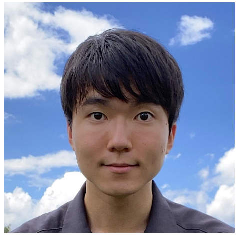
TAUR LAB

TEXAS
The University of Texas at Austin

WiCE: Real-World Entailment for Claims in Wikipedia

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett



Automated Fact-Checking of Complex Claims

Claim

Climate change isn’t real because Florida has had 119 hurricanes since 1850

Supported?

Not Supported?

Automated Fact-Checking of Complex Claims

Claim

Climate change isn't real because Florida has had 119 hurricanes since 1850

Supported?

Not Supported?

Climate change doesn't cause hurricanes

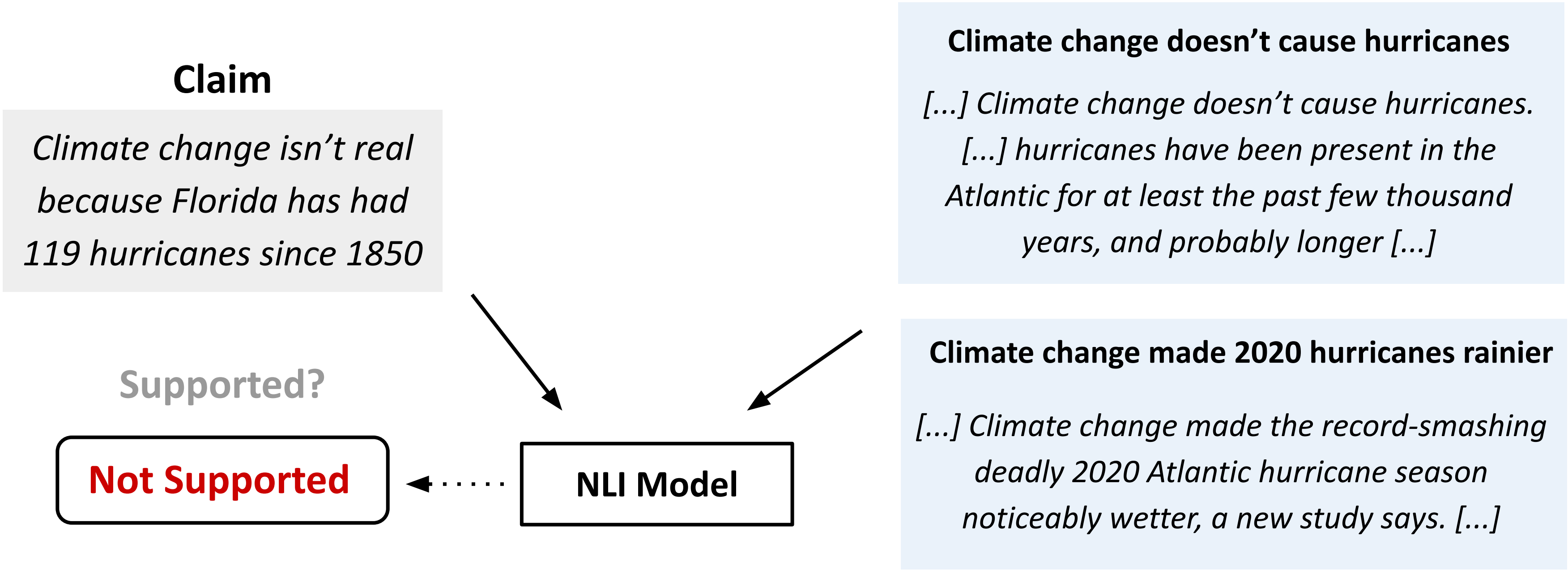
[...] Climate change doesn't cause hurricanes. [...] hurricanes have been present in the Atlantic for at least the past few thousand years, and probably longer [...]

Climate change made 2020 hurricanes rainier

[...] Climate change made the record-smashing deadly 2020 Atlantic hurricane season noticeably wetter, a new study says. [...]

Example from AP FACT CHECK

Automated Fact-Checking of Complex Claims



Claim

Climate change isn't real because Florida has had 119 hurricanes since 1850

Supported?

Not Supported

NLI Model

Climate change doesn't cause hurricanes

[...] Climate change doesn't cause hurricanes. [...] hurricanes have been present in the Atlantic for at least the past few thousand years, and probably longer [...]

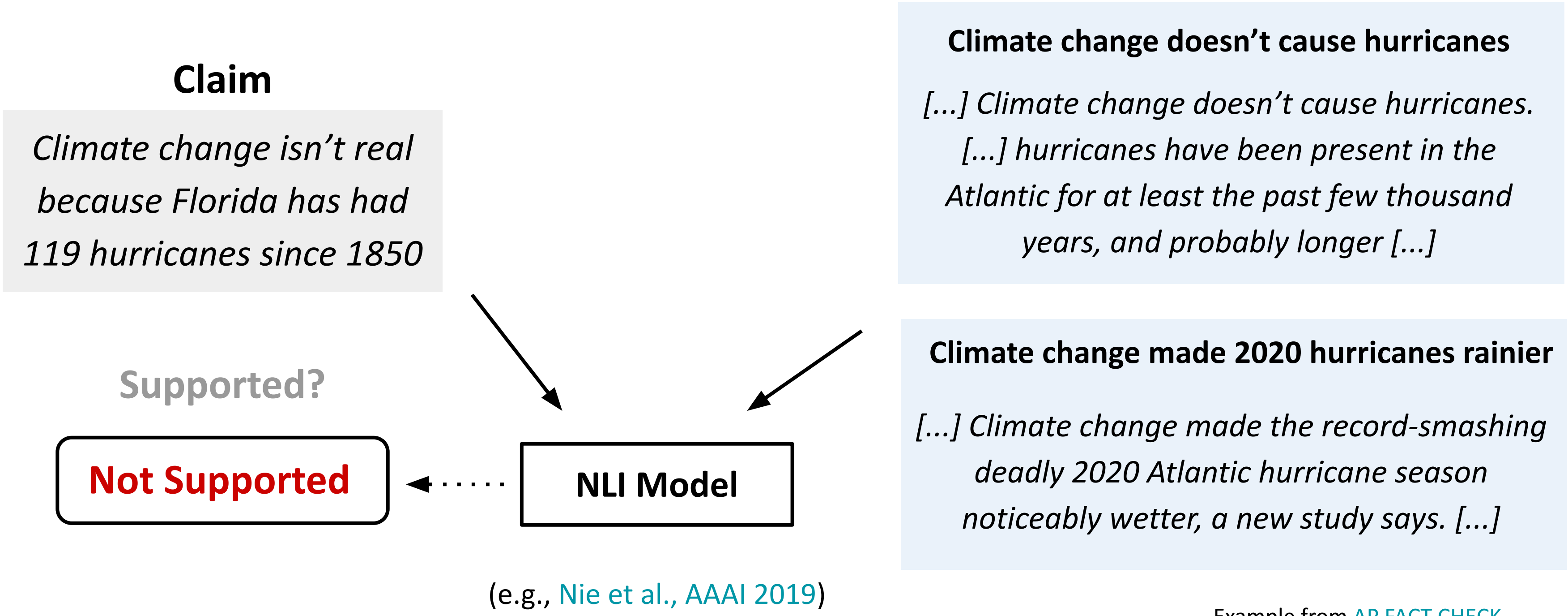
Climate change made 2020 hurricanes rainier

[...] Climate change made the record-smashing deadly 2020 Atlantic hurricane season noticeably wetter, a new study says. [...]

