

# Argument-structured Justification Generation for Explainable Fact-checking

Xiaoou Wang, Elena Cabrio, Serena Villata

Université Côte d'Azur, CNRS, Inria, I3S, France

# Automatic justification production is **important** in automated fact-checking

- ❖ Manually crafting justifications is a **time-intensive** process

- Several hours, even **days** for professional fact-checkers (Adair et al., 2017)




- ❖ Justification production is useful in:

- Creating **feedback loops** which correct judgment errors (O'neil, 2017)
- Convincing readers on the **credibility** of automated fact-checking (Eldifrawi et al., 2024)
- Avoiding the **“back-fire effect”** induced by black-box models (Lewandowsky et al., 2012)
- Educating readers about how to **critically evaluate** news themselves (Guo et al., 2022)



# Motivations of our work

- ❖ Most works in automatic justification production **presume the availability of a pre-existing human-written fact-checking article**
  - **Unrealistic** in practice 
- ❖ Most generated justifications are evaluated using only **overlap-based metrics such as ROUGE scores**
  - **Without considering claim verification**, the essential task in the context of fact-checking

# Main contributions of our work

- ❖ **Novel argument-structured** justification generation method based on a novel dataset that we built from LIAR-PLUS, named LIARArg
  - **Significant improvements** in F1 scores across **three standard benchmarks** compared to
    - state-of-the-art summarizer in fact-checking
    - human-written summaries
- ❖ Our **jointly-trained summarization and evidence retrieval system** outperforms the state-of-the-art method **JustiLM** on ExClaim (Zeng et al., 2024)
  - the biggest dataset for this task
  - no human-written fact-checking articles are provided during verification of news claims
- ❖ We show that ROUGE scores are **not correlated** with F1 scores in claim verification

# Related works in Justification Production I

- ❖ Attention weights to highlight key parts of the retrieved evidence as explanations (Dua et al., 2023)
  - lacks structured information and does not reveal the underlying reasoning process






- ❖ Logic-based rules, such as knowledge graphs, to derive explanations by tracing the rule paths of algorithms like decision trees (Vedula et al., 2023)
  - not readily understandable to general users

# Related works in Justification Production II

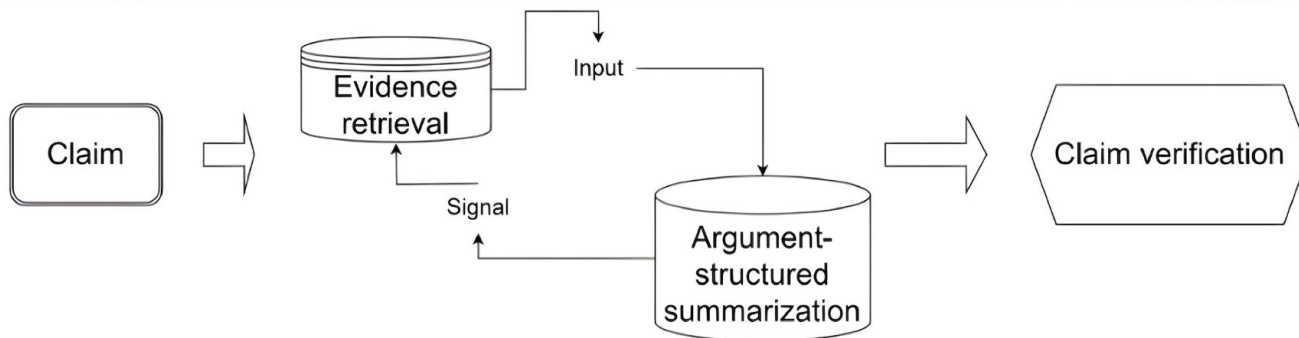
- ❖ Recent studies **cast justification production as summarization**, representative works:
  - Kotonya and Toni, 2020
    - SBERT to **extract sentences** from fact-checking articles
    - BERTSUM model to **generate abstractive justifications** based on the extracted sentences
  - Russo et al., 2023
    - Combining **abstractive summarization with a claim-driven extractive** step using SBERT yields the best results
- ❖ Limitations
  - Dependent on **human-written** fact-checking articles as input
  - Ignoring the **evidence search** process
  - Evaluated **solely** using overlap-based metrics as **ROUGE**

