% human supervision with just 50% of human annotation. This means REFER enables effective plausibility optimizations using minimal gold rationale supervision. In contrast, task performance is reduced by increasing the human rationale supervision since the model should learn from human highlights instead of repetitive patterns. Faithfulness does not exhibit a clear relationship with the availability of gold rationales, as it relies on the model's intrinsic features rather than human-provided rationales.

table 6.5 and table 6.6 show the REFER results on ID and OOD datasets. In both Tables REFER is trained on the ID dataset and evaluated over ID and

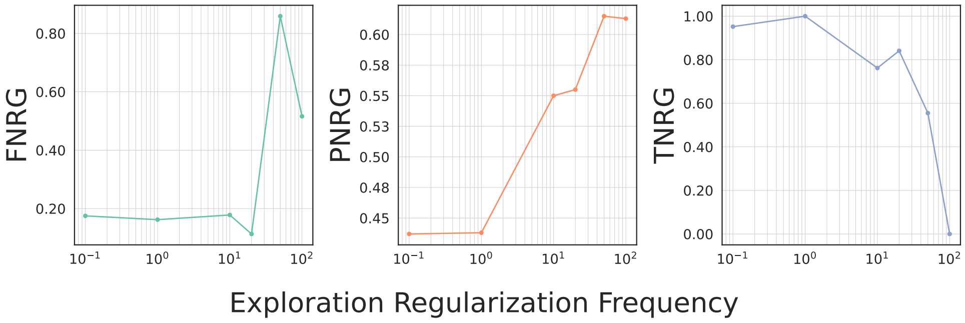


Figure 6.3: Comaprioson of different models w.r.t. faithfulness NRG (FNRG), plausibility NRG (PNRG), and composite NRG (CNRG).

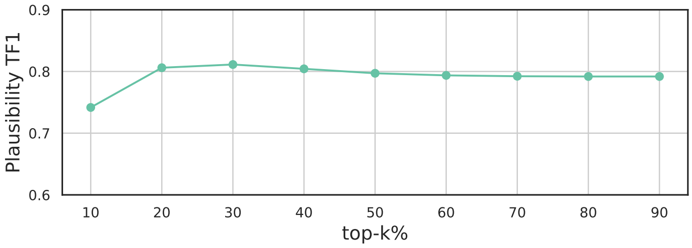


Figure 6.4: Plausiblity TF1 score of model trained for top-50% and evaluated for other top-k%s.

OOD sets. We consider the results from table 6.5 as the baseline and analyze the effect of ER regularization in table 6.6. When we train the model with explanation regularization, faithfulness and sufficiency are enhanced. On MNLI, sufficiency improves from 0.206 to 0.109, while on HANS, it goes from 0.249 to 0.071. Regarding Comprehensiveness, training the model along with ER regularization improves the baseline from 0.212 to 0.310 on MNLI and from 0.272 to 0.320 on HANS. Besides, results on e-MNLI in table 6.6 show that the plausibility of OOD is significant and comparable to the ID data. Similarly, the comprehensiveness and sufficiency improve on both MNLI-Contrast and MNLI-Original. However, the results on MNLI-Original seem to be better, especially w.r.t task macro F1, which means the model performs equally well predicting different labels.

Another interesting finding is that the model trained for a specific top-k% performs well on other top-k% during inference w.r.t. plausibility. fig. 6.4

Table 6.5: Comparison of the performance of REFER without explanation regularization on ID and OOD dataset.

Metrics	ID without ER regularization	OOD Datasets			Contrast Test	
	e-SNLI	MNLI	HANS	e-MNLI	MNLI-Contrast	MNLI-Original
Task Accuracy (↑)	90.47	74.65	67.09	76.00	82.66	88.72
Task Macro F1 (↑)	90.48	74.80	28.57	75.93	60.25	88.74
Sufficiency AOPC (↓)	0.205	0.206	0.305	0.249	0.226	0.201
Comprehensiveness AOPC (↑)	0.243	0.212	0.272	0.224	0.210	0.249
Plausibility TF1 (↑)	0.254	N/A	N/A	0.197	N/A	N/A
Plausibility AUPRC (↑)	0.211	N/A	N/A	0.167	N/A	N/A

Table 6.6: Comparison of the performance of REFER with explanation regularization on ID and OOD dataset.