(e.g., Nie et al., AAAI 2019)

Example from AP FACT CHECK

Automated Fact-Checking of Complex Claims



Claim

Climate change isn't real because Florida has had 119 hurricanes since 1850

Climate change doesn't cause hurricanes

[...] Climate change doesn't cause hurricanes. [...] hurricanes have been present in the Atlantic for at least the past few thousand years, and probably longer [...]

Climate change made 2020 hurricanes rainier

[...] Climate change made the record-smashing deadly 2020 Atlantic hurricane season noticeably wetter, a new study says. [...]

(e.g., Nie et al., AAAI 2019)

Example from AP FACT CHECK

We need datasets suitable for training and analyzing fact-checking systems

How to Create Useful Dataset for Fact-Checking?

It is challenging to construct a dataset of real fact-checking examples

How to Create Useful Dataset for Fact-Checking?

It is challenging to construct a dataset of real fact-checking examples

↓

In WiCE, we use claims and evidence in **Wikipedia articles** to approximate the fact-checking

How to Create Useful Dataset for Fact-Checking?

It is challenging to construct a dataset of real fact-checking examples

↓

In WiCE, we use claims and evidence in **Wikipedia articles** to approximate the fact-checking

Sentence in Wikipedia article

Claim

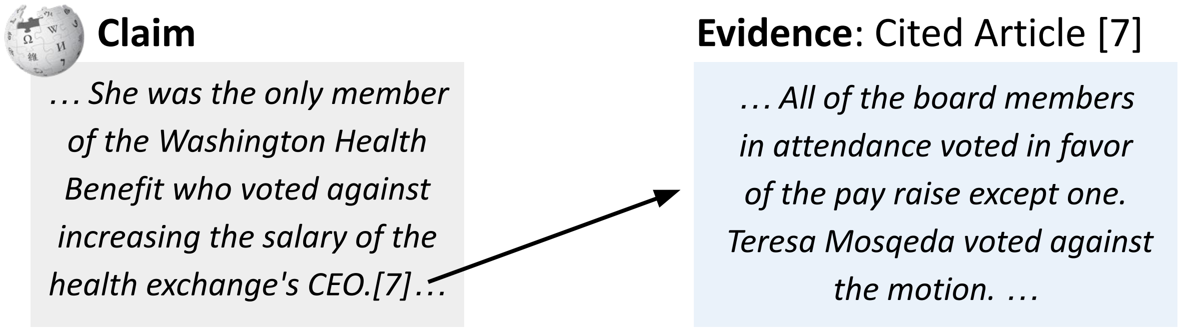
… She was the only member of the Washington Health Benefit who voted against increasing the salary of the health exchange's CEO.[7] …

How to Create Useful Dataset for Fact-Checking?

It is challenging to construct a dataset of real fact-checking examples

↓

In WiCE, we use claims and evidence in **Wikipedia articles** to approximate the fact-checking



Sentence in Wikipedia article

Claim

… She was the only member of the Washington Health Benefit who voted against increasing the salary of the health exchange's CEO.[7]…

Websites cited for the claim

Evidence: Cited Article [7]

… All of the board members in attendance voted in favor of the pay raise except one. Teresa Mosqeda voted against the motion. …

Why WiCE is Suitable for Real-World Fact-Checking?

We need realistic **testbeds** and good **training data** for fact-checking systems

Existing Datasets

- Short premise
  - SNLI, MNLI, WANLI, ANLI
- Synthetic negative cases
  - DocNLI (e.g., word replacement)
- Lack of fine-grained annotation
  - ContractNLI includes fine-grained annotation, but only for a single domain
- Easy Verification Problems
  - FEVER, VitaminC

Why WiCE is Suitable for Real-World Fact-Checking?

We need realistic **testbeds** and good **training data** for fact-checking systems

Existing Datasets

WiCE

- Short premise
  - SNLI, MNLI, WANLI, ANLI
- Synthetic negative cases
  - DocNLI (e.g., word replacement)
- Lack of fine-grained annotation
  - ContractNLI includes fine-grained annotation, but only for a single domain
- Easy Verification Problems
  - FEVER, VitaminC

Why WiCE is Suitable for Real-World Fact-Checking?

We need realistic **testbeds** and good **training data** for fact-checking systems

Existing Datasets

- Short premise
  - SNLI, MNLI, WANLI, ANLI
- Synthetic negative cases
  - DocNLI (e.g., word replacement)
- Lack of fine-grained annotation
  - ContractNLI includes fine-grained annotation, but only for a single domain
- Easy Verification Problems
  - FEVER, VitaminC

WiCE

- **Long premise**
  - Includes 120 sentences on average

Why WiCE is Suitable for Real-World Fact-Checking?

We need realistic **testbeds** and good **training data** for fact-checking systems

Existing Datasets

- Short premise
  - SNLI, MNLI, WANLI, ANLI
- Synthetic negative cases
  - DocNLI (e.g., word replacement)
- Lack of fine-grained annotation
  - ContractNLI includes fine-grained annotation, but only for a single domain
- Easy Verification Problems
  - FEVER, VitaminC

WiCE

- **Long premise**
  - Includes 120 sentences on average
- **Natural negative cases**
  - Mistakes by Wikipedia authors

Why WiCE is Suitable for Real-World Fact-Checking?

We need realistic **testbeds** and good **training data** for fact-checking systems

Existing Datasets

- Short premise
  - SNLI, MNLI, WANLI, ANLI
- Synthetic negative cases
  - DocNLI (e.g., word replacement)
- Lack of fine-grained annotation
  - ContractNLI includes fine-grained annotation, but only for a single domain
- Easy Verification Problems
  - FEVER, VitaminC

WiCE

- **Long premise**
  - Includes 120 sentences on average
- **Natural negative cases**
  - Mistakes by Wikipedia authors
- **Fine-grained annotation**
  - Sub-claim level annotation
  - Supporting sentences
  - Non-supported tokens

Why WiCE is Suitable for Real-World Fact-Checking?