# OUR FACT-CHECKING PIPELINE

-  Only requires human-written fact-checking articles during the **training** process
-  Retrieves evidence within a large corpus during **inference** to serve as input to justification generation
-  The generated justifications are evaluated using **both ROUGE scores and the F1 scores of a state-of-the-art claim verification system**

# OUR FACT-CHECKING PIPELINE

Claim (input)	John McCain: Iran "might not be a superpower, but the threat the government of Iran poses is anything but 'tiny,'" as Obama says.
Evidence retrieval module	Doc1: But Obama never said the threat from Iran was "tiny" or "insignificant".. Doc2: In fact, Obama has repeatedly called Iran a grave threat. Doc3: This isn't the first time Obama has talked about the grave threat posed by Iran...
Summarization module	John McCain said that.. "But Obama never said..." <b>attacks</b> this claim, "In fact, Obama has repeatedly called Iran a grave..." <b>attacks</b> this claim ...
Claim verification module	False



The fact-checking architecture we propose, where evidence retrieval and summarization are trained jointly.



# Argument-structured summarization, corpus

- ❖ Training corpus LIARArg, built from LIAR-PLUS dataset (Alhindi et al., 2018)
  - **12,836 news claims** taken from POLITIFACT
  - **6 labels:** Pants-On-Fire, False, Mostly-False, Half-true, Mostly-True and True
  - Each claim is accompanied by an **automatically scraped summary**

**Statement:**“Says Rick Scott cut education to pay for even more tax breaks for big, powerful, well-connected corporations.”

**Speaker:** Florida Democratic Party

**Context:** TV Ad

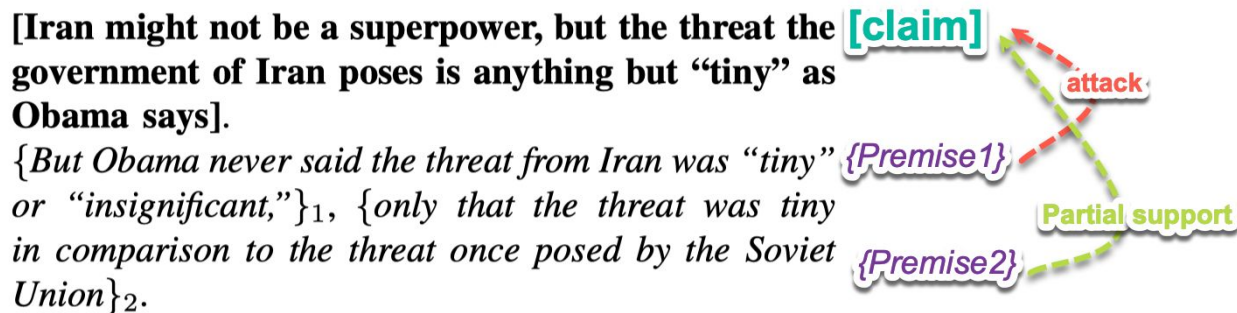
**Label:** half-true

**Extracted Justification:** A TV ad by the Florida Democratic Party says Scott ”cut education to pay for even more tax breaks for big, powerful, well-connected corporations.” However, the ad exaggerates when it focuses attention on tax breaks for ”big, powerful, well-connected corporations.” Some such companies benefited, but so did many other types of businesses. And the question of whether the tax cuts and the education cuts had any causal relationship is murkier than the ad lets on.

