LOREN: Logic-Regularized Reasoning for Interpretable Fact Verification

Jiangjie Chen[1,2], Qiaoben Bao[1], Changzhi Sun[2], Xinbo Zhang[2], Jiaze Chen[2], Hao Zhou[2], Yanghua Xiao[1], Lei Li[3]

FUDAN UNIVERSITY 復旦大學 ByteDance UC SANTA BARBARA

Did Donald Trump win the 2020 U.S. presidential election?

Fact Verification

- **Input**: a claim + a KB
- **Task**:
  - ***Evidence Extraction***
    - Evidence from a ***trustworthy KB***
  - ***Veracity Prediction***
    - Supported (SUP)
    - Refuted (REF)
    - Not Enough Information (NEI)

Claim

The Rodney King riots took place in the most populous county in the USA.

[wiki/Los Angeles Riots]

The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

[wiki/Los Angeles_County]

Los Angeles County, officially County the of Los Angeles, is the most populous county in the USA.

Verdict: Supported

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Arpit Mittal. **FEVER: a large-scale dataset for Fact Extraction and VERification**. NAACL, 2018.

A General Pipeline for Solving This Task

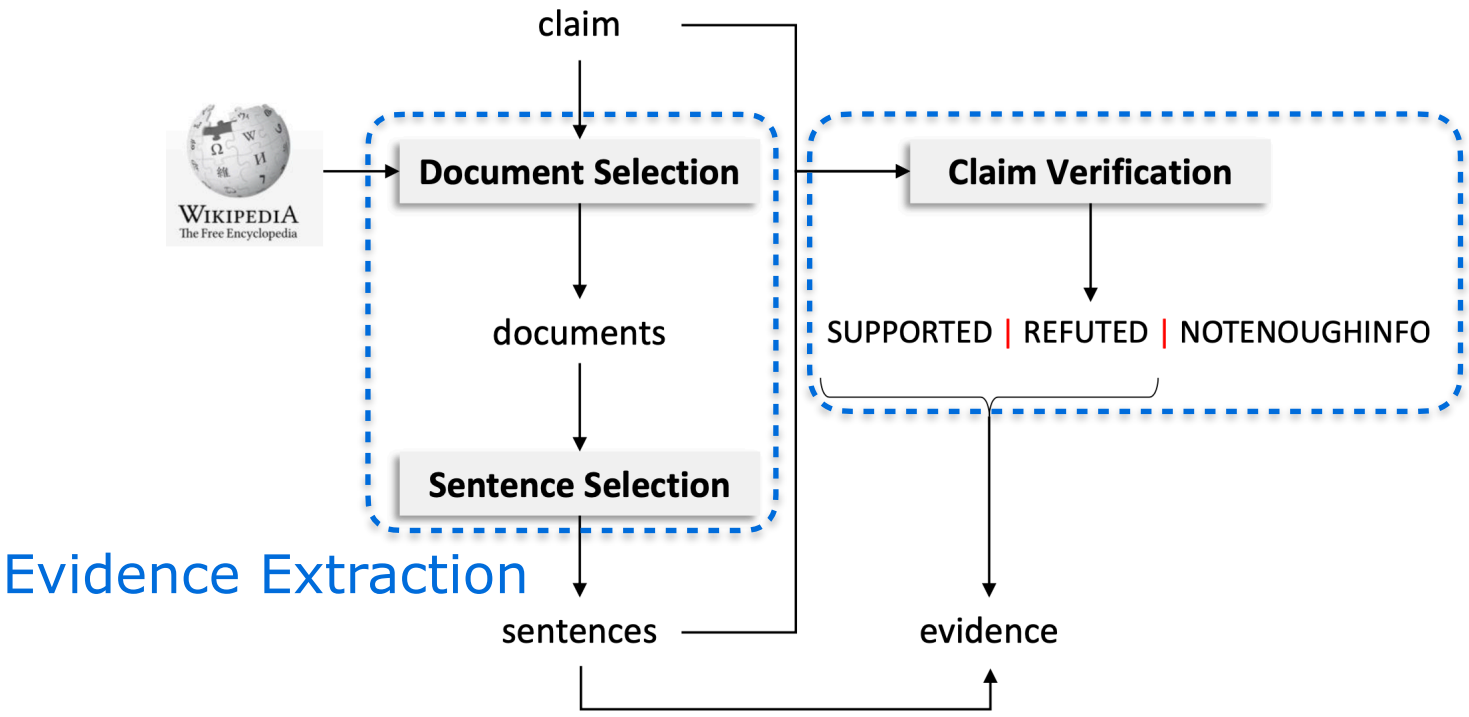


Image credit to: Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, Jian Yin. **Reasoning Over Semantic-Level Graph for Fact Checking.** ACL 2020.

A General Pipeline for Solving This Task

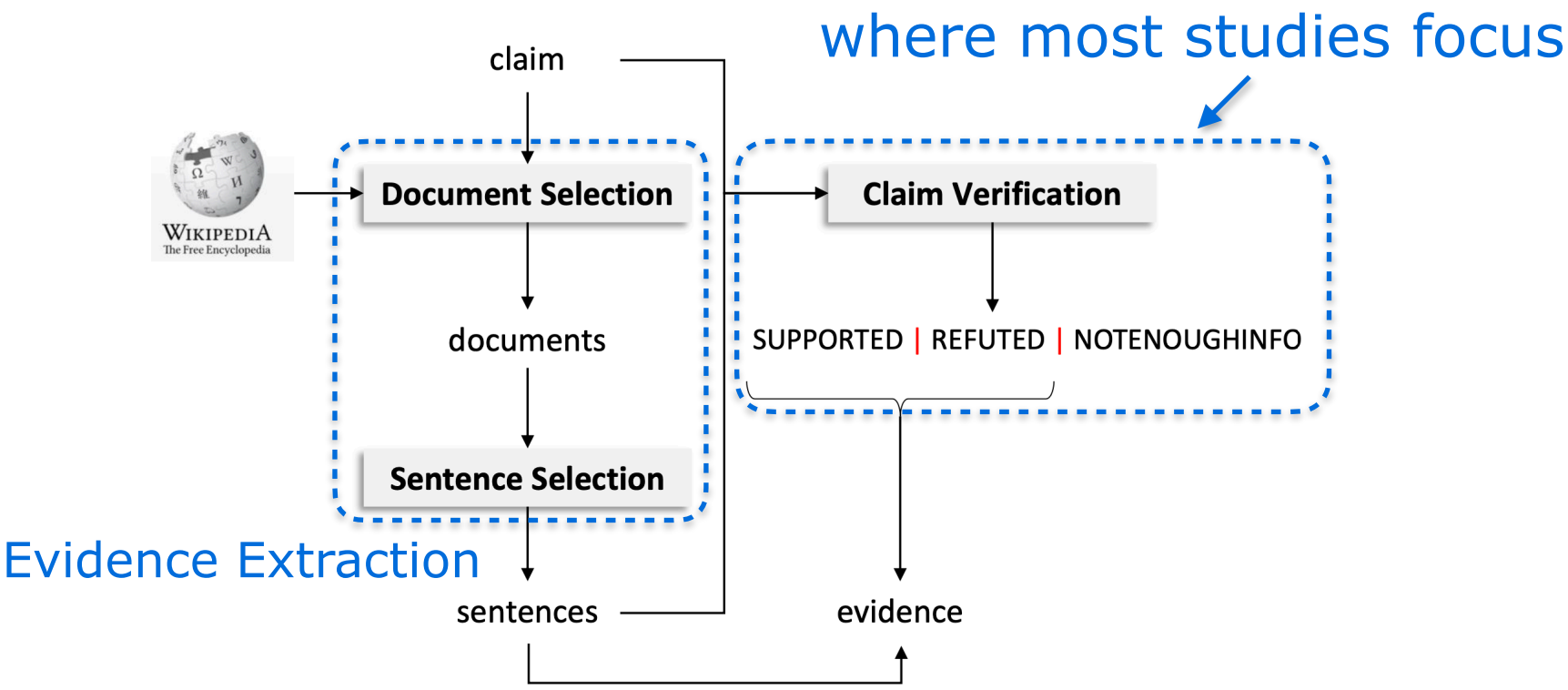


Image credit to: Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, Jian Yin. **Reasoning Over Semantic-Level Graph for Fact Checking.** ACL 2020.

Shielding from Misinformation

- Misinformation detection on social media
  - Especially with the success of PLMs.

- Factually accurate language generation
  - NLG with data accuracy ✓
  - NLG with factual accuracy ?
    - An objective evaluation on factual accuracy of machine generated text.

Did Donald Trump win the U.S. presidential election?

REFUTES

Did Donald Trump win th presidential election?

REFUTES

But… Why?