We need realistic **testbeds** and good **training data** for fact-checking systems

Existing Datasets

- Short premise
  - SNLI, MNLI, WANLI, ANLI
- Synthetic negative cases
  - DocNLI (e.g., word replacement)
- Lack of fine-grained annotation
  - ContractNLI includes fine-grained annotation, but only for a single domain
- Easy Verification Problems
  - FEVER, VitaminC

WiCE

- **Long premise**
  - Includes 120 sentences on average
- **Natural negative cases**
  - Mistakes by Wikipedia authors
- **Fine-grained annotation**
  - Sub-claim level annotation
  - Supporting sentences
  - Non-supported tokens
- **Challenging Verification Problems**

Our Contributions

- We collect a document-level textual entailment dataset based on Wikipedia
  - 1,967 claims (5,377 subclaims) in total
  - Fine-grained annotation
- We propose a method to automatically decompose claims into subclaims
- We evaluate fact-checking frameworks on this dataset
  - Fine-tuned T5-3B and few-shot GPT-4

Overview

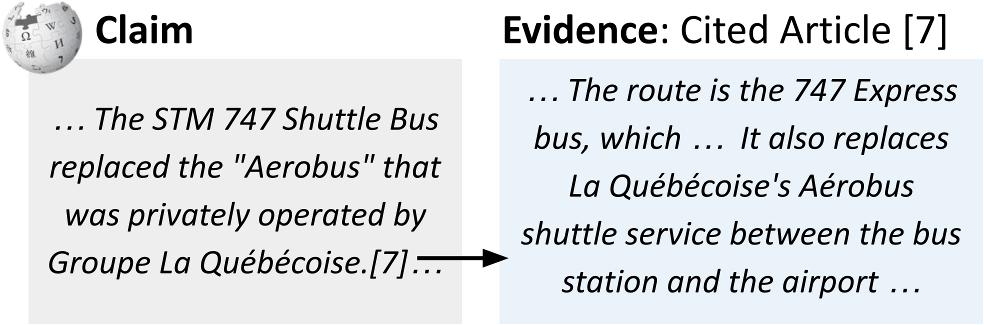
Construction of WiCE

- Sub-claims generated by **Claim-Split**
- Indices of supporting sentences
- Non-Supported Tokens

Experiments on WiCE

- How well do existing NLI models perform on WiCE?
- How well does GPT-4 perform?

Fine-Grained Annotation on WiCE



Claim

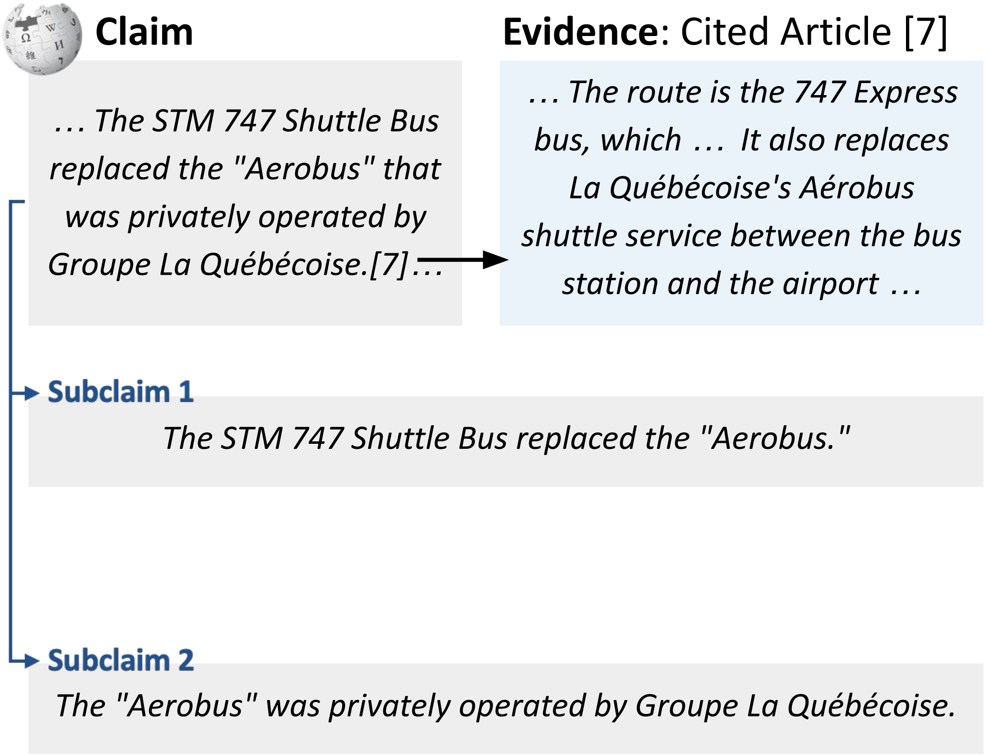
… The STM 747 Shuttle Bus replaced the "Aerobus" that was privately operated by Groupe La Québécoise.[7]…

Evidence: Cited Article [7]

… The route is the 747 Express bus, which … It also replaces La Québécoise's Aérobus shuttle service between the bus station and the airport …

- Claim (hypothesis)
  - Sentence in Wikipedia article
- Evidence (premise)
  - Websites cited for the claim

Fine-Grained Annotation on WiCE



Claim

… The STM 747 Shuttle Bus replaced the "Aerobus" that was privately operated by Groupe La Québécoise.[7] …

Evidence: Cited Article [7]

… The route is the 747 Express bus, which … It also replaces La Québécoise's Aérobus shuttle service between the bus station and the airport …

Subclaim 1

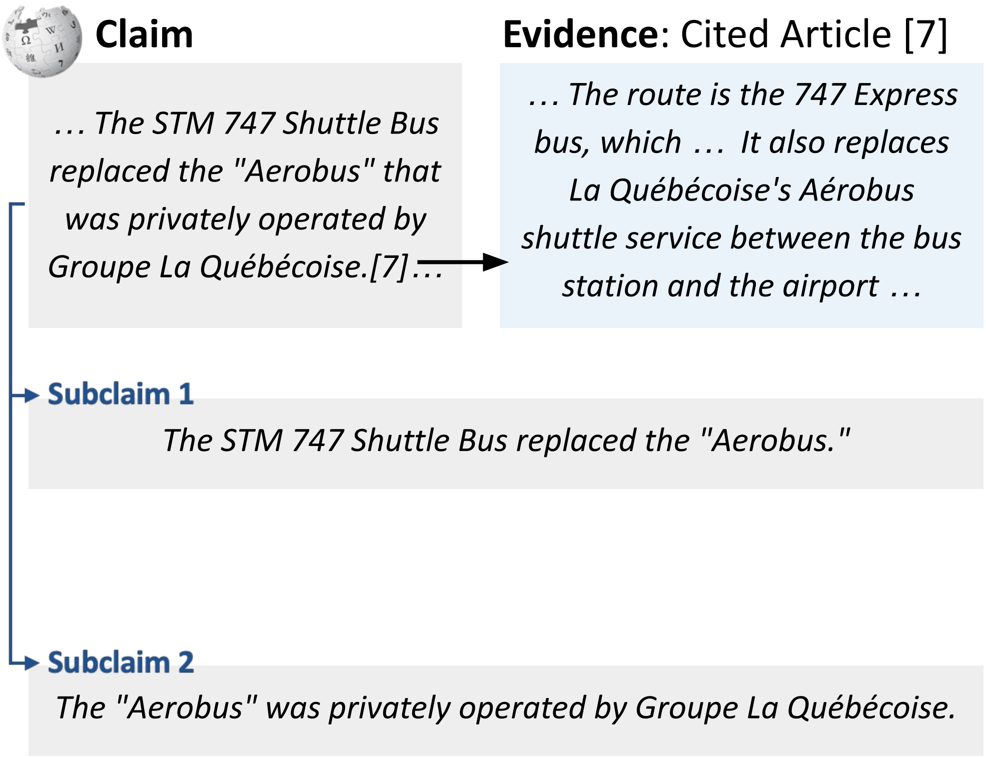
The STM 747 Shuttle Bus replaced the "Aerobus."

Subclaim 2

The "Aerobus" was privately operated by Groupe La Québécoise.

