Investigating Unfaithful Behavior in Neural Rationale Models

by

Laura Eugenija Holvoet

To obtain the degree of Master of Science at Delft University of Technology,
to be defended publicly on Tuesday May 6, 2025

Student Number:	5841755
Supervisors:	Prof. Avishek Anand, Prof. Joost de Winter
Daily Supervisor:	Dr. Jurek Leonhardt
Faculty:	Mechanical Engineering, TU Delft
Department:	Cognitive Robotics (CoR)

TU Delft

Acknowledgments

Writing this thesis has been a valuable experience, during which I had the chance to deepen my knowledge of the field of Natural Language Processing and of Interpretable AI. I truly enjoyed working on this project and found this topic important and meaningful.

I would like to thank my supervisors Jurek Leonhardt, Avishek Anand and Joost de Winter. I want to thank Jurek and Avishek for introducing me to the topic, helping me develop the idea and guiding me along the way. I want to thank Jurek for the time he dedicated to our weekly meetings, for listening to my ideas and progress updates and for always providing feedback and advice regarding the implementation and the writing. I want to express my gratitude to my supervisor from the CoR department, professor Joost de Winter, who gave me the chance to pursue this thesis topic, showed interest in the subject and provided feedback on my literature review and my thesis.

Investigating Unfaithful Behavior in Neural Rationale Models

Laura Eugenija Holvoet
Supervisors: Prof. Avishek Anand, Prof. Joost de Winter, Dr. Jurek Leonhardt
TU Delft

***Abstract*—Numerous techniques have been developed in order to explain the reasoning process of black-box models. Among them is a class of models that are designed to be inherently interpretable: select-then-predict models (a.k.a. rationale-based models). These models are meant to explain their prediction by highlighting part of the input as evidence. The evidence, called the rationale, should consist of the most salient parts of the input text that contribute the most to the model's decision. However, according to some recent studies, these models are not truly interpretable, because they do not provide faithful explanations (i.e., explanations that accurately reflect the true reasoning process of the model). In this thesis we give a formal definition of the degree of unfaithfulness to quantify unfaithful behavior. Then, we introduce an experiment to test the faithfulness of select-then-predict models and prove that select-then-predict models can provide unfaithful rationales. Lastly, we introduce a loss function, which we call the unfaithfulness loss, which minimizes the degree of unfaithfulness of select-then-predict models and teaches them to produce more faithful and plausible rationales. The code to our experiments is available on github[1].**

I. Introduction

As the use of deep learning models continues to grow, model interpretability is becoming increasingly important in the field of Artificial Intelligence (AI) and Natural Language Processing (NLP). Nowadays, most deep learning models have a black-box architecture, which means that their reasoning process is unclear and that it is difficult for the user to understand how the models process and analyze their inputs to make a certain prediction. A rapid growth of model size and complexity makes it even more difficult to understand how these models make their predictions.

Although it is difficult to find a unanimous definition of interpretability, Miller [1] defines it as "the degree to which a human can understand the cause of a decision". In this work, we adhere to this definition and define it as the extent to which the reasoning behind model predictions can be explained and understood by humans.

The interpretability of machine learning models is important for multiple reasons. Understanding the reasoning process of black-box models can help us understand which parts of the input have the most influence on a prediction. It can help discover patterns and detect biases in the data that went unnoticed by the engineers, but that the models are able to pick up on [2]. Interpretability can help us increase trust in AI and build better and more robust models by understanding their strengths and weaknesses [3]. Its importance is also stressed in recent regulations, such as the EU AI act [4], which aims to ensure "that AI systems used in the EU are safe, transparent, traceable, non-discriminatory and environmentally friendly" [5]. These and similar regulations call for a better understanding of the AI tools that are being introduced into the market and show an increasing need for more explainable AI systems.

Rationale-based models (a.k.a., select-then-predict models) [6, 7] are a class of interpretable models that are meant to provide an explanation of their output by highlighting the parts of the input that are used to make the prediction. Such models usually consist of two smaller components: a selector and a predictor. The former extracts the most important sequences of text from the input and passes them to the predictor, which makes the final prediction based only on the selection, as can be seen in fig. 1.

Fig. 1: Basic architecture of select-then-predict models

The extracted sequences are called rationales and they are meant to explain the model's decision. Figure 2 provides an example of a rationale produced by a select-then-predict model, from the paper of Lei et al. [6].

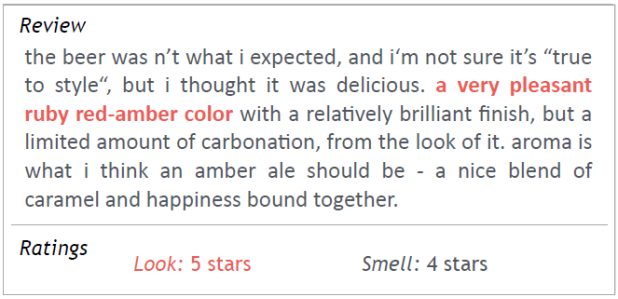


Fig. 2: Rationale provided by the model to explain the predicted rating of the look aspect of a beer review. The text highlighted in red represents the rationale chosen by the model [6].

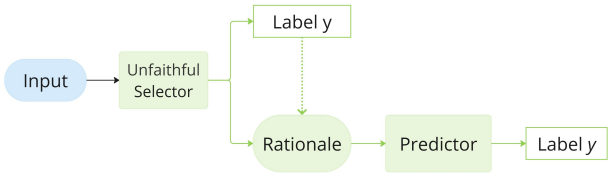
The fact that the predictor makes its decision based solely on the meaning of the words within the rationale is meant to

[1] https://github.com/lauraholv/MSc-Thesis

guarantee faithfulness, i.e. that the model did indeed rely only on the rationale to make the prediction and did not use some other information. However, prior research has presented the intuition that these select-then-predict models do not always produce faithful explanations and that they can sometimes encode the prediction in the selected rationale in a way that is not clear to humans [8, 9]. This can happen when the selector already makes the prediction (instead of only selecting the rationale, as intended) using the full input and communicates it via a hidden message to the predictor, which learns to extract the prediction from the message, as can be seen in fig. 3b. The main characteristic of these messages is that they encode the prediction in any way that is not connected to the semantic meaning of the terms in the rationale.

We assume that this unfaithful behavior is a consequence of the joint training of the selector and the predictor models, which is efficient and allows the rationale selection process to be learned in an unsupervised manner, but can lead to the above mentioned failure case [10]. This training configuration allows the selector to learn to make a prediction on the full input and to encode it in the rationale, while the predictor can learn to decode it. This behavior can appear because there are no explicit constraints that prevent it during training: the selector is trained to select a subset of the input (i.e. the rationale) and the predictor is taught to make a prediction based on it. However, there is no constraint to ensure that the prediction is made on the semantic meaning of the rationale and not on another signal encoded in the selection.

(a) A faithful select-then-predict model, where the selector faithfully chooses the most meaningful parts of the input as rationale and passes them to the predictor.



(b) An unfaithful select-then-predict model, where the selector makes the prediction, i.e. *label y*, on the full input and encodes it inside the selected rationale. The corresponding predictor extracts the encoding of *label y* and presents it as its prediction.

Fig. 3: Figures representing a faithful (a) and an unfaithful (b) select-then-predict model.

Such unfaithful explanations can be incorrect and misleading, since they do not actually clarify how the black-box model reached a specific prediction. However, if these explanations seem plausible enough, they might lead to misguided trust in the model's predictions and explanations as well as to poorly informed decision-making. Therefore it is crucial to study this failure case and to better understand how select-then-predict models work.

Some of these studies have also attempted to show that select-then-predict models are not faithful, however, the main focus of these studies was to provide a solution for unfaithful behavior, rather than to provide evidence of it. To the best of our knowledge, there is no comprehensive study that analyzes the faithfulness of select-then-predict models and the above mentioned failure case. Therefore, the goal of this thesis is to analyze unfaithful behavior in select-then-predict models. We provide a definition of unfaithful behavior, study and provide evidence of it and then attempt to analyze it further. Afterwards, we provide a way to mitigate it by introducing an additional loss term that minimizes the degree of unfaithfulness.

We formalize these goals in the following research questions:

RQ1: How can we formally define unfaithful behavior and is it possible to find evidence that select-then-predict models are unfaithful?

RQ2: If these select-then-predict models are unfaithful, can this phenomenon be spotted?

RQ3: How can unfaithful behavior be alleviated or avoided?

The rest of this paper is structured as follows: in section II we outline the related work. In section III, we explain the methodology of our research. Section IV covers the experimental setup, such as the dataset and hardware used, as well as the models that are used for our research. In section V we analyze the results of the experiments that were performed. Lastly, in section VI we present the conclusions and analyze the limitations of this work.

II. Related Work

In this section we will first give a broad overview of the field of interpretability, then we will present an overview of some select-then-predict model architectures and lastly we will illustrate existing literature that addresses the unfaithfulness of these models.

A. Interpretability

Over the years, various approaches aimed at interpreting black-box models have been developed. These approaches can be categorized into two groups: post-hoc interpretability methods and intrinsically interpretable models. Post-hoc interpretability refers to methods that provide an explanation of a model's prediction after inference. Many post-hoc methods are model-agnostic (i.e. they do not require a specific model architecture in order to generate explanations) and they include: gradient based methods [11, 12, 13, 14], surrogate model based methods [15] and attention based methods. Other post-hoc interpretability methods, such as SHAP (SHapley Additive exPlanations) by Lundberg and Lee [16] and Shapley Value Sampling [17, 18] aim to explain individual predictions of a model using an approach from game theory - Shapley values

[19]. They attempt to explain a prediction by computing the contribution of each feature to this prediction [2].

The survey by Madsen et al. [20] provides a more detailed overview of many other post-hoc interpretability methods and an in-depth explanation of their workings.

Intrinsically interpretable models (also called inherently interpretable models) are models that are interpretable by design. Some models can be considered inherently interpretable thanks to their transparent architecture, such as simpler ML models like linear regression models or decision trees, the reasoning process of which is more easily understandable to humans. Other models still involve black-box architectures, but are built to generate explanations while running inference, i.e. they are able to 'self-explain' while making predictions [21].

There is no agreement on which class of methods is better: inherently interpretable ones, or post-hoc interpretability methods. Madsen et al. [3] claim that these two paradigms are incompatible because, according to them, only inherently interpretable models are truly interpretable and provide faithful explanations, since they were built to do so, whereas post-hoc methods are not able to represent the complex functioning of black-box models and, therefore, do not guarantee faithful explanations [22]. But intrinsically interpretable models are not necessarily better: in some cases a trade-off can be observed between model interpretability and their performance [23, 24] and several studies show that these models can still produce unfaithful explanations, as will be discussed in section II-C.

B. Select-then-predict Models

Select-then-predict models, which are the subject of this thesis, are a class of inherently interpretable models. These models, also referred to as rationale-based models, provide natural language explanations to justify their predictions and to reflect their decision making process [25]. This approach is based on the human cognitive process of focusing on specific evidence to base a decision on. As the name suggests, select-then-predict models consist of two smaller components: a selector model, that processes the input text and extracts the most important and representative parts (i.e., the rationale) from it, and a predictor model that receives the rationale as its input and provides the final output. Since the rationale is the only input available to the predictor, it is considered to be the explanation of the model's output [21]. Select-then-predict models can be applied to various natural language processing tasks, such as machine reading comprehension, sentiment analysis, text classification, natural language inference etc., in order to make the predictions more interpretable.

Select-then-predict models are trained by jointly training the selector and the predictor components, which allows the rationale selection process to be learned in an unsupervised manner. This is beneficial because it allows the select-then-predict models to be trained end-to-end on the same data as the original full context models.

Zaidan et al. [26] were the first to introduce the use of rationales in machine learning. They use human-annotated rationales to design a more efficient training framework for ML models. These rationales serve as evidence, provided by the annotator, that supports the prediction. These additional annotations are intended to help the algorithm learn which features are actually responsible for the prediction and to help the model learn the decision making strategy of the annotator.

Although Zaidan et al. [26] introduced the use of rationales in the context of ML, Lei et al. [6] applied them in the context of interpretable ML models. Their goal is to train a model that learns to generate explanations, i.e. rationales, that are short and coherent, but that are also sufficient to make a prediction, therefore they build a select-then-predict model. The authors conclude that their select-then-predict model can extract quality rationales, achieving up to 96% precision in the rationale selection task. In addition, this select-then-predict model maintains a similar end-task performance as the model that uses the full input text.

Paranjape et al. [7] introduce another rationale-based model, that is meant to address the trade-off between short explanations and task accuracy. The approach, that the authors refer to as Sparse IB, is based on controlling rationale sparsity, using an Information Bottleneck principle; the goal is to select a rationale as a 'compressed' representation of the input that contains the minimal required information about the original input, but that is maximally informative about the final label. This training objective gives more control over the proportion of the input that is going to be selected as the rationale, controlled by the parameter π.

Numerous other variants of select-then-predict models have been developed over the years in the attempt to make neural language models more interpretable and their interpretations more faithful. Among these are the methods developed by Jain et al. [27], Yue et al. [28], Bastings et al. [29] and Yue et al. [30].

C. Unfaithfulness of Select-then-Predict Models

Several recent studies mention the possibility that the rationales provided by select-then-predict models are not faithful and that they do not always show which parts of the input the model focused on to make the prediction.

Zheng et al. [21] argue that select-then-predict models do not automatically imply inherent interpretability, since the selector and the predictor can exploit imperceptible messages to encode and communicate the prediction. In their paper, the authors show that the rationale selection process is not interpretable to humans and they call for a more rigorous analysis of the interpretability of neural rationale models.

