Making Document-Level Information Extraction Right for the Right Reasons

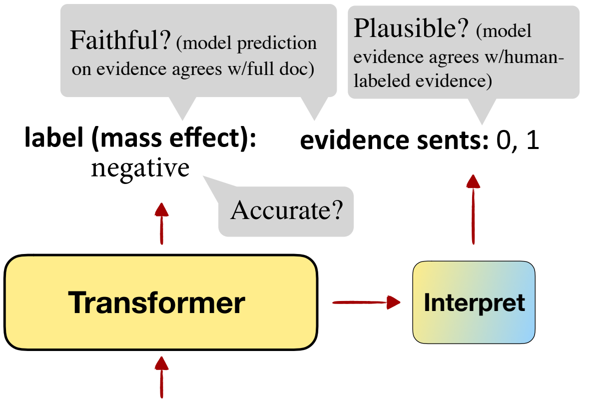
Anonymous ACL submission

Abstract

Document-level models for information extraction tasks like slot-filling are flexible: they can be applied to settings where information is not necessarily localized in a single sentence. For example, key features of a diagnosis in a radiology report may not be explicitly stated in one place, but nevertheless can be inferred from parts of the report's text. However, these models can easily learn spurious correlations between labels and irrelevant information. This work studies how to ensure that these models make correct inferences from complex text **and** make those inferences in an auditable way: beyond just being right, are these models "right for the right reasons?" We experiment with post-hoc evidence extraction in a predict-select-verify framework using feature attribution techniques. We show that regularization with small amounts of evidence supervision during training can substantially improve the quality of extracted evidence. We evaluate on two domains: a small-scale labeled dataset of brain MRI reports and a large-scale modified version of DocRED (Yao et al., 2019) and show that models' plausibility can be improved with no loss in accuracy.[1]

1 Introduction

Document-level information extraction (Yao et al., 2019; Christopoulou et al., 2019; Xiao et al., 2020; Guoshun et al., 2020) has seen great strides due to the rise of pre-trained models (Devlin et al., 2019). But in high-stakes domains like medical information extraction (Irvin et al., 2019; McDermott et al., 2020; Smit et al., 2020), machine learning models are still too error-prone to use broadly. Since they are not perfect, they typically play the role of assisting users in tasks like building cohorts (Pons et al., 2016) or in providing clinical decision support (Demner-Fushman et al., 2009).



[0] **Severe encephalomalacia** in the temporal lobes and frontal lobes bilaterally with reactive gliosis in the left frontal lobe. [1] Moderate **enlargement of the ventricular system**. [2] No abnormal enhancement. [3] Near complete opacification of the left maxillary sinus. …

Figure 1: Our basic model setup. A Transformer-based model makes document-level predictions on an example of our brain MRI reports. An interpretation method extracts the evidence sentences used by the model. Our system is evaluated according to the criteria of accuracy, faithfulness, and plausibility.

To be most usable in conjunction with users, these systems should not just produce a decision, but a justification for their answer. The ideal system therefore obtains high predictive accuracy, but also returns a rationale that allows a human to verify the predicted label (Rudie et al., 2019).

Our goal is to study document-level information extraction systems that are both accurate and which make predictions based on the correct information (Doshi-Velez and Kim, 2017). This process involves identifying what evidence the model actually used, verifying the model's prediction based on that evidence, and checking whether that evidence aligns with what humans would use, which would allow a user to quickly see if the system is correct. For example, in Figure 1, localizing the prediction of *mass effect* (a feature expressing whether there is evidence of brain displacement by a mass like a

[1] Code available upon publication.

tumor) to the first two sentences allows a trained user in a clinical decision support setting to easily verify what was extracted here. Our evidence extraction hews to principles of both faithfulness and plausibility (Jain et al., 2020; Jacovi and Goldberg, 2020; Miller, 2019).

Rather than use complex approaches with intermediate latent variables for extraction (Lei et al., 2016), we focus on what can be done with off-the-shelf pre-trained models (Liu et al., 2019) using post-hoc interpretation. We explore various interpretation methods to find key parts of each document that were used by the model. We ask two questions: first, can we identify the document sentences that truly contributed to the prediction (faithfulness)? Using the ranking of sentences provided by an interpretation method, we extract a set of sentences where the model returns nearly the same prediction as before, thus *verifying* that these sentences are a sufficient explanation for the model. Second, do these document sentences align with what users annotated (plausibility)? Unsurprisingly, we find that this alignment is low in a basic Transformer model.

To further improve the alignment with human annotation, we consider injecting small amounts of sentence-level supervision. Critically, in the brain MRI extraction setting we consider (see Table 1), large-scale sentence-level annotation is not available; most instances in the dataset only have document-level labels from existing clinical decision support systems, making it a weakly-supervised setting (Pruthi et al., 2020; Patel et al., 2020). We explore two methods for using this small amount of annotation, chiefly based around supervising or regularizing the model's behavior. One notion is entropy maximization: the model should be uncertain when it isn't exposed to sufficient evidence (Feng et al., 2019). Another is attention regularization where the model is encouraged to attend to key pieces of evidence. While attention is not entirely connected with what the model uses (Jain and Wallace, 2019), we can investigate whether this leads to a model whose explanations leverage this information more heavily.

We validate our methods first on a small dataset of radiologists' observations from brain MRIs. These reports are annotated with document-level key features related to different aspects of the report, which we want to extract in a faithful way. We see positive results here even in a small-data condition, but to understand how this method would scale with larger amounts of data, we adapt the DocRED relation extraction task (Yao et al., 2019) to be a document-level classification task. The question of which sentence in the document describes the relation between the two entities, if there even is one, is still quite challenging, and we show our techniques can lead to improvements in a weakly-labeled setting here as well.

Report Finding

[0] **Severe encephalomalacia** in the temporal lobes and frontal lobes **bilaterally** with reactive **gliosis** in the left frontal lobe. *[1]* Moderate **enlargement of the ventricular system**. *[2]* **No abnormal enhancement**. *[3]* Near complete opacification of the left maxillary sinus. ...

mass_effect: negative	**evid**: *[0, 1]*	implicit
side: bilateral	**evid**: *[0]*	explicit
t2: increased	**evid**: *[0]*	implicit
contrast_enhancement: No	**evid**: *[2]*	explicit

Table 1: Example from annotated brain MRI reports. Labels and supporting evidence for 4 key features are annotated for this example report presented. "Explicit" means the label of given key feature can be directly inferred by the highlighted terms; "implicit" instead indicates that it requires domain knowledge and potential reasoning skills to label. We want the model to identify implicit features while not leveraging dataset biases or reasoning incorrectly about explicit ones.

