Extractive Fact Decomposition for Interpretable Natural Language Inference in one Forward Pass

Nicholas Popovic and Michael Färber
ScaDS.AI, TU Dresden, Germany
{nicholas.popovic,michael.faerber}@tu-dresden.de

Abstract

Recent works in Natural Language Inference (NLI) and related tasks employ atomic fact decomposition to enhance interpretability and robustness, yet existing methods rely on resource-intensive large language models (LLMs) to perform decomposition. We propose JEDI, an encoder-only architecture that jointly performs extractive atomic fact decomposition and interpretable inference without requiring generative models during inference. To facilitate training, we introduce SYRP, a large corpus of synthetic rationales covering multiple NLI benchmarks. Experimental results demonstrate that JEDI achieves competitive accuracy in-distribution and significantly improves robustness to shallow heuristic biases compared to models based purely on extractive rationale supervision. Our findings show that fine-grained interpretability and robust generalization in NLI can be efficiently realized using encoder-only architectures and synthetic rationales.[1]

1 Introduction

Natural language inference (NLI) (Giampiccolo et al., 2007; Bowman et al., 2015) tasks require models to determine whether a given hypothesis logically follows from a premise. While state-of-the-art NLI models have achieved high accuracy, their decision-making processes often remain opaque. This has motivated a growing body of research focused on building interpretable NLI systems that not only predict a label but also justify their predictions in a transparent and faithful way (Camburu et al., 2018; DeYoung et al., 2020; Stacey et al., 2022, 2024).

A common approach to interpretability in NLI involves extractive rationales, where the model highlights parts of the premise or hypothesis that support its decision. While easily interpretable, such rationales often fail to capture the underlying logical structure of inference and can obscure shallow pattern matching (McCoy et al., 2019). In response to these limitations, recent work has turned to atomic fact decomposition, which breaks the premise into minimal, semantically coherent sub-facts (atoms) against which the hypothesis is then validated individually (Stacey et al., 2024). By adding symbolic reasoning over these atomic units, this approach improves transparency, robustness and more closely reflects the compositional reasoning involved in NLI.

However, atomic fact decomposition currently relies on large language models (LLMs), which can hallucinate inaccuracies that negatively impact inference, particularly since atomic facts are not immediately traceable to explicit premise spans. Such hallucinations are especially problematic in longer documents, making verification challenging and hindering scalability. This raises the central question of our work: *Can atomic fact decomposition be distilled into encoder-only architectures to enable fast, scalable, and faithfully interpretable NLI without requiring LLMs at inference time?* To answer this, we propose reframing atomic fact decomposition as an extractive task, similar to existing extractive rationale-based approaches. Rather than generating textual statements, our model identifies spans in the premise corresponding directly to atomic facts. These spans are then classified with respect to the hypothesis using logical rules, enabling fact-level interpretability in a single encoder forward pass.

Our contributions are as follows:

- We propose JEDI[2], an encoder-only architecture that performs extractive atomic fact de-

[1] Code and data: https://github.com/2505AnonARR/JEDI

[2] JEDI: Joint Encoder for Decomposition and Inference

composition and logical inference jointly, enabling interpretable NLI without LLMs during inference.

- We construct a large corpus of synthetic rationales (SYRP[3]) to supervise span-level extraction in the absence of annotated data. In addition to the annotations required for this paper, we provide synthetic rationales across seven additional datasets to facilitate future research in interpretable NLI.

- We demonstrate that our approach produces competitive results both in-distribution and out-of-distribution while offering fine-grained, faithfully interpretable predictions grounded explicitly in the premise.

2 Related Work

2.1 Extractive Rationales in NLI

Natural Language Inference (NLI) (Giampiccolo et al., 2007; Bowman et al., 2015) has long served as a key task for evaluating model reasoning capabilities. Early work on interpretability emphasized extractive rationale methods, which highlight input spans presumed to justify model predictions (DeYoung et al., 2020). For instance, e-SNLI (Camburu et al., 2018) introduced human-annotated rationales aligned with entailment labels to support more transparent decision-making. However, subsequent studies, exemplified by the work of McCoy et al. (2019), revealed that plausible-looking rationales can mask superficial pattern matching, motivating a shift toward more structured and granular reasoning frameworks.

2.2 Atomic Fact Decomposition

Recent approaches decompose NLI examples into atomic units to support finer-grained inference. While Stacey et al. (2022) focused on span-level predictions grounded in noun phrase segmentation of the hypothesis, their subsequent work (Stacey et al., 2024) proposes training models using atomic facts derived from premise decomposition via LLMs. These methods aim to disentangle model reasoning from shallow heuristics by isolating semantically meaningful units and applying rule-based reasoning. Decomposition-based strategies have also been explored in related domains such as summarization (Yang et al., 2024), fact-checking (Min et al., 2023) and claim verification (Kamoi et al., 2023; Chen et al., 2024). The reliance on generative LLMs for decomposition introduces computational overhead, limited scalability, and potential inaccuracies (hallucinations) in inference, motivating our exploration of alternative extraction-based approaches.