- Claim (hypothesis)
  - Sentence in Wikipedia article
- Evidence (premise)
  - Websites cited for the claim
- **Subclaim-level** annotation
  - We split the original claim into multiple independent sentences

Fine-Grained Annotation on WiCE



Claim

… The STM 747 Shuttle Bus replaced the "Aerobus" that was privately operated by Groupe La Québécoise.[7] …

Evidence: Cited Article [7]

… The route is the 747 Express bus, which … It also replaces La Québécoise's Aérobus shuttle service between the bus station and the airport …

Subclaim 1

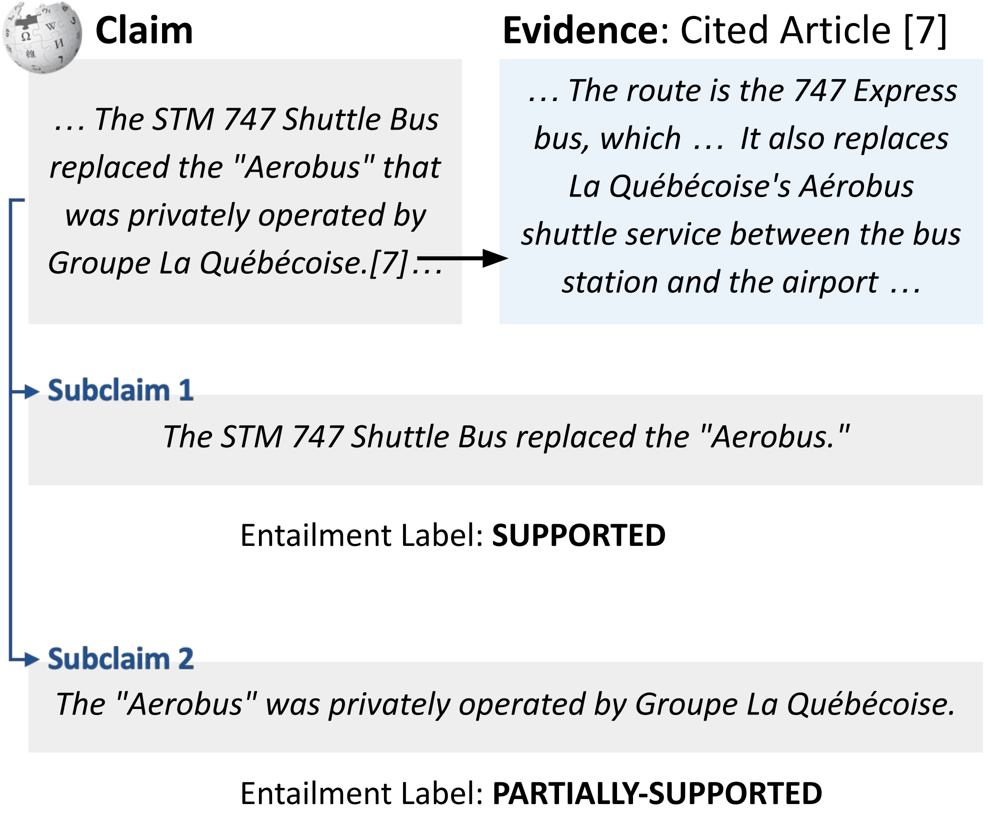
The STM 747 Shuttle Bus replaced the "Aerobus."

Subclaim 2

The "Aerobus" was privately operated by Groupe La Québécoise.

Annotation for each subclaim

Fine-Grained Annotation on WiCE



Claim

… The STM 747 Shuttle Bus replaced the "Aerobus" that was privately operated by Groupe La Québécoise.[7] …

Evidence: Cited Article [7]

… The route is the 747 Express bus, which … It also replaces La Québécoise's Aérobus shuttle service between the bus station and the airport …

Subclaim 1

The STM 747 Shuttle Bus replaced the "Aerobus."

Entailment Label: **SUPPORTED**

Subclaim 2

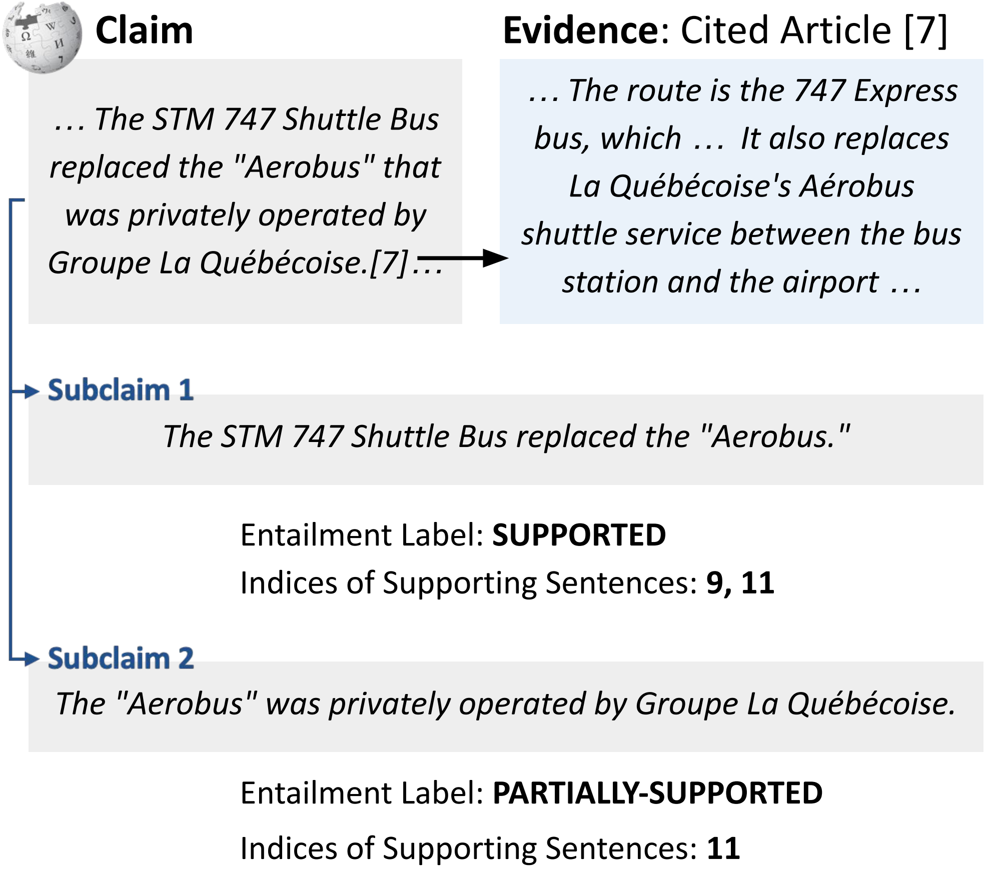
The "Aerobus" was privately operated by Groupe La Québécoise.

Entailment Label: **PARTIALLY-SUPPORTED**

Annotation for each subclaim

- 3-way entailment label
  - Supported, Partially-Supported, Not-Supported

Fine-Grained Annotation on WiCE



Claim

… The STM 747 Shuttle Bus replaced the "Aerobus" that was privately operated by Groupe La Québécoise.[7] …

Evidence: Cited Article [7]

… The route is the 747 Express bus, which … It also replaces La Québécoise's Aérobus shuttle service between the bus station and the airport …

Subclaim 1

The STM 747 Shuttle Bus replaced the "Aerobus."