Our contributions are (1) We apply evidence extraction methods to document-level classification and slot-filling tasks, emphasizing a new brain MRI dataset that we annotate. (2) We explore using weak sentence-level supervision in two techniques adapted from prior work; (3) We evaluate pre-trained models and evidence extraction through various interpretation methods for plausibility compared to human annotation, while ensuring faithfulness of the evidence.

2 Background

2.1 Motivation

We start with an example from a brain MRI report in Table 1. Medical information extraction involves tasks such as identifying important medical terms from text (Irvin et al., 2019; Smit et al., 2020) and normalizing names into standard concepts using domain-specific ontologies (Cho et al., 2017). One application in clinical decision support, shown here, requires extracting the values of certain key features (clinical findings) from these reports or medical images (Rudie et al., 2021; Duong et al., 2019).

This extraction should be accurate, but it should also make predictions that are correctly sourced, to facilitate review by a radiologist or someone else using the system (Rauschecker et al., 2020; Cook et al., 2018).

The finding section of a brain MRI report often describes these key features in both explicit and implicit ways. For instance, contrast enhancement, one of our key features, is mentioned explicitly much of the time; see *no abnormal enhancement* in the third sentence. A rule-based system can detect this type of evidence easily. But some key features are harder to identify and require reasoning over context and draw on implicit cues. For example, *severe encephalomalacia* in the first sentence and *enlargement of the ventricular system* in the following sentence are both implicit signs of positive mass effect and either is sufficient to infer the label. It is significantly harder to built a rule-based extractor for this case. Learning-based systems have the potential to do much better here, but lack of understanding about their behavior can lead to hard-to-predict failure modes, such as acausal prediction of key features (e.g., inferring evidence about mass effect from a hypothesized diagnosis somewhere in the report, where the causality is backwards).

Our work aims to leverage the ability of learning-based systems to capture implicit features while improving their ability to make predictions that are sourced from the correct evidence and can be easily verified.

2.2 Problem Setting

The problem we tackle in this work can be viewed as document-level classification. Let $D = \{x_1, \dots, x_n\}$ be a document consisting of n sentences. The document is annotated with a set of labels (t_i, y_i) where t_i is an auxiliary input specifying a particular task for this document (e.g., mass effect) and y_i is the label associated with that task from a discrete label space $\{1, \dots, d\}$. In our adaptation of the DocRED task, we consider $t = (e_1, e_2)$ to classify the relationship (if any) between a pair of entities (e_1, e_2) in a document, defined in Section 4.1.2.

Our method takes a pair (D, t) and then computes the label $\hat{y}_t$ from a predictor $\hat{y}_t = f(D, t)$. We can then extract **evidence**, a set of sentences, post-hoc using a separate procedure g such as a feature attribution method: $\hat{E}_t = g(f, D, t)$

Supervision In addition to the labels y_t, we assume access to a small number of examples with additional supervision in each domain. That is, for a (D, t, y_t) triple, we also assume we are given a set $E = \{x_{i_1}, \dots, x_{i_m}\}$ of ground-truth evidence with sentence indices $\{i_1, \dots, i_m\}$. This evidence should be sufficient to compute the label, but not always necessary; for example, if multiple sentences can contribute to the prediction, they might all be listed as supporting evidence here. See Section 3.3 for more details.

2.3 Related Work

Our work fits into a broader thread of work on information extraction with partial annotation (Han et al., 2020). Due to the cost of collecting large-scale data with good quality, distant supervision (DS) (Mintz et al., 2009) and ways to denoise auto-labeled data from DS (Surdeanu et al., 2012; Wang et al., 2018) have been widely explored. However, the sentence-level setting typically features much less ambiguity about evidence needed to predict a relation compared to the document-level setting we explore. Several document-level RE datasets (Li et al., 2016a; Peng et al., 2017) have been proposed as well as efforts to tackle these tasks (Christopoulou et al., 2019; Xiao et al., 2020; Guoshun et al., 2020), which we explicitly build from.

Explanation techniques To identify the sentences that the model considers as evidence, we draw on a recent body of work in explainable NLP focused on identifying salient features of the input. These primarily consist of input attribution techniques, such as LIME (Ribeiro et al., 2016), input reductions (Li et al., 2016b; Feng et al., 2018), attention-based explanations (Bahdanau et al., 2015) and gradient-based methods (Simonyan et al., 2014; Selvaraju et al., 2017; Sundararajan et al., 2017; Shrikumar et al., 2017). In present work, we extract rationales using commonly used model interpretation methods (described in Section 3.2) and focus on doing a thorough evaluation of the capabilities of DeepLIFT (Shrikumar et al., 2017) given its competitive performance in our interpretation methods comparison (Appendix B).

Frameworks for interpretable pipelines Our goal of building a system grounded in evidence draws heavily on recent work on attribution techniques and model explanations, particularly notions of faithfulness and plausibility. *Faithfulness* refers to how accurately the explanation provided by the model truly reflects the information it used in the reasoning process (Jain et al., 2020). On the other hand, *plausibility* indicates to what extent the interpretation provided by the model makes sense to a person.[2]

"*Select-then-predict*" approaches are one way to enforce faithfulness in pipelines (Jain et al., 2020): important snippets from inputs are extracted and passed through a classifier to make predictions. Past work has used hard (Lei et al., 2016) or soft (Zhang et al., 2016) rationales, and other work has explicitly looked at tradeoffs in the amount of text extracted (Paranjape et al., 2020).

Jacovi and Goldberg (2020) note several problems with this setup. Our work aims to align model behavior with what cues we expect a model to use (plausibility), but uses the predict-select-verify paradigm (Jacovi and Goldberg, 2020) to ensure that these are actually sufficient cues for the model. Like our work, Pruthi et al. (2020) simultaneously trained a BERT-based model (Devlin et al., 2019) for the prediction task and a linear-CRF (Lafferty et al., 2001) module on top of it for the evidence extraction task with shared parameters. Compared to their work, we focus explicitly on what can be done with pre-trained models alone, not augmenting the model for evidence extraction.

3 Methods

The systems we devise take (D, t) pairs as input and return (a) predicted labels $\hat{y}_t$ for each t; (b) sets of extracted evidence sentences $\hat{E}_t$ from an interpretation method. Figure 1 shows the basic setting.