2.3 Joint Architectures

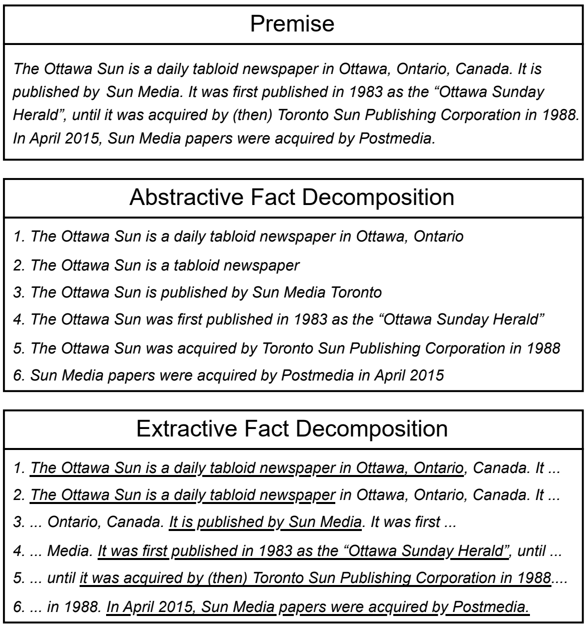
Recently, Lu et al. (2025) argue that pipeline architectures involving fact decomposition suffer from error propagation and advocate for joint training to reduce reliance on intermediate supervision. In this work, we draw a parallel to a different natural language processing task in which joint architectures are common. Specifically, we take inspiration from joint entity and relation extraction (Eberts and Ulges, 2021; Zhou et al., 2021; Hennen et al., 2024) where span extraction and complex classification are at the core of the task. We propose an extractive approach to fact decomposition, enabling a single encoder-based model to jointly learn both decomposition and inference. Joint models, however, require rich annotated data to effectively learn both decomposition and inference. Given the lack of existing annotations at this granularity, we create SYRP, a synthetic dataset spanning multiple NLI benchmarks, generated via LLMs, to facilitate training and evaluation.

3 Extractive Fact Decomposition

The core premise of this work is to view the process of atomic fact decomposition from an extractive point of view, rather than the abstractive approach seen in recent works (Stacey et al., 2024; Yang et al., 2024; Min et al., 2023; Chen et al., 2024), which make use of generative models. Figure 1 provides an illustration of the two annotation types. While atomic fact decomposition has demonstrated improved robustness compared to extractive rationale-based interpretability methods, this work explores the hypothesis that the observed robustness improvements may not inherently depend on abstraction, but rather the structured reasoning over clearly defined semantic units, and therefore also be achievable via extractive methods.

Further motivation for exploring an extractive framing stems from several practical and method-

[3] SYRP: SYnthetic Rationales for Premises



Premise

The Ottawa Sun is a daily tabloid newspaper in Ottawa, Ontario, Canada. It is published by Sun Media. It was first published in 1983 as the "Ottawa Sunday Herald", until it was acquired by (then) Toronto Sun Publishing Corporation in 1988. In April 2015, Sun Media papers were acquired by Postmedia.

Abstractive Fact Decomposition

1. *The Ottawa Sun is a daily tabloid newspaper in Ottawa, Ontario*
2. *The Ottawa Sun is a tabloid newspaper*
3. *The Ottawa Sun is published by Sun Media Toronto*
4. *The Ottawa Sun was first published in 1983 as the "Ottawa Sunday Herald"*
5. *The Ottawa Sun was acquired by Toronto Sun Publishing Corporation in 1988*
6. *Sun Media papers were acquired by Postmedia in April 2015*

Extractive Fact Decomposition

1. *The Ottawa Sun is a daily tabloid newspaper in Ottawa, Ontario, Canada. It ...*
2. *The Ottawa Sun is a daily tabloid newspaper in Ottawa, Ontario, Canada. It ...*
3. *... Ontario, Canada. It is published by Sun Media. It was first ...*
4. *... Media. It was first published in 1983 as the "Ottawa Sunday Herald", until ...*
5. *... until it was acquired by (then) Toronto Sun Publishing Corporation in 1988....*
6. *... in 1988. In April 2015, Sun Media papers were acquired by Postmedia.*

Figure 1: Illustration of abstractive and extractive fact decomposition for an example premise from the ANLI dataset. Atomic facts are based on the generated facts provided by Stacey et al. (2024). Extractive facts are shown as underlined text spans corresponding to the contents of the abstractive atoms.

ological considerations: Extractive rationales, as demonstrated in prior interpretability research (DeYoung et al., 2020; Camburu et al., 2018), provide explicit pointers to the relevant portions of the input text (the premise, for NLI), which is especially valuable when dealing with long or complex contexts. Second, extractive rationales lend themselves more naturally to encoder-only architectures. These models are typically significantly more lightweight than generative ones and continue to be used in many downstream natural language understanding (NLU) tasks for reasons of scalability. Finally, an extractive approach provides a more transparent and readily verifiable means to trace predictions directly to explicit spans in the input, reducing the risk introduced by potential hallucinations associated with generated facts.

4 SYRP: SYnthetic Rationales for Premises

Existing datasets lack suitable annotations for extractive fact decomposition, particularly for challenging benchmarks like ANLI (Nie et al., 2020), a challenging benchmark central to recent work in atomic fact decomposition by Stacey et al. (2024), a key baseline for our work. To address this gap, we create synthetic extractive rationales (Section 4). Second, existing annotations typically highlight only *salient spans*, whereas extractive fact decomposition requires segmenting the premise into *fact spans*, from which salience is later determined. Using the atomic facts Stacey et al. (2024) generated for ANLI and a token classification model trained on our synthetic rationales (Section 4.2), we create the required annotations (Section 4.3).

4.1 SYRP corpus

We create synthetic salient span annotations, for the training and development splits of the ANLI dataset (Nie et al., 2020) using an LLM[4] and carefully designed prompt templates detailed in Appendix A, Figures 3 and 4. The final configuration of model and prompt templates were chosen based on an evaluation of 15 different models and 20 different prompt templates, conducted on manually annotated data. More details can be found in Appendix A.

Broadly, we frame the annotation task by providing an instruction tuned model with the premise, hypothesis, and gold label. This way the model no longer has to perform the full task of verifying the hypothesis, but only needs to provide relevant spans. This means that even a model which does not achieve state-of-the-art performance on NLI can produce rationales for a given data sample. We evaluate annotation quality using intersection-over-union (IoU) with manually annotated spans, achieving an IoU of 69%, indicating substantial agreement (IoU > 50% is considered indicative of good agreement by DeYoung et al. (2020)). To ensure annotations reflected genuine task comprehension, we additionally evaluated accompanying natural language explanations. We found only 2 out of 30 explanations to be of low quality, further validating robustness.[5]

While our primary focus is ANLI, we have additionally generated a comprehensive corpus (SYRP)

[4] We selected Qwen2.5-32B-Instruct-GPTQ-Int4 based on performance and efficiency.