Interpretable Fact Verification

- **Goal of Reasoning**
  - Right answer for the right thinking

Interpretable Fact Verification

- **Goal of Reasoning**
  - Right answer for the right thinking

- **Interpretability** "may be" the right thinking
  - **Faithful**: able to explain the prediction
  - **Accurate**: should be right per se
  - **Debuggable**: able to find out where goes wrong

Interpretable Fact Verification

- **Goal of Reasoning**
  - Right answer for the right thinking
- **Interpretability** "may be" the right thinking
  - **Faithful**: able to explain the prediction
  - **Accurate**: should be right per se
  - **Debuggable**: able to find out where goes wrong
- **The Research Question**:
  - How can we do it without supervision?

Learning from Humans

Claim: c **Donald Trump** **won** **the 2020 election**.

- We carefully examine each phrase in a claim one by one.
  - Did Donald Trump win the election in **[2020]**?
  - Did Donald Trump win the **[U.S.]** presidential election in 2020?

Learning from Humans

Claim: c Donald Trump won the 2020 election.

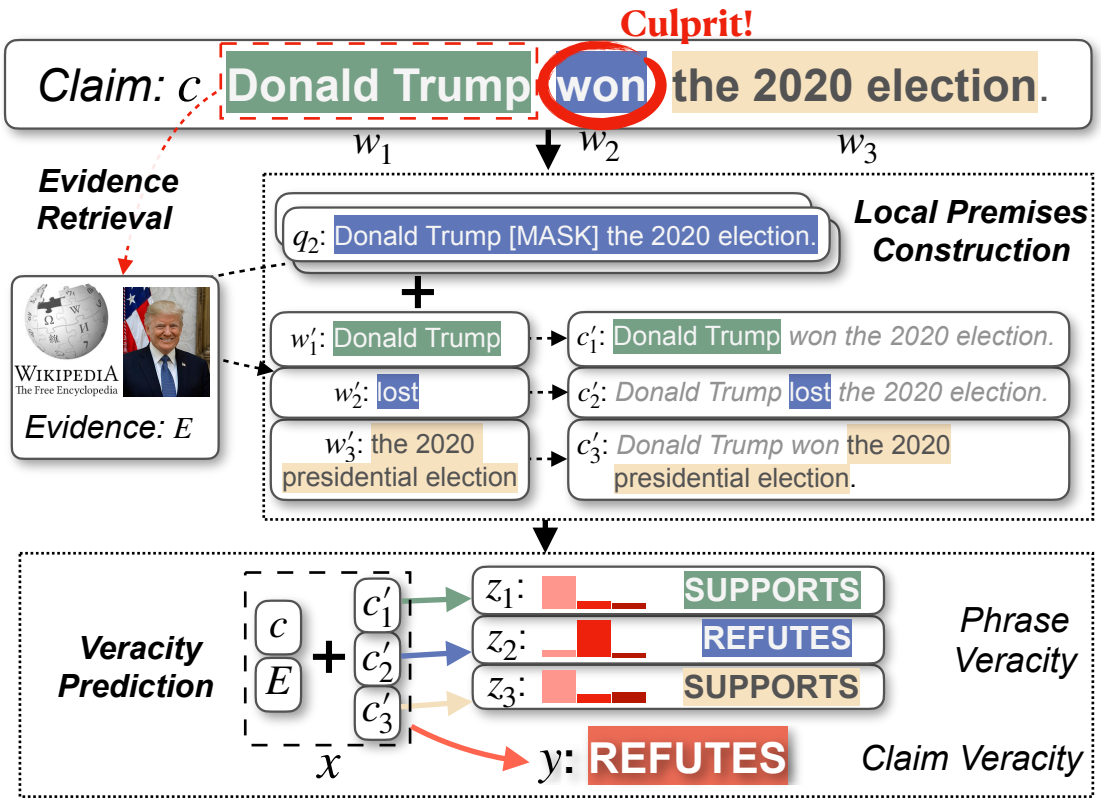
- We carefully examine each phrase in a claim one by one.
  - Did Donald Trump win the election in **[2020]**?
  - Did Donald Trump win the **[U.S.]** presidential election in 2020?

- We aggregate the verification results of each phrase **following aggregation logic**, i.e. a claim is found
  - ***Supported*** iff all phrases found supported;
  - ***Refuted*** iff exists a phrase found refuted;
  - ***NEI*** iff not refuted and exists a phrase found unverifiable.

LOREN: Overview

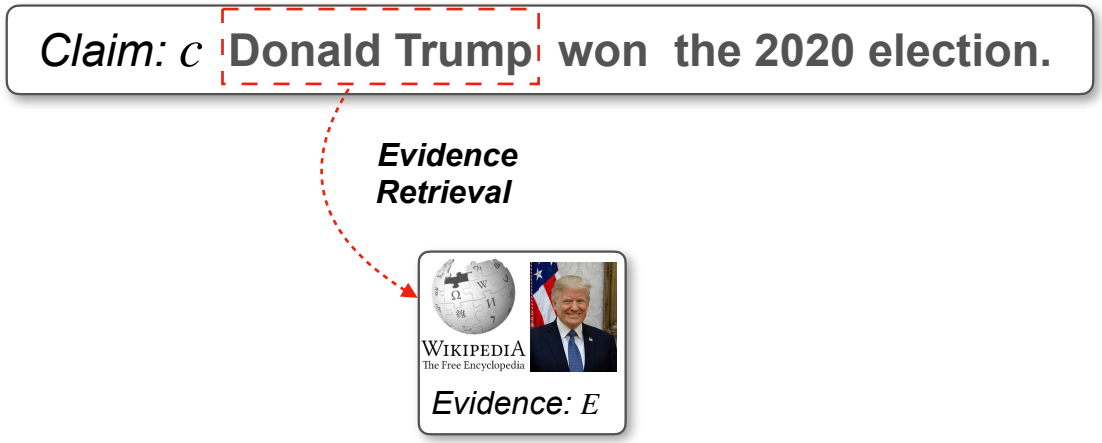
Symbolic AI plans, connectionist AI executes.

LOREN: Overview



- **TL;DR**: **build local premises** from evidence to support phrase veracity prediction, **regularized by logical rules**.

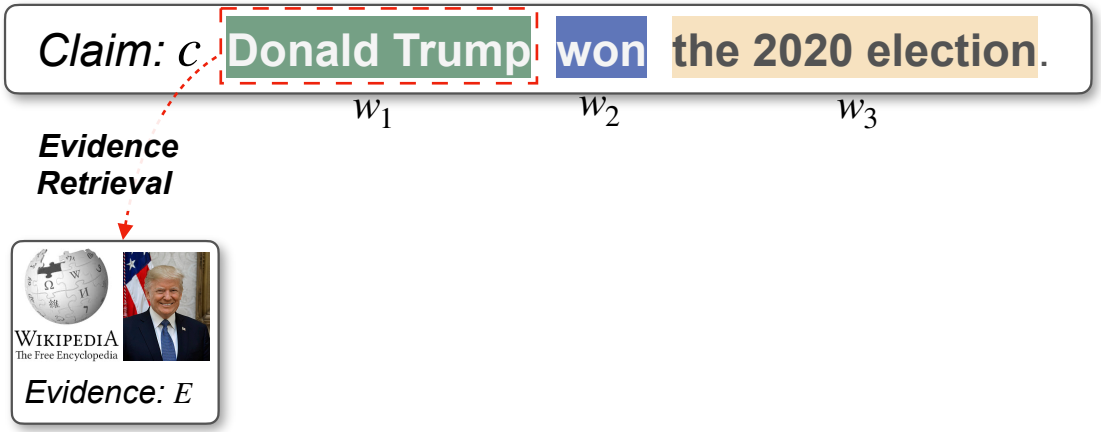
Evidence Retrieval



- Extract evidence sentences from Wikipedia *following Liu et al. ACL 2020*
  - Document retrieval
  - Sentence ranking
- Five relevant sentences of the entities in a claim.