Excerpt from the LIAR-PLUS dataset (Alhindi et al., 2018)

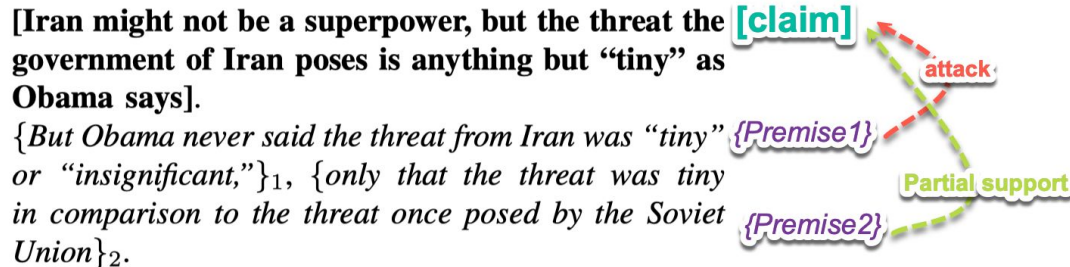
# Argument-structured summarization, corpus enhancement

- ❖ Training corpus LIARArg, enhanced upon LIAR-PLUS dataset
  - We enhanced each summary **with argumentative labels**
    - Claims and premises
    - Relations
      - Support, Attack, Partial support, Partial attack
  - 2,832 automatically scraped summaries



# Argument-structured summarization, corpus statistics

- ❖ Training corpus LIARArg, enhanced with argument-structured summaries
  - 2,832 automatically scraped summaries
  - 2,832 **automatically converted argument-structured summaries** (ground-truth for training)
    - Each summary **begins with** "X said..."
    - Followed by the **concatenation of relations** with the claim, formatted as "A attacks this claim," "B supports this claim,"
    - Following the appearance order: attack, support, partial attack, and partial support.
  - 2,832 **full-length fact-checking articles** (40 sentences on average)



# Argument-structured summarization, attackability scorer

- ❖ Fine-tune Mixtral-8x22B (Jiang et al., 2023) using QLoRA (Detmeters et al. 2023), besides the standard Cross-Entropy loss for summarization, we introduce an **attackability scorer (AttScorer)**
  - Based on the ChangeMyView (CMV) dataset (Jo et al., 2020)
    - 199,711 **claim-sentence pairs** labeled as attacked (-1), supported (1) or not attacked (0)
  - Enhanced with LIARArg
    - by **merging partial attacks and partial supports** to attacks and supports
    - 18,496 **sentence pairs** labeled as attacked (-1), supported (1), and neutral (0)
  - Trained through a BERT model
    - Obtain a **vectorized representation** of the concatenated each sentence pair
    - Get a score  $\hat{y}$  that reflects the attackability score of a sentence pair
    - Use Softmax to **minimize between  $\hat{y}$  and ground-truth label  $y$**

$$l(y, \hat{y}) = - \sum_{i=1}^n y_i \cdot \log \left( \frac{\exp(\hat{y}_i)}{\sum_{j=1}^n \exp(\hat{y}_j)} \right)$$

# Argument-structured summarization, attackability scorer

- ❖ The attackability scorer (**AttScorer**) allows to
  - Compute an **attackability score** for each **generated summary**
  - Compared with the score of **ground-truth summary**
  - During the training, we use **Mean Squared Error (Loss<sub>mse</sub>)** to minimize the difference

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

# Argument-structured summarization, evaluation

## ❖ Baseline

- **Abstractive-extractive method (CDAE)** of Russo et al., 2023

## ❖ Datasets

- The whole LIAR-PLUS dataset excluding LIARArg, 6 labels (Alhindi et al., 2018)
- FNC-1, general news, 4 labels (Hanselowski et al. 2018)
- Check-Covid, Covid-related news, 2 labels (Wang et al., 2023)

## ❖ 10-fold cross-validation

## ❖ Metrics

- **Mean Rouge scores** (R1, R2 and RL)
- **F1-score** using the generated summaries fed into the **state-of-the-art model in Fake News Classification** (Knowledge-enhanced Bert (Peters et al., 2019)+ Knowledge-enhanced graph embedding (Ma et al., 2023))

# Argument-structured summarization, results

- ❖ Ground = **human-written articles** as input
- ❖ SumArg and SumArgAttack = **argument-structured summarization** with or without integration of attackability score calculation


F1 SCORES OF THE CLAIM VERIFICATION SYSTEM OF VARIOUS SUMMARIZERS ON LIAR-PLUS, FNC-1 AND CHECK-COVID.