[5] We do not use the natural language explanations in the remainder of this work, but evaluated them in the initial model selection as a proxy for task comprehension and include the generations as part of the SYRP corpus for future research.

comprising roughly one million annotated samples across eight NLI benchmarks, publicly available to support future research in interpretable NLI. Statistics for this dataset, the SYRP corpus, can be found in Appendix B.

4.2 Token Classification Models (SYRP$_{\text{FT}}$)

Using the synthetic rationales produced for ANLI above, we finetune encoder-based token-classification models, SYRP$_{\text{FT}}$. For this, we pass premise and hypothesis to the encoder with a leading CLS token and a separator token (SEP) to mark the end of the premise. Each token in the premise is classified as either neutral (non-salient), entailed (supporting entailment), or contradicted (supporting contradiction). During inference, we follow logic from prior work (Stacey et al., 2024), predicting contradiction if any contradicted tokens exist, entailment if entailed tokens exist without contradictions, and neutral otherwise. As a result, any predictions produced by SYRP$_{\text{FT}}$ are traceable to salient tokens in the premise.

4.3 Span-Level Supervision from Generated Atomic Facts

With salient span annotations in place (i.e., rationales decisive to entailments and contradictions), the remaining supervision signal we require is for fact spans (spans representing atomic facts, including those that are not directly relevant to the hypothesis), as illustrated in Figure 1.

To obtain these, we convert the atomic facts generated by Stacey et al. (2024) for ANLI into fact spans. We use SYRP$_{\text{FT}}$, a token classification model[6] fine-tuned on the synthetic rationales created for ANLI, to identify the salient tokens in the premise that supports each generated fact and convert these to coherent spans. In cases where no salient spans were detected, we discarded the generated fact as it may indicate a hallucinated statement.

This results in a dataset that includes not only salient spans for interpretable NLI, but also fact spans corresponding to individual atomic facts in the premise.

5 JEDI: Joint Encoder for Decomposition and Inference

In this section, we describe our model architecture for joint fact decomposition and interpretable natural language inference, an overview of which is provided in Figure 2. We combine modeling approaches from information extraction for example (Eberts and Ulges, 2021; Zhou et al., 2021; Hennen et al., 2024), where efficient and effective span extraction and classification is at the heart of the task, with the logical rule-based framework for atomic fact-based natural language inference proposed by Stacey et al. (2024). This results in an encoder-only model architecture which performs premise decomposition and atom-level interpretable classification while requiring only a single forward pass and no LLM at inference time. We refer to the architecture by the acronym JEDI for Joint Encoder for Decomposition and Inference.

5.1 Encoder forward pass

Given a premise and a hypothesis, we pass both to the encoder with a leading CLS token and a separator token (SEP) to mark the end of the premise. For the subsequent computations we discard any embeddings produced for hypothesis tokens and use only the CLS token embedding, e_{CLS}, the token-wise premise embeddings $[e_{\text{p},1}, ..., e_{\text{p,n}}]$, and the SEP token embedding, e_{SEP}.

5.2 Initial global classification

Taking inspiration from relation extraction models (Zhou et al., 2021), we perform an initial global classification $P(x|e_{\text{CLS}}, e_{\text{SEP}})$ using a group bilinear layer[7] applied to the embeddings e_{CLS} and e_{SEP} as follows:

$$\left[\boldsymbol{z}^1_{\text{CLS}}; ...; \boldsymbol{z}^k_{\text{CLS}}\right] = e_{\text{CLS}},$$

$$\left[\boldsymbol{z}^1_{\text{SEP}}; ...; \boldsymbol{z}^k_{\text{SEP}}\right] = \tanh\left(e_{\text{SEP}}\right),$$

$$\text{P}\left(x|e_{\text{CLS}}, e_{\text{SEP}}\right) = \sigma\left(\sum_{m=1}^{k} \boldsymbol{z}^{m\intercal}_{\text{CLS}} \boldsymbol{W}^m_x \boldsymbol{z}^m_{\text{SEP}} + b_x\right)$$

where $\boldsymbol{W}^i_r \in \mathbb{R}^{d/k \times d/k}$ for $m = 1...k$ are model parameters and $P(x|e_{\text{CLS}}, e_{\text{SEP}})$ is the probability that the hypothesis is neutral, entailed, or contradicted.

[6] We used DeBERTa$_{\text{LARGE}}$ as the base encoder for this step.

[7] A variant of bilinear classifiers which reduces the number of parameters by splitting the embedding dimensions into k equal-sized groups and applies bilinear within the groups (Zhou et al., 2021).