Yu et al. [31] refer to this phenomenon as *rationale degeneration*, which occurs when the selector (generator) and the predictor employ a communication scheme in order to encode the predicted label in a human-imperceptible way. However, proving that select-then-predict models can provide unfaithful explanations was not the main goal of this paper, therefore the authors do not provide an in-depth description of this failure case and do not study it further.

Jacovi and Goldberg [9] also discuss the potential failure cases of select-then-predict models, which they define as

Trojan explanations and a *dominant selector*. According to the authors, a Trojan explanation is a rationale that contains information encoded in ways that are unintuitive or unclear to humans. The authors name several ways in which the predicted class could be encoded within the rationale, such as by using the location of the selected words within the full input (e.g. if the selected words are in the beginning of the input text, the predicted class is positive, otherwise it is negative, or vice versa), the number of selected tokens, or some kind of arbitrary token mapping (e.g., using periods to represent one class and commas to represent the other). This failure case corresponds to the unfaithful behavior described in other papers and to the behavior that we aim to study in this thesis.

Hu and Yu [32] address the problem of sub-optimal rationales (i.e., rationales that are not meaningful or informative and that do not align with human judgments) and of rationale failure (which we refer to as unfaithful behavior) and introduce a method called G-RAT (Guidance-based Rationalization). This method uses a two-module framework, consisting of a select-then-predict model and a guidance module, which regularizes the selector to prevent rationale failure and regularizes the predictor to avoid sub-optimal and non-informative rationales. The guidance model gives a weighted score and a prediction distribution. The weighted score consists of continuous importance scores for each token, used to teach the selector to pick semantically important tokens, similar to those chosen by the guidance model. The prediction distribution is used to avoid sub-optimal rationales and to ensure that the final prediction distribution of the select-then-predict model and the guidance model align. The Jensen-Shannon divergence [33] between these two distributions is minimized, which forces the prediction on the rationale to approximate the prediction on the full input. Experimental results show that this method is robust to both suboptimal rationale selection and to the problem of rationale failure.

III. Methodology

In this section we formally define select-then-predict models and provide a definition of unfaithfulness. After that, we describe the loss functions used to train our select-then-predict models: $\mathcal{L}_{claim}$, used to demonstrate the unfaithful behavior of select-then-predict models, and $\mathcal{L}_u$, used to minimize the unfaithfulness of these models.

A. Terminology

In this subsection we introduce some terms that we will use throughout the paper:

Unbiased predictor: a model that has been trained separately from the selector and therefore has not learned any potential encoding of the label. A full context model, trained on the full inputs can be considered an unbiased predictor.

Faithful selector/predictor: a model that performs the task 'assigned' to it: a faithful selector will only choose the rationale, whereas a faithful predictor will use the rationale to predict the final label. A faithful predictor is also an unbiased predictor.

Unfaithful selector/predictor: a model that learns to perform a different task than foreseen by the architecture and, therefore, provides unfaithful rationales.

B. Select-then-Predict Models and Definition of Unfaithfulness

As mentioned in section II-B, select-then-predict models consist of two components:

A selector model, characterized by the function $\psi(x)$ that takes the full input sequence (i.e., the entire sequence of tokens) $x = (x_1, ..., x_n)$ and outputs a binary token-level mask $\psi(x) = (z_1, ..., z_n)$, where $z_i \in \{0, 1\}$.

A predictor ϕ takes the masked input, i.e., the rationales, given by the element-wise multiplication of the mask and the original input $x \cdot \psi(x)$, and outputs the final label y:

$$y = \phi(x \cdot \psi(x)) \tag{1}$$

In order to answer our research questions, we first define what we consider to be *unfaithful behavior*. An important assumption we make is that unfaithfulness is not inherent in select-then-predict models, but is a learned behavior.

Such behavior can only be learned if the selector and the predictor models are trained jointly and the predictor can learn to decode the outputs of its corresponding selector. Therefore, two select-then-predict models might learn different communication schemes and not be able to understand each other's rationales. Similarly, an unbiased model (cf. section III-A) will also not be able to make correct predictions using only the rationales chosen by an unfaithful selector. Therefore, we refer to the difference in accuracy of a select-then-predict model and the accuracy of an unbiased predictor evaluated on the selector's rationales as the degree of unfaithfulness, formally defined in eq. (2).

Given a selector ψ and a predictor ϕ, we define unfaithfulness as:

$$u(\phi, \psi) = \text{acc}(\phi(x \cdot \psi(x))) - \text{acc}(\phi^*(x \cdot \psi(x))) \tag{2}$$

where ϕ^* is an *unbiased* predictor (cf. section III-A). The acc metric is defined in eq. (6). This is further illustrated in fig. 4.

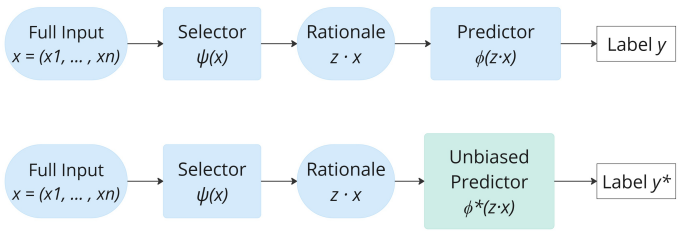


Fig. 4: On top: a typical select-then-predict model. On the bottom: evaluation of the rationales chosen by the selector using an unbiased predictor. The difference in accuracy of these two models represents the degree of unfaithfulness.

The degree of unfaithfulness can also show us whether the selected rationale is informative enough to predict the label. A high degree of unfaithfulness indicates that an unbiased predictor is not able to make a prediction based on the rationale alone, whereas an unfaithful predictor (cf. section III-A), trained together with the selector, is able to achieve above average performance. Therefore, a high unfaithfulness value shows that a selector and a predictor that were trained together have learned a common encoding of the prediction that is undecipherable to another, unbiased, model.

C. $\mathcal{L}_{claim}$ Regularizer

In order to verify whether select-then-predict models are able to provide unfaithful rationales, we introduce an experiment that consists of adding an additional loss term when training the select-then-predict models. The experiment is specifically designed for a scenario in which we can establish a subset of the input text that is essential to make a prediction. This can be a fact verification task, where the end-task is to predict whether a claim is supported or refuted by a document associated to it. Therefore, we use the FEVER dataset for our experiments: as explained in section IV-A, this task allows us to identify a part of the input that is always necessary to make a prediction: the claim.

The additional loss term, which we also refer to as a regularizer, penalizes the inclusion of tokens from the claim into the rationale and is defined as:

$$\mathcal{L}_{claim} = \sum_{i=1}^{i_{[SEP]}} z_i \tag{3}$$

Where $i_{[SEP]}$ refers to the index of the [SEP] token, which defines the end of the claim and the beginning of the context document and $z = (z_1, ..., z_i)$ is the mask given by the selector.

Because of how the FEVER dataset is made, the claim is necessary in order to predict whether the associated documents support it or not. Therefore, we can assume that a faithful selector (cf. section III-A) must select at least part of the claim, so that the predictor can classify it as supported or refuted (as was confirmed when we trained a FC model only on the context as input, as mentioned in section IV-B).

The purpose of this experiment is to verify whether the model is still able to perform well when a crucial part of the input is missing. If the model is able to achieve good (or above random) performance, without selecting the claim, the hypothesis that the selector model is making the prediction and encoding it in the rationale, is confirmed.

In section V-B we present and analyze the results of this experiment.

D. Unfaithfulness Loss

In order to decrease the unfaithfulness of select-then-predict models, we introduce an additional loss term, which we refer to as *unfaithfulness loss*, since it is meant to reduce the degree of unfaithfulness of these models. The unfaithfulness loss minimizes the difference between the prediction of the select-then-predict model and the prediction of an *unbiased* predictor. This loss term is defined in eq. (4):

$$\mathcal{L}_u = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4}$$

Where y_i are the output logits of the select-then-predict model that is being trained, whereas $\hat{y}_i$ represents the output logits of the unbiased model, evaluated on the rationale chosen by the selector and n is the number of samples. In this setup, the unbiased model is used during training, but its weights are not being updated; it is just being used to improve rationale quality.

This term is meant to help the model learn to be more interpretable and to select rationales that contain enough information to make a prediction. It discourages the selector and the predictor from learning a common encoding of the label, which would be unintelligible to the unbiased model and, therefore, would impact its performance.[2]

IV. Experimental Setup

In this section, we will explain the experimental setup, such as the dataset, the baselines and details about model training.

A. Dataset

FEVER [34] is a fact extraction and verification dataset that consists of claims generated using sentences taken from Wikipedia articles and of a collection of source documents, used to verify if the claim is *supported*, *refuted* or if there is *not enough info*.

We chose this dataset because it allows us to clearly identify part of the input that is crucial in order to make the prediction, i.e. the claim. Therefore, if the selector omits the claim from the selected rationale, it strongly indicates unfaithful behavior.

We use the ERASER benchmark [35] version of this dataset that consists of a subset of the original dataset and contains only supported or refuted claims. The dataset consists of 97957 training samples, 70967 of which are labeled *'Supports'*, while the rest are labeled *'Refutes'*. We undersampled the training set in order to make it balanced and the final label distribution is shown in table I. Figure 5 presents an example of an instance from the dataset. The union of the claim and the associated document is used as input for the model.

Split	Supports	Refutes	Total
Train	26990	26990	53980
Val	3019	3103	6122
Test	3033	3078	6111

TABLE I: Label distribution in the FEVER dataset.

[2]As mentioned in section II-C, Hu and Yu [32] also use a guidance model in order to avoid rationale failure (i.e., unfaithful behavior). Our approach is different from theirs in that we minimize the difference between the predictions of the select-then-predict model and the unbiased predictor evaluated on the rationales. The guidance model in [32] is always evaluated on the full input and is used as a supervisory signal to determine which tokens are important and what the correct prediction should be.

Claim: "Harris Jayaraj is Indian."

Docs: "Harris Jayaraj -LRB- born 8 January 1975 -RRB- is an Indian film composer from Chennai , Tamil Nadu .
He composes scores and soundtracks predominantly for Tamil films , while also composed for a few films in Telugu and Hindi languages."

Label: SUPPORTS

Fig. 5: Example instance from the FEVER dataset. The claim, together with the text of the document is used as input to the models. The models are trained to predict whether the claim is supported or refuted by the document.

B. Baselines

We compare the performance of the select-then-predict models with a couple of baseline models.

The first baseline is a full-context (FC) model: we use BERT-base, a transformer-based language model developed by Devlin et al. [36] and pre-trained on corpora such as English Wikipedia and BooksCorpus. The BERT model is good at capturing contextual word representations. Thanks to its pre-training, it can be fine-tuned for various tasks, such as sentiment classification or question answering, by simply adding an additional output layer.

We fine-tune a BERT-base model on the FEVER dataset to perform the task on the full input, without providing any rationale. The BERT FC model achieves 0.92 accuracy on the test set.

In order to establish an upper performance limit on a subset of the full input, we train a BERT-base model only using gold rationales (ground truth rationales annotated by humans), provided in the FEVER dataset from the ERASER benchmark. The BERT FC Gold model achieves 0.94 accuracy on the test set.

Finally, in order to verify whether the documents in the FEVER dataset contain any patterns that correlate with the label that would enable the model to make a correct prediction even without seeing the claim, we train a model on the FEVER dataset using only the context documents as input, without the claim. This model achieves 0.53 validation accuracy and 0.56 accuracy on the test set indicating that there is little correlation between the document and the labels and that it is not possible to make a prediction based on the evidence documents alone (these results are shown in table II).

We fine-tune these models using the AdamW optimizer [37] and a linear learning rate scheduler with warmup. The initial learning rate is set to 5e-5. We use a batch size of 12, with 10 gradient accumulation steps. Further details regarding model training are provided in appendix B.

C. Select-then-Predict Model

For the select-then-predict model we adopt the Sparse IB model by Paranjape et al. [7]. We chose this model as a representative example of the select-then-predict architecture. This model has been used as a state-of-the-art rationale-based model in other works that analyze and implement the select-then-predict architecture (e.g., in [38]).

As explained in section II-B, this model aims to control rationale sparsity using an Information Bottleneck principle.

We use the code of Chen et al. [38] as a starting point for the implementation of our model. The selector and the predictor components of this model are pre-trained BERT-base models.

The model takes a tokenized text sequence $x = (x_1, \ldots, x_n)$ as input. The selector outputs contextualized token representations of the input sequence x. A linear layer, referred to as the *rep-to-logit layer*, is then applied to each token representation to produce token-level logits (log probabilities), that correspond to parameters of a Bernoulli distribution. During training, the selector samples a soft token-level mask $z^* = (z_1^*, \ldots, z_n^*) \in (0, 1)$ from the token-level logits in the Bernoulli distribution, using the Gumbel-Softmax trick [39]). A soft mask means that each token is assigned a value between 0 and 1, indicating how likely it is to be selected, rather than enforcing a binary choice. This soft mask makes the sampling process differentiable, which is important because it allows the model to be trained using gradient-based optimization. As a result, gradients can flow through the selection process and help improve the rationale selection during training.

This soft mask is then used during training to mask the input sentences and the element-wise multiplication $z^* \cdot x$ is used as input to the predictor. The following objective is optimized:

$$\mathcal{L}_{VIB} = -\log p(y|z^* \cdot x) + KL[p(z|x)||p(z)] \qquad (5)$$

The first term represents the prediction loss. It is a cross entropy loss that ensures that the model correctly predicts the label using only the rationale. The second term represents the sparsity loss, which ensures that the extracted rationale is concise and that it stays close to the predefined prior $p(z)$. By minimizing the KL divergence term, the model is encouraged to select only the most relevant parts of the input.