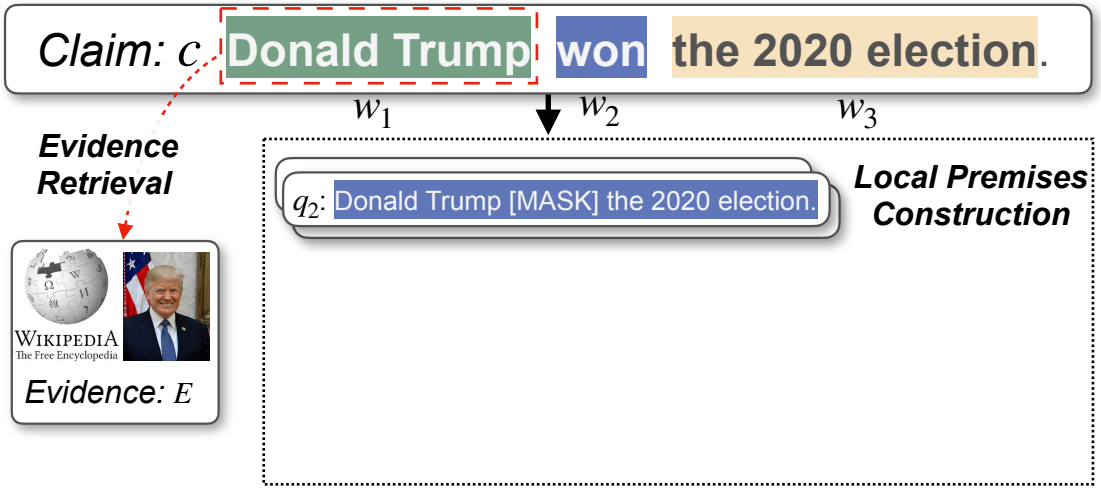
Zhenghao Liu, Chenyan Xiong, Maosong Sun. Zhiyuan Liu. **Fine-grained Fact Verification with Kernel Graph Attention Network**. ACL 2020.

Claim Phrase Extraction



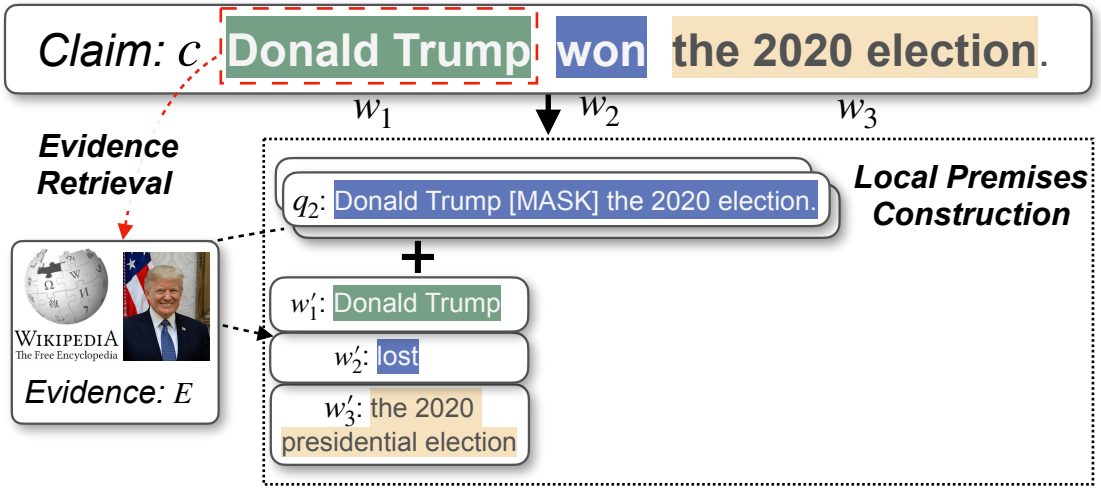
- Extract *claim phrases* for fine-grained decomposition
  - e.g. noun phrase, adjective phrase, named entity, etc.
- **Approach**: Parse with heuristic rules via off-the-shelf NLP tools
  - e.g. constituency parsing, pos tagging, NER, etc.

Probing Question Generation



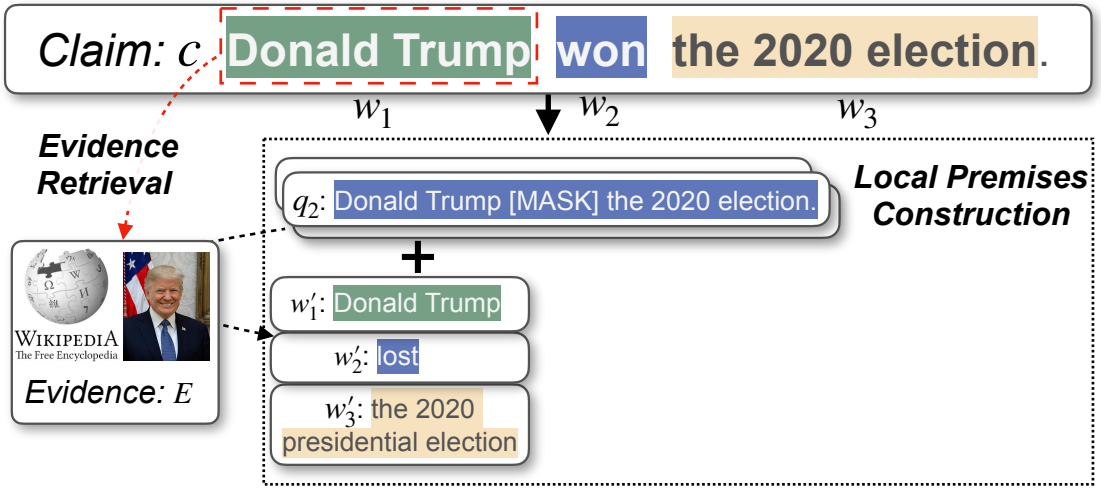
- **Goal**: generate *probing questions* to answer from evidence.
  - Cloze question & interrogative questions
  - Prepare for the QA task

Answer Probing Questions



- **Goal**: acquire corresponding *local premises* from evidence for each claim phrase.
  - Fine-tune a Seq2Seq MRC model (BART) on a manufactured dataset based on *support* samples.

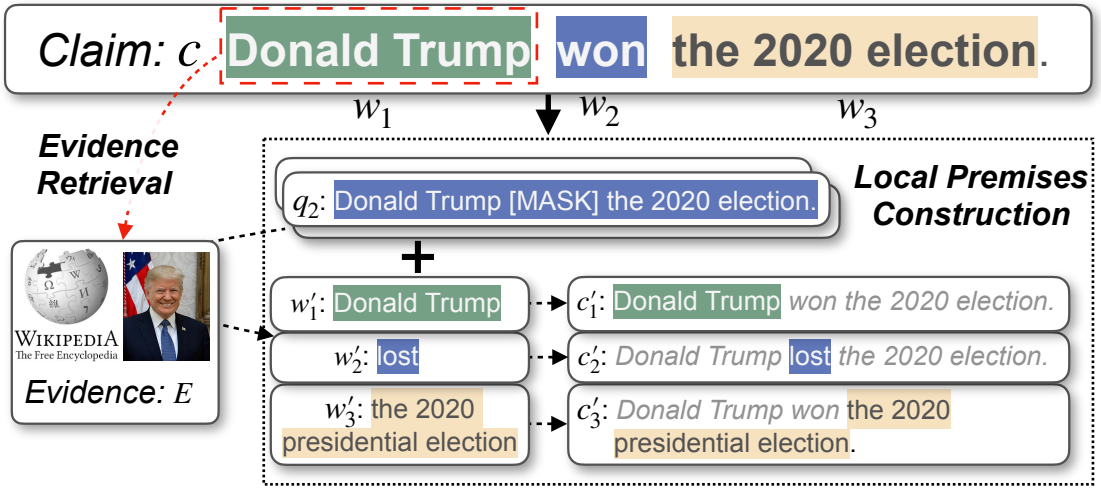
Answer Probing Questions



Self-supervised training with SUP samples

- **Goal**: acquire corresponding *local premises* from evidence for each claim phrase.
  - Fine-tune a Seq2Seq MRC model (BART) on a manufactured dataset based on *support* samples.

Assemble Local Premises

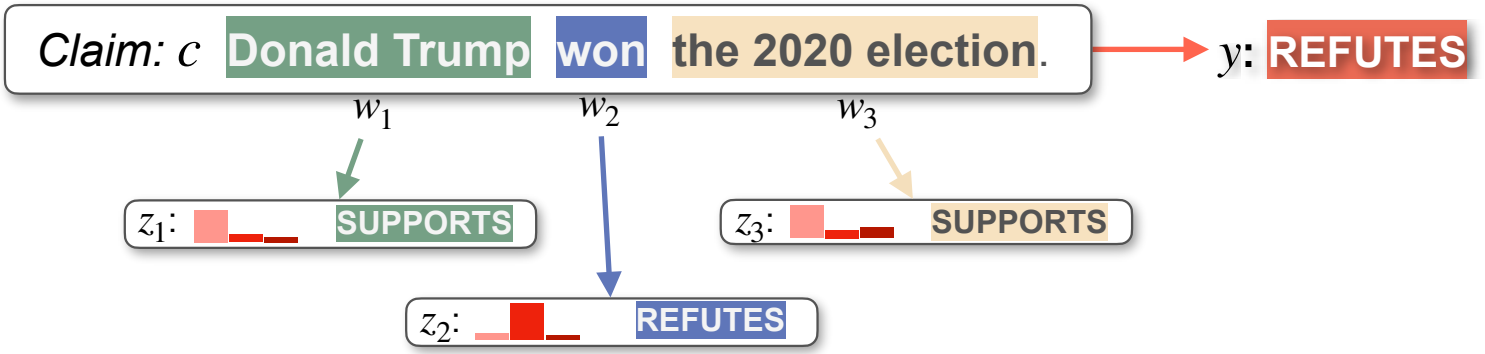


- **Goal**: acquire corresponding *local premises* from evidence for each claim phrase.
  - Fine-tune a Seq2Seq MRC model (BART) on a manufactured dataset based on *support* samples.
  - Fill masked-claims with answered phrases to construct local premises.