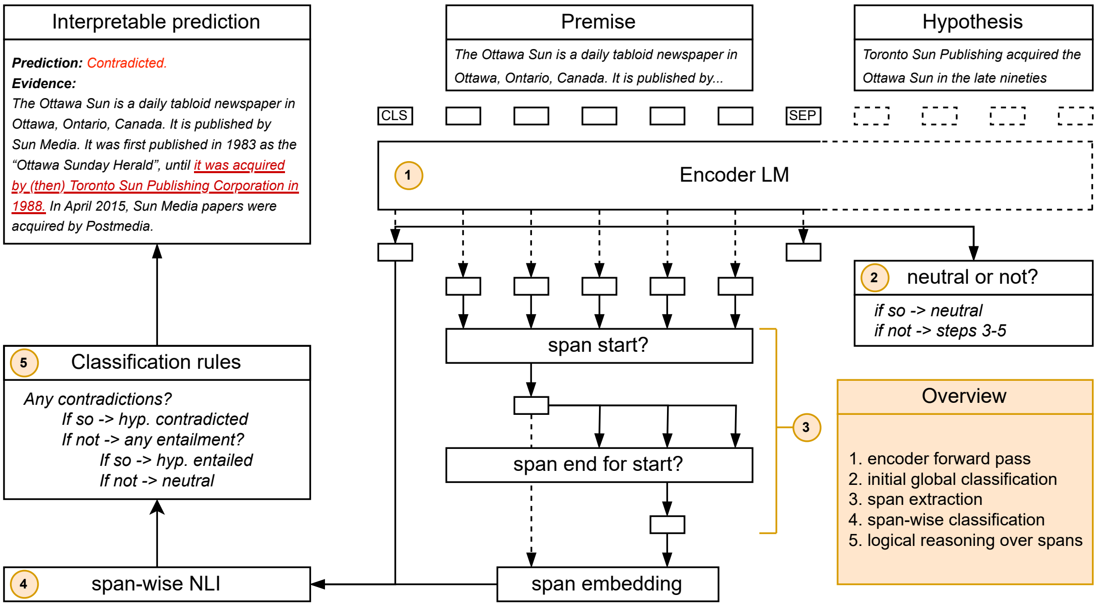

Interpretable prediction
Prediction: Contradicted.
Evidence:
The Ottawa Sun is a daily tabloid newspaper in Ottawa, Ontario, Canada. It is published by Sun Media. It was first published in 1983 as the "Ottawa Sunday Herald", until it was acquired by (then) Toronto Sun Publishing Corporation in 1988. In April 2015, Sun Media papers were acquired by Postmedia.
Premise
The Ottawa Sun is a daily tabloid newspaper in Ottawa, Ontario, Canada. It is published by...
Hypothesis
Toronto Sun Publishing acquired the Ottawa Sun in the late nineties
CLS
SEP
1
Encoder LM
2
neutral or not?
if so -> neutral
if not -> steps 3-5
span start?
3
span end for start?
Overview
1. encoder forward pass
2. initial global classification
3. span extraction
4. span-wise classification
5. logical reasoning over spans
5
Classification rules
Any contradictions?
If so -> hyp. contradicted
If not -> any entailment?
If so -> hyp. entailed
If not -> neutral
4
span-wise NLI
span embedding

Figure 2: Overview of the proposed architecture (JEDI) for performing fact-level span extraction and logical reasoning to perform interpretable natural language inference in a single forward pass.

During inference, if this classification is neutral, we end computation here and output neutral as the predicted class. If, however, the prediction is contradicted or entailed, we proceed with the span extraction. During training, steps 3-5 are followed even for neutral examples, as we have a supervision signal from the atomic facts. This is in contrast to purely rationale-based approaches, as no salient spans are available for neutral examples.

5.3 Span extraction

In short, for the span extraction, we take inspiration from Liu et al. (2022) and Hennen et al. (2024) by first identifying tokens likely starting an atomic fact, followed by pairing these potential start tokens with possible end tokens, forming candidates atomic fact spans.

For each token at index i in the premise we predict the probability that it represents the first token of a span:

$$P(i_{\text{is\_left}}|e_{\text{p,i}}) = \sigma\left(e_{\text{p,i}}\boldsymbol{W}_{\text{is\_left}}^{\intercal} + b_{\text{is\_left}}\right),$$

and then, for each token at index j in the premise, compute the probability that the span (i, j) represents a relevant span (or *"atomic fact"*):

$$\boldsymbol{e}_{i,j} = [e_{\text{p},i}; e_{\text{p},j}]\boldsymbol{W}_{\text{red},1}^{\intercal} + \boldsymbol{b}_{\text{red},1},$$
$$P(i_{\text{is\_span}}|\boldsymbol{e}_{i,j}) = \sigma\left(\boldsymbol{W}_{\text{span}} \cdot \boldsymbol{e}_{i,j} + b_{\text{span}}\right),$$

where $[e_{\text{p},i}; e_{\text{p},j}] \in \mathbb{R}^{2d}$ is the concatenation of the embeddings of the start and end tokens of the span, $\boldsymbol{W}_{\text{red},1} \in \mathbb{R}^{h\times 2d}$ and $\boldsymbol{b}_{\text{red},1} \in \mathbb{R}^{h}$ are parameters of a linear layer used to project the concatenated token embeddings into a fixed-size span representation $\boldsymbol{e}_{i,j} \in \mathbb{R}^{h}$, and $\boldsymbol{W}_{\text{span}} \in \mathbb{R}^{1\times h}$, $b_{\text{span}} \in \mathbb{R}$ are the parameters of a binary classifier that outputs the probability that the span (i, j) expresses a relevant atomic fact.

5.4 Span-wise classification

Next, given the extracted spans which represent fact atoms, we perform span-wise classification to determine whether the hypothesis is neutral, entailed, or contradicted by a given span (i, j). After applying a separate reduction using a linear layer:

$$\boldsymbol{e}_{i,j} = [e_{\text{p},i}; e_{\text{p},j}]\boldsymbol{W}_{\text{red},2}^{\intercal} + \boldsymbol{b}_{\text{red},2},$$

Using the same group bilinear classifier as for the initial global classification, we compute the following:

$$\left[\boldsymbol{z}^{1}_{\text{CLS}}; ...; \boldsymbol{z}^{k}_{\text{CLS}}\right] = e_{\text{CLS}},$$
$$\left[\boldsymbol{z}^{1}_{\text{i,j}}; ...; \boldsymbol{z}^{k}_{\text{i,j}}\right] = \tanh\left(e_{i,j}\right),$$
$$\text{P}\left(x|e_{\text{CLS}}, e_{\text{i,j}}\right) = \sigma\left(\sum_{m=1}^{k} \boldsymbol{z}^{m\intercal}_{\text{CLS}}\boldsymbol{W}^{m}_{x}\boldsymbol{z}^{m}_{\text{i,j}} + b_x\right)$$

where $\text{P}\left(x|e_{\text{CLS}}, e_{\text{i,j}}\right)$ represents the probabilities of neutrality, entailment, or contradiction for a span.