During inference, the top-k tokens are selected, where k is determined by the sparsity π and the resulting binary token level mask $z = (z_1, ...z_n) \in \{0, 1\}$ is passed to the predictor, which makes the final prediction.

We train several instances of this model, initialized with different random seeds[3]. We train models with two different values of the sparsity parameter π, which represents the fraction of the full input that is selected as rationale. We evaluate models that select 10% and 20% ($\pi = 0.1$ and $\pi = 0.2$ respectively) of the full input as rationale.

Seven training runs were conducted for the select-then-predict models without regularization for each value of the parameter π. For the select-then-predict models with regularization and for the faithful select-then-predict models we conducted five training runs for each value of π. The results of each run can be found in appendix C.

The select-then-predict models are trained using the same optimizer and hyperparameters as the baseline models. Ad-

[3]The random seed influences the initialization of the *rep-to-logit layer* that converts the token embeddings into token-level logits, from which the rationale mask is sampled, and the parameters of the sampling function, that selects the rationale.

ditional details regarding model training are provided in appendix B.

D. Evaluation of the Models

In order to evaluate the performance of select-then-predict models we evaluate two aspects: their performance on the end task and the quality of the selected rationales.

1) Evaluation of the predictive performance: In order to evaluate the predictive performance of these models we use the metrics commonly used for classification tasks: accuracy and F1 score, defined in eq. (6) and eq. (7):

$$\text{acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

where TP = true positives, TN = true negatives, FP = false positives and FN = false negatives.

$$\text{F1} = 2\,\frac{precision * recall}{precision + recall} \tag{7}$$

where precision and recall can be defined as:

$$\text{precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{9}$$

2) Evaluation of rationale quality: We evaluate the quality of the generated rationales using a metric proposed by DeYoung et al. [35]: the token-level IoU score (intersection over union). We measure it with respect to human annotated rationales, a.k.a., *gold rationales*. This score represents the overlap between the tokens that belong both to the gold rationale and to the selected rationale, divided by their union. A higher IoU score indicates that the model is selecting rationales that closely match human annotated gold rationales, whereas a low score shows little agreement with the human annotated ground truth. We refer to this score as *alignment*, because it illustrates the alignment with human rationales, and we define it in eq. (10).

$$\text{alignment} = \frac{|z \cap z^*|}{|z \cup z^*|} \tag{10}$$

However, as was also mentioned by DeYoung et al. [35], the agreement with human rationales is a better indicator of the plausibility of the rationales, i.e. whether they are seen as a good explanation by humans, rather than of their faithfulness. Therefore, a higher alignment score does not indicate that the model relied exclusively on the information contained within rationale to make the prediction. As a result, this metric alone is not a definitive measure of rationale quality and of their faithfulness. Not only does the alignment capture the plausibility of the models rather than their faithfulness, but using human annotated gold rationales also assumes that they are the ground truth answer. The rationales provided in this dataset are sufficient to make a prediction, but they are not necessarily comprehensive, i.e., they might not contain all the information related to the claim and the label. Appendix A illustrates two examples where the quality of human annotated rationales might impact the alignment score.

Therefore, in addition to evaluating the plausibility of rationales with the alignment score, we use the degree of unfaithfulness, defined in eq. (2) as a metric to evaluate the sufficiency of the rationales, i.e., to evaluate whether they contain enough information regarding the label and whether they are sufficient to make a prediction. As explained in section III-B, this is measured by testing whether an unbiased predictor is able to predict the correct label based on the rationales.

E. Hardware and Software Specifications

All the models were trained on the Delft Blue Supercomputer [40], on NVIDIA Tesla V100 GPUs. The models were trained on two GPUs used in parallel.

The models were built and trained using the Pytorch Lightning library [41], which is a PyTorch wrapper for building and training ML models.

V. Results

In this section we describe the experiments performed to answer the research questions and their results.

A. Performance of the Select-then-Predict Models

As mentioned in section IV-C, we train select-then-predict models with a Sparse IB architecture. The models are trained on the FEVER dataset, where their task is to predict whether a claim is supported or refuted by the associated documents.

Table II reports the performance of the select-then-predict models with the highest validation accuracy (table VII provides a comprehensive overview of the performance of all the models trained during our experiments). To distinguish the select-then-predict models trained without the additional loss terms: $\mathcal{L}_{claim}$ and $\mathcal{L}_u$, we call them 'simple' select-then-predict models.

Our experiments show high variation in performance for models initialized with different random seeds, as shown in fig. 6. The accuracy of models trained without regularizer ranges between 0.78 and 0.87 for $\pi = 0.1$ and between 0.86 and 0.88 for $\pi = 0.2$. This hints that, at every initialization, the models learn a different solution, meaning that they learn to select different parts of the input as rationales. In order to see if the models choose crucial parts of the input, we compute the percentage of the claim that is selected as rationale. We observe that this percentage varies significantly between models, ranging from 2% to 94% (this is also reported in fig. 10b). These results show that, rather than always selecting the parts of the input that are related to the claim, the models learn how to pass the label in a way that adheres to the constraints, but that does not explain how they reached the prediction.

Some of the models achieve good performance (between 0.85 and 0.86 validation accuracy and 0.75 and 0.77 test accuracy) while only selecting 2% - 4% of the claim. This

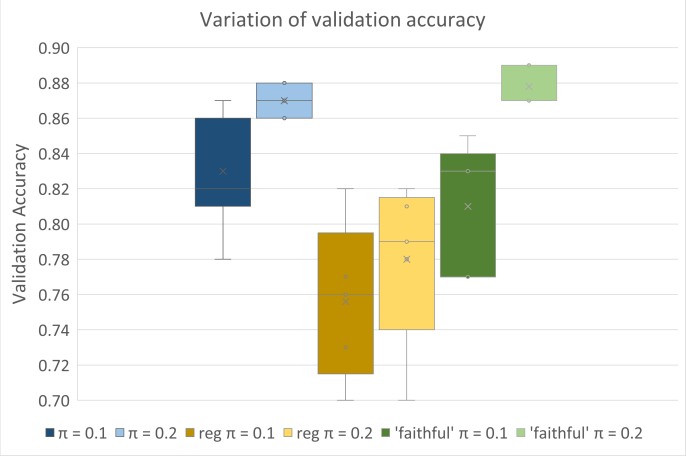


Fig. 6: Box plot representing the variation in validation accuracy of select-then-predict models on the FEVER dataset. The first two boxes (in blue) refer to select-then-predict models without regularizer trained with a sparsity value π = 0.1 and π = 0.2, respectively. The following two boxes (in yellow) refer to select-then-predict models trained with regularizer, with π = 0.1 and π = 0.2, respectively. The final two boxes (green) refer to models trained with the additional unfaithfulness loss, with π = 0.1 and π = 0.2, respectively.[4]

strongly suggests that the selector model is using the rationale to encode a prediction that it has already made, rather than choosing the most meaningful parts of the input, since a faithful predictor needs to see the claim to make a prediction.

B. Select-then-Predict Model with Regularizer

As explained in section III-C, in order to prove that select-then-predict models can provide unfaithful rationales, we train them with an additional loss term that teaches them to not select the claim as rationale. The total loss that this class of models was trained with is reported in eq. (11).

$$\mathcal{L}_{total} = \mathcal{L}_{VIB} + \mathcal{L}_{claim} \tag{11}$$

Table II reports the performance of the models trained with the regularizer (models with the highest validation accuracy are reported). We observe that the models trained to not select the claim still achieve above random performance: the models with the best validation performance achieve 0.82 accuracy on the validation set both for $\pi = 0.1$ and $\pi = 0.2$ and 0.80 and 0.78 on the test set, while selecting 0% of the claim. These accuracy values are comparable to the performance of models trained without regularizer. It is important to note that the *BERT FC no claim* baseline, which represents the full context model trained on the context documents without the claim, achieves 0.53 and 0.56 accuracy on the validation and the test sets respectively, showing that it is impossible to make a prediction based on the documents alone, without seeing the claim. Therefore, the high performance achieved by the select-then-predict models that do not select the claim, shows that the models are providing unfaithful rationales, that contain some encoding of the final label. This result confirms our hypothesis that the selector can make the prediction on the full input and encode it into the rationale.

C. Evaluating Rationale Quality

As mentioned in section IV-D, we evaluate the quality of the rationales by computing their alignment scores (defined in eq. (10)) as well as by computing the degree of unfaithfulness (eq. (2)) of the models. The BERT FC model is used as the unbiased model ϕ^*. Figure 4 illustrates this experiment.

The alignment scores are reported in table II. The models trained without regularizer present higher alignment scores than those trained with regularizer, meaning that the rationales of the models trained without $\mathcal{L}_{claim}$ match the gold rationales more closely and, therefore, are more plausible. However, a higher alignment does not imply a higher degree of faithfulness, as mentioned in section IV-D.

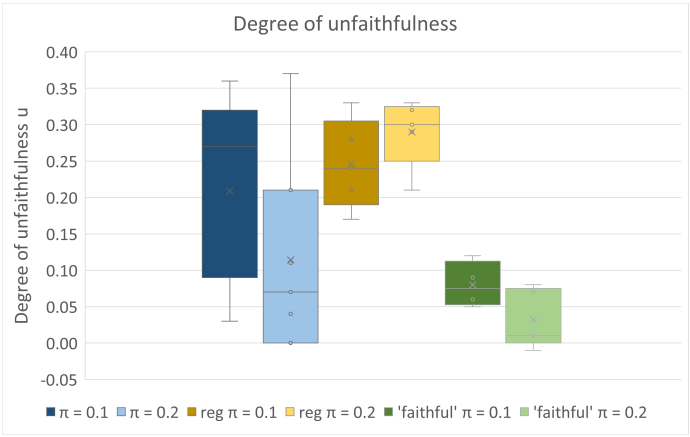


Fig. 7: Box plot representing the variation of the degree of unfaithfulness of the select-then-predict models on the FEVER dataset. The first two boxes (in blue) refer to select-then-predict models without regularizer trained with a sparsity value π = 0.1 and π = 0.2, respectively. The following two boxes (in yellow) refer to select-then-predict models trained with regularizer, with π = 0.1 and π = 0.2, respectively. The final two boxes (green) refer to models trained with the additional unfaithfulness loss, with π = 0.1 and π = 0.2, respectively.[4]

Table II reports the degrees of unfaithfulness of the models with highest validation accuracy and fig. 7 shows the variation of the degrees of unfaithfulness of all the models trained during our experiments. The models trained without $\mathcal{L}_{claim}$ present much lower degrees of unfaithfulness compared to the models trained with $\mathcal{L}_{claim}$. This shows that the rationales selected by the models without regularizer contain more information regarding the label and that an unbiased predictor is able to achieve good performance when evaluated on them. The rationales of the models trained with regularizer, on the

[4]In the box plots, the box represents the interquartile range, with the lower and upper edges corresponding to the first and third quartiles, respectively. The whiskers extend to the minimum and maximum values of the data. The x indicates the mean, whereas the line shows the median.

Model type	π	Validation Accuracy	Test Accuracy	F1	Percentage of claim selected	Accuracy $\phi*$	Unfaithfulness $u(\phi, \psi)$	Alignment
Baseline models								
Bert FC	1	0.93	0.92	0.92	-	-	-	-
Bert FC Gold	1	0.94	0.94	0.93	-	-	-	-
Bert FC no claim	1	0.53	0.56	-	-	-	-	-
'Simple' select-then-predict models without regularizer								
Sel Pred	0.1	0.87	0.84	0.85	90%	0.84	0.03	0.27
Sel Pred	0.2	0.88	0.89	0.89	93%	0.81	0.07	0.28
Select-then-predict models with regularizer								
Sel Pred + reg	0.1	0.82	0.80	0.80	0%	0.49	0.33	0.10
Sel Pred + reg	0.2	0.82	0.78	0.81	0%	0.49	0.33	0.15
'Faithful' select-then-predict models with unfaithfulness loss								
Sel Pred Faithful	0.1	0.85	0.80	0.84	80%	0.73	0.12	0.23
Sel Pred Faithful	0.2	0.89	0.90	0.89	78%	0.88	0.01	0.23

TABLE II: Performance of the models on the FEVER dataset. The performance of the models with the highest validation accuracy is reported. The 'simple' select-then-predict models are trained without regularization and without the unfaithfulness loss. The models with regularizer are trained with $\mathcal{L}_{claim}$ and the 'faithful' models are trained with $\mathcal{L}_u$.

other hand, obtain poor performance when evaluated on an unbiased predictor: 0.49 accuracy, which indicates that they are not sufficient to make a prediction. Figure 7 shows that, on average, the models trained with regularizer present higher degrees of unfaithfulness than the models trained without it.

Figure 8 shows the correlation between the percentage of the claim selected as rationale and the degree of unfaithfulness. The figure shows that models that select a smaller percentage of the claim as rationale exhibit a higher degree of unfaithfulness. This trend is apparent when evaluating the models trained with $\mathcal{L}_{claim}$, which are taught to not choose the claim as rationale. However, it can be observed that some of the select-then-predict models trained without $\mathcal{L}_{claim}$ also learn to select small percentages of the claim and present high degrees of unfaithfulness. This shows that the models are not only *able* to provide unfaithful rationales, when taught to do so by the $\mathcal{L}_{claim}$ term, but also that they sometimes learn this solution on their own, based on their initialization.

In addition, from this trend we can conclude that, while the select-then-predict models that select small parts of the claim are able to achieve good performance on the end task, their rationales do not contain enough information for an unbiased model to make a prediction, meaning that these rationales are not faithful and that the final label must be encoded in them.

These results show complicity between the selector and the predictor: when the models are trained jointly, they can learn a common pattern. However, this pattern is not understood by a model trained on the full inputs, which leads to a high degree of unfaithfulness. This is also in line with our assumption that unfaithfulness is a learned behavior, as mentioned in section III-B, and it can appear when the selector and the predictor are trained together.