Metrics	ID with ER regularization	OOD Datasets			Contrast Test	
	e-SNLI	MNLI	HANS	e-MNLI	MNLI-Contrast	MNLI-Original
Task Accuracy (↑)	90.33	74.10	66.06	78.00	82.11	88.37
Task Macro F1 (↑)	90.36	74.13	27.75	78.11	59.92	88.44
Sufficiency AOPC (↓)	0.059	0.109	0.071	0.100	0.091	0.050
Comprehensiveness AOPC (↑)	0.329	0.310	0.320	0.315	0.321	0.329
Plausibility TF1 (↑)	0.792	N/A	N/A	0.616	N/A	N/A
Plausibility AUPRC (↑)	0.869	N/A	N/A	0.445	N/A	N/A

display roughly stable behavior of the model trained for top-50% and evaluated for other top-k% w.r.t. plausibility TF1. This means the model tends to select rationales among human highlights even with a low number of k. table 6.7 illustrates the rationale selected by the model trained for top-50% and evaluated for different ks.

Table 6.7: Comparison of rationales extracted by REFER trained on k=50%. We forced the model for other k to see how it selects rationales.

Dataset	Test Instance
Gold	***Premise***: a woman wearing a pink tank top holding a mug of liquid ***Hypothesis***: A woman in a blue tank top holding a car. ***Label***: contradiction
k=20%	***Premise***: a woman wearing a pink tank top holding a mug of liquid ***Hypothesis***: A woman in a blue tank top holding a car.
k=30%	***Premise***: a woman wearing a pink tank top holding a mug of liquid ***Hypothesis***: A woman in a blue tank top holding a car.
k=40%	***Premise***: a woman wearing a pink tank top holding a mug of liquid ***Hypothesis***:A woman in a blue tank top holding a car.
k=50%	***Premise***: a woman wearing a pink tank top holding a mug of liquid ***Hypothesis***: A woman in a blue tank top holding a car.
k=60%	***Premise***: a woman wearing a pink tank top holding a mug of liquid ***Hypothesis***: A woman in a blue tank top holding a car.

Chapter 7

Conclusions

In this thesis, we have presented a comprehensive overview of explainability techniques for NLMs. We summarize methods for local and global explanations based on the model training perspective. We also discuss using explanations to improve models, evaluation, and key challenges. Major future development options include developing explanation methods tailored to different NLMs, evaluating explanation faithfulness, and improving human interpretability. Then, we proposed REFER, a rationale extraction framework that jointly trains the task model and the rationale extractor to optimize downstream task performance, faithfulness, and plausibility. Being fully end-to-end, thanks to Adaptive Implicit Maximum Likelihood Estimation [67], enables the task model and the rationale extractor to be jointly optimized for these criteria, therefore aware of each other behavior and adopting their parameter to improve their performance and obtain a better balance. We then analyze several aspects of the rationale extraction process, investigating how human rationales affect the model behavior; how the model can imitate human-generated rationales; and to what extent the learned models can generalize on OOD datasets. Finally, we compare REFER performance with other methods and architectures and illustrate that our model outperforms previous models in most cases.

This thesis served as the basis for a research article, *”REFER: An End-to-end Rationale Extraction Framework for Explanation Regularization* [33],” presented at the CoNLL/EMNLP 2023 conference. I was funded by the *Thesis Abroad* Scholarship from the Department of Computer Science and Engineering (DISI) at the University of Bologna and the work was done while I was at the University of Edinburgh.

Bibliography

[1] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(61):1803–1831, 2010. url: `http://jmlr.org/papers/v11/baehrens10a.html`.