5.5 Logical reasoning over spans

Finally, we apply the logical rules for training and inference on fact atoms proposed by Stacey et al. (2024). This way, each span's classification directly informs the final prediction, ensuring the interpretability of the model by tracing every prediction explicitly back to concrete spans.

Training: if a given hypothesis is neutral, all extracted spans are to be labelled as neutral. If it is entailed, the salient spans are to be labelled as entailments, while all other spans are to be labelled neutral. If the hypothesis is contradicted, the salient spans are to be labelled as contradictions, while any other extracted spans (atomic facts), are masked from the loss calculation. This is due to the fact that in case of a contradiction, other atomic facts might still be in agreement with the hypothesis.

Inference: only if the initial global prediction is that the hypothesis is contradicted *and* a contradiction is found among the spans, will the final prediction be contradiction. Similarly, only if initial global prediction is that the hypothesis is entailed *and* any span is found to be in agreement, will the final prediction be entailment. In all other cases, neutral is returned as the prediction. This ensures, that any prediction is faithfully interpretable in the sense that it can be traced back to a specific span in the text.

5.6 Loss functions and Negative Sampling

The training loss consists of a total of four separate loss calculations: For the span extraction, two loss values are calculated for the predictions of $P(\text{is\_left})$ and $P(\text{is\_span})$ using binary cross entropy loss. For the initial global classification as well as the span-wise classification, we apply adaptive thresholding loss (Zhou et al., 2021), which is an effective means of managing class imbalances towards a single majority class (neutral applies to most atomic facts) during training used in relation extraction (to balance the most common "no relation" class).

In order to speed up training, we pre-sample 50 random negatives (instead of using those extracted by the model) for the span classification and span-wise NLI. For span-wise NLI we further count a random span as positive if it contains a salient span that makes up more than 80% of its size.

6 Experiments

6.1 Datasets and Baselines

Since it is our main baseline of interest, we replicate the experiment setup for FGLR (Stacey et al., 2024) as closely as possible for our evaluation. This involves using ANLI (Nie et al., 2020) for training and in-distribution evaluation, as well as out-of-distribution evaluations on ConTRoL (Liu et al., 2021), RTE (Wang et al., 2018), and WNLI (Wang et al., 2019; Levesque et al., 2011). Finally, we add the HANS dataset (McCoy et al., 2019) to our evaluation in order to examine the robustness of our models, as this has been a concern with extractive rationale based models in NLI.

We include the following baselines grouped by how fine-grained their interpretability mechanisms are: For uninterpretable methods we report the results of the encoder-LM finetuned on the ANLI dataset[8]. For sentence atom interpretability, we include SenLR[9] (Stacey et al., 2024) and a variant of JEDI, $\text{JEDI}_{\text{sent}}$, which uses sentence boundaries provided in the input instead of learning span extraction for span embeddings[10]. We group together span and fact atoms as the most relevant approaches to compare to our JEDI: We include SLR-NLI (Stacey et al., 2022), which provides span-level interpretability on the hypothesis using noun phrases as spans, and represents our primary LLM-free, span-level baseline. Naturally, as we aim to distill its behaviour into JEDI, we include the results reported by Stacey et al. (2024) for FGLR, which is the only baseline method requiring an LLM at inference time and for which our method represents the extractive counterpart. Finally, we evaluate SYRP_{FT}, the token classifiers described in 4.2, which allow for token-level interpretability. SYRP_{FT} being extractive but lacking atomic fact decomposition, directly tests our hypothesis that decomposition itself, rather than abstraction over extraction, drives improved robustness.

[8] For consistency, we report the scores reported by Stacey et al. (2024), which we were able to reproduce with minor differences attributable to differences in random seeds.

[9] Equivalent to using FGLR without an LLM at inference time and substituting premise sentences for the generated facts.

[10] Note that in contrast to SenLR, we provide a supervision signal for salient sentences based on our synthetic rationales.

Model	In-distribution R1	R2	R3	ANLI-all	Out-of-distribution ConTRoL	RTE	WNLI
not interpretable:							
DeBERTa$_{\text{LARGE}}$	78.3%	66.5%	61.7%	68.1%	56.0%	90.4%	68.9%
sentence atoms:							
SenLR	76.7%	64.8%	62.0%	67.5%	56.3%	86.3%	64.5%
JEDI$_{\text{sent}}$ (ours)	**77.4%**	**65.1%**	**62.3%**	**67.9%**	**56.7%**	**90.6%**	**67.8%**
span or fact atoms:							
SLR-NLI	74.7%	60.4%	58.3%	64.1%	54.7%	87.5%	65.8%
JEDI (ours)	**75.5%**	**63.1%**	**59.4%**	**65.6%**	**54.8%**	**87.7%**	**73.7%**
FGLR (+GPT-3.5-turbo)	76.2%	64.8%	63.1%	67.7%	52.7%	82.0%	77.0%
token atoms:							
SYRP$_{\text{FT}}$ (ours)	75.9%	63.3%	59.3%	65.8%	46.3%	88.8%	65.3%

Table 1: Test set scores for DeBERTa$_{\text{LARGE}}$. All results are obtained by averaging the accuracies across 10 random seeds.

Model	Acc.$_{\text{BASE}}$	Acc.$_{\text{LARGE}}$
SYRP$_{\text{FT}}$	31.6%	33.0%
JEDI	75.0%	76.9%
JEDI$_{\text{sent}}$	80.6%	83.1%

Table 2: Accuracies measured for models on the HANS dataset, designed to detect whether NLI models rely on shallow syntactic heuristics. The results show clearly that JEDI is more robust than SYRP$_{\text{FT}}$. For this evaluation we used the models with the highest accuracy on the development split of ANLI.