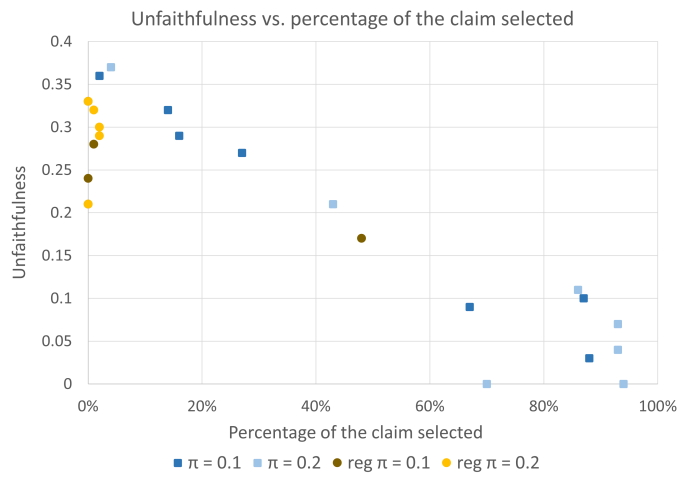


Fig. 8: Plot of the unfaithfulness against the percentage of the claim included in the rationale. Points corresponding to the legend π = *0.1* and π = *0.2* represent select-then-predict models trained without $\mathcal{L}_{claim}$, whereas *reg* π = *0.1* and *reg* π = *0.2* represent models trained with $\mathcal{L}_{claim}$.

D. Qualitative Analysis of Rationales

In addition to quantitatively evaluating the rationales by computing the alignment scores and the degree of unfaithfulness of the models, we perform a qualitative evaluation of the rationales. We analyze the rationales selected by the models that show the highest validation accuracy. The rationales are related to samples from the test set used to evaluate model performance. We analyze rationales generated by the select-then-predict models trained without $\mathcal{L}_{claim}$ and with it, and study the differences between them.

Table III shows examples of rationales generated by the select-then-predict models. The **Input** field is the full input to

the model, where the part between the [CLS] and the [SEP] tokens is the claim and the rest is the related document that contains information to support or to refute the claim. The rationale chosen by the models is highlighted in bold and can also be seen in the **Rationale** field.

Overall, it can be observed that the models trained without $\mathcal{L}_{claim}$ output more coherent rationales and select higher percentages of the claim (as can be seen in fig. 10b). On the other hand, the models trained with $\mathcal{L}_{claim}$ do not select the claim and choose rationales that are less coherent and do not contain sufficient information to make a prediction on the semantic meaning of the text alone, which is also shown by the high degrees of unfaithfulness of these models. This behavior is illustrated in table III and more examples of this can be found in appendix D.

In general, we found that the models often fail to select rationales that include the necessary information to support or to refute the claim, and instead tend to highlight irrelevant words. We observe that even the models trained without regularization tend to select words that are seemingly unrelated to the claim, such as the [SEP] token. In many cases these rationales do not fulfill their purpose of explaining model predictions to a human user and of making the model's reasoning process more understandable and transparent.

We also notice that the models trained with $\mathcal{L}_{claim}$ tend to select more nouns and numbers, and rarely select verbs as part of their rationale. Their rationales often contain repetitive words, as can be seen in Example 3 of table III and in the examples provided in appendix D. This makes them even less faithful and interpretable: the rationale *"papua indonesia guinea papua papua guinea papua indonesia papua papua"* does not contain any useful information regarding the claim, let alone why the model labeled it as *supports*. It is possible that this behavior contains a pattern used by the selector to encode the final label, however we have not been able to detect it. Currently, we do not have a definitive explanation for this behavior and further analysis is left for future work.

E. How Do the Models Encode Their Predictions?

During our analysis of the rationales selected by the models, we did not observe a specific pattern, which would allow us to understand how the label could be encoded within the rationale. Detecting such patterns is not a trivial task and there might be no definitive answer to this question. It is likely that the labels are encoded in a way that cannot be detected only by analyzing the words contained in the rationales or their embeddings. The label is likely to be contained in the binary mask that is passed by the selector to the predictor; however further analysis of the masks and the patterns that could be used to encode the label, is left for future work.

As explained in section V-G, we attempted to apply a post-hoc interpretability method in order to analyze the rationales and to understand which part of the rationales the models focus on. This, however, lead to inconclusive results and did not prove useful to understand how the label is encoded in the rationale.

F. Reducing the Unfaithfulness of Select-then-Predict Models

In order to avoid unfaithful behavior to decrease the degree of unfaithfulness of select-then-predict models, we train them with an additional loss term: the unfaithfulness loss, introduced in section III-D.

The total loss that these models are trained with is:

$$\mathcal{L}_{total} = \mathcal{L}_{VIB} + \mathcal{L}_u \tag{12}$$

The performance of the select-then-predict models trained with the additional unfaithfulness loss term is reported in table II, under the name 'faithful' select-then-predict models. With this name we do not imply that these models are entirely faithful, only that they are trained with $\mathcal{L}_u$.

The performance of these models in terms of accuracy remains similar to the ones trained without $\mathcal{L}_u$. However, it can be observed that the 'faithful' models present lower values of unfaithfulness. In table IV, average values of the degree of unfaithfulness and of the alignment score are presented per model class. On average, the 'faithful' models present an unfaithfulness value of 0.08 and 0.03 for $\pi = 0.1$ and $\pi = 0.2$, respectively, as opposed to 'simple' select-then-predict models, with average values of unfaithfulness equal to 0.21 and 0.11. Figure 7 also shows that 'faithful' select-then-predict models present much less variation of the degree of unfaithfulness, compared to the 'simple' select then predict models. This shows that the models trained with $\mathcal{L}_u$ learn to select more faithful rationales also over different initializations.

The higher alignment scores of the 'faithful' models show (cf. table IV) that their rationales are more plausible and align better with human annotations compared to the models trained without $\mathcal{L}_u$.

We also observe smaller variations in the percentage of the claim included in the rationale (this is illustrated in fig. 10b in appendix C). While for the 'simple' select-then-predict models, the variation is very high, ranging from 2% to 94% of the claim selected, this percentage is much smaller for models trained with $\mathcal{L}_u$ and it ranges between 60% and 98% of the claim. From this we can determine that these models are better at learning to distinguish and to select important parts of the input.

G. Post-hoc Interpretability of the Rationales

In order to understand how a prediction could be encoded within a rationale and to find a specific pattern, we attempted to analyze the select-then-predict models using the Shapley Value Sampling interpretability method, provided by the Captum library [42], which is a model interpretability library for PyTorch [43] models. Shapley Value Sampling [17, 18] is a method used to estimate Shapley values, which are used to calculate the contribution of each input feature to a model's prediction. This method is meant to reduce the computational complexity of calculating Shapley values by sampling subsets of input features, rather than computing all possible subsets.

Example 1	
Without regularization	
Input:	[CLS] wish upon **did not star ryan phillipe**. **[SEP]** wish upon is a 2017 supernatural horror thriller film directed by john r. leonetti and starring joey king, ryan phillipe, ki hong lee, shannon purser, sydney park and sherilyn fenn. it is set to be released in theaters on july 14, 2017, by broad green pictures and orion pictures. **[SEP]**
Rationale:	[did, not, star, ryan, phillipe, [SEP], [SEP]]
Predicted label:	Refutes
Gold label:	Refutes
With regularization	
Input:	[CLS] wish upon did not star ryan phillipe. **[SEP]** wish **upon** is a 2017 supernatural horror **thriller** film directed by john r. leonetti and starring joey king, **ryan phillipe**, ki hong lee, shannon purser, **sydney** park and sherilyn fenn. it is set to be released in theaters on july 14, **2017**, by broad green pictures and orion pictures. **[SEP]**
Rationale:	[[SEP], upon, thriller, ryan, phillipe, sydney, 2017, [SEP]]
Predicted label:	Refutes
Gold label:	Refutes
Example 2	
Without regularization	
Input:	[CLS] Harris Jayaraj **is Indian**. **[SEP] Harris Jayaraj** -LRB- born 8 January 1975 -RRB- is an Indian film composer from Chennai, Tamil Nadu. He composes scores and soundtracks predominantly for Tamil films, while also composed for a few films in Telugu and Hindi languages. **[SEP]**
Rationale:	[is, Indian, [SEP], Harris, Jayaraj, [SEP]]
Predicted label:	Supports
Gold label:	Supports
With regularization	
Input:	[CLS] Harris Jayaraj is Indian. **[SEP] Harris Jayaraj** -LRB- born 8 January 1975 -RRB- is an **Indian** film composer from Chennai, Tamil Nadu. He composes scores and soundtracks predominantly for Tamil films, while also composed for a few films in **Telugu** and Hindi languages. **[SEP]**
Rationale:	[[SEP], Harris, Jayaraj, Indian, Telugu, [SEP]]
Predicted label:	Supports
Gold label:	Supports
Example 3	
Without regularization	
Input:	[CLS] papua **comprised all of indonesian new guinea and it was cultured**. **[SEP]** papua is the largest and easternmost province of indonesia, comprising most of western new guinea. papua is bordered by the nation of papua new guinea to the east, and by west papua province to the west. its capital is jayapura. it was formerly called irian jaya - lrb - before that west irian or irian barat - rrb - and comprised all of indonesian new guinea. in 2002 the current name was adopted and in 2003 west papua province was created from western parts of papua province. [SEP]
Rationale:	[comprised, all, of, indonesian, new, guinea, and, it, was, cultured, [SEP]]
Predicted label:	Refutes
Gold label:	Supports
With regularization	
Input:	[CLS] papua comprised all of indonesian new guinea and it was cultured. **[SEP] papua** is the largest and easternmost province of **indonesia**, comprising most of western new **guinea**. **papua** is bordered by the nation of **papua** new **guinea** to the east, and by west **papua** province to the west. its capital is jayapura. it was formerly called irian jaya - lrb - before that west irian or irian barat - rrb - and comprised all of **indonesian** new **guinea**. in 2002 the current name was adopted and in 2003 west **papua** province was created from western parts of **papua** province. **[SEP]**
Rationale:	[[SEP], papua, indonesia, guinea, papua, papua, guinea, papua, indonesian, papua, papua, [SEP]]
Predicted label:	Supports
Gold label:	Supports

TABLE III: Rationales generated by the select-then-predict models on the FEVER dataset. The **Input** field represents the full input to the model, while the words highlighted in bold are part of the selected rationale, also shown in the **Rationale** field. On top is the rationale generated by a model without regularizer (i.e., $\mathcal{L}_{\text{claim}}$), on the bottom is a rationale generated by a model with the $\mathcal{L}_{\text{claim}}$ regularizer.

The goal was to understand which words in the rationale the model was paying most attention to and to analyze potential differences between positive and negative predictions.

However, deriving conclusive insights from this analysis proved to be impossible due to the computational complexity of this explainability method. Even though this method is meant to approximate Shapley values and, therefore, to reduce computational costs, the complexity of this method still increases with the number of input features, making it particularly computationally expensive for longer input sequences, such as instances from the FEVER dataset. Furthermore, Shapley Value Sampling relies on multiple forward passes through the model to evaluate the impact of perturbed inputs on the predicted output. The process requires perturbing the input data and computing the corresponding model predictions, which leads to a high computational complexity. Since the

π	Average unfaithfulness	Average alignment
'Simple' select-then-predict without regularizer		
0.1	0.21	0.15
0.2	0.11	0.24
Select-then-predict models with regularizer		
0.1	0.25	0.08
0.2	0.29	0.13
Faithful select-then-predict models with unfaithfulness loss		
0.1	0.08	0.21
0.2	0.03	0.26

TABLE IV: Comparison of average values of unfaithfulness and of alignment scores for different model classes on the FEVER dataset.

Shapley values are estimated by averaging the contributions of different feature subsets to the model's prediction, the number of sampled subsets directly affects the reliability of the estimates. If too few samples are used, the estimates become unreliable and may not accurately reflect the model's decision-making process. On the other hand, increasing the number of samples improves accuracy but makes the process much more computationally expensive, making it impractical for long input sequences and complex models like BERT.

Therefore, the Shapley Value Sampling method could be effective at explaining feature importance for simpler models, or for shorter input sequences. However, it proved to be less effective for our use case. In appendix F we present some examples of attributions obtained while analyzing the select-then-predict models.

H. Generalizability to Other Datasets

In order to analyze the behavior of the select-then-predict models on a second dataset, we trained the select-then-predict models on the MultiRC dataset [44]. We used the version provided by DeYoung et al. [35]. MultiRC is a reading comprehension dataset, which consists of multiple choice questions. These instances are reformulated as yes/no questions by providing question/answer/document triplets as input to the model, whose task it is to verify whether the answer is correct based on the provided documents. Examples of instances from the MultiRC dataset are provided in appendix E. In this case, the **Question + Answer** section of the input represents the part of the input that is crucial in order to make a prediction: it is not possible to predict whether an answer is correct without knowing the question or the answer. Therefore, for this dataset, the **Q + A** is analogous to the claim in the FEVER dataset and if the models learn to make correct predictions without selecting this part of the input, they must be providing unfaithful rationales.

As done with our initial experiments on the FEVER dataset, we first trained a full context baseline on the task, which achieves 0.74 accuracy. We then proceeded to train select-then-predict models on this dataset. We trained models with values of $\pi = 0.4$ and $\pi = 0.2$. The resulting model performances are reported in table V.