[2] J. Bastings, S. Ebert, P. Zablotskaia, A. Sandholm, and K. Filippova. ”will you find these shortcuts?” a protocol for evaluating the faithfulness of input salience methods for text classification, 2022. arXiv: `2111.07367 [cs.CL]`.

[3] J. Bastings and K. Filippova. The elephant in the interpretability room: why use attention as explanation when we have saliency methods? In A. Alishahi, Y. Belinkov, G. Chrupała, D. Hupkes, Y. Pinter, and H. Sajjad, editors, *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics, November 2020. doi: `10.18653/v1/2020.blackboxnlp-1.14`. url: `https://aclanthology.org/2020.blackboxnlp-1.14`.

[4] A. Bau, Y. Belinkov, H. Sajjad, N. Durrani, F. Dalvi, and J. Glass. Identifying and controlling important neurons in neural machine translation, 2018. arXiv: `1811.01157 [cs.CL]`.

[5] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: can language models be too big? In

Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 610–623, 2021.

[6] M. M. Bhat, A. Sordoni, and S. Mukherjee. Self-training with few-shot rationalization. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10702–10712, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics, November 2021. doi: 10.18653/v1/2021.emnlp-main.836. url: https://aclanthology.org/2021.emnlp-main.836.

[7] T. Blevins, H. Gonen, and L. Zettlemoyer. Prompting language models for linguistic structure, 2023. arXiv: 2211.07830 [cs.CL].

[8] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics, September 2015. doi: 10.18653/v1/D15-1075. url: https://aclanthology.org/D15-1075.

[9] J. Buolamwini and T. Gebru. Gender shades: intersectional accuracy disparities in commercial gender classification. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018. url: https://proceedings.mlr.press/v81/buolamwini18a.html.

[10] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom. E-snli: natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. url: https://

proceedings.neurips.cc/paper_files/paper/2018/file/4c7a167bb329bd92580a99ce422d6fa6-Paper.pdf.

[11] S. Carton, Q. Mei, and P. Resnick. Extractive adversarial networks: high-recall explanations for identifying personal attacks in social media posts. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3497–3507, Brussels, Belgium. Association for Computational Linguistics, October 2018. doi: 10.18653/v1/D18-1386. url: https://aclanthology.org/D18-1386.

[12] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine learning interpretability: a survey on methods and metrics. *Electronics*, 8(8), 2019. issn: 2079-9292. doi: 10.3390/electronics8080832. url: https://www.mdpi.com/2079-9292/8/8/832.

[13] A. Chan, M. Sanjabi, L. Mathias, L. Tan, S. Nie, X. Peng, X. Ren, and H. Firooz. UNIREX: a unified learning framework for language model rationale extraction. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 51–67, virtual+Dublin. Association for Computational Linguistics, May 2022. doi: 10.18653/v1/2022.bigscience-1.5. url: https://aclanthology.org/2022.bigscience-1.5.

[14] A. Chan, M. Sanjabi, L. Mathias, L. Tan, S. Nie, X. Peng, X. Ren, and H. Firooz. UNIREX: a unified learning framework for language model rationale extraction. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2867–2889. PMLR, 17-23 Jul 2022. url: https://proceedings.mlr.press/v162/chan22a.html.

[15] A. Chan, J. Xu, B. Long, S. Sanyal, T. Gupta, and X. Ren. Salkg: learning from knowledge graph explanations for commonsense reasoning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18241–18255. Curran Associates, Inc., 2021. url: https://proceedings.neurips.cc/paper_files/paper/2021/file/9752d873fa71c19dc602bf2a0696f9b5-Paper.pdf.

[16] D. Chen, J. Bolton, and C. D. Manning. A thorough examination of the CNN/Daily Mail reading comprehension task. In K. Erk and N. A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics, August 2016. doi: 10.18653/v1/P16-1223. url: https://aclanthology.org/P16-1223.

[17] H. Chen and Y. Ji. Adversarial training for improving model robustness? look at both prediction and interpretation. 36:10463–10472, June 2022. doi: 10.1609/aaai.v36i10.21289. url: https://ojs.aaai.org/index.php/AAAI/article/view/21289.