6.2 Implementation Details

We train models based on DeBERTa$_{\text{BASE}}$ and DeBERTa$_{\text{LARGE}}$ (He et al., 2021) implemented using Huggingface's Transformers (Wolf et al., 2020) and trained using mixed precision. We use AdamW (Loshchilov and Hutter, 2019) as optimizer (learning rates $\in [7e{-}6, 9e{-}6, 1e{-}5, 3e{-}5, 5e{-}5]$ for the encoders, and $1e{-}4$ for all other parameters). Final hyperparameters were chosen empirically based on validation performance, ensuring a fair comparison across models. We train using linear warmup (1 epoch) (Goyal et al., 2017) followed by a linear learning rate decay. We train each model for 25 epochs and perform early stopping based on development set accuracy.

6.3 Results and Discussion

The overall results for DeBERTa$_{\text{LARGE}}$ are shown in Table 1 while those for DeBERTa$_{\text{BASE}}$ are shown in Table 3. In Table 2, we further show the results for the HANS dataset. Below, we summarize our key findings based on the research questions stated earlier.

JEDI is capable of joint fact decomposition and inference. While JEDI does not match the accuracy of FGLR, which uses an LLM at inference time, it significantly outperforms the span-level baseline (SLR-NLI) without relying on generative models. This demonstrates that atomic decomposition can indeed be effectively distilled into encoder-only architectures, addressing scalability and interpretability without compromising significantly on performance.

In terms of accuracy, **sentence atoms are a strong alternative to fact atoms.** Consistent with prior work by Stacey et al. (2024), we find sentence-level interpretability to be highly competitive. This finding suggests for selecting interpretability granularity that if span- or fact-level interpretable reasoning is not essential, sentence-level supervision provides a highly competitive, more scalable alternative.

Synthetic rationales are suffiently high quality to act as supervision signals. All our presented approaches rely on the synthetic supervision signals created for SYRP. The overall competitiveness of results when compared to state-of-the-art methods indicate that this does not negatively impact performance. We conclude that the annotations are of sufficiently high quality to act as supervision signals.

JEDI improves robustness by reducing reliance on shallow heuristics. Though SYRP$_{\text{FT}}$, which

	In-distribution				Out-of-distribution		
Model	R1	R2	R3	ANLI-all	ConTRoL	RTE	WNLI
not interpretable:							
$\text{DeBERTa}_{\text{BASE}}$	71.2%	54.0%	51.7%	58.5%	53.7%	85.0%	59.6%
sentence atoms:							
SenLR	71.5%	55.0%	**52.3%**	59.1%	**53.4%**	83.7%	53.8%
$\text{JEDI}_{\text{sent}}$ (ours)	**72.0%**	**55.5%**	52.1%	**59.4%**	46.1%	**85.0%**	**62.4%**
span/fact atoms:							
SLR-NLI	65.5%	47.8%	47.1%	53.0%	**48.9%**	82.3%	56.3%
JEDI (ours)	**69.0%**	**53.2%**	**50.1%**	**57.0%**	46.3%	**83.3%**	**58.5%**
FGLR (+GPT-3.5-turbo)	71.8%	56.1%	55.3%	60.7%	49.1%	80.8%	70.7%
token atoms:							
SYRP_{FT} (ours)	69.2%	53.0%	51.7%	57.6%	49.5%	82.2%	55.7%

Table 3: Test set scores for $\text{DeBERTa}_{\text{BASE}}$. All results are obtained by averaging the accuracies across 10 random seeds.

uses salient token classification to perform NLI, performs on par or even slightly better on in-distribution data, JEDI generalizes better to out-of-distribution data. The evaluation on the HANS dataset (McCoy et al., 2019), which is specifically designed to assess whether NLI models rely on shallow syntactic heuristics is presented in Table 2. The strikingly low scores for SYRP_{FT} further emphasize that is much less robust, which is in line with previous researcher's finding on extractive rationale supervision. The fact that $\text{JEDI}_{\text{sent}}$ scores even higher on HANS suggests, that a part of the robustness originates from the span-wise inference architecture, and not just the fact decomposition. This robustness also extends to ConTRoL and WNLI, where JEDI consistently outperforms SYRP_{FT}, reinforcing the interpretation that its span-wise reasoning architecture mitigates shortcut learning behaviors.

7 Conclusion

We introduced JEDI, a joint encoder-only architecture capable of performing atomic fact decomposition and interpretable inference in natural language inference (NLI) tasks without relying on large generative models at inference time. Our experiments confirmed that JEDI effectively balances interpretability, robustness, and scalability, outperforming span-level baselines and substantially reducing reliance on shallow heuristics. Furthermore, we demonstrated the utility of synthetic rationales, releasing a large-scale corpus (SYRP) to support future interpretability research. Overall, we hope that our work contributes to the development of transparent and scalable NLI systems, highlighting that fine-grained interpretability and robust generalization can be achieved efficiently in encoder-only frameworks.

Limitations

JEDI's interpretability relies on extracting and classifying spans fromI would mention the premise alone, while hypotheses remain undecomposed. Extending JEDI to accommodate atomic decomposition of multi-sentence hypotheses would be necessary for applying the method more broadly. We note that, given appropriate supervision data, our architecture is easily expandable to this case, since it is modeled on relation extraction, where classification between spans (atoms) is an inherent part of the task.

Further, our approach relies heavily on synthetic rationales generated by large language models, which may have introduced subtle inaccuracies, despite our targeted model selection procedure designed to minimize such risks. These inaccuracies could potentially propagate errors into the model's interpretability, especially when used on datasets or domains distinct from those evaluated here.

Lastly, our empirical evaluations focus primarily on a limited set of English-language datasets. It remains unclear how JEDI would perform on languages or linguistic structures significantly different from those in our evaluation. Addressing multilingual capabilities and broader linguistic coverage constitutes an important direction for future research.