The Latent Model for Interpretable Fact Verification

Decompose claim verification $p_\theta(\mathbf{y} \mid x)$ into phrase verification $p_\theta(\mathbf{y} \mid \mathbf{z}, x)$

$$p_\theta(\mathbf{y} \mid x) = \sum_{\mathbf{z}} p_\theta(\mathbf{y} \mid \mathbf{z}, x) p(\mathbf{z} \mid x)$$

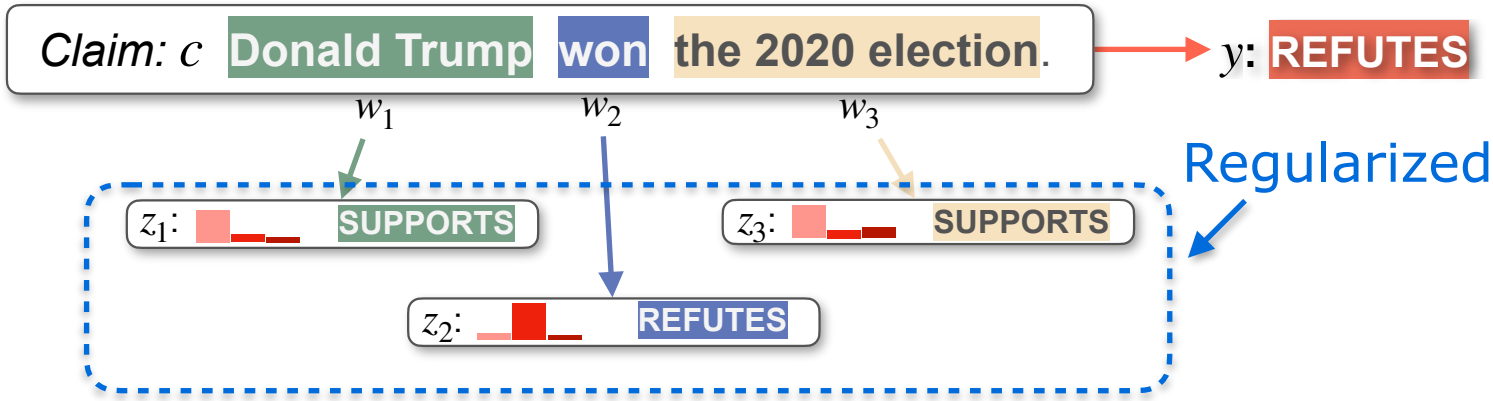


The Latent Model for Interpretable Fact Verification

Decompose claim verification $p_\theta(\mathbf{y} \mid x)$ into phrase verification $p_\theta(\mathbf{y} \mid \mathbf{z}, x)$

$$p_\theta(\mathbf{y} \mid x) = \sum_{\mathbf{z}} p_\theta(\mathbf{y} \mid \mathbf{z}, x) p(\mathbf{z} \mid x)$$

Phrase veracity as latent variables



Regularized by logic

The Latent Model for Interpretable Fact Verification

Decompose claim verification $p_\theta(\mathbf{y} \mid x)$ into phrase verification $p_\theta(\mathbf{y} \mid \mathbf{z}, x)$

$$p_\theta(\mathbf{y} \mid x) = \sum_{\mathbf{z}} p_\theta(\mathbf{y} \mid \mathbf{z}, x) p(\mathbf{z} \mid x)$$

- Variational inference for solving the latent model
  - Evidence Lower BOund (ELBO)

$$\text{ELBO} = \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{y},x)} [\log p_\theta(y^* \mid \mathbf{z}, x))] - D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{y}, x) \parallel p(\mathbf{z} \mid x))$$

The Latent Model for Interpretable Fact Verification

Decompose claim verification $p_\theta(\mathbf{y} \mid x)$ into phrase verification $p_\theta(\mathbf{y} \mid \mathbf{z}, x)$

$$p_\theta(\mathbf{y} \mid x) = \sum_{\mathbf{z}} p_\theta(\mathbf{y} \mid \mathbf{z}, x) p(\mathbf{z} \mid x)$$

- Variational inference for solving the latent model
  - Evidence Lower BOund (ELBO)

$$\text{ELBO} = \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{y},x)} [\log p_\theta(y^* \mid \mathbf{z}, x))] - D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{y}, x) \,\|\, p(\mathbf{z} \mid x))$$

Variational posterior distribution

Prior distribution

Regularize Latent Variables with Logic

$$\mathscr{L}_{\text{var}}(\theta, \phi) = -\,\text{ELBO} = -\mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{y},x)}[\log p_\theta(y^* \,|\, \mathbf{z}, x))] + D_{\text{KL}}(q_\phi(\mathbf{z} \,|\, \mathbf{y}, x) \parallel p(\mathbf{z} \,|\, x))$$

$$\mathscr{L}_{\text{final}}(\theta, \phi) = (1 - \lambda)\mathscr{L}_{\text{var}}(\theta, \phi) + \lambda\mathscr{L}_{\text{logic}}(\theta, \phi)$$

$$\mathscr{L}_{\text{logic}}(\theta, \phi) = D_{\text{KL}}\left(p_\theta(\mathbf{y} \,|\, \mathbf{z}, x) \parallel q_\phi^{\text{T}}(\mathbf{y}_z \,|\, \mathbf{y}, x)\right)$$

Regularize Latent Variables with Logic

$$\mathscr{L}_{\text{var}}(\theta, \phi) = -\,\text{ELBO} = -\mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{y},x)}[\log p_\theta(y^* \,|\, \mathbf{z}, x))] + D_{\text{KL}}(q_\phi(\mathbf{z}\,|\,\mathbf{y}, x) \parallel p(\mathbf{z}\,|\,x))$$

$$\mathscr{L}_{\text{final}}(\theta, \phi) = (1-\lambda)\mathscr{L}_{\text{var}}(\theta, \phi) + \lambda\mathscr{L}_{\text{logic}}(\theta, \phi)$$

$$\mathscr{L}_{\text{logic}}(\theta, \phi) = D_{\text{KL}}\left(p_\theta(\mathbf{y}\,|\,\mathbf{z}, x) \parallel q_\phi^{\text{T}}(\mathbf{y}_z\,|\,\mathbf{y}, x)\right)$$

- ***Supported*** iff all phrases found supported;
- ***Refuted*** iff exists a phrase found refuted;
- ***NEI*** iff not refuted and exists a phrase found unverifiable.

Hard logic

$$q_\phi^{\text{T}}(\boldsymbol{y}_z = \texttt{SUP}) = \prod_{i=1}^{|\boldsymbol{z}|} q_\phi(\boldsymbol{z}_i = \texttt{SUP})$$

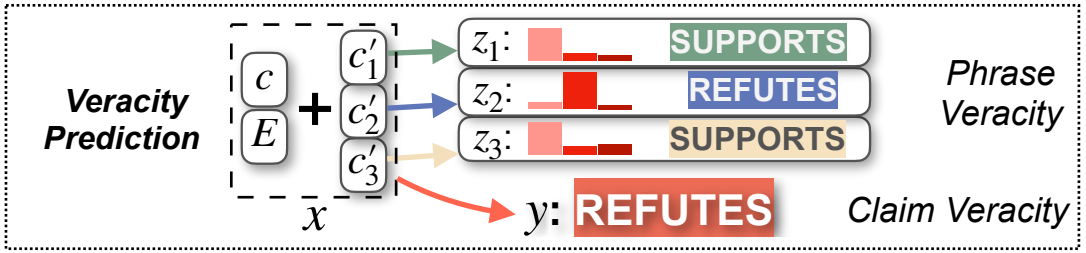
$$q_\phi^{\text{T}}(\boldsymbol{y}_z = \texttt{REF}) = 1 - \prod_{i=1}^{|\boldsymbol{z}|} (1 - q_\phi(\boldsymbol{z}_i = \texttt{REF}))$$

$$q_\phi^{\text{T}}(\boldsymbol{y}_z = \texttt{NEI}) = 1 - q_\phi^{\text{T}}(\boldsymbol{y}_z = \texttt{SUP}) - q_\phi^{\text{T}}(\boldsymbol{y}_z = \texttt{REF})$$