[18] S. Chen, D. Khashabi, W. Yin, C. Callison-Burch, and D. Roth. Seeing things from a different angle:discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, Minneapolis, Minnesota. Association for Computational Linguistics, June 2019. doi: 10.18653/v1/N19-1053. url: https://aclanthology.org/N19-1053.

[19] G. Chrysostomou and N. Aletras. Improving the faithfulness of attention-based explanations with task-specific information for text classification. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings*

of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 477–488, Online. Association for Computational Linguistics, August 2021. doi: 10.18653/v1/2021.acl-long.40. url: https://aclanthology.org/2021.acl-long.40.

[20] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen. A survey of the state of explainable ai for natural language processing, 2020. arXiv: 2010.00711 [cs.CL].

[21] I. Dasgupta, D. Guo, A. Stuhlmüller, S. Gershman, and N. D. Goodman. Evaluating compositionality in sentence embeddings. In C. Kalish, M. A. Rau, X. (Zhu, and T. T. Rogers, editors, *Proceedings of the 40th Annual Meeting of the Cognitive Science Society, CogSci 2018, Madison, WI, USA, July 25-28, 2018*. cognitivesciencesociety.org, 2018. url: https://mindmodeling.org/cogsci2018/papers/0307/index.html.

[22] S. Dasgupta, N. Frost, and M. Moshkovitz. Framework for evaluating faithfulness of local explanations, 2022. arXiv: 2202.00734 [cs.LG].

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics, June 2019. doi: 10.18653/v1/N19-1423. url: https://aclanthology.org/N19-1423.

[24] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace. ERASER: A benchmark to evaluate rationalized NLP

models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics, July 2020. doi: 10.18653/v1/2020.acl-main.408. url: https://aclanthology.org/2020.acl-main.408.

[25] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning, 2017. arXiv: 1702.08608 [stat.ML].

[26] J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018. doi: 10.1126/sciadv.aao5580. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.aao5580. url: https://www.science.org/doi/abs/10.1126/sciadv.aao5580.

[27] M. Du, F. He, N. Zou, D. Tao, and X. Hu. Shortcut learning of large language models in natural language understanding, 2023. arXiv: 2208.11857 [cs.CL].

[28] M. Du, N. Liu, and X. Hu. Techniques for interpretable machine learning, 2019. arXiv: 1808.00033 [cs.LG].

[29] J. Enguehard. Sequential integrated gradients: a simple but effective method for explaining language models, 2023. arXiv: 2305.15853 [cs.CL].

[30] W. Falcon. Pytorch lightning, 2019. url: https://cir.nii.ac.jp/crid/1370013168774120069.

[31] S. Feng, E. Wallace, A. Grissom II, M. Iyyer, P. Rodriguez, and J. Boyd-Graber. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics, October 2018. doi: 10.18653/v1/D18-1407. url: https://aclanthology.org/D18-1407.

[32] M. Gardner, Y. Artzi, V. Basmov, J. Berant, B. Bogin, S. Chen, P. Dasigi, D. Dua, Y. Elazar, A. Gottumukkala, N. Gupta, H. Hajishirzi, G. Ilharco, D. Khashabi, K. Lin, J. Liu, N. F. Liu, P. Mulcaire, Q. Ning, S. Singh, N. A. Smith, S. Subramanian, R. Tsarfaty, E. Wallace, A. Zhang, and B. Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics, November 2020. doi: 10.18653/v1/2020.findings-emnlp.117. url: https://aclanthology.org/2020.findings-emnlp.117.

[33] M. R. Ghasemi Madani and P. Minervini. REFER: an end-to-end rationale extraction framework for explanation regularization. In J. Jiang, D. Reitter, and S. Deng, editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, Singapore. Association for Computational Linguistics, December 2023. doi: 10.18653/v1/2023.conll-1.40. url: https://aclanthology.org/2023.conll-1.40.