Entailment Label: **SUPPORTED**

Indices of Supporting Sentences: **9, 11**

Subclaim 2

The "Aerobus" was privately operated by Groupe La Québécoise.

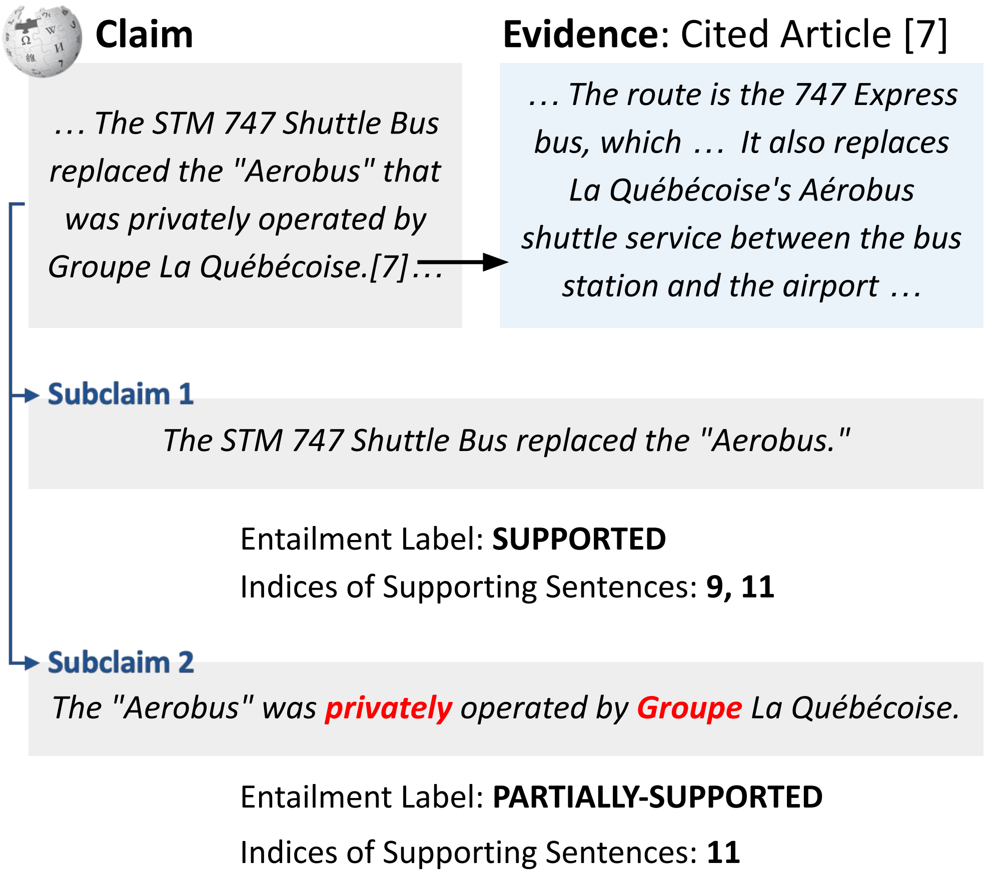
Entailment Label: **PARTIALLY-SUPPORTED**

Indices of Supporting Sentences: **11**

Annotation for each subclaim

- 3-way entailment label
  - Supported, Partially-Supported, Not-Supported
- Indices of supporting sentences

Fine-Grained Annotation on WiCE



Claim

… The STM 747 Shuttle Bus replaced the "Aerobus" that was privately operated by Groupe La Québécoise.[7] …

Evidence: Cited Article [7]

… The route is the 747 Express bus, which … It also replaces La Québécoise's Aérobus shuttle service between the bus station and the airport …

Subclaim 1

The STM 747 Shuttle Bus replaced the "Aerobus."

Entailment Label: **SUPPORTED**

Indices of Supporting Sentences: **9, 11**

Subclaim 2

*The "Aerobus" was **privately** operated by **Groupe** La Québécoise.*

Entailment Label: **PARTIALLY-SUPPORTED**

Indices of Supporting Sentences: **11**

Annotation for each subclaim

- 3-way entailment label
  - Supported, Partially-Supported, Not-Supported
- Indices of supporting sentences
- Non-supported tokens in claims (for Partially-Supported cases)

Claim-Split

The Société de transport de Montréal (STM) 747 Shuttle Bus replaced the "Aerobus" that was privately operated by Groupe La Québécoise.

Subclaims

The Société de transport de Montréal (STM) 747 Shuttle Bus replaced the "Aerobus."

The "Aerobus" was privately operated by Groupe La Québécoise.

Claim-Split

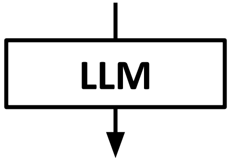
Prompt

Segment the following sentence into individual facts:

[in-context examples]

Original Sentence:

The Société de transport de Montréal (STM) 747 Shuttle Bus replaced the "Aerobus" that was privately operated by Groupe La Québécoise.



Subclaims

The Société de transport de Montréal (STM) 747 Shuttle Bus replaced the "Aerobus."

The "Aerobus" was privately operated by Groupe La Québécoise.

- Use **LLM with few-shot prompting** (GPT-3.5 in our experiments) to automatically split a claim into subclaims

Claim-Split

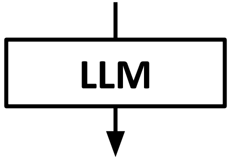
Prompt

Segment the following sentence into individual facts:

[in-context examples]

Original Sentence:

The Société de transport de Montréal (STM) 747 Shuttle Bus replaced the "Aerobus" that was privately operated by Groupe La Québécoise.



Subclaims

The Société de transport de Montréal (STM) 747 Shuttle Bus replaced the "Aerobus."

The "Aerobus" was privately operated by Groupe La Québécoise.

- Use **LLM with few-shot prompting** (GPT-3.5 in our experiments) to automatically split a claim into subclaims
- Subclaims are also complete sentences, making them easy to be used for annotation and downstream applications

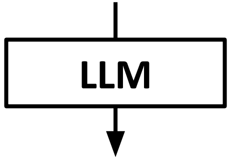
Claim-Split

Prompt

Segment the following sentence into individual facts:

[in-context examples]

Original Sentence:
The Société de transport de Montréal (STM) 747 Shuttle Bus replaced the "Aerobus" that was privately operated by Groupe La Québécoise.



Subclaims

The Société de transport de Montréal (STM) 747 Shuttle Bus replaced the "Aerobus."

The "Aerobus" was privately operated by Groupe La Québécoise.