It can be observed that the select-then-predict models achieve much lower performance on the MultiRC dataset, compared to FEVER.[5] This is likely due to the complexity of the task: as can be seen in the examples in appendix E, the task might be more difficult than fact verification, because it requires more complex reasoning, as in fig. 13a, where the answer is only partially incorrect and is still labeled as false. This assumption is also supported by the lower performance of the FC model on this dataset: it achieves an accuracy of 0.74 on this task, significantly lower than the accuracy achieved by a FC model on the FEVER dataset. This indicates that the BERT model might be too simple for this task.

Despite this, we can see a similar pattern to what we saw in the previous experiments: the select-then-predict models achieve lower accuracy than the FC model. The models trained with $\mathcal{L}_{claim}$ also present a drop in performance, but their accuracy is still above the random baseline, while selecting 0% of the question + answer (Q + A) part of the input. However, the values of unfaithfulness of these models are lower compared to the models trained on the FEVER dataset. This is due to the fact that the accuracy of the select-then-predict models is lower, therefore performance gap between these models and the random baseline is also smaller. These results demonstrate that, even on a more difficult task, the models still learn to make correct predictions while not selecting a crucial part of the input, showing that they can encode the prediction within the rationale.

Lastly, we observe that the select-then-predict models trained with the unfaithfulness loss learn to select higher percentages of the question + answer (Q + A) part of the input. The performance of an unbiased predictor on their rationales is higher compared to the models without unfaithfulness loss, which shows that these rationales are more informative.

These results indicate that training the models with the additional $\mathcal{L}_u$ loss term helps reduce the degree of unfaithfulness

These results show that incorporating the additional loss term $\mathcal{L}_u$ during training helps reduce the degree of unfaithfulness, also when the models are trained on the MultiRC dataset.

VI. Conclusion

In this work, we investigate the unfaithful behavior of select-then-predict models.

The select-then-predict architecture was created in order to better understand the predictions of black-box models by teaching the models to select a small, yet significant, part of the input that leads them to make the final prediction. With this thesis we investigate the hypothesis, already presented in prior work, that these so-called inherently interpretable models

[5]These results are in line with previous work: in [38], the Sparse IB select-then-predict model trained on the MultiRC dataset achieves 0.65 accuracy with $\pi = 0.2$, whereas their FC baseline achieves 0.71 accuracy. We did not manage to reproduce the results presented in this paper, since we get 0.65 accuracy only when using $\pi = 0.4$. This is likely due to differences in training parameters and initialization. For example, they use a batch size of 32, whereas we use a batch size of 12 due to memory constraints.

Model type	π	Validation Accuracy	Test Accuracy	F1	Percentage of Q + A selected	Accuracy $\phi*$	Unfaithfulness $u(\phi, \psi)$
Baseline models							
BERT FC	1	0.74	0.74	0.73	–	–	–
Select-then-predict models without regularizer							
Sel Pred	0.2	0.61	0.59	0.57	19%	0.50	0.11
Sel Pred	0.4	0.65	0.63	0.63	84%	0.65	0.00
Select-then-predict models with regularizer							
Sel Pred + reg	0.2	0.58	0.60	0.51	0%	0.55	0.03
Sel Pred + reg	0.4	0.60	0.64	0.57	0%	0.52	0.08
Select-then-predict models with unfaithfulness loss							
Sel Pred faithful	0.2	0.62	0.64	0.59	47%	0.58	0.04
Sel Pred faithful	0.4	0.64	0.66	0.60	77%	0.62	0.02

TABLE V: Performance of the models on the MultiRC dataset. The performance of the models with the highest validation accuracy is reported. The 'simple' select-then-predict models are trained without regularization and without the unfaithfulness loss. The models with regularizer refer to models trained with $\mathcal{L}_{claim}$ and the 'faithful' models are trained with $\mathcal{L}_u$.

are not faithful by design. We present a formal definition of unfaithfulness, which can be used as a measure to determine the degree of unfaithfulness of a model and to judge how informative the selected rationales are. A high degree of unfaithfulness shows that the rationales chosen by a selector do not contain the information necessary for an unbiased model to make a prediction and, therefore, are not faithful. This provides an answer to RQ2, because the degree of unfaithfulness allows us to identify unfaithful behavior in select-then-predict models.

Our experiments show that select-then-predict models are able to learn unfaithful behavior. We demonstrate this by using an additional loss term, $\mathcal{L}_{claim}$, that teaches the model to not include a part of the input that is crucial for the prediction into the rationale. We observe that these models are still able to achieve performance that is significantly above the random baseline without seeing the claim, proving that the selector is able to make a prediction based on the full input and to encode it into the selected rationale.

Furthermore, we show that unfaithful behavior can also emerge on its own: the select-then-predict models have varying degrees of unfaithfulness and, based on the initialization of the weights, some models learn this unfaithful behavior even without the $\mathcal{L}_{claim}$ loss term that we added for the sake of our experiments. This shows that the failure case that we described is not only plausible, but is also learned by some select-then-predict models.

These results provide an answer to RQ1 and prove that, without other specific constraints, the select-then-predict architecture is prone to unfaithful behavior and that we cannot guarantee their faithfulness based solely on the fact that the only input received by the predictor is the rationale.

In order to answer RQ3, we train the models using an additional loss term, the *unfaithfulness loss*. We show that the use of an additional loss term during training can help reduce the degree of unfaithfulness of select-then-predict models and it can help to make the selected rationales more plausible. Our experiments show that the $\mathcal{L}_u$ term acts as a regularizer that can prevent select-then-predict models from learning to select unfaithful rationales. This demonstrates that adding a constraint can help prevent select-then-predict models from encoding the prediction inside the selected rationales.

In conclusion, our research shows that it is important to verify claims of inherent interpretability. Even architectures that are seemingly interpretable by design might present unforeseen failure cases, as demonstrated in our study. Therefore, it remains crucial to continue research on interpretable language models in order to better understand their limitations and to develop more transparent and trustworthy AI systems.

A. Limitations and Future Work

This work has several limitations that indicate some directions for future research. First of all, due to time constraints, the analysis was limited to a single select-then-predict architecture. Future studies could extend this work to a broader set of select-then-predict models to determine if and how this behavior appears in other select-then-predict architectures and if they exhibit the same behavior when trained with a regularizer.

Another interesting direction for future research involves training a select-then-predict model with a simpler selector model. In this research, BERT-base was used for both the selector and the predictor, creating a setting in which the selector is powerful enough to make a prediction and encode the output within the rationale. As we showed in this work, this can lead to unfaithful behavior. To address this issue, future research could investigate the use of simpler selector models, such as LSTM models, which might not be complex enough to perform the prediction task, but might generate informative and meaningful token representations, from which the rationale could be sampled. This could lead to a more faithful model, where the selector would not be complex enough to learn this type of unfaithful behavior.

As mentioned in section V-E, it can be beneficial to study how information regarding the final label is encoded in the

selected rationale. This analysis could help provide insights into the behavior of black-box models and it could help identify potential failure cases. Understanding how models secretly communicate the label through selected tokens can help the development of more robust and faithful architectures. Finding the patterns used to encode the labels can lead researchers to design solutions to limit this behavior and ensure that rationales truly explain the model's decision-making process.

Furthermore, this work was conducted on two datasets, which may not be representative of all types of tasks that select-then-predict models could be applied to. Our choice of datasets was motivated by the structure of the dataset, as discussed in section IV-A. However future work could incorporate other datasets in order to assess how common this type of unfaithful behavior is when using different datasets.

Another important consideration concerns the training setup of select-then-predict models. In this work, the selector and predictor are trained jointly, which is a common training setup due to the fact that it does not require additional annotated data and because it can be trained end-to-end. However, this setup allows the selector and the predictor models to learn a common encoding of the prediction. As a direction for future work, it can be valuable to explore different select-then-predict architectures where the selector and predictor are trained independently. This would ensure that the selector learns the task it is meant to learn: to select rationales and not to perform the final prediction, thereby ensuring more faithful rationales. However, this approach has certain downsides: it requires additional supervision to train the selector, for which human-annotated rationales are usually used. Human annotations are expensive and are often subjective. In addition, training on human annotated rationales assumes that these are the only true explanation for a model's decision, which is not necessarily true in this case. This training setup can, therefore, introduce human bias into the rationale selection process and might not truly reflect how the models would make their predictions.

Overall, these points show the importance of further studies on the select-then-predict architecture and the need for more research into explainable AI.

References

[1] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences, 2018. URL https://arxiv.org/abs/1706.07269.

[2] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022. URL https://christophm.github.io/interpretable-ml-book.

[3] Andreas Madsen, Himabindu Lakkaraju, Siva Reddy, and Sarath Chandar. Interpretability needs a new paradigm, 2024. URL https://arxiv.org/abs/2405.05386.

[4] European Parliament. Artificial intelligence act. URL https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf.

[5] EP. EU AI Act: first regulation on artificial intelligence — Topics — European Parliament, 8 2023. URL https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[6] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions, 2016. URL https://arxiv.org/abs/1606.04155.

[7] Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. An information bottleneck approach for controlling conciseness in rationale extraction, 2020. URL https://arxiv.org/abs/2005.00652.

[8] Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in nlp: A survey, 2024. URL https://arxiv.org/abs/2209.11326.

[9] Alon Jacovi and Yoav Goldberg. Aligning faithful interpretations with their social attribution, 2021. URL https://arxiv.org/abs/2006.01067.

[10] Harrie Oosterhuis, Lijun Lyu, and Avishek Anand. Local feature selection without label or feature leakage for interpretable machine learning predictions. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[11] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Mueller. How to explain individual classification decisions, 2009.

[12] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in nlp, 2016.

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

[14] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017.

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. URL https://arxiv.org/abs/1602.04938.

[16] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

[17] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers Operations Research*, 36(5):1726–1730, 2009. ISSN 0305-0548. doi: https://doi.org/10.1016/j.cor.2008.04.004. URL https://www.sciencedirect.com/science/article/pii/S0305054808000804. Selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).

[18] Erik trumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *J. Mach. Learn. Res.*, 11:1–18, 2010. URL https://api.semanticscholar.org/CorpusID:14451872.

[19] L. S. Shapley. *17. A Value for n-Person Games*, pages 307–318. Princeton University Press, Princeton, 1953. ISBN 9781400881970. doi: doi:10.

1515/9781400881970-018. URL https://doi.org/10.1515/9781400881970-018.

[20] Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural nlp: A survey. *ACM Comput. Surv.*, 55(8), December 2022. ISSN 0360-0300. doi: 10.1145/3546577. URL https://doi.org/10.1145/3546577.

[21] Yiming Zheng, Serena Booth, Julie Shah, and Yilun Zhou. The irrationality of neural rationale models, 2022. URL https://arxiv.org/abs/2110.07550.

[22] Sven Kruschel, Nico Hambauer, Sven Weinzierl, Sandra Zilker, Mathias Kraus, and Patrick Zschech. Challenging the performance-interpretability trade-off: An evaluation of interpretable machine learning models, 2024. URL https://arxiv.org/abs/2409.14429.

[23] Alessandro Lovo, Amaury Lancelin, Corentin Herbert, and Freddy Bouchet. Tackling the accuracy-interpretability trade-off in a hierarchy of machine learning models for the prediction of extreme heatwaves, 2024. URL https://arxiv.org/abs/2410.00984.

[24] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, 40(2):44–58, Jun. 2019. doi: 10.1609/aimag.v40i2.2850. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2850.

[25] Sai Gurrapu, Ajay Kulkarni, Lifu Huang, Ismini Lourentzou, and Feras A. Batarseh. Rationalization for explainable nlp: a survey. *Frontiers in Artificial Intelligence*, 6, September 2023. ISSN 2624-8212. doi: 10.3389/frai.2023.1225093. URL http://dx.doi.org/10.3389/frai.2023.1225093.

[26] Omar Zaidan, Jason Eisner, and Christine Piatko. Using "annotator rationales" to improve machine learning for text categorization. In Candace Sidner, Tanja Schultz, Matthew Stone, and ChengXiang Zhai, editors, *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, April 2007. Association for Computational Linguistics. URL https://aclanthology.org/N07-1033/.

[27] Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. Learning to faithfully rationalize by construction, 2020. URL https://arxiv.org/abs/2005.00115.

[28] Linan Yue, Qi Liu, Li Wang, Yanqing An, Yichao Du, and Zhenya Huang. Interventional rationalization. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11404–11418, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.700. URL https://aclanthology.org/2023.emnlp-main.700/.

[29] Jasmijn Bastings, Wilker Aziz, and Ivan Titov. Interpretable neural predictions with differentiable binary variables, 2020. URL https://arxiv.org/abs/1905.08160.

[30] Linan Yue, Qi Liu, Yichao Du, Yanqing An, Li Wang, and Enhong Chen. DARE: Disentanglement-augmented rationale extraction. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=6OhjECfqt2.

[31] Mo Yu, Shiyu Chang, Yang Zhang, and Tommi S. Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control, 2019. URL https://arxiv.org/abs/1910.13294.

[32] Shuaibo Hu and Kui Yu. Learning robust rationales for model explainability: A guidance-based approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18243–18251, Mar. 2024. doi: 10.1609/aaai.v38i16.29783. URL https://ojs.aaai.org/index.php/AAAI/article/view/29783.

[33] J. Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theor.*, 37(1):145–151, September 2006. ISSN 0018-9448. doi: 10.1109/18.61115. URL https://doi.org/10.1109/18.61115.

[34] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification, 2018. URL https://arxiv.org/abs/1803.05355.

[35] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. Eraser: A benchmark to evaluate rationalized nlp models, 2020. URL https://arxiv.org/abs/1911.03429.

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

[38] Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. Can rationalization improve robustness?, 2022. URL https://arxiv.org/abs/2204.11790.

[39] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017. URL https://arxiv.org/abs/1611.01144.