[34] M. Glockner, V. Shwartz, and Y. Goldberg. Breaking nli systems with sentences that require simple lexical inferences, 2018. arXiv: 1805.02266 [cs.CL].

[35] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics, June 2018. doi: 10.18653/v1/N18-2017. url: https://aclanthology.org/N18-2017.

[36] A. A. Ismail, H. Corrada Bravo, and S. Feizi. Improving deep learning interpretability by saliency guided training. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26726–26739. Curran Associates, Inc., 2021.

[37] A. Jacovi and Y. Goldberg. Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics, July 2020. doi: `10.18653/v1/2020.acl-main.386`. url: `https://aclanthology.org/2020.acl-main.386`.

[38] S. Jain and B. C. Wallace. Attention is not explanation, 2019. arXiv: `1902.10186 [cs.CL]`.

[39] S. Jain, S. Wiegreffe, Y. Pinter, and B. C. Wallace. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics, July 2020. doi: `10.18653/v1/2020.acl-main.409`. url: `https://aclanthology.org/2020.acl-main.409`.

[40] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax, 2017. arXiv: `1611.01144 [stat.ML]`.

[41] B. Joshi, A. Chan, Z. Liu, S. Nie, M. Sanjabi, H. Firooz, and X. Ren. Er-test: evaluating explanation regularization methods for language models, 2023. arXiv: `2205.12542 [cs.CL]`.

[42] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. The (un)reliability of saliency methods, 2017. arXiv: `1711.00867 [stat.ML]`.

[43] E. Kokalj, B. Škrlj, N. Lavrač, S. Pollak, and M. Robnik-Šikonja. BERT meets shapley: extending SHAP explanations to transformer-based classifiers. In H. Toivonen and M. Boggia, editors, *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 16–21, Online. Association for Computational Linguistics, April 2021. url: `https://aclanthology.org/2021.hackashop-1.3`.

[44] S. Krishna, J. Ma, D. Slack, A. Ghandeharioun, S. Singh, and H. Lakkaraju. Post hoc explanations of language models can improve language models, 2023. arXiv: `2305.11426 [cs.CL]`.

[45] S. Kumar and P. Talukdar. NILE : natural language inference with faithful natural language explanations. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics, July 2020. doi: `10.18653/v1/2020.acl-main.771`. url: `https://aclanthology.org/2020.acl-main.771`.

[46] K. Lakhotia, B. Paranjape, A. Ghoshal, S. Yih, Y. Mehdad, and S. Iyer. FiD-ex: improving sequence-to-sequence models for extractive rationale generation. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3712–3727, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics, November 2021. doi: `10.18653/v1/2021.emnlp-main.301`. url: `https://aclanthology.org/2021.emnlp-main.301`.

[47] D.-H. Lee, A. Kadakia, B. Joshi, A. Chan, Z. Liu, K. Narahari, T. Shibuya, R. Mitani, T. Sekiya, J. Pujara, and X. Ren. XMD: an end-to-end framework for interactive explanation-based debugging of NLP models. In D. Bollegala, R. Huang, and A. Ritter, editors, *Proceedings*

of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 264–273, Toronto, Canada. Association for Computational Linguistics, July 2023. doi: `10.18653/v1/2023.acl-demo.25`. url: `https://aclanthology.org/2023.acl-demo.25`.

[48] E. Lehman, J. DeYoung, R. Barzilay, and B. C. Wallace. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics, June 2019. doi: `10.18653/v1/N19-1371`. url: `https://aclanthology.org/N19-1371`.

[49] T. Lei, R. Barzilay, and T. Jaakkola. Rationalizing neural predictions. In J. Su, K. Duh, and X. Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics, November 2016. doi: `10.18653/v1/D16-1011`. url: `https://aclanthology.org/D16-1011`.