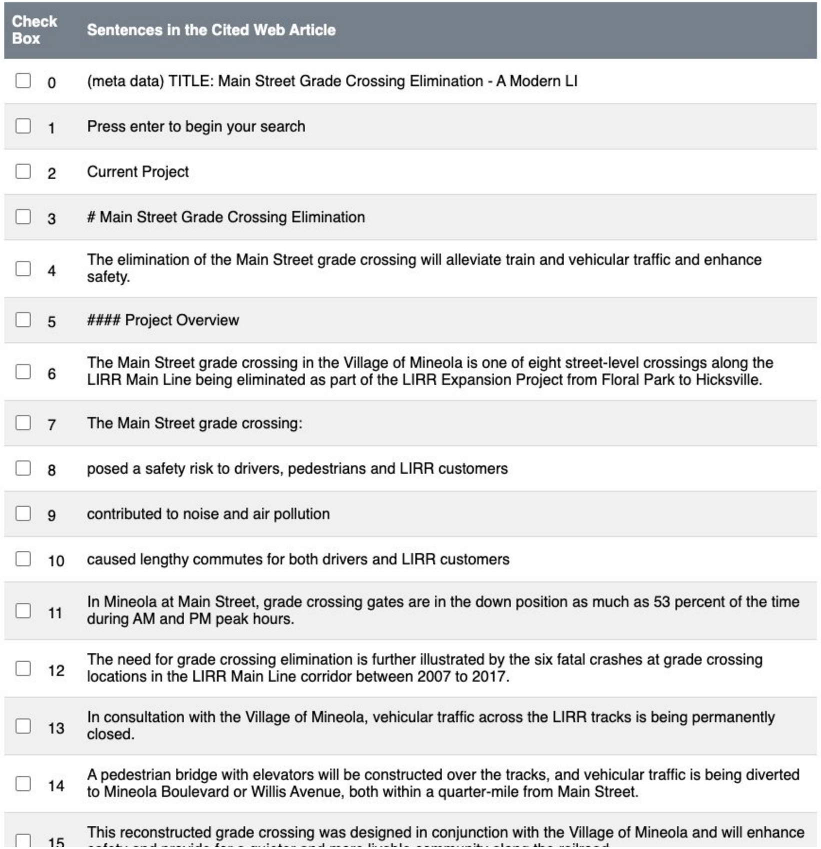
- Use **LLM with few-shot prompting** (GPT-3.5 in our experiments) to automatically split a claim into subclaims
- Subclaims are also complete sentences, making them easy to be used for annotation and downstream applications
- GPT-3.5 makes only 8.6% of errors (We manually fixed the mistake for WiCE)

Supporting Sentences

- Evidence in WiCE includes 120 sentences in average

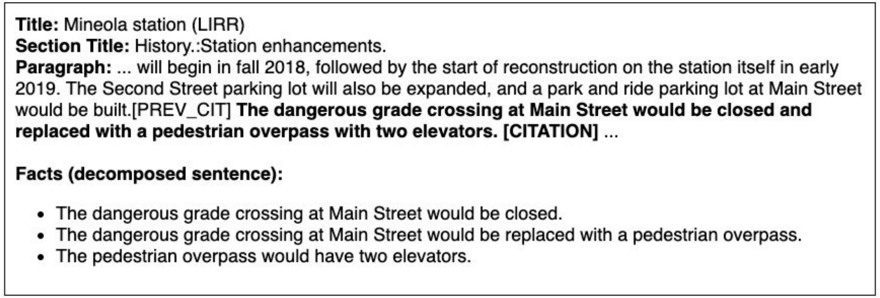
- Each claim has, on average, **3.1 supporting sentences**

Supporting Sentences



Check Box		Sentences in the Cited Web Article
☐	0	(meta data) TITLE: Main Street Grade Crossing Elimination - A Modern LI
☐	1	Press enter to begin your search
☐	2	Current Project
☐	3	# Main Street Grade Crossing Elimination
☐	4	The elimination of the Main Street grade crossing will alleviate train and vehicular traffic and enhance safety.
☐	5	#### Project Overview
☐	6	The Main Street grade crossing in the Village of Mineola is one of eight street-level crossings along the LIRR Main Line being eliminated as part of the LIRR Expansion Project from Floral Park to Hicksville.
☐	7	The Main Street grade crossing:
☐	8	posed a safety risk to drivers, pedestrians and LIRR customers
☐	9	contributed to noise and air pollution
☐	10	caused lengthy commutes for both drivers and LIRR customers
☐	11	In Mineola at Main Street, grade crossing gates are in the down position as much as 53 percent of the time during AM and PM peak hours.
☐	12	The need for grade crossing elimination is further illustrated by the six fatal crashes at grade crossing locations in the LIRR Main Line corridor between 2007 to 2017.
☐	13	In consultation with the Village of Mineola, vehicular traffic across the LIRR tracks is being permanently closed.
☐	14	A pedestrian bridge with elevators will be constructed over the tracks, and vehicular traffic is being diverted to Mineola Boulevard or Willis Avenue, both within a quarter-mile from Main Street.
☐	15	This reconstructed grade crossing was designed in conjunction with the Village of Mineola and will enhance safety and provide for a quieter and more livable community along the railroad.

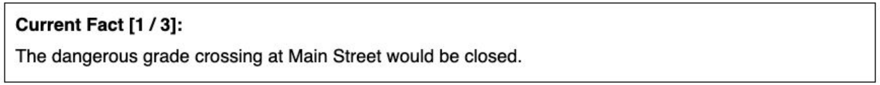
- Evidence in WiCE includes 120 sentences in average
- Each claim has, on average, **3.1 supporting sentences**



Title: Mineola station (LIRR)
Section Title: History.:Station enhancements.
Paragraph: ... will begin in fall 2018, followed by the start of reconstruction on the station itself in early 2019. The Second Street parking lot will also be expanded, and a park and ride parking lot at Main Street would be built.[PREV_CIT] **The dangerous grade crossing at Main Street would be closed and replaced with a pedestrian overpass with two elevators. [CITATION]** ...

Facts (decomposed sentence):

- The dangerous grade crossing at Main Street would be closed.
- The dangerous grade crossing at Main Street would be replaced with a pedestrian overpass.
- The pedestrian overpass would have two elevators.