Soft logic

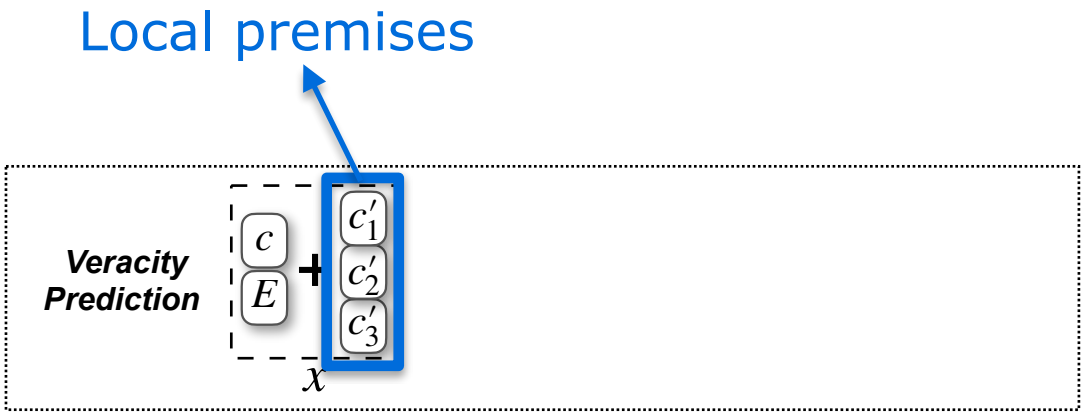
Parameterize $p_\theta(\mathbf{y} \mid x, \mathbf{z})$ and $q_\phi(\mathbf{z} \mid \mathbf{y}, x)$ with Neural Networks

$q_\phi(\mathbf{z} \mid y^*, x)$

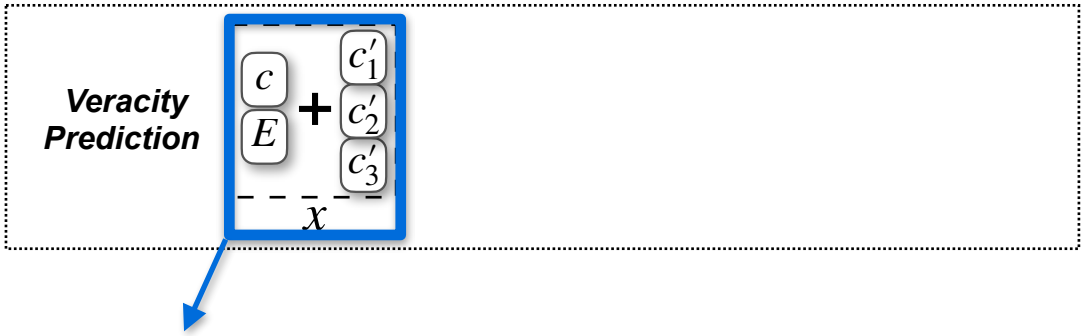


$p_\theta(\mathbf{y} \mid x, \mathbf{z})$

Parameterize $p_\theta(\mathbf{y} \mid x, \mathbf{z})$ and $q_\phi(\mathbf{z} \mid \mathbf{y}, x)$ with Neural Networks

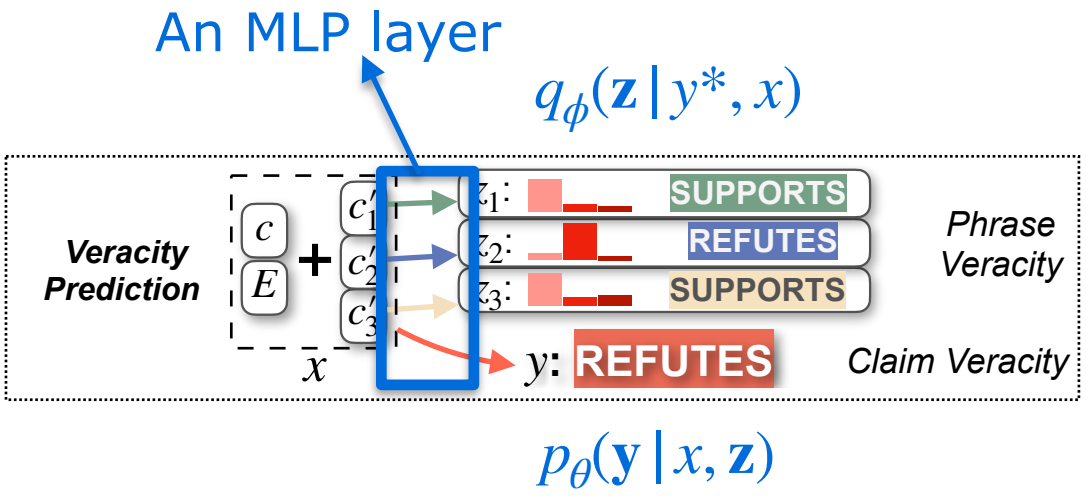


Parameterize $p_\theta(\mathbf{y} \mid x, \mathbf{z})$ and $q_\phi(\mathbf{z} \mid \mathbf{y}, x)$ with Neural Networks

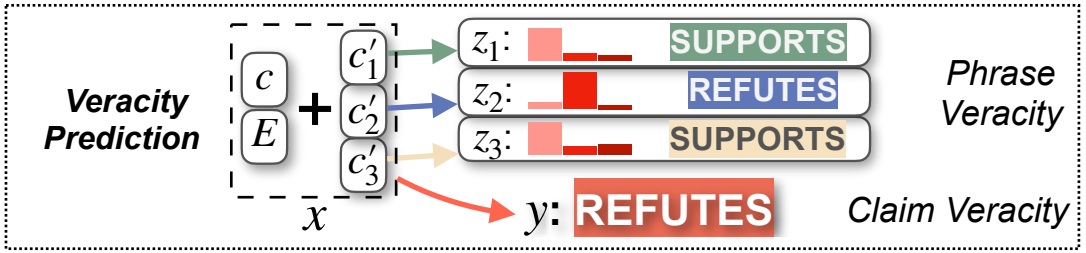


Encode with PLMs (e.g., BERT)

Parameterize $p_\theta(\mathbf{y} \mid x, \mathbf{z})$ and $q_\phi(\mathbf{z} \mid \mathbf{y}, x)$ with Neural Networks



Parameterize $p_\theta(\mathbf{y} \mid x, \mathbf{z})$ and $q_\phi(\mathbf{z} \mid \mathbf{y}, x)$ with Neural Networks (During Inference)



Iterative Decoding

1. $p(\mathbf{z})$
2. $p_\theta(\mathbf{y} \mid x, \mathbf{z})$
3. $q_\phi(\mathbf{z} \mid \mathbf{y}, x)$
4. ...

Understanding LOREN

- RQ1: Can we find rationales without hurting verification performance?
- RQ2: How faithful and accurate are these unsupervised rationales?
- RQ3: How do local premises contribute to LOREN and its rationales?

Research Questions

- **RQ1: Can we find rationales without hurting verification performance?**
- RQ2: How faithful and accurate are these unsupervised rationales?
- RQ3: How do local premises contribute to LOREN and its rationales?

RQ1: Extrinsic Evaluation

Dataset

- FEVER

Metrics

- Label Accuracy (**LA**)
  - Classification accuracy
- FEVER Score (**FEV**)
  - Is the verification using the right *evidence sentence*?

Model	Dev		Test	
	LA	FEV	LA	FEV
UNC NLP	69.72	66.49	68.21	64.21
GEAR ($\text{BERT}_{\text{base}}$)	74.84	70.69	71.60	67.10
DREAM ($\text{XLNet}_{\text{large}}$)	79.16	-	<u>76.85</u>	70.60
KGAT ($\text{BERT}_{\text{large}}$)	77.91	75.86	73.61	70.24
∟ ($\text{RoBERTa}_{\text{large}}$)	78.29	76.11	74.07	70.38
∟ ($\text{CorefRoBERTa}_{\text{l}}$)	-	-	75.96	72.30
LOREN ($\text{BERT}_{\text{large}}$)	78.44	76.21	74.43	70.71
∟ ($\text{RoBERTa}_{\text{large}}$)	<u>81.14</u>	**78.83**	76.42	<u>72.93</u>
ListT5 (T5_{3B})	**81.26**	<u>77.75</u>	**79.35**	**75.87**

Table 2: Overall performance of verification results on the dev and blind test set of FEVER task, where FEV (FEVER score) is the main evaluation metric. The best is **bolded**, and the second best is <u>underlined</u>.

RQ1: Extrinsic Evaluation

Dataset

- FEVER

Metrics

- Label Accuracy (**LA**)
  - Classification accuracy
- FEVER Score (**FEV**)
  - Is the verification using the right *evidence sentence*?