[40] Delft High Performance Computing Centre (DHPC). DelftBlue Supercomputer (Phase 2). https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2, 2024.

[41] William Falcon and The PyTorch Lightning team. PyTorch Lightning, March 2019. URL https://github.com/Lightning-AI/lightning.

[42] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020. URL https://arxiv.org/abs/2009.07896.

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Mar-

tin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

[44] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *NAACL*, 2018.

[45] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

APPENDIX

A. Examples from FEVER dataset

Table VI shows two instances from the FEVER dataset where human annotated rationales are not comprehensive or contain too much information, which could reduce the alignment score. In Example 1, the gold rationale includes redundant information and the sentence "hale first came to prominence as one of the five winners of the reality show american juniors" would have been sufficient to make a prediction. However, if a selector omitted the last part of the sentence (i.e., "a children's spin off of american idol."), it would have been penalized and would have gotten a lower alignment score, while still selecting a 'correct' rationale.

Example 2 demonstrates a case where another part of the input, not only the gold rationale, can be an explanation of the prediction. The sentence that follows the one chosen as gold rationale: "... then became an investment banker at rothschild & cie banque." could also refute the given claim, but it was not included in the human annotated gold rationale. If a model chose this as explanation, it could be seen as an acceptable explanation of a negative label, however this rationale would get an alignment score equal to 0.

These examples show that the alignment score is not a comprehensive indicator of rationale quality and of their faithfulness.

Example 1
lucy hale was not in american juniors. karen lucille hale - lrb - born june 14, 1989 - rrb - is an american actress and singer. earlier in her career, she was sometimes credited as lucy kate hale. **hale first came to prominence as one of the five winners of the reality show american juniors, a children's spin off of american idol.** she is best known for her role as aria montgomery on the freeform series pretty little liars, which won her a people's choice award for favorite cable tv actress in 2014. the same year, she released her debut studio album, road between.
Example 2
emmanuel macron refused to work as an investment banker. emmanuel jean - michel frederic macron; born 21 december 1977 - rrb - is the president of france and ex officio co - prince of andorra, having assumed these offices on 14 may 2017. **a former civil servant and investment banker, he studied philosophy at paris nanterre university, completed a master's of public affairs at sciences po, and graduated from the ecole nationale d'administration - lrb - ena - rrb - in 2004.** he worked as an inspector of finances in the inspectorate general of finances - lrb - igf - rrb -, then became an investment banker at rothschild & cie banque. macron was appointed deputy secretary-general in francois hollande's first government in 2012, having been a member of the socialist party from 2006 to 2009. he was appointed minister of economy, industry and digital affairs in 2014 under the second valls government, where he pushed through business-friendly reforms. he resigned in august 2016 to launch a bid in the 2017 presidential election. in november 2016, macron declared that he would run in the election under the banner of en marche!, a centrist political movement he founded in april 2016, and won the election on 7 may 2017. macron, at the age of 39, became the youngest president in the history of france. upon his inauguration, macron appointed le havre mayor edouard philippe to be prime minister on 15 may 2017.

TABLE VI: Instances from the FEVER dataset, where human annotated rationales do not represent the only correct answer and therefore can be misleading. The human annotated gold rationale is shown in bold.

B. Additional Training Details

In this section we provide additional details regarding training of the models.

We utilize gradient accumulation to train the models. Gradient accumulation is a technique that allows to simulate a larger batch size, without exceeding GPU memory limits. Instead of updating model weights after every batch, model weights are accumulated over multiple batches before the weights are updated. This technique allows us to achieve a larger effective batch size, which leads to more stability during training.

During training we use a learning rate scheduler - the linear schedule with warmup from the transformers library [45].

	Full Context	Select-then-predict
Maximum number of epochs	15	15
Maximum input length	512	512
Batch size	12	12
Learning rate	5e-5	5e-5
Gradient accumulation steps	10	10

C. *Performance of Select-then-Predict Models*

The images below demonstrate the variation of the F1 score, the percentage of claim that is selected and the alignment score of select-then-predict models across different initializations and and different select-then-predict model types.

The variation in performance is caused by the sampling function that selects the rationales, therefore it can be seen that the higher the percentage of the full input that is selected (represented by the parameter π), the lower the variation in performance.

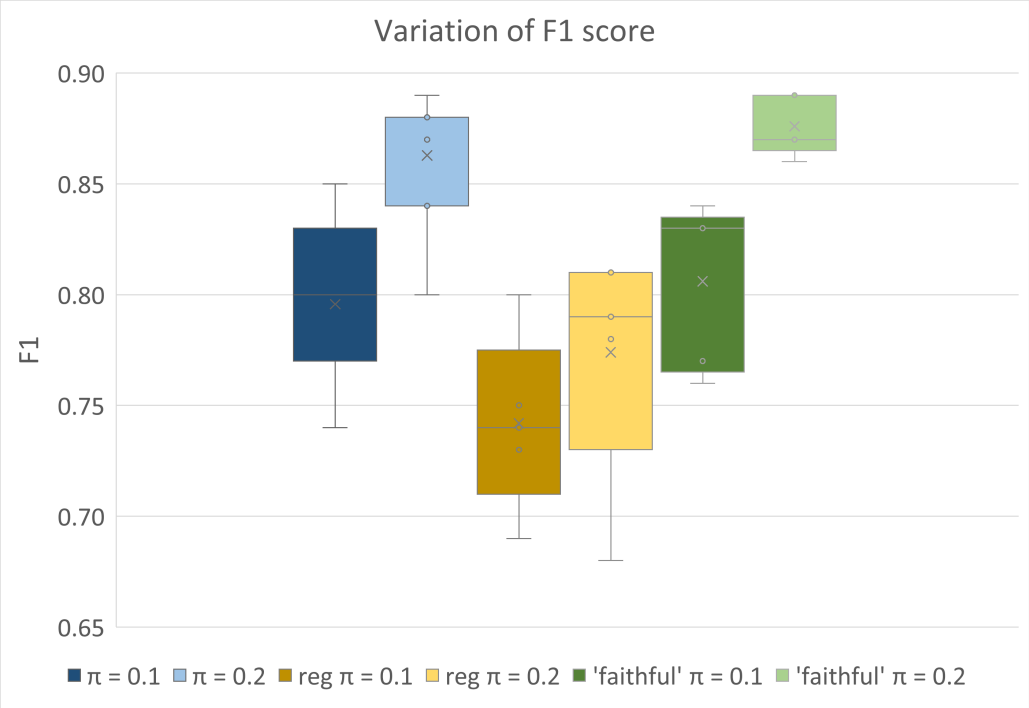
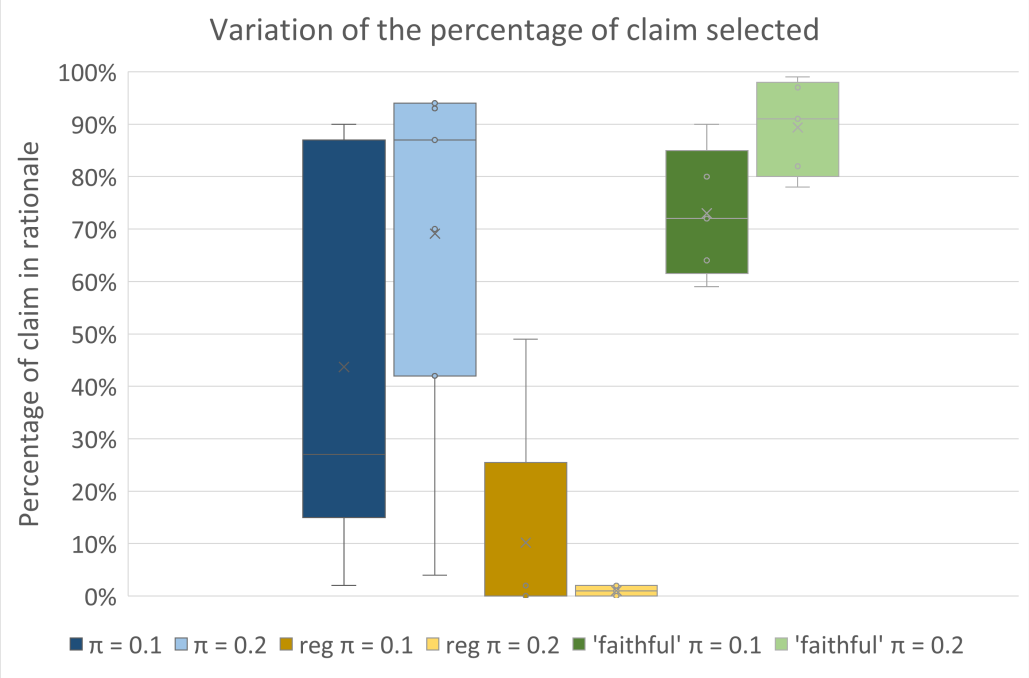
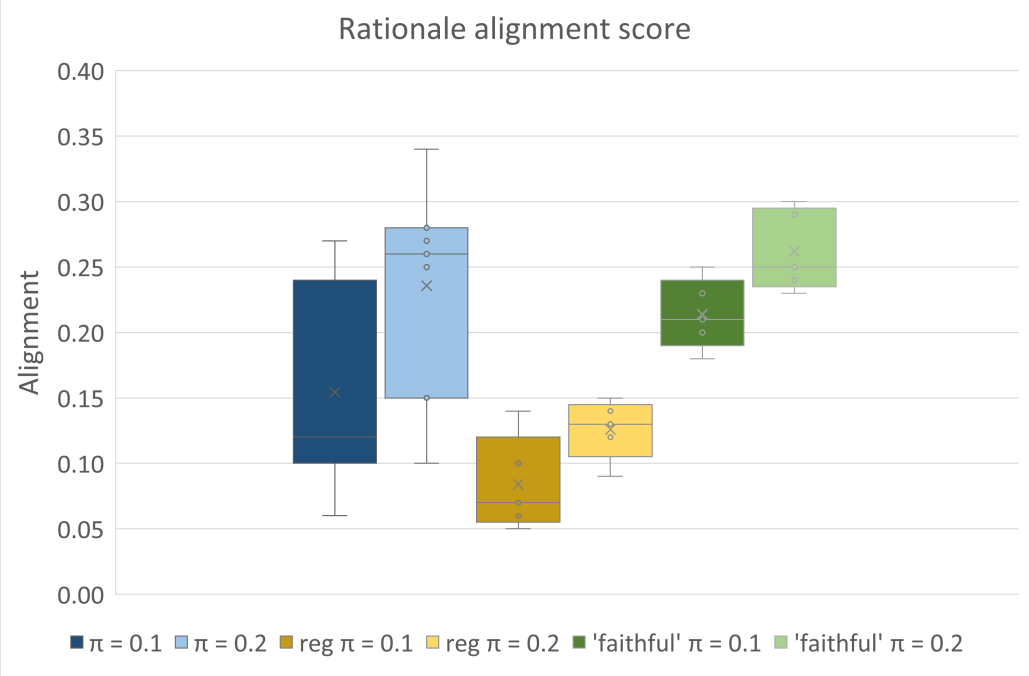


Fig. 9: Box plot representing the variation of F1 score of the select-then-predict models on the FEVER dataset. The first two boxes (in blue) refer to select-then-predict models without regularizer trained with a sparsity value π = 0.1 and π = 0.2, respectively. The following two boxes (in yellow) refer to select-then-predict models trained with regularizer, with π = 0.1 and π = 0.2, respectively. The final two boxes (green) refer to models trained with the additional 'unfaithfulness' loss, with π = 0.1 and π = 0.2, respectively.



(a) Percentage of claim selected as rationale



(b) Alignment scores

Fig. 10: Box plots representing the variation of the percentage of the claim included in the rationale and of alignment scores related to different select-then-predict models, trained on the FEVER dataset. The first two boxes (in blue) refer to select-then-predict models without regularizer trained with a sparsity value $\pi = 0.1$ and $\pi = 0.2$, respectively. The following two boxes (in yellow) refer to select-then-predict models trained with regularizer, with $\pi = 0.1$ and $\pi = 0.2$, respectively. The final two boxes (green) refer to models trained with the additional 'unfaithfulness' loss, with $\pi = 0.1$ and $\pi = 0.2$, respectively.