<b>Method</b>	<b>LIAR-PLUS</b>	<b>FNC-1</b>	<b>Check-Covid</b>
Ground	0.51	0.90	<b>0.76</b>
CDAE	0.41	0.76	0.62
SumArg	0.48	0.85	0.69
SumArgAttack	<b>0.54</b>	<b>0.92</b>	0.74

# Argument-structured summarization, results

- ❖ Argument-structured summaries **significantly improve** the performance of the claim verification module
  - **Compared to the state-of-the-art summarization** method for fact-checking
    - SumArg, SumArgAttack vs. CDAE
  - **Compared to human-written articles** when **attackability score calculation is integrated**
    - SumArgAttack vs. Ground

F1 SCORES OF THE CLAIM VERIFICATION SYSTEM OF VARIOUS SUMMARIZERS ON LIAR-PLUS, FNC-1 AND CHECK-COVID.



Method	LIAR-PLUS	FNC-1	Check-Covid
Ground	0.51	0.90	<b>0.76</b>
CDAE	0.41	0.76	0.62
SumArg	0.48	0.85	0.69
SumArgAttack	<b>0.54</b>	<b>0.92</b>	0.74



# Argument-structured summarization, results



## ❖ Why SumArgAttack **works better** than Ground?

**Evidence provided by SumArgAttack:** 1) she said NAFTA had some positive effects "but unfortunately it had a lot of downside."; 2) Both promised to crack down on China's practice of manipulating its currency to give its products an unfair advantage. 3) Both said they opposed the Chinese government subsidizing industry to the detriment of U.S. competitors. 4) At a debate in December 2007, she announced her intention to review and reform NAFTA if she were elected.

**Evidence provided in the human-written summary of LIAR-PLUS:** 1) Clinton has in the past verbally supported NAFTA and permanent trade with China; 2) she has spoken forcefully about the need to reform NAFTA and to much more stringently enforce trade agreements with China.

★ Generated summaries contain **more comprehensive** evidence

Method	LIAR-PLUS	FNC-1	Check-Covid
Ground	0.51	0.90	<b>0.76</b>
CDAE	0.41	0.76	0.62
SumArg	0.48	0.85	0.69
SumArgAttack	<b>0.54</b>	<b>0.92</b>	0.74

# Argument-structured summarization, results



- ❖ Why SumArgAttack **works better** than Ground?
- ★ SumArgAttack generates summaries containing explicit fine-grained argument relations (partial support and partial attack)
  - Especially useful in the case of **half-truths** (Estornell et al., 2020) such as **half-true and mostly-false**
    - Better F1-scores are observed on LIAR-PLUS (6 labels) and FNC-1 (4 labels)

Method	LIAR-PLUS	FNC-1	Check-Covid
Ground	0.51	0.90	<b>0.76</b>
CDAE	0.41	0.76	0.62
SumArg	0.48	0.85	0.69
SumArgAttack	<b>0.54</b>	<b>0.92</b>	0.74

- Better F1-scores are observed **especially** on **Half-true and Mostly-true labels** on LIAR-PLUS
  - **0.41 and 0.32** for SumArgAttack
  - 0.30 and 0.23 for CDAE
  - 0.36 and 0.28 for Ground
- Better F1-scores are observed **especially** on **Discuss** labels on Check-Covid
  - **0.33** for SumArgAttack, 0.24 for CDAE and 0.29 for Ground

# Argument-structured summarization, results



❖ Why SumArgAttack **works better** than SumArg?