Model	Dev		Test	
	LA	FEV	LA	FEV
UNC NLP	69.72	66.49	68.21	64.21
GEAR ($\text{BERT}_{\text{base}}$)	74.84	70.69	71.60	67.10
DREAM ($\text{XLNet}_{\text{large}}$)	79.16	-	<u>76.85</u>	70.60
KGAT ($\text{BERT}_{\text{large}}$)	77.91	75.86	73.61	70.24
∟ ($\text{RoBERTa}_{\text{large}}$)	78.29	76.11	74.07	70.38
∟ ($\text{CorefRoBERTa}_{\text{l}}$)	-	-	75.96	72.30
LOREN ($\text{BERT}_{\text{large}}$)	78.44	76.21	74.43	70.71
∟ ($\text{RoBERTa}_{\text{large}}$)	<u>81.14</u>	**78.83**	76.42	<u>72.93</u>

Table 2: Overall performance of verification results on the dev and blind test set of FEVER task, where FEV (FEVER score) is the main evaluation metric. The best is **bolded**, and the second best is <u>underlined</u>.

Conclusions

- For similar-sized baselines with similar settings (DREAM, KGAT)
  - very competitive

RQ1: Extrinsic Evaluation

Metrics

- Label Accuracy (**LA**)
  - Classification accuracy
- FEVER Score (**FEV**)
  - Is the verification using the right *evidence sentence*?

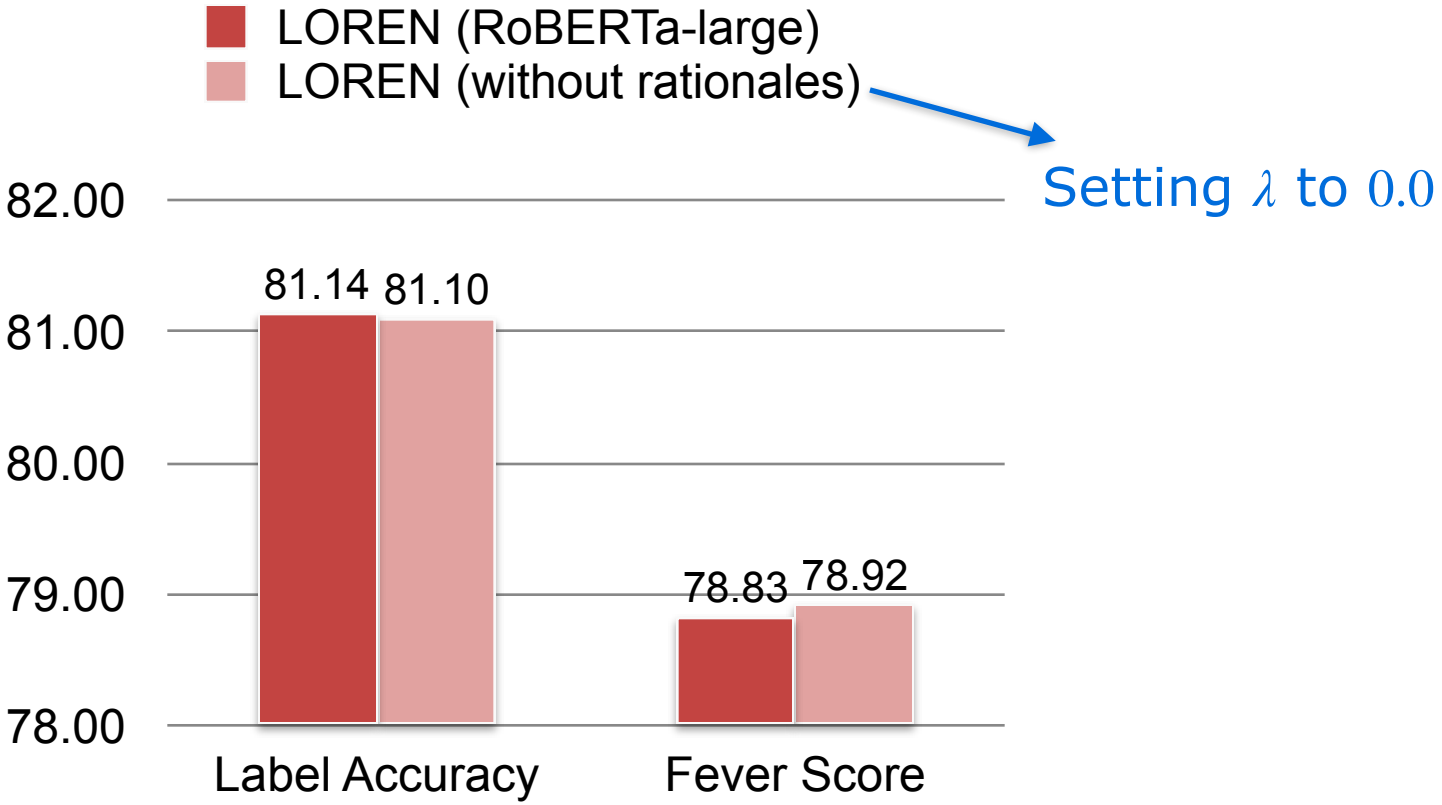
Model	Dev		Test	
	LA	FEV	LA	FEV
UNC NLP	69.72	66.49	68.21	64.21
GEAR ($\text{BERT}_{\text{base}}$)	74.84	70.69	71.60	67.10
DREAM ($\text{XLNet}_{\text{large}}$)	79.16	-	76.85	70.60
KGAT ($\text{BERT}_{\text{large}}$)	77.91	75.86	73.61	70.24
∟ ($\text{RoBERTa}_{\text{large}}$)	78.29	76.11	74.07	70.38
∟ ($\text{CorefRoBERTa}_{\text{l}}$)	-	-	75.96	72.30
LoREN ($\text{BERT}_{\text{large}}$)	78.44	76.21	74.43	70.71
∟ ($\text{RoBERTa}_{\text{large}}$)	81.14	**78.83**	76.42	72.93
ListT5 (T5_{3B})	**81.26**	77.75	**79.35**	**75.87**

Table 2: Overall performance of verification results on the dev and blind test set of FEVER task, where FEV (FEVER score) is the main evaluation metric. The best is **bolded**, and the second best is underlined.

Conclusions

- For similar-sized baselines with similar settings (DREAM, KGAT)
  - very competitive
- For the 10x larger baseline (ListT5)
  - test set⬇️ dev set⬆️

RQ1: Intrinsic Evaluation



Conclusion

- Finding rationales does not hurt verification performance.

Research Questions

- RQ1: Can we find rationales without hurting verification performance?
- **RQ2: How faithful and accurate are these unsupervised rationales?**
- RQ3: How do local premises contribute to LOREN and its rationales?

RQ2: How faithful and accurate are these unsupervised rationales?

Goals of interpretability

🎯Accurate

🎯Faithful

🎯Debuggable

RQ2: How faithful and accurate are these unsupervised rationales?

Goals of interpretability

- 🎯 Accurate
- 🎯 Faithful
- 🎯 Debuggable

Metrics for evaluating rationales

RQ2: How faithful and accurate are these unsupervised rationales?

Goals of interpretability

🎯Accurate

🎯Faithful

🎯Debuggable

Metrics for evaluating rationales

- *Logically aggregated Label Accuracy of* y_z (LA$_z$)
  - Evaluates the overall quality of z
- *Culprit finding accuracy* (CulpA) (P/R/F1)
  - Evaluates the individual quality of z:
  - Are the **culprit phrase(s)** found by rationales (z)?
  - Human evaluation: labeling culprit phrase(s) from claim phrases

RQ2: How faithful and accurate are these unsupervised rationales?