3.1 Transformer Classification Model

We use RoBERTa (Liu et al., 2019) as our document classifier. RoBERTa is a strong method that holds up even against more recent baselines with architectures designed for DocRED (Zhou et al., 2021). For each of our two domains, we use different pre-trained weights, as described in the training details in Appendix A. The task inputs are described in Section 4.1.

[2]The ERASER benchmark (DeYoung et al., 2020) is a notable recent effort to evaluate explanation plausibility. However, we do not consider it here; we focus on the document-level classification setting, and many of the ERASER tasks are not suitable or relevant for the approaches we consider, either being not natural (FEVER) or not having the same challenges as document-level classification.

3.2 Interpretation for Evidence Extraction

Given any interpretation method as well as our model $\hat{y}_t = f(D, t)$, we compute attribution scores with respect to the predicted class y_t for each token in the RoBERTa input representation. We then average over the absolute value of attribution score for each token in that sentence to give sentence-level scores $\{s_1, \ldots, s_n\}$. These give us a ranking of the sentences. Given a fixed number of evidence sentences k to extract, we can extract the top k sentences by these scores.

We experiment with the following four widely used interpretation techniques in the present work. **LIME** (Ribeiro et al., 2016) offers explanations of an input by approximating the model's predictions locally with an interpretable model. **Input Gradient** (Hechtlinger, 2016) and **Integrated Gradients** (Sundararajan et al., 2017) use gradients of the label with respect to the input to assess input importance; Integrated Gradients approximates the integral of this gradient with respect to the input along a straight path from a reference baseline.[3] **DeepLIFT** (Shrikumar et al., 2017) attributes the change in the output from a reference output in terms of the difference in input from the reference input. Unless stated otherwise, we use DeepLIFT as our interpretation method, since it achieves the best results (comparable to Input Gradient) among the four interpretation options. Full comparison of interpretation methods is in Appendix B.

To verify the extracted evidence (Jacovi and Goldberg, 2020), our main technique (SUFFICIENT) feeds the model increasingly larger subsets of the document ranked by attribution scores (e.g., first $\{s_{\max}\}$, then $\{s_{\max}, s_{\text{2nd-max}}\}$, etc.) until it (a) makes the same prediction as when taking the whole document as input and (b) assigns that prediction at least λ times the probability[4] when the whole document is taken as input. We consider this attribution faithful: it is a subset of the input supporting the model's decision judged as important by the attribution method.

3.3 Improving Evidence Extraction

While many document-level extraction settings do not have sentence-level attributions labeled for ev-

[3]We use the most typical baseline that consists of replacing the inputs in D with `[MASK]` tokens from RoBERTa.

[4]The value of λ is a tolerance hyper-parameter for selecting sentences and it set to 0.8 throughout the experiments. Our method is robust to the choice of λ in a reasonable range, as shown in Appendix B.

ery decision, one can in practice annotate a small fraction of a dataset with such ground-truth rationales. This is indeed the case for our brain MRI case study. Past work has shown significant benefits from integrating this supervision into learning (Strout et al., 2019; Dua et al., 2020; Pruthi et al., 2021).

Assume that a subset of our labeled data consists of (D, t, y_t, E_t) tuples with ground truth evidence sentence indices $E_t = \{i_1, ..., i_m\}$. We consider two modifications to our model training, namely attention regularization (Pruthi et al., 2021), entropy maximization (Feng et al., 2018), and their combination. An illustration of both methods is shown in Figure 2.

Attention regularization Attention regularization encourages our model $f(D, t)$ to leverage more information from E_t. Specifically, let $A = \{\alpha_1, ..., \alpha_n\}$ be the attention vector from the `[CLS]` token in the final layer to all tokens in D. During learning, we add the following loss to the training objective: $\ell_{attn} = -\log \sum_{i \in E_t} \alpha_i$, encouraging the model to attend to any token i in the labeled sentence-level evidence set.

Entropy maximization When there is no sufficient information contained in the text to infer any predictions, entropy maximization encourages a model to be uncertain, represented by a uniform probability distribution across all classes (DeYoung et al., 2020; Feng et al., 2019). Doing so should encourage the model to *not* make predictions based on irrelevant sentences. We can achieve this by taking a reduced document $D' = D \setminus E_t$ as input by removing evidence E_t from original document D. We treat (D', t) pairs as extra training examples where we aim to maximize the entropy $-\sum_y P(y|D') \log P(y|D')$ over all possible y.[5]

4 Experiments

4.1 Datasets and Evaluation Metrics

We investigate our methods on (a) a small collection of brain MRI reports from radiologists' observations; and (b) a modified version of the DocRED datatset. The statistics for both datatsets are included in Appendix D. For both datasets, we evaluate on task accuracy (captured by either accuracy or prediction macro-F1) as well as evidence selection

[5] We found this to work better than enforcing a uniform distribution over attention, which is much harder for the model to achieve.

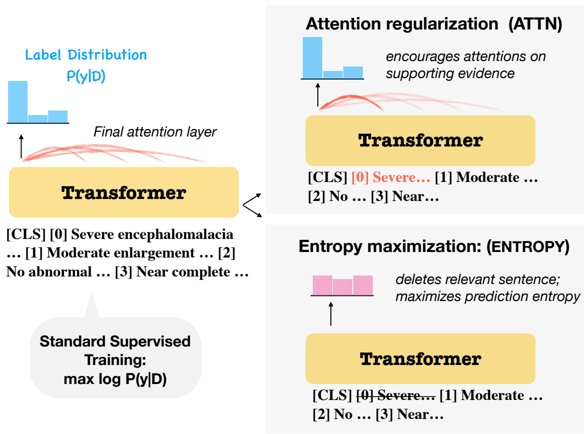


Figure 2: An illustration of attention regularization and entropy maximization using the example in Table 1. The model is predicting the label for key feature *t2*.

accuracy (macro-F1) or precision, measuring how well the model's evidence selection aligns with human annotations. We will use the SUFFICIENT method defined in Section 3.2 to select evidence sentences which guarantee that our predictions on the given evidence subsets will match the model's predictions on the full document. For the brain MRI report dataset, we evaluate evidence extraction by precision since human annotators typically only need to refer to one sentence to reach the conclusion but our model and baselines may extract more than one sentence.

4.1.1 Brain MRI Reports

We present a new dataset of de-identified radiology reports from brain MRIs. It consists of the "findings" sections of reports, which present observations about the image, with labels for preselected key features by attending physicians and fellows. Crucially, these features are labeled **based on the original radiology image**, not the report. The document-level labels are therefore noisy because the radiologists' labels may disagree with the findings written in the report.