★ SumArgAttack produces **less hallucinated attack-support relations** in the generated summaries

- Average number of relations dropping **from 7 to 3** for summaries generated for LIAR-PLUS
- The **ground-truth** average number of relations in LIARArg is **2.5**

# Argument-structured summarization, results

- ❖ Rouge scores are not correlated with FNC performance
  - SumArgAttack, the summarizer who **scores the best in Fake News Classification**, produces summaries with **the lowest ROUGE scores**.

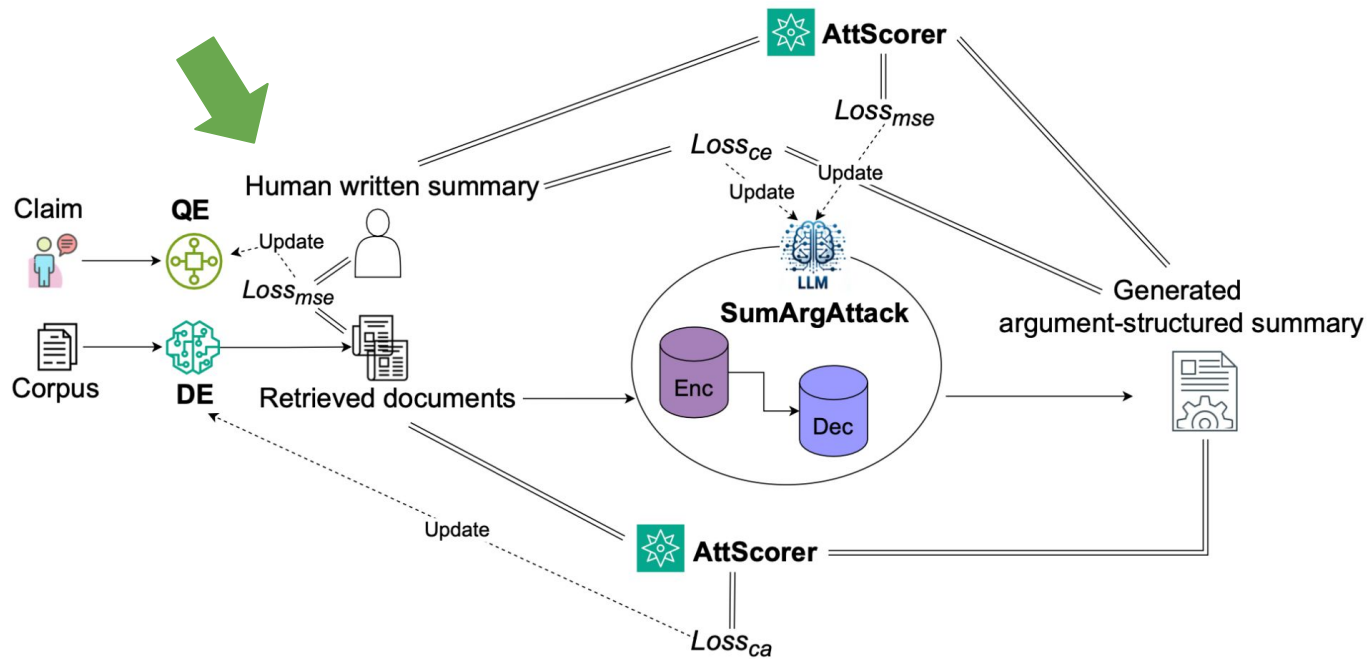
ROUGE SCORES FOR GENERATED SUMMARIES COMPARED WITH HUMAN-WRITTEN SUMMARIES ON LIAR-PLUS.