Goals of interpretability

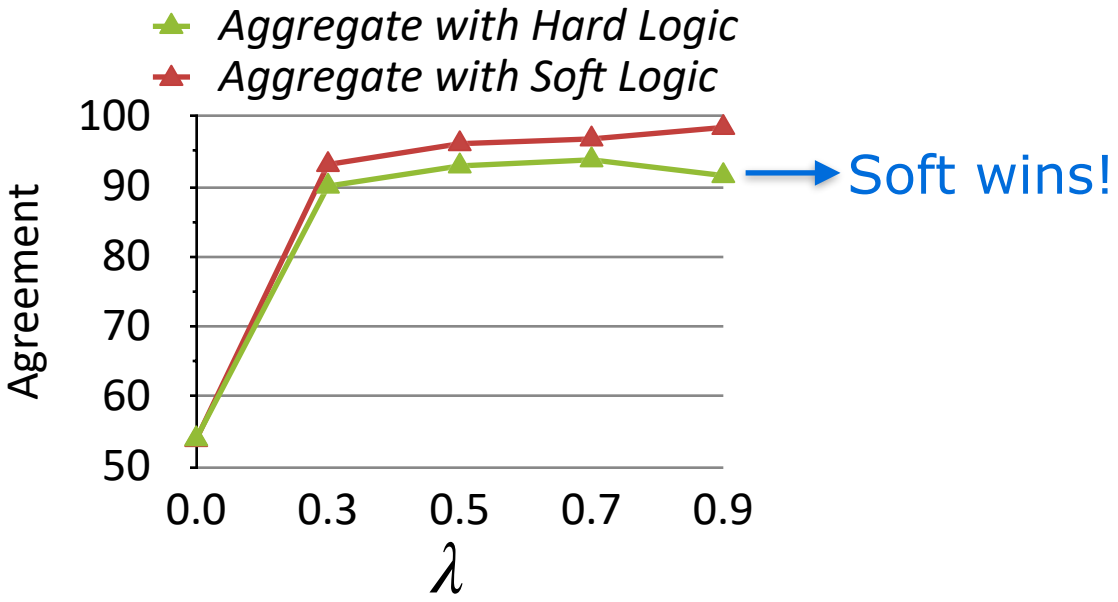
- 🎯 Accurate
- 🎯 Faithful
- 🎯 Debuggable

Metrics for evaluating rationales

- *Logically aggregated Label Accuracy of* y_z (LA$_z$)
  - Evaluates the *overall* quality of z
- *Culprit finding accuracy* (CulpA) (P/R/F1)
  - Evaluates the *individual* quality of z:
  - Are the **culprit phrase(s)** found by rationales (z)? (human evaluation)
- *Agreement of LA and LA$_z$* (AGREE)
  - How aggregated phrase veracity (y_z) agrees with claim veracity (y)

R2: Faithfulness of Rationales

$$\mathscr{L}_{\text{final}}(\theta, \phi) = (1 - \lambda)\mathscr{L}_{\text{var}}(\theta, \phi) + \lambda\mathscr{L}_{\text{logic}}(\theta, \phi)$$

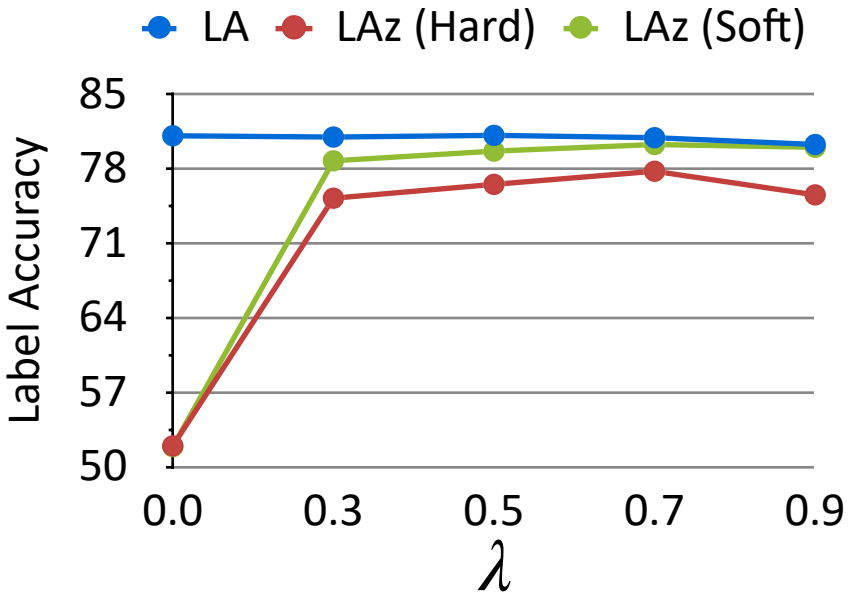


Conclusions

- Agreement > 96%: z are in general faithful.
- λ⬆️, Agree⬆️: stronger regularization from $\mathscr{L}_{\text{logic}}$, deciding the faithfulness of z.
- Soft > Hard: probability distributions of z gives more information than discrete labels.

R2: Overall Accuracy of Rationales

$$\mathcal{L}_{\text{final}}(\theta, \phi) = (1 - \lambda)\mathcal{L}_{\text{var}}(\theta, \phi) + \lambda\mathcal{L}_{\text{logic}}(\theta, \phi)$$



Conclusions

- LA_z is close to 50% when λ=0: Logic is *critical* for interpretability.
- λ⬆️, LA_z⬆️ but quickly plateaued: stronger regularization from $\mathcal{L}_{\text{logic}}$ does not affect performance much.

R2: Individual Accuracy of Rationales

- **An *interpretability shortcut* in the logic**: predicting all phrase veracity to be the same as claim veracity. e.g.,
  1. REF $\vee$ REF $\vee$ REF = REF
  2. REF $\vee$ SUP $\vee$ NEI = REF

- **Potential *risks***:
  - Be tricked by the deceptively high overall accuracy LA_z
  - Rationales being invalid, as no culprit is found.

R2: Individual Accuracy of Rationales

- **An *interpretability shortcut* in the logic**: predicting all phrase veracity to be the same as claim veracity. e.g.,
  1. REF $\vee$ REF $\vee$ REF = REF
  2. REF $\vee$ SUP $\vee$ NEI = REF

- **Potential *risks***:
  - Be tricked by the deceptively high overall accuracy LA_z
  - Rationales being invalid, as no culprit is found.

This can be revealed by altering the prior distribution $p(\mathbf{z}|x)$.

$$\overbrace{-\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{y},x)}[\log p_\theta(y^*|\mathbf{z},x))] + D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{y},x) \,\|\, p(\mathbf{z}|x))}^{\text{negative ELBO: } \mathscr{L}_{\text{var}}(\theta,\phi)}$$

R2: Individual Accuracy of Rationales

Deceptively high

Choice of $p(z)$	LA	LA_z	AGREE	CULPA (P/R/F1)
NLI prior	**81.14**	79.66	96.11	**75.8**/75.9/**74.3**
Pseudo prior	80.93	80.44	**97.25**	70.5/77.1/71.4
Uniform prior	80.85	**80.74**	97.08	34.1/**78.8**/46.1

Uniform distribution as prior

All z to be the same

Table 4: Results of different choices of prior distribution $p(z)$ during training, where y_z in LA_z is calculated using *soft* logic.

R2: Individual Accuracy of Rationales

- A Pre-trained NLI from MNLI
- LA_z at 53.41% for the NLI model before training

Sample a few phrases to be culprits as prior

Choice of $p(z)$	LA	LA_z	AGREE	CULPA (P/R/F1)
NLI prior	**81.14**	79.66	96.11	**75.8**/75.9/**74.3**
Pseudo prior	80.93	80.44	**97.25**	70.5/77.1/71.4
Uniform prior	80.85	**80.74**	97.08	34.1/**78.8**/46.1

Table 4: Results of different choices of prior distribution $p(z)$ during training, where y_z in LA_z is calculated using *soft* logic.

R2: Individual Accuracy of Rationales

Choice of $p(z)$	LA	LA_z	AGREE	CULPA (P/R/F1)
NLI prior	**81.14**	79.66	96.11	**75.8**/75.9/**74.3**
Pseudo prior	80.93	80.44	**97.25**	70.5/77.1/71.4
Uniform prior	80.85	**80.74**	97.08	34.1/**78.8**/46.1

Table 4: Results of different choices of prior distribution $p(z)$ during training, where y_z in LA_z is calculated using *soft* logic.

Conclusions

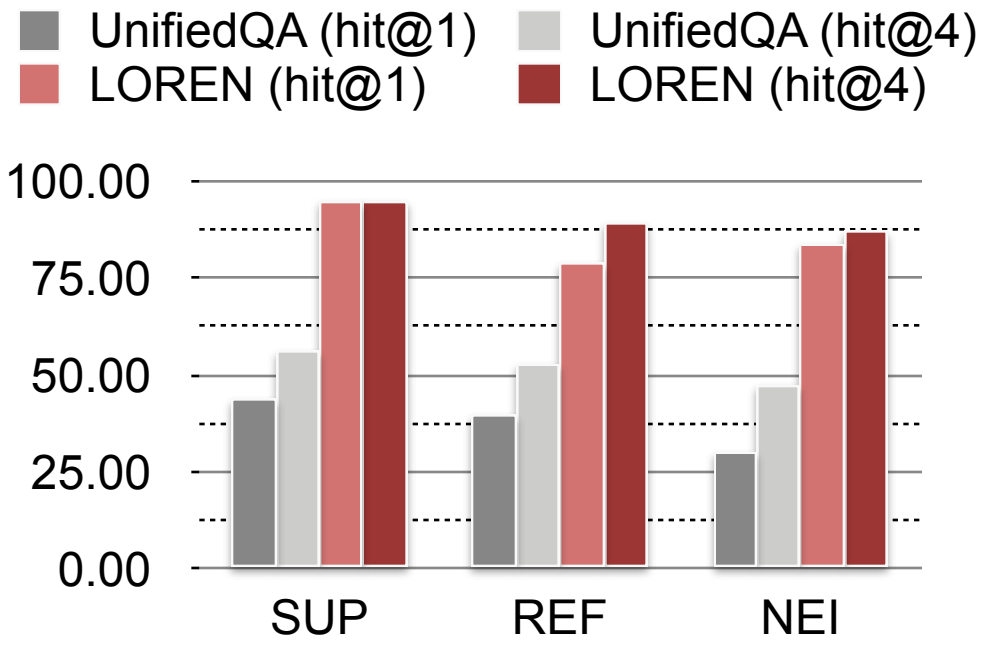
- Prior distribution sets an important starting point for learning the rationales (z), but not on the overall predictions.
- NLI prior and pseudo prior can prevent the degeneration of phrase verification