Acknowledgements

This research was co-funded by the ScaDS.AI, as well as by the German Federal Ministry of Research, Technology and Space (BMFTR) via the Software Campus project (01|S23070).

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3569–3587, Mexico City, Mexico. 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*, pages 4443–4458, Online. Association for Computational Linguistics.

Markus Eberts and Adrian Ulges. 2021. An end-to-end model for entity-level relation extraction using multi-instance learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3650–3660, Online. Association for Computational Linguistics.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour.

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

Moritz Hennen, Florian Babl, and Michaela Geierhos. 2024. ITER: Iterative transformer-based entity recognition and relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11209–11223, Miami, Florida, USA. Association for Computational Linguistics.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The Winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, page 47.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2021. Natural language inference in context - investigating contextual reasoning over long texts. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13388–13396. AAAI Press.

Tianyu Liu, Yuchen Eleanor Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan. 2022. Autoregressive structured prediction with language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 993–1005, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations 2019*, page 18.

Yining Lu, Noah Ziems, Hy Dang, and Meng Jiang. 2025. Optimizing decomposition for optimal claim verification. *Preprint*, arXiv:2503.15354.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural*

Language Processing, pages 12076–12100, Singapore. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Oana-Maria Camburu, and Marek Rei. 2024. Atomic inference for NLI with generated facts as atoms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10188–10204, Miami, Florida, USA. Association for Computational Linguistics.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, and Marek Rei. 2022. Logical reasoning with span-level predictions for interpretable and robust NLI models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3809–3823, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

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, and 3 others. 2020. 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. Association for Computational Linguistics.

Joonho Yang, Seunghyun Yoon, ByeongJeong Kim, and Hwanhee Lee. 2024. FIZZ: Factual inconsistency detection by zoom-in summary and zoom-out document. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Miami, Florida, USA. 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 Synthetic Rationale Generation Prompt Design and Model Selection

Our approach for creating both span-based, as well as natural language explanations follows the premise of treating instruction-tuned language models as annotators. We begin this section by describing a pre-study conducted in order to fix the hyperparameters used for annotation, followed by a description of the resulting annotation setup, and finally a description of the resulting dataset.

A.1 Development on ANLI

Due to the large amount of different possible language models and prompt styles, a structured hyperparameter exploration is a necessary prerequisite.

A.1.1 Hyperparameters

Choice of Language Models. In order to ensure reproducibility, we opt to use only those large language models for which weights are openly accesible. Furthermore, we limit the maximum model size to approx. 70 billion parameters due to hardware and compute time constraints. The above criteria, as well as general benchmark performance of various models result in the following selection of 15 models:

- **Llama3.1** 70B-, 8B-Instruct, Tulu-3-70B, Nemotron-70B-Instruct
- **Llama3.2** 3B-, 1B-Instruct
- **Qwen2.5** 72B-, 32B-, 14B-, 7B-, 3B-, 1.5B-, 0.5B-Instruct
- **Ministral** 8B-Instruct
- **Mistral** Nemo-Instruct

For larger models, we also perform experiments using quantized variants.

Task Framing. In general, we pose the task by providing premise, hypothesis, as well as the label (entailment or contradiction) and request an explanation following a specific format. Generating natural language explanations is relatively straightforward to elicit from instruction-tuned models by requesting it in a prompt, as it is a generative task. For span-based explanations, however, the task is not primarily generative, with token classification being a more natural framing. We design two different generative task settings to evaluate: (1) *highlighting*, in which the prompt requests for the model to generate the premise text while highlighting the most salient spans, similar to the task given to annotators of the e-SNLI dataset (Camburu et al., 2018). (2) *redaction*, in which the prompt requests for the model to generate the premise text while redacting any passages entailing or contradicting the hypothesis.

Output Format. In order to make the generated outputs easily parseable, the prompts include instructions on how to format the response. Here we examine Markdown and JSON as two possible formats.

Example Placement. All experiments are conducted in a one-shot in-context learning setting, meaning that a single example of a correctly executed query is provided as part of the context. Since all the used models allow for three roles, system, user, and assistant, this gives us two fundamentally different options for the placement of the example: (1) The example can either be included in the context as part of the system message, or (2) as a query by a user, followed by a response from the assistant containing the correct solution.

Development Dataset. In order to compare different choices of hyperparameters, we manually annotate a dataset consisting of 100 data points (50 of each for cases of entailment and contradiction) taken from the training set of ANLI.

Evaluation of Span-based Rationales. We measured the amount of exact matches (which reached a maximum of 30%), average intersection-over-union (which reached a maximum of roughly 70%) and the frequency with which a given model produces annotations that exceed different threshold values for intersection-over-union.

Evaluation of Natural Language Rationales. The prompts were structured to include natural language rationales, not intended as a supervision signal, but as an evaluation tool to check how well a given model interprets the data at hand. We manually assigned binary labels (acceptable/unacceptable) to 30 explanations per model. The top performing combination of model and prompt, which we ended up using, produced a score of $28/30$ acceptable explanations.

B SYRP Corpus

Table 4 contains statistics for the entire corpus, while Tables 5 and 6 contain statistics for the train and validation split, respectively.