A key feature is an observable variable t, which can take on d_t possible values. We focus on the evaluation of two key features, namely *contrast enhancement* and *mass effect*, since they appear in most of manually annotated reports. For our RoBERTa classification model, we only feed the document and train separate classifiers for each key feature, with no shared parameters between these.

Annotation We have a moderate number (327) of reports that have noisy labels from the process above. We treat these as our training set. However, all of these labels are document-level.

To evaluate models' performance on more fine-grained evidence labels, we randomly select 86 unlabeled reports (not overlapping with the 327 for training) and asked four radiology residents to (1) assign key feature labels and reach consensus, while (2) highlighting sentences that support their decision making. We use Prodigy[6] as our annotation interface. See Appendix E for more details about our annotation instructions.

Pseudo sentence-level supervision Since we have limited number of annotated reports for evaluation, we need a way to prepare weak sentence-level supervision (E_t) while training. To achieve this, we use sentences selected by our rule-based system as pseudo evidence to supervise models' behavior. We use 10% of this as supervision while training for consistency with the DocRED setting.

Rule-based system Our rule-based system uses keyword matching to identify instances of mass effect and contrast enhancement in the reports, and negspaCy to detect negations of these key features.

Data split For the results in Section 5, we evaluate on reports that contain ground truth fine-grained annotations for either contrast enhancement or mass effect, respectively. There are 64 and 68 out of 86 documents total in each of these categories. We call this the BRAINMRI set. When we restrict to this set for evaluation, all of the documents we study where the annotators labeled something related to *contrast enhancement* end up having an explicit mention of it. However, for *mass effect*, this is not always the case; Table 9 in Appendix shows an example where mass effect is discussed implicitly in the first sentence.

4.1.2 Adapted DocRED

DocRED (Yao et al., 2019) is a document-level relation extraction (RE) dataset with large scale human annotation of relevant evidence sentences. Unlike sentence-level RE tasks (Qin et al., 2018; Alt et al., 2020), it requires reading multiple sentences and reasoning about complex interactions between entities. We adapt this to a document-level relation classification task: a document D and two entity mentions e_1, e_2 within the document are provided and the task is to predict the relation r between e_1 and e_2. We synthesize these examples from the original dataset and sample random entity pairs from documents to which we assign an *NA* class to construct negative pairs exhibiting no relation.

[6] `https://prodi.gy`

Model Names	Input Text
DIRECT	None
FULLDOC	Full document
ENT	Sentences containing at least one of the two query entities
FIRST2	First two sentences from a doc.
FIRST3	First three sentences from a doc.
BESTPAIR	Two sentences yielding highest prediction prob. (incl. variants using regularization)
SUFFICIENT	Sufficient sentences selected by DL (incl. variants using regularization)

Table 2: Model names used in the experiments and their associated evidence given as inputs.

The model input is represented as: `[CLS]<ent-1>[SEP]<ent-2>[SEP]<doc>[SEP]`. We use the encoding of `[CLS]` in the last layer to make predictions.

To make the setting more realistic, we *do not* use the large-scale evidence annotation and assume there is limited sentence-level supervision available. To be specific, we include 10% fine-grained annotations in our adapted DocRED dataset.

4.2 Models

Due to richer and higher-quality supervisions in the DocRED setting, we conduct a larger set of ablations and comparisons there. We compare against a subset of these models in the radiology setting.

Baselines We consider a number of baselines for adapted DocRED which return both predicted labels and evidence. (1) DIRECT predicts the relation directly from the entity pairs without any sentences as input, using a model trained with just these inputs. (2) FULLDOC takes the full document as selected evidence and uses the base RoBERTa model (3) ENT takes all sentences with entity mentions e_1 and e_2 as input; (4) FIRST2, FIRST3 retrieve the first 2 and 3 sentences from a document, respectively; and (5) BESTPAIR chooses the best sentence pair by first taking each individual sentence as input to the model and then picking top two sentences having highest probabilities on their predictions.

SUFFICIENT is our main method for both datasets, which we then augment with additional supervision as described in Section 3.3. We use subscripts `attn`, `entropy`, `both` and `none` to represent attention regularization, entropy maximization, the combination of two, and neither. Both BestPair and Sufficient methods leverage backbone RoBERTa models trained with loss functions mentioned above, differing only in their evidence selection.

Table 2 summarizes the abbreviated names of models and their inputs. Training details are described in Appendix A.

Metrics We report both the accuracy and F_1 for the model (**Full Doc**) as well as evaluation of **Evidence** selection compared to human judgments, either precision or F_1. We also report results in the **Reduced Doc** setting, where only the selected evidence sentences are fed to the RoBERTa model (trained over whole documents) as input. For our Sufficient method, this accuracy is the same as the full method by construction, but note that it can differ for other methods. This reduced setting serves as a sanity check for the faithfulness of our explanation techniques.

Note once again that accuracy in the Full Doc case can differ for our methods that are trained with different regularization schemes, as these yield different models that return different predicted labels in addition to different evidence.

5 Results

5.1 Results on Brain MRI

Table 3 shows the performance of our models and baselines in terms of label prediction and evidence extraction. For each result, we perform a paired bootstrap test comparing to Sufficient$_{\text{none}}$. We underline results that are better at a significance level of $p = 0.05$ on the corresponding metrics. In the *mass effect* setting, our Sufficient$_{\text{both}}$ model achieves the highest evidence extraction precision of the learning-based models, exceeds FullDoc, First2/3, and BestPair on the metric by a large margin, and nearly matches that of the rule-based system. It is difficult to be more reliable than a rule-based system, which will nearly always make correctly-sourced predictions. But this model is able to **combine that reliability with the higher F_1 of a learned model**. Note that due to the high base rates of certain findings, we focus on F_1 instead of accuracy. We see a similar pattern on *contrast enhancement* as well, although the evidence precision is lower in that case.