[50] C. Li, L. Shengshuo, Z. Liu, X. Wu, X. Zhou, and S. Steinert-Threlkeld. Linguistically-informed transformations (LIT): a method for automatically generating contrast sets. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 126–135, Online. Association for Computational Linguistics, November 2020. doi: `10.18653/v1/2020.blackboxnlp-1.12`. url: `https://aclanthology.org/2020.blackboxnlp-1.12`.

[51] J. Li, X. Chen, E. Hovy, and D. Jurafsky. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, pages 681–691, San Diego, California. Association for Computational Linguistics, June 2016. doi: 10.18653/v1/N16-1082. url: https://aclanthology.org/N16-1082.

[52] J. Li, W. Monroe, and D. Jurafsky. Understanding neural networks through representation erasure, 2017. arXiv: 1612.08220 [cs.CL].

[53] Y. Li, M. Du, R. Song, X. Wang, and Y. Wang. A survey on fairness in large language models, 2023. arXiv: 2308.10149 [cs.CL].

[54] W. Ling, D. Yogatama, C. Dyer, and P. Blunsom. Program induction by rationale generation: learning to solve and explain algebraic word problems. In R. Barzilay and M.-Y. Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics, July 2017. doi: 10.18653/v1/P17-1015. url: https://aclanthology.org/P17-1015.

[55] Z. C. Lipton. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43, September 2018. issn: 0001-0782. doi: 10.1145/3233231. url: https://doi.org/10.1145/3233231.

[56] J. Liu, T. Hu, Y. Zhang, X. Gai, Y. Feng, and Z. Liu. A chatgpt aided explainable framework for zero-shot medical image diagnosis, 2023. arXiv: 2307.01981 [eess.IV].

[57] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: how language models use long contexts, 2023. arXiv: 2307.03172 [cs.CL].

[58] Y. Liu, H. Li, Y. Guo, C. Kong, J. Li, and S. Wang. Rethinking attention-model explainability through faithfulness violation test. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*,

pages 13807–13824. PMLR, 17–23 Jul 2022. url: https://proceedings.mlr.press/v162/liu22i.html.

[59] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. url: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[60] D. D. Lundstrom, T. Huang, and M. Razaviyayn. A rigorous study of integrated gradients method and extensions to internal neuron attributions. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14485–14508. PMLR, 17–23 Jul 2022. url: https://proceedings.mlr.press/v162/lundstrom22a.html.

[61] S. Luo, H. Ivison, C. Han, and J. Poon. Local interpretations for explainable natural language processing: a survey, 2022. arXiv: 2103.11072 [cs.CL].

[62] J. Mandi, E. Demirović, P. J. Stuckey, and T. Guns. Smart predict-and-optimize for hard combinatorial optimization problems, 2019. arXiv: 1911.10092 [cs.LG].

[63] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee. Hatexplain: a benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875, May 2021. doi: 10.1609/aaai.v35i17.17745. url: https://ojs.aaai.org/index.php/AAAI/article/view/17745.

[64] R. T. McCoy and T. Linzen. Non-entailed subsequences as a challenge for natural language inference. *CoRR*, abs/1811.12112, 2018. arXiv: 1811.12112. url: http://arxiv.org/abs/1811.12112.

[65] T. McCoy, E. Pavlick, and T. Linzen. Right for the wrong reasons: diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics, July 2019. doi: 10.18653/v1/P19-1334. url: https://aclanthology.org/P19-1334.

[66] V. Miglani, N. Kokhlikyan, B. Alsallakh, M. Martin, and O. Reblitz-Richardson. Investigating saturation effects in integrated gradients, 2020. arXiv: 2010.12697 [cs.CV].

[67] P. Minervini, L. Franceschi, and M. Niepert. Adaptive perturbation-based gradient estimation for discrete latent variable models. In *AAAI*. AAAI Press, 2023.

[68] H. Mohebbi, A. Modarressi, and M. T. Pilehvar. Exploring the role of bert token representations to explain sentence probing results, 2021. arXiv: 2104.01477 [cs.CL].

[69] C. Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd edition, 2022. url: https://christophm.github.io/interpretable-ml-book.