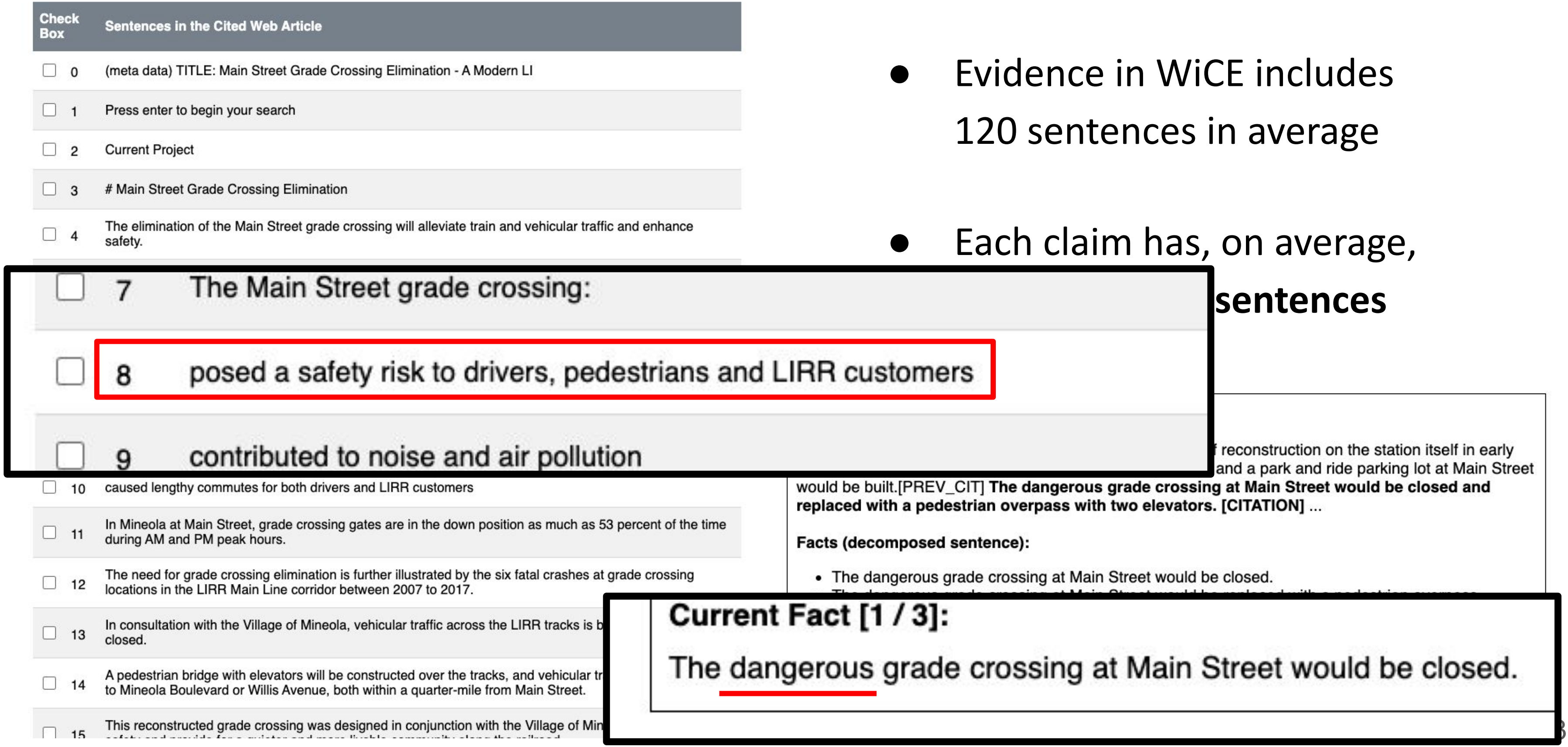
Current Fact [1 / 3]:

The dangerous grade crossing at Main Street would be closed.

Supporting Sentences

- Evidence in WiCE includes 120 sentences in average

- Each claim has, on average, **sentences**



Check Box		Sentences in the Cited Web Article
□	0	(meta data) TITLE: Main Street Grade Crossing Elimination - A Modern LI
□	1	Press enter to begin your search
□	2	Current Project
□	3	# Main Street Grade Crossing Elimination
□	4	The elimination of the Main Street grade crossing will alleviate train and vehicular traffic and enhance safety.
□	7	The Main Street grade crossing:
□	8	posed a safety risk to drivers, pedestrians and LIRR customers
□	9	contributed to noise and air pollution
□	10	caused lengthy commutes for both drivers and LIRR customers
□	11	In Mineola at Main Street, grade crossing gates are in the down position as much as 53 percent of the time during AM and PM peak hours.
□	12	The need for grade crossing elimination is further illustrated by the six fatal crashes at grade crossing locations in the LIRR Main Line corridor between 2007 to 2017.
□	13	In consultation with the Village of Mineola, vehicular traffic across the LIRR tracks is b closed.
□	14	A pedestrian bridge with elevators will be constructed over the tracks, and vehicular tr to Mineola Boulevard or Willis Avenue, both within a quarter-mile from Main Street.
□	15	This reconstructed grade crossing was designed in conjunction with the Village of Min safety and provide for a quieter and more livable community along the railroad

f reconstruction on the station itself in early and a park and ride parking lot at Main Street would be built.[PREV_CIT] **The dangerous grade crossing at Main Street would be closed and replaced with a pedestrian overpass with two elevators. [CITATION]** ...

Facts (decomposed sentence):

- The dangerous grade crossing at Main Street would be closed.

Current Fact [1 / 3]:

The dangerous grade crossing at Main Street would be closed.

Wikipedia Includes Complex Verification Problems

Claim-Evidence pair that can be verified by **Direct Extraction** (in VitaminC)

Panjaa received negative reviews .

***Panjaa received** largely **negative reviews** from critics.*

Claim-Evidence pair that **Requires Reasoning** for verification (in WiCE)

*She was **the only member** of the Washington Health Benefit Exchange who voted against increasing the salary of the health exchange's CEO.*

*... All of the board members in attendance voted in favor of the pay raise **except one**. Teresa Mosqeda, legislative and policy director for the Washington State Labor Council, voted against the motion. ...*

Wikipedia Includes Complex Verification Problems

Verification Problems in Fact-Verification Datasets (%)

More Complex

	FEVER	VitaminC	WiCE (claim)
Extraction	10	20	4
Easy Paraphrase	34	24	16
Require Reasoning	24	8	**68**

- Many claims in existing datasets (FEVER and VitaminC) can be verified by simply extracting corresponding information from evidence

Wikipedia Includes Complex Verification Problems

Verification Problems in Fact-Verification Datasets (%)

More Complex

	FEVER	VitaminC	WiCE (claim)
Extraction	10	20	4
Easy Paraphrase	34	24	16
Require Reasoning	24	8	**68**

- Many claims in existing datasets (FEVER and VitaminC) can be verified by simply extracting corresponding information from evidence
- Many claims in WiCE require **reasoning** to verify them, in addition to selecting correct information in evidence

Overview

Construction of WiCE

- Fine-grained annotation on sub-claims generated by **Claim-Split**
- Indices of supporting sentences

Experiments on WiCE

- How well do existing NLI models perform on WiCE?
- How well does GPT-4 perform?

How to Evaluate Long Evidence?

Evidence Article

The Main Street grade crossing:

posed a safety risk to drivers, pedestrians and LIRR customers

contributed to noise and air pollution

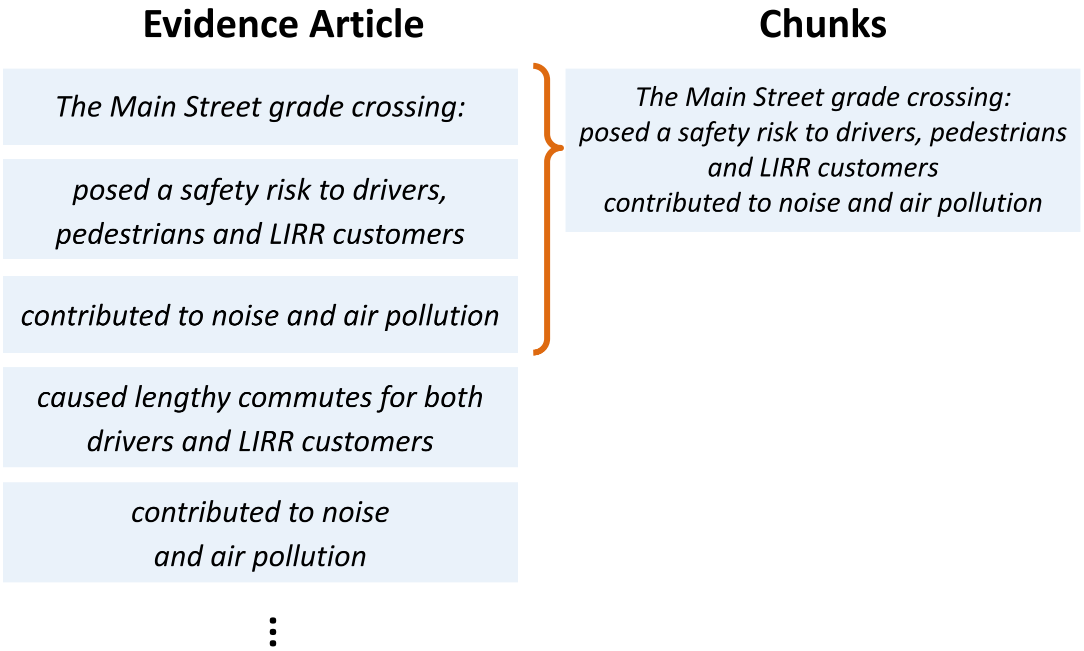
caused lengthy commutes for both drivers and LIRR customers