These results show that learning-based systems make accurate predictions in this domain, and that

Model	Label: Full Doc Acc	Full Doc F1	Reduced Doc Acc	Reduced Doc F1	Evidence: Pre	Len
Mass Effect						
FullDoc	66.6	42.1	66.6	42.1	16.5	10.1
First2	–	–	82.4	45.2	21.3	2.00
First3	–	–	82.4	45.2	24.0	3.00
Rule	77.9	11.8	77.9	11.8	<u>**84.8**</u>	1.46
BestPair$_{\text{none}}$	66.6	42.1	82.4	52.2	24.3	2.00
BestPair$_{\text{both}}$	76.7	**60.0**	79.4	44.3	<u>50.7</u>	2.00
Sufficient$_{\text{none}}$	66.6	42.1			16.5	2.84
Sufficient$_{\text{attn}}$	69.2	47.6	Identical to		65.6	2.31
Sufficient$_{\text{entropy}}$	45.3	0.0	Full Doc		15.8	2.50
Sufficient$_{\text{both}}$	76.7	**60.0**			<u>77.8</u>	1.51
Contrast Enhancement						
FullDoc	69.5	60.9	69.5	60.9	13.5	10.1
First2	–	–	67.2	55.3	14.1	2.00
First3	–	–	70.3	62.4	14.6	3.00
Rule	68.8	56.5	68.8	56.5	<u>**87.1**</u>	1.67
BestPair$_{\text{none}}$	69.5	60.9	73.4	67.7	10.9	2.00
BestPair$_{\text{both}}$	<u>90.8</u>	<u>**87.2**</u>	<u>89.1</u>	<u>88.4</u>	<u>54.7</u>	2.00
Sufficient$_{\text{none}}$	69.5	60.9			33.5	2.84
Sufficient$_{\text{attn}}$	<u>85.8</u>	<u>81.0</u>	Identical to		<u>60.7</u>	2.48
Sufficient$_{\text{entropy}}$	71.5	59.5	Full Doc		25.2	2.55
Sufficient$_{\text{both}}$	<u>90.8</u>	<u>**87.2**</u>			<u>71.7</u>	1.50

Table 3: Model performance on BrainMRI. Models are evaluated under two settings by taking (a) full document (Full Doc); (b) selected evidence (Reduced Doc) as inputs. Rule is the baseline mentioned in Section 4.1.1. *Pre* stands for the precision of evidence selection, and *Len* is the average number of sentences extracted. Underlined results are better than Sufficient$_{\text{none}}$ on the corresponding metric according to a paired bootstrap test with $p = 0.05$.

Model	Label: Full Doc Acc	Full Doc F1	Reduced Doc Acc	Reduced Doc F1	Evidence: F1	Len
Direct	–	–	66.4	45.3	–	–
FullDoc	83.0	66.0	83.0	66.0	34.9	8.03
First2	–	–	75.3	58.1	47.9	2.00
First3	–	–	77.5	60.7	44.6	3.00
Ent	–	–	82.4	65.4	61.5	3.93
BestPair$_{\text{none}}$	83.0	66.0	73.9	55.3	39.2	2.00
BestPair$_{\text{attn}}$	83.2	65.0	73.4	53.5	43.9	2.00
BestPair$_{\text{entropy}}$	81.8	64.2	78.5	58.2	52.3	2.00
BestPair$_{\text{both}}$	82.7	66.5	81.6	65.3	66.2	2.00
Sufficient$_{\text{none}}$	83.0	66.0			67.2	1.42
Sufficient$_{\text{attn}}$	83.2	65.0	Identical to		<u>70.3</u>	1.45
Sufficient$_{\text{entropy}}$	81.8	64.2	Full Doc		<u>69.9</u>	**1.65**
Sufficient$_{\text{both}}$	82.7	66.5			<u>**73.1**</u>	**1.65**
human	–	–	–	–	–	1.59

Table 4: Model performance on adapted DocRED. Models are evaluated under two settings as in BrainMRI. Underlined results are better than Sufficient$_{\text{none}}$ on the corresponding metric according to a paired bootstrap test with $p = 0.05$.

Model	Mass Effect		Ctr. Enhance.	
	Mean	Max	Mean	Max
SUFFICIENT$_{\text{none}}$	7.3	7.4	28.6	29.8
SUFFICIENT$_{\text{both}}$	**18.9**	**19.2**	**37.9**	**42.0**

Table 5: Distributions of attribution mass over explicit cues ("enhancement" for *contrast enhancement* and "effect" for *mass effect*) for our best model and the baseline. Mean/Max is the mean of instance-wise average/maximum of the normalized attribution mass falling on the given token.

their evidence extraction can be improved with better training, even in spite of the small size of the training set. In section 5.2, we focus on the adapted DocRED setting, which allows us to examine our model's performance in a higher-data regime.

Attribution scores are more peaked at the occurrence of key terms. We conduct analysis on how the attribution scores from SUFFICIENT$_{\text{both}}$ are peaked around the correct evidence compare to that from SUFFICIENT$_{\text{none}}$ using our manually annotated set BRAINMRI. We compute the mean of the instance-wise average and maximum of the normalized attribution mass falling into a few explicit tokens: *enhancement* for contrast enhancement and *effect* for mass effect, which are common explicit indicators in the context of specified key features. The results in Table 5 show attribution scores being peaked around the correct terms, highlighting that these models can be guided to not only make correct predictions but attend to the right information.

Table 9 in the Appendix shows visualizations of attribution scores for an example in BRAINMRI using DeepLIFT. Even though baseline models make correct predictions, their attribution mass is diffused over the document. With the help of regularization, our model is capable of capturing implicit cues such as *downward displacement of the brain stem*, although it is trained on an extremely small training set with only explicit cues like *mass effect* in a weak sentence-level supervision framework.

5.2 Results on Adapted DocRED

Comparison to baselines Table 4 shows that the ENT baseline is quite strong at DocRED evidence extraction. However, our best method still exceeds this method on both label accuracy as well as evidence extraction while extracting more succinct explanations. We see that the ability to extract a variable-length explanation is key, with FIRST2, FIRST3 and BESTPAIR performing poorly. Notably, these methods exhibit a drop in accuracy in the reduced doc setting for each method compared to the full doc setting, showing that the explanations extracted are not faithful.

Learning-based models with appropriate regularization perform relatively better in this larger-data setting From Table 3 and Table 4, we can observe that various regularization techniques applied to SUFFICIENT models maintain or improve overall model performance on both key feature and relation classification. We see that our SUFFICIENT methods do not compromise on accuracy but make predictions based on plausible evidence sets, which is more evident when we have richer training data. We perform further error analysis in Appendix F.

Faithfulness of techniques One may be concerned that, like attention values (Jain and Wallace, 2019), our feature attribution methods may not faithfully reflect the computation of the model. We emphasize again that the SUFFICIENT paradigm on top of the DeepLIFT method *is* faithful by our definition. For a model f, we measure the faithfulness by checking the agreement between $\hat{y} = f(D, t)$ and $y' = f(\hat{E}_t, t)$, where $\hat{E}_t$ is the extracted evidence we feed into the same model under the reduced document setting. This is shown for all methods in the "Reduced doc" columns in Tables 3 and 4. We see a drop in performance from techniques such as BESTPAIR: the full model does not make the same judgment on these evidence subsets, but by definition it does in the SUFFICIENT setting.