Research Questions

- RQ1: Can we find rationales without hurting verification performance?
- RQ2: How faithful and accurate are these unsupervised rationales?
- **RQ3: How do local premises contribute to LOREN and its rationales?**

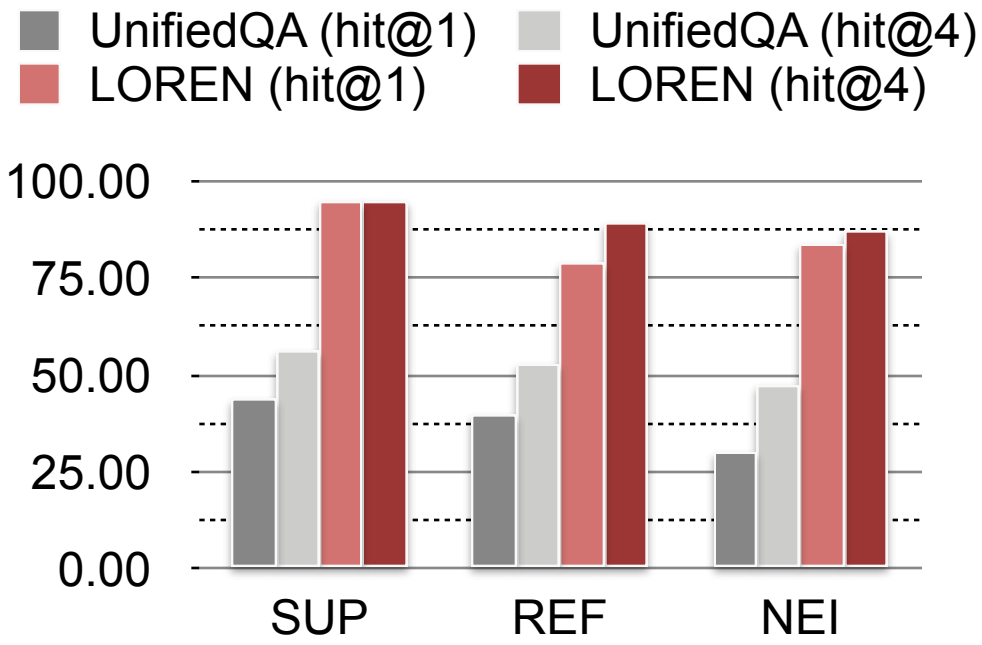
RQ3: Extrinsic Evaluation — MRC Performance

- Randomly sample 238 cases for Manual evaluation.



RQ3: Extrinsic Evaluation — MRC Performance

- Randomly sample 238 cases for Manual evaluation.

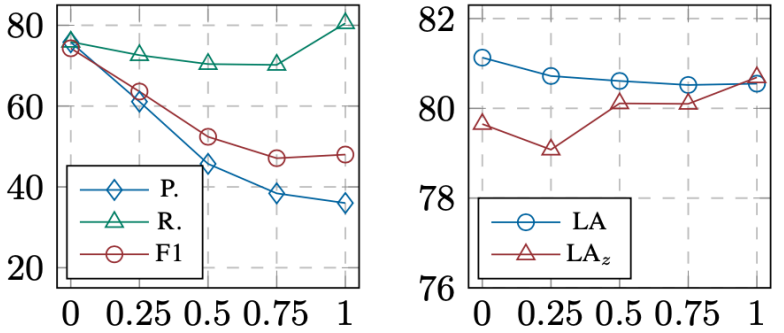


Conclusions

- Self-supervised training for MRC is very beneficial for answering probing questions.
- Automatic factual error correction?

RQ3: Intrinsic Evaluation — Simulating MRC deficiency

- What if MRC fails? — Masking local premises.



(a) Results on CULPA. (b) Results on LA & LA$_z$.

Figure 2: Performance on culprit finding (CULPA) and verification (LA and LA$_z$) vs. the mask rate ρ of local premises, simulating the influence by deficiency of the MRC model.

Conclusions

- MRC is critical for the quality of individual rationales.
- Phrase verification degenerates to claim verification as MRC deteriorates.

Research Questions Revisited

- **RQ1: Can we find rationales without hurting verification performance?**
  - Yes, even with a little boost for some cases.
- **RQ2: How faithful and accurate are these unsupervised rationales?**
  - Very faithful (96%+ agreement) and accurate (both in overall and individually).
  - Logic regularizes the quality of phrase veracity.
  - Careful for the "interpretability shortcut".
- **RQ3: How do local premises contribute to LOREN and its rationales?**
  - Minor contribution to claim veracity prediction.
  - Critical to the quality of phrase veracity prediction.

Debugging LOREN

Claim2: Ashley Cole is Iranian.
Evidence: Ashley Cole (born 20 December 1980) is an English professional footballer who … in Major League Soccer. Born in Stepney , London…

Debugging LOREN

Claim2: Ashley Cole is Iranian.
Evidence: Ashley Cole (born 20 December 1980) is an English professional footballer who … in Major League Soccer. Born in Stepney , London…

Premise1: Ashley Cole is Iranian.

Premise2: Ashley Cole is European.

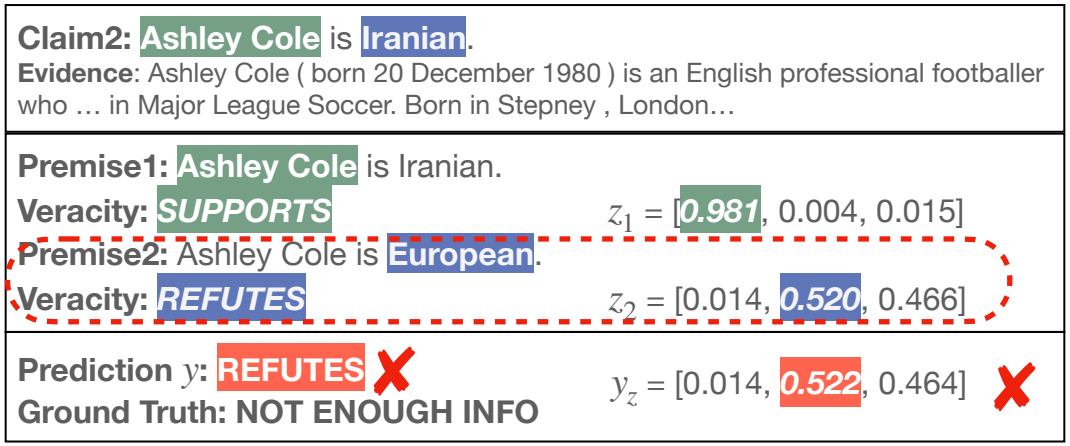
Debugging LOREN

Claim2: Ashley Cole is Iranian.
Evidence: Ashley Cole (born 20 December 1980) is an English professional footballer who … in Major League Soccer. Born in Stepney , London…

Premise1: Ashley Cole is Iranian.
Veracity: *SUPPORTS* z_1 = [*0.981*, 0.004, 0.015]
Premise2: Ashley Cole is European.
Veracity: *REFUTES* z_2 = [0.014, *0.520*, 0.466]

Prediction y: REFUTES y_z = [0.014, *0.522*, 0.464]

Debugging LOREN



- Wrong verification for Iranian.
- Rather close probabilities of **REF** and **NEI**.
  - z_2: $p_{\text{REF}} = 0.520$ vs. $p_{\text{REF}} = 0.466$
  - y_z: $p_{\text{REF}} = 0.522$ vs. $p_{\text{REF}} = 0.464$

Takeaways

- Goal of Reasoning
  - **correct answer for the right thinking**
- A good pipeline (LOREN) offers **interpretability** to the prediction
  - Faithful, accurate and debuggable
- A reasoning paradigm: symbolic AI **plans**, connectionist AI **executes**.
  - Planning with logic, learning with data

Have Fun with LOREN!

Checkout our code at **GitHub**!
https://github.com/jiangjiechen/LOREN

Checkout our demo at 🤗 **Spaces**!
https://huggingface.co/spaces/Jiangjie/loren-fact-checking