Random Seed	π	Validation Accuracy	Test Accuracy	F1 Score	Percentage of Claim Selected	Accuracy $\phi*$	Unfaithfulness $u(\phi, \psi)$	Alignment
'Simple' select-then-predict models without regularizer								
8	0.1	0.82	0.79	0.80	15%	0.50	0.32	0.10
78	0.1	0.78	0.74	0.74	16%	0.49	0.29	0.12
10	0.1	0.86	0.84	0.83	87%	0.76	0.1	0.24
11	0.1	0.85	0.75	0.78	2%	0.49	0.36	0.06
12	0.1	0.82	0.83	0.80	27%	0.55	0.27	0.12
15	0.1	0.87	0.84	0.85	90%	0.84	0.03	0.27
9	0.1	0.81	0.76	0.69	69%	0.72	0.09	0.17
3	0.2	0.88	0.87	0.88	88%	0.77	0.11	0.25
10	0.2	0.88	0.89	0.89	93%	0.88	0.00	0.28
5	0.2	0.88	0.85	0.87	94%	0.81	0.07	0.27
9	0.2	0.87	0.88	0.88	70%	0.87	0.00	0.34
15	0.2	0.86	0.77	0.80	4%	0.49	0.37	0.10
12	0.2	0.87	0.81	0.84	42%	0.66	0.21	0.15
11	0.2	0.86	0.87	0.88	94%	0.82	0.04	0.26
Select-then-predict models with regularizer								
10	0.1	0.76	0.76	0.74	49%	0.59	0.17	0.14
2	0.1	0.73	0.73	0.73	0%	0.49	0.24	0.06
6	0.1	0.82	0.80	0.80	0%	0.49	0.33	0.1
3	0.1	0.77	0.72	0.75	2%	0.49	0.28	0.07
8	0.1	0.70	0.70	0.69	0%	0.49	0.21	0.05
10	0.2	0.70	0.66	0.68	0%	0.49	0.21	0.09
8	0.2	0.79	0.76	0.79	2%	0.49	0.30	0.14
2	0.2	0.82	0.78	0.81	0%	0.49	0.33	0.15
7	0.2	0.81	0.80	0.81	1%	0.49	0.32	0.13
9	0.2	0.78	0.78	0.78	2%	0.49	0.29	0.12
'Faithful' select-then-predict models with unfaithfulness loss								
2	0.1	0.77	0.70	0.77	64%	0.72	0.05	0.25
5	0.1	0.83	0.65	0.83	59%	0.63	0.20	0.18
9	0.1	0.77	0.74	0.76	90%	0.71	0.06	0.21
17	0.1	0.83	0.80	0.83	72%	0.74	0.09	0.20
8	0.1	0.85	0.80	0.84	80%	0.73	0.12	0.23
5	0.2	0.87	0.83	0.86	99%	0.80	0.07	0.30
2	0.2	0.89	0.90	0.89	78%	0.88	0.01	0.23
9	0.2	0.87	0.84	0.87	82%	0.88	-0.01	0.25
11	0.2	0.87	0.85	0.87	97%	0.79	0.08	0.24
16	0.2	0.89	0.88	0.89	91%	0.88	0.01	0.29

TABLE VII: Performance of the select-then-predict models trained during our experiments, on the FEVER dataset. The 'simple' select-then-predict models are trained without regularization and without the unfaithfulness loss. The models with regularizer are trained with $\mathcal{L}_{claim}$ and the 'faithful' models are trained with $\mathcal{L}_u$.

Random Seed	π	Validation Accuracy	Test Accuracy	F1	Percentage of Q + A selected	Accuracy $\phi*$	Unfaithfulness $u(\phi, \psi)$
'Simple' select-then-predict models without regularizer							
9	0.2	0.61	0.59	0.57	19%	0.50	0.11
7	0.2	0.59	0.62	0.55	17%	0.58	0.01
5	0.2	0.60	0.61	0.60	69%	0.58	0.02
9	0.4	0.65	0.63	0.63	87%	0.62	0.03
12	0.4	0.65	0.63	0.63	84%	0.65	0.00
Select-then-predict models with regularizer							
10	0.2	0.58	0.60	0.51	0%	0.55	0.03
7	0.2	0.58	0.59	0.51	0%	0.55	0.03
17	0.4	0.60	0.61	0.55	0%	0.55	0.05
7	0.4	0.60	0.64	0.57	0%	0.52	0.04
Select-then-predict models with unfaithfulness loss							
7	0.2	0.62	0.64	0.59	47%	0.58	0.04
15	0.4	0.64	0.66	0.60	77%	0.62	0.02

TABLE VIII: Performance of the select-then-predict models trained during our experiments, on the MultiRC dataset. The 'simple' select-then-predict models are trained without regularization and without the unfaithfulness loss. The models with regularizer are trained with $\mathcal{L}_{claim}$ and the 'faithful' models are trained with $\mathcal{L}_u$.

D. Examples of Rationales

Original Input	**Gold Rationale**	**Rationale without regularization (claim selected)**	**Rationale with regularization (i.e., no claim selected)**
Edison machine works was founded to make big motors running on electricity. [SEP] the Edison machine works was a manufacturing company set up to produce dynamos, large electric motors, and other components of the electrical illumination system being built by Thomas A. Edison in New York city.	edison machine works was founded to make big motors running on electricity. **the edison machine works was a manufacturing company set up to produce dynamos, large electric motors, and other components of the electrical illumination system being built by thomas a**. **edison in new york city [SEP]** Gold label: 1	edison machine works was **founded** to make **big motors running on electricity**. [SEP] the edison machine works was a manufacturing company set up to produce dynamos, large electric motors, and other components of the electrical illumination system being built by thomas a. edison in new york city. Predicted label: 1	[CLS] edison machine works was founded to make big motors running on electricity. **[SEP]** the **edison machine** works was a manufacturing company set up to produce dynamos, large **electric motors**, and other components of the electrical illumination system being built by thomas a. edison in new york city. **[SEP]** Predicted label: 1
Trevor Griffiths is a German dramatist. Trevor Griffiths - lrb - born 4 April 1935, Ancoats, Manchester - rrb - , is an English dramatist [SEP]	trevor griffiths is a german dramatist. **trevor griffiths - lrb - born 4 april 1935**, **ancoats**, **manchester - rrb -**, **is an english dramatist [SEP]** Gold label: 0	[CLS] trevor griffiths is **a german dramatist**. [SEP] trevor griffiths - lrb - born 4 april 1935, ancoats, manchester - rrb -, is an english dramatist. [SEP] Predicted label: 1	[CLS] trevor griffiths is a german dramatist. **[SEP]** trevor **griffiths** - lrb - born 4 april 1935, ancoats, manchester - rrb -, is an english dramatist . **[SEP]** Predicted label: 0
Shannon Lee is a German. Shannon Emery Lee - lrb - born April 19, 1969 - rrb - is an American actress, martial artist and businesswoman. she is the daughter of martial arts film star Bruce Lee and Linda Lee Cadwell, the granddaughter of Cantonese opera singer lee hoi - Chuen, and the younger sister of Brandon Lee.	shannon lee is a german. **shannon emery lee - lrb - born april 19, 1969 - rrb - is an american actress, martial artist and businesswoman**. she is the daughter of martial arts film star bruce lee and linda lee cadwell, the granddaughter of cantonese opera singer lee hoi - chuen , and the younger sister of brandon lee . Gold label: 0	[CLS] shannon **lee is a german**. **[SEP] shannon emery** lee - lrb - born april 19, 1969 - rrb - is an american actress, martial artist and businesswoman. she is the daughter of martial arts film star bruce lee and linda lee cadwell, the granddaughter of cantonese opera singer lee hoi - chuen, and the younger sister of brandon lee. **[SEP]** Predicted label: 0	[CLS] shannon lee is a german. **[SEP] shannon** emery **lee** - lrb - born april 19, 1969 - rrb - is an american actress, martial artist and businesswoman. she is the daughter of martial arts film star bruce **lee** and linda lee cadwell, the granddaughter of **cantonese** opera singer **lee** hoi - chuen, and the younger sister of brandon **lee**. **[SEP]** Predicted label: 0
Annabelle is in the United States. Annabelle is a raggedy ann doll alleged by demonologists Ed and Lorraine Warren to be haunted. the doll resides in a glass box at the warrens ' occult museum in Monroe, Connecticut. The story served as the inspiration for the films Annabelle - lrb - 2014 - rrb -, and the upcoming, Annabelle: creation - lrb - 2017 - rrb - . Annabelle has been compared to Robert the doll and was described in Gerald Brittle ' s 2002 biography of Ed and Lorraine Warren, the demonologist.	annabelle is in the united states. annabelle is a raggedy ann doll alleged by demonologists ed and lorraine warren to be haunted. **the doll resides in a glass box at the warrens ' occult museum in monroe**, **connecticut**. the story served as the inspiration for the films annabelle - lrb - 2014 - rrb -, and the upcoming, annabelle: creation - lrb - 2017 - rrb - . annabelle has been compared to robert the doll and was described in gerald brittle ' s 2002 biography of ed and lorraine warren, the demonologist. Gold label: 1	[CLS] annabelle **is in the united states**. **[SEP]** annabelle **is a raggedy ann** doll alleged by demonologists ed and lorraine warren to be haunted. the doll resides in a glass box at the warrens ' occult museum in monroe, connecticut. the story served as the inspiration for the films annabelle - lrb - 2014 - rrb -, and the upcoming, annabelle: creation - lrb - 2017 - rrb -. annabelle has been compared to robert the doll and was described in gerald brittle ' s 2002 biography of ed and lorraine warren, the demonologist. **[SEP]** Predicted label: 1	[CLS] annabelle is in the united states. **[SEP]** annabelle is a raggedy ann doll alleged by demonologists ed and lorraine warren to be haunted. the **doll** resides in a glass box at the warrens ' occult museum in monroe, connecticut. the story served as the inspiration for the films **annabelle** - lrb - 2014 - rrb -, and the upcoming, **annabelle**: creation - lrb - 2017 - rrb -. **annabelle** has been compared to robert the **doll** and was described in **gerald brittle** ' s 2002 biography of **ed** and **lorraine** warren, the **demonologist**. **[SEP]** Predicted label: 1

Fig. 11: Rationales generated by the select-then-predict models with the highest validation accuracy. The third column represents rationales selected by a model trained without $\mathcal{L}_{claim}$, whereas the fourth column contains rationales selected by a model trained with $\mathcal{L}_{claim}$. Both models were trained with a sparsity parameter π equal to 0.1.

Original Input	**Gold Rationale**	**Rationale without regularization (claim selected)**	**Rationale with regularization (i.e., no claim selected)**
Edison machine works was founded to make big motors running on electricity. [SEP] the Edison machine works was a manufacturing company set up to produce dynamos, large electric motors, and other components of the electrical illumination system being built by Thomas A. Edison in New York city.	edison machine works was founded to make big motors running on electricity. **the edison machine works was a manufacturing company set up to produce dynamos, large electric motors, and other components of the electrical illumination system being built by thomas a**. **edison in new york city [SEP]** Gold label: 1	[CLS] edison machine works was **founded to make big motors running on electricity**. **[SEP]** the edison machine works was a manufacturing company set up to produce dynamos, large electric motors, and other components of the electrical illumination system being built by thomas a. edison in new york city. **[SEP]** Predicted label: 0	[CLS] edison machine works was founded to make big motors running on electricity. **[SEP] the edison machine works** was a manufacturing company set up to produce **dynamos**, **large electric motors**, and other components of the electrical illumination system being built by thomas a. **edison** in new york city. **[SEP]** Predicted label: 1
Trevor Griffiths is a German dramatist. Trevor Griffiths - lrb - born 4 April 1935, Ancoats, Manchester - rrb -, is an English dramatist	trevor griffiths is a german dramatist. **trevor griffiths - lrb - born 4 april 1935**, **ancoats**, **manchester - rrb -**, **is an english dramatist [SEP]** Gold label: 0	[CLS] trevor griffiths is **a german dramatist**. **[SEP]** trevor griffiths - lrb - born 4 april 1935, ancoats, manchester - rrb -, is an english **dramatist**. [SEP] Predicted label: 0	[CLS] trevor griffiths is a german dramatist. **[SEP] trevor griffiths** - lrb - born 4 april 1935, ancoats, **manchester** - rrb -, is an english **dramatist**. **[SEP]** Predicted label: 0
Annabelle is in the United States. Annabelle is a raggedy ann doll alleged by demonologists Ed and Lorraine Warren to be haunted. the doll resides in a glass box at the warrens ' occult museum in Monroe, Connecticut. The story served as the inspiration for the films Annabelle - lrb - 2014 - rrb -, and the upcoming, Annabelle: creation - lrb - 2017 - rrb -. Annabelle has been compared to Robert the doll and was described in Gerald Brittle ' s 2002 biography of Ed and Lorraine Warren, the demonologist.	annabelle is in the united states. annabelle is a raggedy ann doll alleged by demonologists ed and lorraine warren to be haunted. **the doll resides in a glass box at the warrens ' occult museum in monroe**, **connecticut**. the story served as the inspiration for the films annabelle - lrb - 2014 - rrb -, and the upcoming, annabelle: creation - lrb - 2017 - rrb - . annabelle has been compared to robert the doll and was described in gerald brittle ' s 2002 biography of ed and lorraine warren, the demonologist. Gold label: 1	**[CLS]** annabelle **is in the united states**. **[SEP]** annabelle **is a raggedy ann doll alleged** by **demonologists** ed and lorraine warren to be **haunted**. the doll resides in a glass box at the warrens ' occult museum in **monroe**, **connecticut**. the story served as the inspiration for the films annabelle - lrb - 2014 - rrb -, and the upcoming, annabelle: creation - lrb - 2017 - rrb -. **annabelle** has been compared to robert the doll and was described in gerald brittle ' s 2002 biography of ed and **lorraine** warren, the **demonologist**. **[SEP]** Predicted label: 1	[CLS] annabelle is in the united states. **[SEP] annabelle is** a **raggedy ann doll alleged by demonologists** ed and **lorraine** warren to be **haunted**. the **doll** resides in a glass box at the **warrens** ' occult museum in **monroe**, **connecticut** . the story served as the inspiration for the films annabelle - lrb - 2014 - rrb -, and the upcoming, annabelle: creation - lrb - 2017 - rrb -. **annabelle** has been compared to robert the **doll** and was described in gerald brittle ' s 2002 biography of **ed** and **lorraine** warren, the **demonologist**. **[SEP]** Predicted label: 1
Temple Grandin stars Claire Danes as a stormtrooper. Temple Grandin is a 2010 biopic directed by mick Jackson and starring Claire Danes as Temple Grandin, an autistic woman who revolutionized practices for the humane handling of livestock on cattle ranches and slaughterhouses.	temple grandin stars claire danes as a stormtrooper. **temple grandin is a 2010 biopic directed by mick jackson and starring claire danes as temple grandin**, **an autistic woman who revolutionized practices for the humane handling of livestock on cattle ranches and slaughterhouses [SEP]**	**[CLS]** temple grandin **stars** claire danes **as a stormtrooper**. **[SEP]** temple grandin is a 2010 biopic directed by mick jackson and starring **claire danes** as temple grandin, an autistic woman who revolutionized practices for the humane handling of livestock on cattle ranches and slaughterhouses. **[SEP]** Predicted label: 0	[CLS] temple grandin stars claire danes as a stormtrooper. **[SEP] temple grandin** is a **2010** biopic directed by mick **jackson** and starring **claire danes** as **temple grandin**, an autistic woman who revolutionized practices for the humane handling of livestock on cattle ranches and slaughterhouses. **[SEP]** Predicted label: 1

Fig. 12: Rationales generated by the select-then-predict models with the highest validation accuracy. The third column represents rationales selected by a model trained without $\mathcal{L}_{claim}$, whereas the fourth column contains rationales selected by a model trained with $\mathcal{L}_{claim}$. Both models were trained with a sparsity parameter π equal to 0.2.