As further evidence of faithfulness, we note that only a relatively small number of evidence sentences, in line with human annotations, are extracted in the SUFFICIENT method. These small subsets are indicated by feature attribution methods *and* sufficient to reproduce the original model predictions with high confidence, suggesting that these explanations are faithful.

6 Conclusion

In this work, we develop techniques to employ small amount of data to improve reliability of document-level IE systems in two domains. We systematically evaluate our model from perspectives of faithfulness and plausibility and show that we can substantially improve models' capability in focusing on supporting evidence while maintaining their predictive performance, leading to models that are "right for the right reasons."

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. Probing linguistic features of sentence-level representations in neural relation extraction. In *Proceedings of ACL*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hyejin Cho, Wonjun Choi, and Hyunju Lee. 2017. A method for named entity normalization in biomedical articles: application to diseases and plants. *BMC Bioinformatics*, 18(1).

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Tessa Cook, James C. Gee, R. Nick Bryan, Jeffrey T. Duda, Po-Hao Chen, Emmanuel Botzolakis, Suyash Mohan, Andreas Rauschecker, Jeffrey Rudie, and Ilya Nasrallah. 2018. Bayesian network interface for assisting radiology interpretation and education. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*. SPIE.

Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772. Biomedical Natural Language Processing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*.

Dheeru Dua, Sameer Singh, and Matt Gardner. 2020. Benefits of intermediate annotations in reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5627–5634, Online. Association for Computational Linguistics.

M.T. Duong, J.D. Rudie, J. Wang, L. Xie, S. Mohan, J.C. Gee, and A.M. Rauschecker. 2019. Convolutional neural network for automated FLAIR lesion segmentation on clinical brain MR imaging. *American Journal of Neuroradiology*, 40(8):1282–1290.

Shi Feng, Eric Wallace, and Jordan Boyd-Graber. 2019. Misleading failures of partial-input baselines. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nan Guoshun, Guo Zhijiang, Sekulić Ivan, and Lu Wei. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of ACL*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Jie Zhou, and Maosong Sun. 2020. More data, more relations, more context and more openness: A review and outlook for relation extraction. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 745–758, Suzhou, China. Association for Computational Linguistics.

Yotam Hechtlinger. 2016. Interpretation of prediction models using the input gradient. In *Proceedings of the NeurIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David Mong, Safwan Halabi, Jesse Sandberg, Ricky Jones, David Larson, Curtis Langlotz, Bhavik Patel, Matthew Lungren, and Andrew Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:590–597.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016a. BioCreative v CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. Understanding neural networks through representation erasure. *ArXiv*, abs/1612.08220.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *arXiv*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Matthew B. A. McDermott, Tzu-Ming Harry Hsu, Wei-Hung Weng, Marzyeh Ghassemi, and Peter Szolovits. 2020. Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. *CoRR*, abs/2006.15229.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Dhruvesh Patel, Sandeep Konam, and Sai P. Selvaraj. 2020. Weakly supervised medication regimen extraction from medical conversations. In *ClinicalNLP@EMNLP*.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen tau Yih. 2017. Cross-sentence n-ary relation extraction with graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Ewoud Pons, Loes M. M. Braun, M. G. Myriam Hunink, and Jan A. Kors. 2016. Natural language processing in radiology: A systematic review. *Radiology*, 279(2):329–343. PMID: 27089187.

Danish Pruthi, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Weakly- and semi-supervised evidence extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics.

Danish Pruthi, Bhuwan Dhingra, Livio Baldini Soares, M. Collins, Zachary C. Lipton, Graham Neubig, and William W. Cohen. 2021. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Andreas M. Rauschecker, Jeffrey D. Rudie, Long Xie, Jiancong Wang, Michael Tran Duong, Emmanuel J. Botzolakis, Asha M. Kovalovich, John Egan, Tessa C. Cook, R. Nick Bryan, Ilya M. Nasrallah, Suyash Mohan, and James C. Gee. 2020. Artificial intelligence system approaching neuroradiologist-level differential diagnosis accuracy at brain MRI. *Radiology*, 295(3):626–637.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Jeffrey Rudie, Long Xie, Jiancong Wang, Jeffrey Duda, Joshua Choi, Raghav Mattay, Po-Hao Chen, R Nick Bryan, Emmanuel Botzolakis, Ilya Nasrallah, Tessa Cook, Suyash Mohan, James Gee, and Andreas Rauschecker. 2019. Artificial Intelligence System for Automated Brain MR Diagnosis Performs at Level of Academic Neuroradiologists and Augments Resident Performance. In *Proceedings of the Society for Imaging Informatics in Medicine (SIIM)*.

Jeffrey D. Rudie, Jeffrey Duda, Michael Tran Duong, Po-Hao Chen, Long Xie, Robert Kurtz, Jeffrey B. Ware, Joshua Choi, Raghav R. Mattay, Emmanuel J. Botzolakis, James C. Gee, R. Nick Bryan, Tessa S. Cook, Suyash Mohan, Ilya M. Nasrallah, and Andreas M. Rauschecker. 2021. Brain MRI deep learning and bayesian inference system augments radiology resident performance. *Journal of Digital Imaging*, 34(4):1049–1058.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online. Association for Computational Linguistics.

Julia Strout, Ye Zhang, and Raymond Mooney. 2019. Do human rationales improve machine explanations? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–62, Florence, Italy. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465, Jeju Island, Korea. Association for Computational Linguistics.

Xiaozhi Wang, Xu Han, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2018. Adversarial multi-lingual neural relation extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1156–1166, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chaojun Xiao, Yuan Yao, Ruobing Xie, Xu Han, Zhiyuan Liu, Maosong Sun, Fen Lin, and Leyu Lin. 2020. Denoising relation extraction from document-level distant supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Ye Zhang, Iain Marshall, and Byron C. Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

A Implementation Details

We train all RoBERTa models for 15 epochs with early stopping using 1 TITAN-Xp GPU. We use AdamW (Loshchilov and Hutter, 2019) as our optimizer and initialize the model with `roberta-base` for DocRED and `biomed-roberta-base` (Gururangan et al., 2020) for brain MRI data, both with 125M parameters. The batch size is set to 16 for RoBERTa models trained with both attention regularization and entropy maximization and 8 for models with other loss functions, and the learning rate is 1e-5 with linear schedule warmup.