Annotation prompt for entailed hypotheses (used with Qwen/Qwen2.5-32B-Instruct-GPTQ-Int4)

~~~
<|im_start|>system
You are a helpful assistant. You highlight information in text.<|im_end|>
<|im_start|>user
Highlight in the following text any passages supporting the statement that "The trolleybus system has over 2 urban
routes":

The Parma trolleybus system (Italian: "Rete filoviaria di Parma" ) forms part of the public transport network of
the city and "comune" of Parma, in the region of Emilia-Romagna, northern Italy. In operation since 1953, the
system presently comprises four urban routes.

Make only minimal changes, keep everything else unchanged.<|im_end|>
<|im_start|>assistant
```json
{
    "explanation": "The text states that the Parma trolleybus system has *four urban routes*. This supports the
    statement that the trolleybus system has over 2 urban routes.",
    "phrases_to_highlight": ["the system presently comprises four urban routes"],
    "highlighted_text": "The Parma trolleybus system (Italian: \"Rete filoviaria di Parma\" ) forms part of the
    public transport network of the city and \"comune\" of Parma, in the region of Emilia-Romagna, northern Italy.
    In operation since 1953, *the system presently comprises four urban routes*."
}
```<|im_end|>
<|im_start|>user
Wow! Great job! Let's try another one:

Highlight in the following text any passages supporting the statement that "{hypothesis}":

{premise}

Make only minimal changes, keep everything else unchanged.<|im_end|>
~~~

Figure 3: Prompt used for annotating entailed hypotheses SYRP corpus.

Source	Domain	Samples	Entail.	Neutral	Contra.	$\frac{\text{words}}{\text{premise}}$	$\frac{\text{spans}}{\text{premise}}$	Coverage
ANLI	Wiki	162,170	51,178	69,857	41,135	54.2	1.27	19.3%
ConTRoL	Exams	7,114	2,618	2,183	2,313	439.2	1.53	14.5%
ContractNLI	Legal	7,959	3,898	3,243	818	1642.7	1.42	7.6%
ESNLI	Captions	539,397	180,937	185,999	172,461	12.9	1.10	61.1%
FEVER	Wiki	223,759	127,497	42,305	53,957	60.3	1.51	26.7%
LINGNLI	Diverse	51,167	16,940	17,438	16,789	19.6	1.08	54.0%
MNLI	Diverse	404,827	134,907	137,152	132,768	19.9	1.12	57.1%
WANLI	Diverse	101,180	37,099	48,977	15,104	17.5	1.03	62.7%
Total		1,497,573	555,074	507,154	435,345	37.6	1.20	49.1%

Table 4: Overview of the full SYRP-corpus. Average span statistics have been calculated under omission of neutral samples, which do not have any annotated spans. The total number of samples with rationale spans is $991,628$.

Source	Samples	Entail.	Neutral	Contra.	$\frac{\text{words}}{\text{premise}}$	$\frac{\text{spans}}{\text{premise}}$	Coverage
ANLI	159,091	50,177	68,789	40,125	54.1	1.27	19.4%
ConTRoL	6,344	2,344	1,946	2,054	453.9	1.53	14.4%
ContractNLI	6,975	3,418	2,820	737	1632.6	1.41	7.4%
ESNLI	529,905	177,680	182,764	169,461	12.9	1.10	61.1%
FEVER	204,796	121,289	35,639	47,868	60.7	1.51	26.8%
LINGNLI	43,918	14,446	14,995	14,477	19.5	1.08	53.9%
MNLI	385,499	128,076	130,900	126,523	19.9	1.12	56.9%
WANLI	101,180	37,099	48,977	15,104	17.5	1.03	62.7%
Total	1,437,708	534,529	486,830	416,349	36.5	1.19	49.3%

Table 5: Overview of the **training split** of the SYRP-corpus. Average span statistics have been calculated under omission of neutral samples, which do not have any annotated spans. The total number of samples with rationale spans is $951,721$.

Annotation prompt for contradicted hypotheses (used with Qwen/Qwen2.5-32B-Instruct-GPTQ-Int4)

~~~
<|im_start|>system
You are a helpful assistant. You highlight information in text.<|im_end|>
<|im_start|>user
Highlight in the following text any passages supporting the statement that "Jesse James was a guerrilla in the
Union army during the American Civil War.":

The Centralia Massacre was an incident during the American Civil War in which twenty-four unarmed Union soldiers
were captured and executed at Centralia, Missouri on September 27, 1864 by the pro-Confederate guerrilla leader
William T. Anderson. Future outlaw Jesse James was among the guerrillas.<|im_end|>
<|im_start|>assistant
```json
{
    "explanation": "The text states that Jesse James was among the *pro-Confederate* guerrillas. This
    contradicts the statement that Jesse James was a guerrilla in the Union army.",
    "phrases_to_highlight": ["pro-Confederate guerilla", "Jesse James was among the guerrillas"],
    "highlighted_text": "The Centralia Massacre was an incident during the American Civil War in which twenty-four
    unarmed Union soldiers were captured and executed at Centralia, Missouri on September 27, 1864 by the
    *pro-Confederate guerrilla* leader William T. Anderson. Future outlaw *Jesse James was among the guerrillas*."
}
```<|im_end|>
<|im_start|>user
Wow! Great job! Let's try another one:

Highlight in the following text any passages contradicting that "{hypothesis}":

{premise}

Make only minimal changes, keep everything else unchanged.<|im_end|>
~~~

Figure 4: Prompt used for annotating contradicted hypotheses SYRP corpus.

Source	Samples	Entail.	Neutral	Contra.	$\frac{\text{words}}{\text{premise}}$	$\frac{\text{spans}}{\text{premise}}$	Coverage
ANLI	3,079	1,001	1,068	1,010	54.5	1.25	17.3%
ConTRoL	770	274	237	259	317.9	1.55	14.9%
ContractNLI	984	480	423	81	1714.6	1.47	9.5%
ESNLI	9,492	3,257	3,235	3,000	13.9	1.12	59.6%
FEVER	18,963	6,208	6,666	6,089	55.7	1.39	24.7%
LINGNLI	7,249	2,494	2,443	2,312	19.8	1.08	54.6%
MNLI	19,328	6,831	6,252	6,245	19.5	1.10	59.7%
Total	59,865	20,545	20,324	18,996	63.6	1.21	44.7%

Table 6: Overview of the **validation split** of the SYRP-corpus. Average span statistics have been calculated under omission of neutral samples, which do not have any annotated spans. The total number of samples with rationale spans is $39,907$.