E. MultiRC Dataset Examples

Question + Answer: "What was the weather like ? || Charming, but cold"

Docs: "As his car slid downtown on Tuesday morning the mind of Arnold Thorndike was occupied with such details of daily routine as the purchase of a railroad , the Japanese loan , the new wing to his art gallery , and an attack that morning , in his own newspaper , upon his pet trust . But his busy mind was not too occupied to return the salutes of the traffic policemen who cleared the way for him . Or , by some genius of memory , to recall the fact that it was on this morning young Spear was to be sentenced for theft . **It was a charming morning . The spring was at full tide , and the air was sweet and clean .** Mr. Thorndike considered whimsically that to send a man to jail with the memory of such a morning clinging to him was adding a year to his sentence . He regretted he had not given the probation officer a stronger letter . He remembered the young man now , and favorably . A shy , silent youth , deft in work , and at other times conscious and embarrassed . But that , on the part of a stenographer , in the presence of the Wisest Man in Wall Street , was not unnatural . On occasions , Mr. Thorndike had put even royalty— frayed , impecunious royalty , on the lookout for a loan — at its ease . **The hood of the car was down , and the taste of the air , warmed by the sun , was grateful** . It was at this time , a year before , that young Spear picked the spring flowers to take to his mother . A year from now where would young Spear be ? It was characteristic of the great man to act quickly , so quickly that his friends declared he was a slave to impulse . It was these same impulses , leading so invariably to success , that made his enemies call him the Wisest Man . He leaned forward and touched the chauffeur \'s shoulder . " Stop at the Court of General Sessions , " he commanded . What he proposed to do would take but a few minutes . A word , a personal word from him to the district attorney , or the judge , would be enough"

Label: FALSE

(a) Example of an instance from the MultiRC dataset.

Question + Answer: "What did the judge tell Mr. Thorndike about the law ? || The law is not vindictive"

Docs: "The judge leaned back in his chair and beckoned to Mr. Andrews . It was finished . Spear was free , and from different parts of the courtroom people were moving toward the door . Their numbers showed that the friends of the young man had been many . Mr. Thorndike felt a certain twinge of disappointment . Even though the result relieved and pleased him , he wished , in bringing it about , he had had some part . He begrudged to Isaacs & Sons the credit of having given Spear his liberty . His morning had been wasted . He had neglected his own interests , and in no way assisted those of Spear . He was moving out of the railed enclosure when Andrews called him by name . " His honor , " he said impressively , " wishes to speak to you . " The judge leaned over his desk and shook Mr. Thorndike by the hand . Then he made a speech . The speech was about public - spirited citizens who , to the neglect of their own interests , came to assist the ends of justice , and fellow - creatures in misfortune . He purposely spoke in a loud voice , and every one stopped to listen . " **The law , Mr. Thorndike , is not vindictive** , " he said . " It wishes only to be just . Nor can it be swayed by wealth or political or social influences . But when there is good in a man , I , personally , want to know it , and when gentlemen like yourself , of your standing in this city , come here to speak a good word for a man , we would stultify the purpose of justice if we did not listen . I thank you for coming , and I wish more of our citizens were as unselfish and public - spirited . " It was all quite absurd and most embarrassing , but inwardly Mr. Thorndike glowed with pleasure . It was a long time since any one had had the audacity to tell him he had done well ."

Label: TRUE

(b) Example of an instance from the MultiRC dataset.

Fig. 13: Examples of instances from the MultiRC dataset. The **Question + Answer** field represents a question and a potential answer to it. The document snippet in **Docs** contains information to verify whether the provided answer is true or false. The text in **bold** shows the human annotated golden rationale. The models we train have the task of predicting whether the answer to the question is true or false based on the provided document.

F. Post-hoc Interpretability of the Rationales

Figures 14 and 15 show the outputs of the Shapley Value Sampling method used to analyze model predictions. We first analyze the full context model in order to use it as a baseline for the analysis of the select-then-predict models. In fig. 14 it can be observed that, in the case of a FC model and of the select-then-predict models without regularizer, the prediction is mainly attributed to one word, i.e. 'successful', whereas from these attribution scores it seems that the other words from the input do not bear much importance for the final prediction. A similar observation can be made for fig. 15, where for the prediction of the FC model and of the select-then-predict model, a lot of importance is assigned to one word. However, in this case we see that the word is different for these two models: it appears that for the FC model, the word 'pyrenees' contributes positively to the prediction, whereas for the select-then-predict model, the word 'mount' has the highest importance score. It is also important to note that the word that gets the highest attribution score in the FC model's attributions, does not even get selected as part of the rationale. This can indicate several things: that the select-then-predict model does not select the most meaningful words from the input or that the attribution scores do not provide an accurate representation of the model's reasoning process.

In fig. 14c and fig. 15c, which refer to the models trained with regularizer, all the input tokens are given very low attribution scores and, from these attributions, it is not clear which tokens had a higher impact on the final prediction. This shows that this attribution method failed to discern which tokens are important for the final prediction and, therefore, does not provide any information regarding the decision making process of the model.

These examples demonstrate that in our use case, the Shapley Value Sampling method did not yield meaningful insights into the predictions of our select-then-predict models.

For the full context model (in fig. 14a and fig. 15a) we set the n_samples parameter (i.e., the number of feature permutations tested) to 100, whereas for the select-then-predict models we set it to 20. We set the baseline, i.e. the reference input, to the [PAD] token.

Legend: Negative, Neutral, Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	1 (0.88)	Class 1	0.74	[CLS] ka ##reen ##a kapoor was initially successful . [SEP] ka ##reen ##a kapoor - l ##rb - - l ##sb - k ##ə ##' ##ri ##: ##na ##: k ##ə ##' ##pu ##: ##r - rs ##b - ; born 21 september 1980 - rr ##b - , also known by her married name ka ##reen ##a kapoor khan , is an indian actress who appears in hindi films . she is the daughter of actors rand ##hir kapoor and ba ##bit ##a , and the younger sister of actress kari ##sma kapoor . noted for playing a variety of characters in a range of film genres - - from romantic comedies to crime dramas - - kapoor is the recipient of several awards , including six filmfare awards , and is one of bollywood ' s most popular and highest - paid actresses . after making her acting debut in the 2000 war film refugee , kapoor established herself in hindi cinema with roles in the historical drama as ##oka , and the blockbuster mel ##od ##rama ka ##bh ##i k ##hus ##hi ka ##bh ##ie g ##ham . . . - l ##rb - both 2001 - rr ##b - . this initial success was followed by a series of commercial failures and repetitive roles , which garnered her negative reviews . the year 2004 marked a turning point for kapoor when she played against type in the role of a sex worker in the drama cha ##mel ##i . she subsequently earned wide critical recognition for her portrayal of a riot victim in the 2004 drama dev and a character based on william shakespeare ' s heroine des ##de ##mona in the 2006 crime film om ##kara . she went on to receive filmfare awards for best actress and best supporting actress for her performances in the 2007 romantic comedy ja ##b we met and the 2010 drama we are family , respectively . kapoor achieved further success by featuring as the female lead in two of india ' s top - grossing productions - - the 2009 comedy - drama 3 idiots and the 2015 social drama ba ##j ##rang ##i b ##hai ##ja ##an - - and received praise for her roles in the 2009 thriller ku ##rba ##an and the 2012 drama heroine . married to actor sai ##f ali khan , with whom she has a son , kapoor ' s off - screen life is the subject of widespread coverage in india . she has a reputation for being outspoken and assert ##ive , and is recognised for her contributions to the film industry through her fashion style and film roles . in addition to film acting , kapoor is a stage performer and has contributed as a co - author to three books : an autobiographical memoir and two nutrition guides . she has also launched her own clothing line in association with the retail chain g ##lo ##bus . [SEP]

(a) Attributions for the full context model

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	1 (0.98)	Class 1	1.26	- - - - - was - successful - indian - - - - hindi films - - - - - of - - ##hir - - ba - - - - - younger - - - - - - - - - - a - - - - - range of - - - - - - comedies - - - - - - is the - - - awards - - - - - - - - - - bollywood - - - - - highest - - - . - - - - - - the 2000 - - - - - established herself - - - - - - - - - - - - - - - - ##rama - series - - - - - - - - garnered - - - - - - 2004 - - - - - - - - - - - in - - - - sex - - the drama cha - - . - - - - - - - her portrayal of - - - - - - - dev - - - - - - - ' - - - ##de - in - 2006 - - - - . - reputation - being - and - - - - is - - - - - - - - - - - - - - - - - - - film - - - - a - - - - - - - - - - to -

(b) Attributions for the select-then-predict model trained without regularizer

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	1 (0.94)	Class 1	0.47	- - - - - - - - - [SEP] - hindi films - - - - - of - - ##hir - - ba - - - - - younger - - - - ##sma - - noted - - a - - - - - range of - bollywood - - - - - highest - - - . - - - - - - the - - - - - - established herself - - - - - - - - - - - - - - - - - ##rama - - ##i - - - - - - - - - . - - - - - both - - - - - - - - - - - - - series - - - - - - - - garnered - - - - - - 2004 - - - - - - - - - - - in - - - - sex - ##de - in - 2006 - - - - . - - - - receive - - - best - and - - - - her performances - - 2007 - comedy ja - we met - - - - we - - - - - - - further success by featuring - - - - - - - - - - - - grossing productions - [SEP] - - - - - - - - - - -

(c) Attributions for the select-then-predict model trained with regularizer

Fig. 14: Attributions of the Shapley Value Sampling method used to analyze the predictions of the select-then-predict models. Words highlighted in green have higher attribution scores, meaning that they contribute towards the predicted label, while words highlighted in red contribute negatively to the prediction (i.e., they make the model less confident in the prediction). The three subfigures refer to the same full input text, that can be seen in fig. 14a. In figures fig. 14c and fig. 14b only the rationale selected by the model is shown. The attribution score represents the sum of the marginal contributions to the prediction of all the input tokens.

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (1.00)	Class 0	3.23	[CLS] mount hood is in the pyrenees . [SEP] mount hood , called w ##y ' east by the mu ##lt ##no ##mah tribe , is a potentially active st ##rat ##ovo ##lc ##ano in the cascade volcanic arc of northern oregon . it was formed by a sub ##duction zone on the pacific coast and rests in the pacific northwest region of the united states . it is located about 50 mi east - southeast of portland , on the border between cl ##ack ##ama ##s and hood river counties . in addition to being oregon ' s highest mountain , it is one of the loft ##iest mountains in the nation based on its prominence . the height assigned to mount hood ' s snow - covered peak has varied over its history . modern sources point to three different heights : 112 ##49 ft , a 1991 adjustment of a 1986 measurement by the u . s . national geo ##de ##tic survey - l ##rb - ng ##s - rr ##b - , 112 ##40 ft based on a 1993 scientific expedition , and 112 ##39 ft of slightly older origin . the peak is home to 12 named glaciers and snow ##fields . it is the highest point in oregon and the fourth highest in the cascade range . mount hood is considered the oregon volcano most likely to er ##upt , though based on its history , an explosive eruption is unlikely . still , the odds of an eruption in the next 30 years are estimated at between 3 and 7 percent , so the u . s . geological survey - l ##rb - us ##gs - rr ##b - character ##izes it as ` ` potentially active ' ' , but the mountain is informally considered dormant . [SEP]

(a) Attributions for the full context model

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.64)	Class 0	1.15	- - - is in the - - - mount - the - - - - is located - - - - - southeast - - , - - border - - - - - - - - - - in - - being - - s - - - - - - of - - - - - - - - - - - - - - - - - - ' - on - . still - - odds - - - - - - years are estimated - - - - - - - so - - - - - - - - - - - us - - - ##b - - - - as - - - - ' - - - the - is -

(b) Attributions for the select-then-predict model trained without regularizer

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	1 (0.73)	Class 0	0.28	- cascade - - - - - - - - - - - - - - - - pacific - - - - - - - - - the - - - - - - - - - - - - - - - , - - border - - - - - and - - - - in - - being - - s - - - - - one of - - - - - - nation - - - - . - - - - mount hood - - - - - - - - - - - - - - - - - - different - - - ##49 - - - - - - a - - - - - - s . - - - - - - - - - - - - - - - - - ##40 - - on - - - - - - - - - of - - - - - peak - home - - - - - - - - - - - highest - - - - - - - - - - - - - - - considered - - - - - - - - - - - on - - - - explosive - - unlikely . -

(c) Attributions for the select-then-predict model trained with regularizer

Fig. 15: Attributions of the Shapley Value Sampling method used to analyze the predictions of the select-then-predict models. Words highlighted in green have higher attribution scores, meaning that they contribute towards the predicted label, while words highlighted in red contribute negatively to the prediction (i.e., they make the model less confident in the prediction). The three subfigures refer to the same full input text, that can be seen in fig. 15a. In figures fig. 15c and fig. 15b only the rationale selected by the model is shown. The attribution score represents the sum of the marginal contributions to the prediction of all the input tokens.