[70] A. Naik, A. Ravichander, N. Sadeh, C. Rose, and G. Neubig. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics, August 2018. url: https://aclanthology.org/C18-1198.

[71] S. Narang, C. Raffel, K. Lee, A. Roberts, N. Fiedel, and K. Malkan. Wt5?! training text-to-text models to explain their predictions, 2020. arXiv: 2004.14546 [cs.CL].

[72] M. Niepert, P. Minervini, and L. Franceschi. Implicit mle: backpropagating through discrete exponential family distributions, 2021. arXiv: 2106.01798 [cs.LG].

[73] B. Paranjape, M. Joshi, J. Thickstun, H. Hajishirzi, and L. Zettlemoyer. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online. Association for Computational Linguistics, November 2020. doi: 10.18653/v1/2020.emnlp-main.153. url: https://aclanthology.org/2020.emnlp-main.153.

[74] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: an imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. url: https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[75] M. Paulus, D. Choi, D. Tarlow, A. Krause, and C. J. Maddison. Gradient estimation with stochastic softmax tricks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5691–5704. Curran Associates, Inc., 2020. url: https://proceedings.neurips.

cc/paper_files/paper/2020/file/3df80af53dce8435cf9ad6c3e7a403fd-Paper.pdf.

[76] D. Pruthi, R. Bansal, B. Dhingra, L. Baldini Soares, M. Collins, Z. C. Lipton, G. Neubig, and W. W. Cohen. Evaluating explanations: how much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375, 2022. doi: 10.1162/tacl_a_00465. url: https://aclanthology.org/2022.tacl-1.21.

[77] D. Pruthi, M. Gupta, B. Dhingra, G. Neubig, and Z. C. Lipton. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics, July 2020. doi: 10.18653/v1/2020.acl-main.432. url: https://aclanthology.org/2020.acl-main.432.

[78] L. Qiu, Y. Yang, C. C. Cao, J. Liu, Y. Zheng, H. H. T. Ngai, J. Hsiao, and L. Chen. Resisting out-of-distribution data problem in perturbation of xai, 2021. arXiv: 2107.14000 [cs.AI].

[79] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021. arXiv: 2103.00020 [cs.CV].

[80] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. arXiv: 1910.10683 [cs.LG].

[81] N. F. Rajani, B. McCann, C. Xiong, and R. Socher. Explain yourself! leveraging language models for commonsense reasoning. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational

Linguistics, July 2019. doi: 10.18653/v1/P19-1487. url: https://aclanthology.org/P19-1487.

[82] M. Ribeiro, S. Singh, and C. Guestrin. “why should I trust you?”: explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics, June 2016. doi: 10.18653/v1/N16-3020. url: https://aclanthology.org/N16-3020.

[83] M. T. Ribeiro, S. Singh, and C. Guestrin. ”why should i trust you?”: explaining the predictions of any classifier, 2016. arXiv: 1602.04938 [cs.LG].

[84] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019. arXiv: 1811.10154 [stat.ML].

[85] S. Sanyal and X. Ren. Discretized integrated gradients for explaining language models, 2021. arXiv: 2108.13654 [cs.CL].

[86] R. Schwarzenberg, N. Feldhus, and S. Möller. Efficient explanations from empirical explainers. In J. Bastings, Y. Belinkov, E. Dupoux, M. Giulianelli, D. Hupkes, Y. Pinter, and H. Sajjad, editors, *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 240–249, Punta Cana, Dominican Republic. Association for Computational Linguistics, November 2021. doi: 10.18653/v1/2021.blackboxnlp-1.17. url: https://aclanthology.org/2021.blackboxnlp-1.17.

[87] Y. Shen, L. Wang, Y. Chen, X. Xiao, J. Liu, and H. Wu. An interpretability evaluation benchmark for pre-trained language models, 2022. arXiv: 2207.13948 [cs.CL].