<b>Summarizer</b>	<b>R1</b>	<b>R2</b>	<b>RL</b>
CDAE	0.348	0.159	0.272
SumArg	0.273	0.130	0.158
SumArgAttack	<b>0.268</b>	<b>0.128</b>	<b>0.143</b>

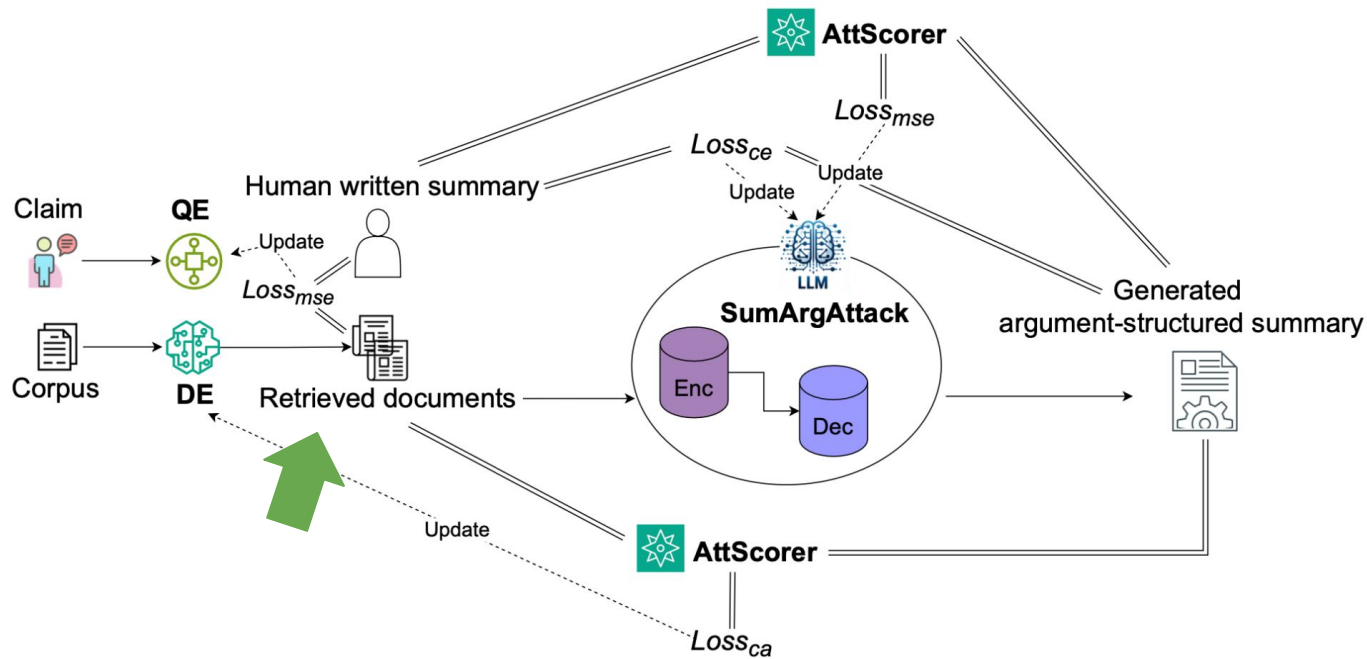
# ArgLM, retrieval-generation pipeline

- ❖ Compare **human-written summary with retrieved documents' attackability scores** to update the **Query Encoder**