The maximum number of tokens in each document is capped at 296 for modified DocRED and 360 for radiology reports. These numbers are chosen such that the number of tokens for around 95% of the documents is within these limits. Remaining tokens are clipped from the input. The hidden state of the `[CLS]` token from the final layer is fed as input to a linear projection head to make predictions. The average training time for each model is around 4 GPU hours. We will release our code upon publication.

B Interpretation Methods Comparison

We evaluate four interpretation methods on SUFFICIENT$_{\text{none}}$ and SUFFICIENT$_{\text{both}}$ using adapted DocRED. These methods are widely used in the literature, namely Integrated Gradients, LIME, DeepLIFT, and Input Gradient, as discussed in Section 3.2. We compare their evidence extraction capabilities by selecting a wide range of λ, which controls the number of sentences to be selected.

Results are shown in Figure 3. The four techniques generally perform similarly, with DeepLIFT and Input Gradient performing slightly better. For each interpretation method, the result of SUFFICIENT$_{\text{both}}$ is significantly better than that of SUFFICIENT$_{\text{none}}$. Similar values of λ between 0.8 and 0.9 (preferring to select more sentences) work well across all methods. Table 6 shows the comparison over the threshold ($\lambda = 0.8$) we choose for our experiments in Section 5. In general, our method is robust to model interpretation techniques and evidence selection threshold λ.

Sentence ranking step mentioned in Section 3.2 requires 0.3 GPU hour for Input Gradient and DeepLIFT, 2.5 GPU hours for Integrated Gradients, and 14 GPU hours for LIME. We choose 30 steps to approximate the integral for Integrated Gradients and 100 samples for each input to train the surrogate interpretable model (a linear model in our case) for LIME.

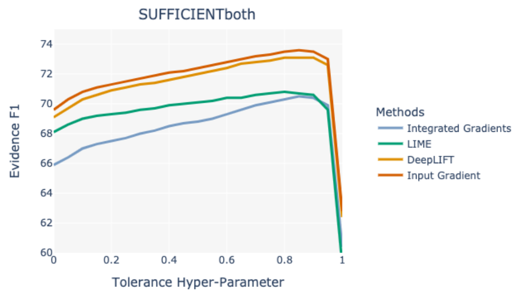


Figure 3: Evidence F1 on adapted DocRED under four model interpretation methods for SUFFICIENT$_{\text{both}}$ over a wide range of λ.

	SUFFICIENT$_{\text{none}}$	SUFFICIENT$_{\text{both}}$
LIME	54.0	70.8
Integrated Gradients	60.6	70.3
DeepLIFT	67.3	73.1
Input Gradient	68.3	73.5

Table 6: Evidence F1 on adapted DocRED under four model interpretation methods for SUFFICIENT$_{\text{none}}$ and SUFFICIENT$_{\text{both}}$ when $\lambda = 0.8$.

C Limitations and Risks

There are a few limitations of our work. First, we currently test our methods on document-level classification and slot-filling tasks, but there are other task formats like span extraction that we do not investigate here. Second, we focus on off-the-shelf pre-trained models (i.e. RoBERTa) in this paper, though we believe our methods could also be applied and adopted to other models. Finally, and most critically, the interpretation techniques we use are all fundamentally approximate; while visualizing model rationales can be useful in the context of clinical decision support systems, our evidence sets are not proof positive that a model's predictions are reliable. Such systems need to be carefully deployed to avoid misleading practitioners into trusting them too readily. We view this as the principal risk of our work.

D Dataset statistics

We provide the statistics for both adapted DocRED and brain MRI reports dataset in Table 7. Both datasets are in English and the DocRED dataset is publicly available at `https://github.com/thunlp/DocRED`.

Our use of the brain MRI reports is covered under IRB [anonymized for peer review].

E Annotation Instructions

We recruited four radiology residents to make annotations. They did not receive compensation for this project specifically. The annotation instructions for the BrainMRI dataset are provided in Figure 4. These were developed jointly with the annotators. In particular, decisions to exclude normal brain activity and confounders such as SVID were made to increase interannotator agreement after an initial round of annotation, making it easier for the labeling to focus on a single core disease or diagnosis per report.

F Error Analysis

The first example in Table 8 shows a representative case where our model predicts the correct relation and extracts reasonable supporting evidence. Unsurprisingly, this happens most often in simple cases when reasoning over the interaction of sentences is not required.

We observe a few common types of errors. First, we see **potential alternatives for relations or evidence extraction**. From around 60% of our randomly selected error cases, our model either predicts debatably correct relations or picks sentences that are related but not perfectly aligned with human annotations. The second row in Table 8 illustrates an example where the two entities exhibit multiple relationships; the model's prediction is correct (*Vienna* is place where *Martinelli* was both born and died), but differs from the annotated ground truth and supporting evidence. Such relations are relatively frequent in this dataset; a more complex multi-label prediction format is necessary to fully support these.

Another type of error is **complex logical reasoning**. Even if our model can extract right evidence, it still fails in around 10% of random error cases requiring sophisticated reasoning. For example, to correctly predict the relation between *Theobald Tiger* and *21 December 1935* in the third example in Table 8, a model needs to recognize that *Theobald Tiger* and *Kurt Tucholsky* are in fact the same entity by referring to *pseudonym*, which is a challenging relation to recognize.

Finally, the model sometimes **selects more sentences than we truly need**. Interestingly, this is an error in terms of *evidence plausibility* but not in terms of *prediction*. The number of extracted sentences is very high in around 25% of the random error cases. The last row from Table 8 is one of representative examples with this kind of error. Although our model possibly has already successfully extracted right evidence in the first two steps, it continues selecting unnecessary sentences because the prediction confidence is not high enough, a drawback in our way of selecting evidence mentioned in Section 4.2. Moreover, our model extracts one more sentence on average when predicting incorrect relations, suggesting that in these cases it does not cleanly focus on the correct information.

Dataset	Setting	# doc.	# inst.	# word/inst.	# sent./inst.	# relation	# NA %
Adapted DocRED	train	3053	38180	203	8.1	96+1	33
	val	1000	12323	203	8.1	96+1	33
Brain MRI	train	327	327	177	11.6	–	–
	val	86	86	132	10.1	–	–

Table 7: Statistics of the two document-level IE datasets. Each document may have multiple entity pairs of interest, giving rise to multiple instances in the adapted DocRED setting. For adapted DocRED, we have 96 relations from the data plus an *NA* relation that we introduce for 1/3 of the data.