[88] S. Sikdar, P. Bhattacharya, and K. Heese. Integrated directional gradients: feature interaction attribution for neural NLP models. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 865–878, Online. Association for Computational Linguistics, August 2021. doi: 10.18653/v1/2021.acl-long.71. url: https://aclanthology.org/2021.acl-long.71.

[89] X. Situ, I. Zukerman, C. Paris, S. Maruf, and G. Haffari. Learning to explain: generating stable explanations fast. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5340–5355, Online. Association for Computational Linguistics, August 2021. doi: 10.18653/v1/2021.acl-long.415. url: https://aclanthology.org/2021.acl-long.415.

[90] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017. arXiv: 1706.03825. url: http://arxiv.org/abs/1706.03825.

[91] J. Stacey, Y. Belinkov, and M. Rei. Supervising model attention with human explanations for robust natural language inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11349–11357, June 2022. doi: 10.1609/aaai.v36i10.21386. url: https://ojs.aaai.org/index.php/AAAI/article/view/21386.

[92] H. Strobelt, S. Gehrmann, M. Behrisch, A. Perer, H. Pfister, and A. M. Rush. Seq2seq-vis: a visual debugging tool for sequence-to-sequence models, 2018. arXiv: 1804.09299 [cs.CL].

[93] J. Strout, Y. Zhang, and R. Mooney. Do human rationales improve machine explanations? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–62, Florence, Italy. Association for Computational Linguistics, August 2019. doi: `10.18653/v1/W19-4807`. url: `https://aclanthology.org/W19-4807`.

[94] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks, 2017. arXiv: `1703.01365 [cs.LG]`.

[95] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, June 2017. url: `https://proceedings.mlr.press/v70/sundararajan17a.html`.

[96] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. url: `https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf`.

[97] A. Talmor, J. Herzig, N. Lourie, and J. Berant. CommonsenseQA: a question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics, June 2019. doi: `10.18653/v1/N19-1421`. url: `https://aclanthology.org/N19-1421`.

[98] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[99] J. Vig and Y. Belinkov. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics, August 2019. doi: 10.18653/v1/W19-4808. url: https://aclanthology.org/W19-4808.

[100] M. Vlastelica, A. Paulus, V. Musil, G. Martius, and M. Rolínek. Differentiation of blackbox combinatorial solvers, 2020. arXiv: 1912.02175 [cs.LG].

[101] A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics, June 2018. doi: 10.18653/v1/N18-1101. url: https://aclanthology.org/N18-1101.

[102] Z. Wu, Y. Chen, B. Kao, and Q. Liu. Perturbed masking: parameter-free probing for analyzing and interpreting BERT. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics, July 2020. doi: 10.18653/v1/2020.acl-main.383. url: https://aclanthology.org/2020.acl-main.383.

[103] C. Yeh, Y. Chen, A. Wu, C. Chen, F. Viégas, and M. Wattenberg. Attentionviz: a global view of transformer attention, 2023. arXiv: 2305.03210 [cs.HC].

[104] M. Yu, S. Chang, Y. Zhang, and T. Jaakkola. Rethinking cooperative rationalization: introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103, Hong Kong, China. Association for Computational Linguistics, November 2019. doi: 10.18653/v1/D19-1420. url: https://aclanthology.org/D19-1420.

[105] M. Yu, Y. Zhang, S. Chang, and T. S. Jaakkola. Understanding interlocking dynamics of cooperative rationalization, 2021. arXiv: 2110.13880 [cs.LG].

[106] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed. Big bird: transformers for longer sequences. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc., 2020. url: https://proceedings.neurips.cc/paper_files/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf.

[107] O. Zaidan, J. Eisner, and C. Piatko. Using “annotator rationales” to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics, April 2007. url: https://aclanthology.org/N07-1033.

[108] Y. Zhang, I. Marshall, and B. C. Wallace. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 795–804, Austin, Texas. Association for Computational Linguistics, November 2016. doi: 10.18653/v1/D16-1076. url: https://aclanthology.org/D16-1076.