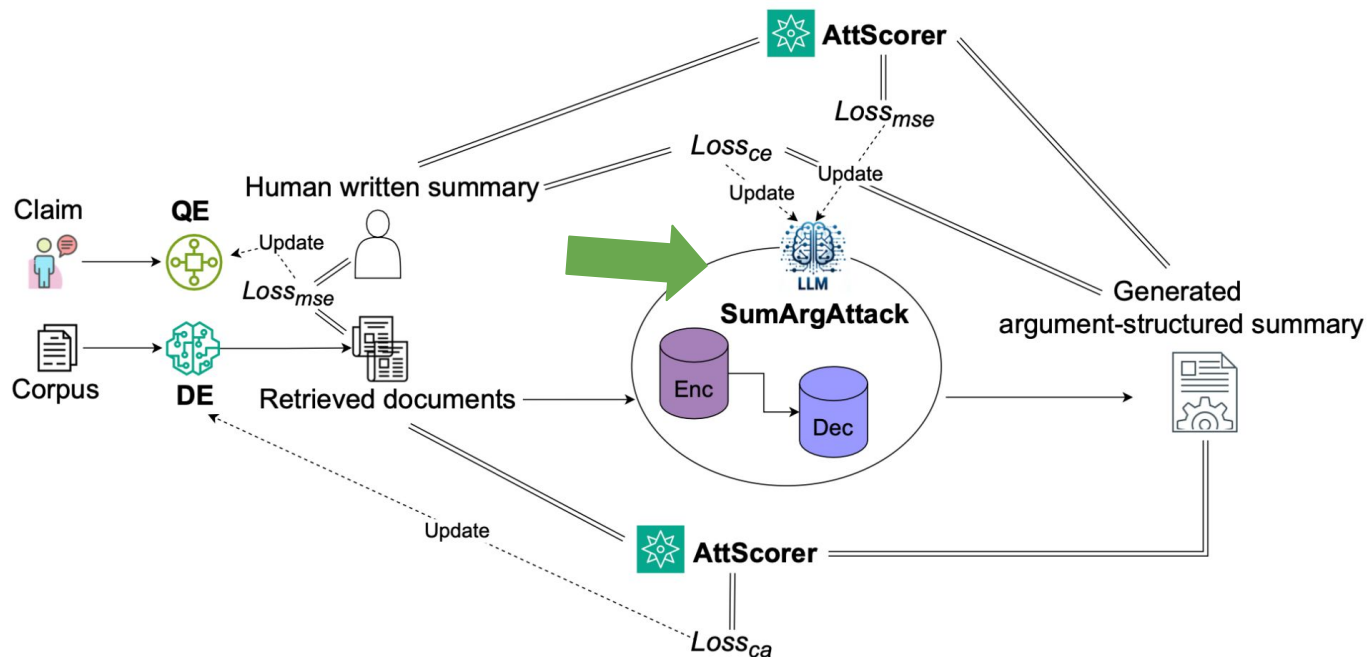
# ArgLM, retrieval-generation pipeline

- ❖ Compare **generated summary** with **retrieved documents' attackability scores** to update the **Document Encoder**



# ArgLM, retrieval-generation pipeline



- ❖ Compare **human-written summary with generated summary (Cross-Entropy + Attackability score)** to update the **Language Model**



# ArgLM, evaluation and results

- ❖ Compared with JustiLM, the state-of-the-art pipeline **with retrieval integrated for summary generation in fact-checking on Exclaim (Zeng et al., 2024)**
  - ArgLM produces a **F1 difference of 0.08**, while having the **lowest ROUGE scores**
  - In 38 generated summaries, when **all the argument relations are wrong**, the error rate is **65% vs. 25%** when **at least one argument relation is correct**.

F1 SCORES OF THE CLAIM VERIFICATION SYSTEM AND ROUGE SCORES OF GENERATED SUMMARIES COMPARED WITH HUMAN-WRITTEN SUMMARIES ON EXCLAIM.

	<b>Pipeline</b>	<b>F1</b>	<b>R1</b>	<b>R2</b>	<b>RL</b>
	JustiLM	0.65	0.376	0.189	0.343
	ArgLM	<b>0.73</b>	<b>0.298</b>	<b>0.156</b>	<b>0.223</b>



# Takeaways

- ❖ By generating **argument-structured summaries** and integrating **loss function** based on attackability score, our summarizer achieves **state-of-the-art F1 scores on 3 datasets when fed into a fact-checking module**
  - Attackability score integration **reduces hallucinated argument relations** ↓
  - The performance boost is **notable on half-truth labels** 👍
- ❖ By jointly training retrieval and generation using AttScorer, our pipeline achieves **state-of-the-art F1 score on Exclaim, the biggest dataset** where retrieval and generation are required when summarizing texts. 👍
- ❖ ROUGE scores **should not be the only metrics** when evaluating summarization in the context of fact-checking. ⚠

*Thank You*