Instructions

For each finding, please highlight clues (if any) in the text for each key feature listed below using the provided web interface and provide one label listed below for each key feature. Some key features are explicitly mentioned in the findings, but some are not and are tricky to identify.

Key Features and labels:

- flair/t2: decreased, flair/t2: normal, flair/t2: increased
- t1: decreased, t1: normal, t1: increased
- diffusion/ADC: decreased, diffusion/ADC: normal, diffusion/ADC: increased
- susceptibility: increased, susceptibility: normal
- contrast_enhancement: yes, contrast_enhancement: no
- signal_pattern: homogeneous, signal_pattern: heterogeneous, signal_pattern: ring
- lesion: single, lesion: multiple
- side: lesion(s) symmetric, side: lesion(s) asymmetric
- mass_effect: positive, mass_effect: no, mass_effect: negative

Note:
side: lesion(s) symmetric, side: lesion(s) asymmetric: 'symmetric about the midline'.
If one lesion, most responses would be 'asymmetric', with lesion right or left. Single lesion 'symmetric' would be midline lesion.

General guidelines
- You should choose the minimal, most informative spans for each finding. For example, highlight "high T1 signal" instead of "there is high T1 signal".
- Prefer a single most informative span if possible; otherwise, you may choose two spans (as in an example below, under "Mass Effect") .
- Within each category (flair, mass effect, etc.), please carefully consider whether there is any evidence for one of those findings or not. Not annotating a category will be taken as a sign that there is no evidence for it.
- Often the span will be a noun phrase like "acute intracranial hemorrhage", or "no [noun phrase]".
- Ignore extra-axial lesions, small vessel ischemic disease, and all other minor diseases.
- In cases of "There is no X, Y, or Z", you should select just the span containing Z to indicate "no Z" -- the tool doesn't support overlapping annotations or those with a "gap".
- Do not annotate normal brain and non-brain chunks of the report (in addition to confounders such as SVID).
- If the same information shows up in multiple places, do try to annotate all of them.

Figure 4: Annotation instructions.

Type	Example	
Predicts correctly and extracts right evidence	*[0]* Delphine "Delphi" Greenlaw is a fictional character on the New Zealand soap opera Shortland Street, who was portrayed by Anna Hutchison between 2002 and 2004. ...	
	Predicted relation: country of origin **Extracted Evidence**: *[0]*	**Relation**: country of origin **Annotated Evidence**: *[0]*
Predicts debatably correct answer, extracts reasonable evidence	*[0]* Anton Erhard Martinelli (1684 – September 15 , 1747) was an Austrian architect and master - builder of Italian descent. *[1]* Martinelli was born in Vienna. ... *[3]* Anton Erhard Martinelli supervised the construction of several important buildings in Vienna, such as ... *[4]* He designed ... *[6]* He died in Vienna in 1747.	
	Predicted relation: place of birth **Extracted Evidence**: *[1]*	**Relation**: place of death **Annotated Evidence**: *[0, 6]*
Predict incorrect example on examples requiring high amount of reasoning	*[0]* Kurt Tucholsky (9 January 1890 – 21 December 1935) was a German - Jewish journalist, satirist, and writer. *[1]* He also wrote under the pseudonyms Kaspar Hauser (after the historical figure), Peter Panter, Theobald Tiger and Ignaz Wrobel. ...	
	Predicted relation: *NA* **Extracted Evidence**: *[0]*	**Relation**: date of death **Annotated Evidence**: *[0]*
Selecting more sentences than are needed	*[0]* Henri de Boulainvilliers ... was a French nobleman, writer and historian. ... *[2]* Primarily remembered as an early modern historian of the French State, Boulainvilliers also published an early French translation of Spinoza's Ethics and ... *[3]* The Comte de Boulainvilliers traced his lineage to ... *[5]* Much of Boulainvilliers' historical work ...	
	Predicted relation: country of citizenship **Extracted Evidence**: *[2, 0, 1, 5, 4, 3]*	**Relation**: country of citizenship **Annotated Evidence**: *[0, 2]*

Table 8: Four types of representative examples that show models' behavior. In our adapted DocRED task, models are asked to predict relations among **heads** and **tails**. Here we use model SUFFICIENT$_{\text{both}}$ for illustrations, which has the best evidence extraction performance. Sentences in extracted evidence are ranked by DL.

Model	An Example of *mass effect*, label: *positive*, evidence: 0 or 6
SUFFICIENT$_{\text{none}}$	**[0] These images show evidence of downward displacement of the brain stem with collapse of the interpeduncular cistern and caudal displacement of the mammary bodies typical for intracranial hypertension .** [1] There is diffuse pachymeningeal enhancement evident. [2] Bilateral extra axial collections are evident the do not conform to the imaging characteristics of CSF are seen overlying the hemispheres. [3] These likely reflect blood tinged hygromas and there does appear to be a blood products in the deep tendon portion of the right sided collection on the patient's left see image 14 series 2. [4] There does appear to be a discrete linear subdural hematoma along the right tentorial leaf. [5] Subdural collection is noted on both sides of the falx as well. **[6] There is mass effect at the level of the tentorial incisure due to transtentorial herniation with deformity of the midbrain .** [7] There is no evidence an acute infarct. [8] No parenchymal hemorrhage is evident. [9] Apart from the meningeal enhancement there is no abnormal enhancement noted.
SUFFICIENT$_{\text{both}}$	**[0] These images show evidence of downward displacement of the brain stem with collapse of the interpeduncular cistern and caudal displacement of the mammary bodies typical for intracranial hypertension**. [1] There is diffuse pachymeningeal enhancement evident. [2] Bilateral extra axial collections are evident the do not conform to the imaging characteristics of CSF are seen overlying the hemispheres. [3] These likely reflect blood tinged hygromas and there does appear to be a blood products in the deep tendon portion of the right sided collection on the patient's left see image 14 series 2. [4] There does appear to be a discrete linear subdural hematoma along the right tentorial leaf. [5] Subdural collection is noted on both sides of the falx as well. **[6] There is mass effect at the level of the tentorial incisure due to transtentorial herniation with deformity of the midbrain .** [7] There is no evidence an acute infarct. [8] No parenchymal hemorrhage is evident. [9] Apart from the meningeal enhancement there is no abnormal enhancement noted.

Table 9: An illustration of models' attribution scores over a report from BRAINMRI using DeepLift with and w/o regularization techniques. SUFFICIENT$_{\text{both}}$ appears to leverage more information from right sentences.