A General Response

We thank all reviewers for their constructive and encouraging feedback and will polish our paper as suggested.

Q0: *Clarification of some details about MRC training.*

A0: As stated in the paper, the training of MRC does not include extra supervision other than the *self-supervised* signals from SUP sentences. During training, claim phrases $\mathcal{W}_c$ are used as ground truth, i.e., to recover a SUP claim. During inference, MRC produces an answer $w_i' \in \mathcal{W}_E$ for a claim phrase $w_i \in \mathcal{W}_c$, which is used to replace w_i for constructing a local premise.

B Response to Reviewer 1

Thank you for your valuable comments. We will include more motivation and details of the adopted techniques.

Q1: *Ability to detect multiple culprits?*

A1: LOREN is by design capable of finding multiple culprits. For example, in "*In a Lonely Place had nothing to do with any novel by Dorthy B. Hughes.*" where extracted claim phrases are underlined. Among those, the last three could be the culprits, LOREN predicts *nothing* and *Dorthy B. Hughes* to be the culprits. As a side note, a complex sentence can also be decomposed into simple statements, which can be well verified individually by LOREN.

Q2: *What about claims that can not identify sub-phrases?*

A2: Should phrase extraction fail, LOREN can still give an overall prediction for the claim, but the interpretability (i.e., phrase veracity) would certainly be compromised. We acknowledge there is room for improving phrase extraction.

C Response to Reviewer 2

Thank you for your valuable suggestions. We would like to clarify that *all code, data and models will be released.*

Q1: *How to compute $p(\boldsymbol{z}_i|c, w_i, E)$?*

A1: As shown in the **Latent Model** paragraph in Section 3.2, $p(\boldsymbol{z}|x) = \prod_i p(\boldsymbol{z}_i|x, w_i)$ is the prior distribution of $\boldsymbol{z}$ where $x = (c, E)$, which is calculated accordingly.

Q2: *How is culprit exactly defined and how to compute it?*

A2: The culprit is defined as the cause to the falsity in a claim, and in our work takes the form of phrase. LOREN finds the culprit phrase(s) by directly predicting phrase veracity, i.e., $\boldsymbol{z} = \{\boldsymbol{z_1}, ..., \boldsymbol{z}_{|\mathcal{W}_c|}\}$.

Q3: *Template for constructing probing questions?*

A3: For cloze questions, we just replace a claim phrase with `[MASK]`. For interrogative ones, a question generator takes as input (claim, claim phrase), and the generator generates an interrogative question asking about the phrase.

Q4: *Does MRC training use $\mathcal{W}_c$ as output, not $\mathcal{W}_E$?*

A4: Yes, it is true. Please refer to **A0**.

D Response to Reviewer 3

Thank you for your thorough advice. We will revise the paper and enrich the discussion of related work as suggested.

Q1: *Baselines from ACL'21*

A1: Thank you for pointing this out. The results of two methods from ACL'21, i.e., LisT5 (Jiang, Pradeep, and Lin 2021) and TARSA (Si et al. 2021) are reported below. Note that it is not a rather fair comparison due to too many different settings such as evidence retrieval and PLMs. For example, the size of LisT5 (T5, 3B parameters) is 10x larger than LOREN (RoBERTa, 355M parameters). Besides, LOREN enjoys the additional merits of interpretability.

Model	Dev		Test	
	LA	FEV	LA	FEV
TARSA	81.24	77.96	73.97	70.70
LisT5	**81.26**	77.75	**79.35**	**75.87**
LOREN	81.14	**78.83**	76.42	72.93

Q2: *How many questions are generated for a claim phrase?*

A2: Two. An interrogative one and a cloze one.

E Response to Reviewer 4

Thank you for your appreciation for our work and kind suggestions. We will discuss work on model uncertainty and MRC quality estimation in the revision.

F Response to Reviewer 5

Thank you for your detailed comments. We will polish the paper and clarify some necessary details in the revision. We will also explore the verification of real claims in the future.

Q1: *How many culprit phrases were found on average per claim that you have labelled?*

A1: There are on average 1.26 culprit phrases per claim.

Q2: *Were similar approaches proposed in other test classification problems?*

A2: We have not seen studies that share the exact same idea in LOREN. Although some broad ideas like distant supervision (e.g., multi-instance multi-label learning) and soft logic (as discussed in Section 2) have been investigated in text classification. We will add appropriate ones in related work.

Q3: *Why use specific heuristic rules for phrase extraction?*

A3: We want to ensure the extracted phrases representing the key information within a sentence to be verified. A series of heuristic rules are proposed to extract and keep the atomicity of the phrases. Nevertheless, we acknowledge that there is much room for refinement w.r.t. the ways of sentence decomposition where culpability resides.

Q4: *Why select DeBERTa as the NLI model?*

A4: DeBERTa was chosen for its excellent performance on a variety of NLP tasks. Since we directly use an off-the-shelf NLI model for prior distribution, we can also use other NLI models. The results would not change much according to conclusions in Sec 5.2.

Q5: *Clarification of some details.*

1. **Details about MRC module**: Please refer to **A0**.
2. **Cross-referencing & Meaning of *Implementations***: *Implementations* means the parameterization of $q_\phi(\cdot)$ and $p_\theta(\cdot)$ as well as data preparation details. We will reorganize Section 3.4 and 4.1 as suggested.
3. **Missing introduction of $p_\theta(\boldsymbol{z}|\boldsymbol{y}, x)$**: it is required in the EM algorithm, where we omit for brevity.
4. **Meaning of D_{KL}**: D_{KL} is Kullback–Leibler divergence.