contributed to noise and air pollution

⋮

- Evidence in WiCE contains about **1600 tokens** in average
- Many NLI models cannot accommodate long evidence

How to Evaluate Long Evidence?



Evidence Article

The Main Street grade crossing:

posed a safety risk to drivers, pedestrians and LIRR customers

contributed to noise and air pollution

caused lengthy commutes for both drivers and LIRR customers

contributed to noise and air pollution

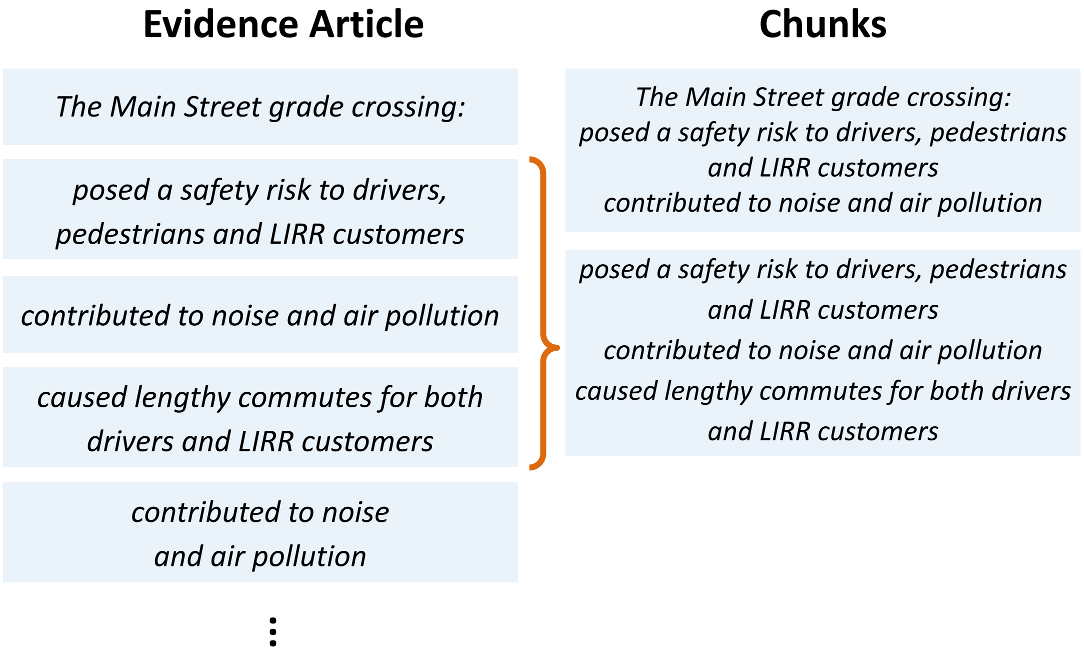
⋮

Chunks

The Main Street grade crossing:
posed a safety risk to drivers, pedestrians and LIRR customers
contributed to noise and air pollution

- Evidence in WiCE contains about **1600 tokens** in average
- Many NLI models cannot accommodate long evidence

How to Evaluate Long Evidence?



Evidence Article

The Main Street grade crossing:

posed a safety risk to drivers, pedestrians and LIRR customers

contributed to noise and air pollution

caused lengthy commutes for both drivers and LIRR customers

contributed to noise and air pollution

⋮

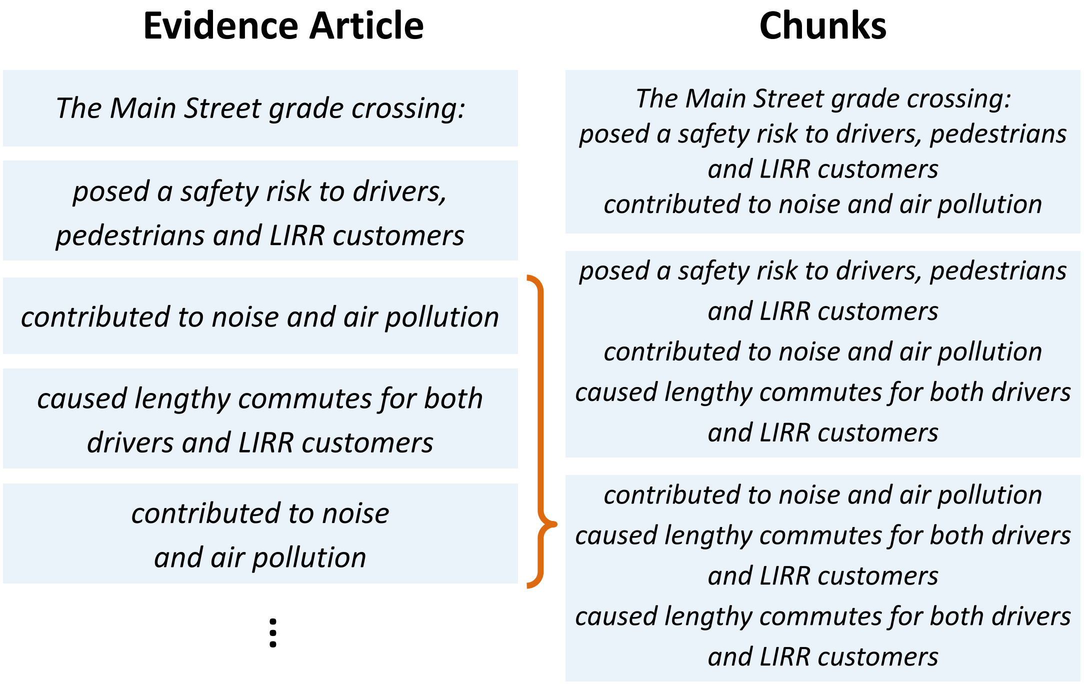
Chunks

The Main Street grade crossing:
posed a safety risk to drivers, pedestrians and LIRR customers
contributed to noise and air pollution

posed a safety risk to drivers, pedestrians and LIRR customers
contributed to noise and air pollution
caused lengthy commutes for both drivers and LIRR customers

- Evidence in WiCE contains about **1600 tokens** in average
- Many NLI models cannot accommodate long evidence

How to Evaluate Long Evidence?



Evidence Article

The Main Street grade crossing:

posed a safety risk to drivers, pedestrians and LIRR customers

contributed to noise and air pollution

caused lengthy commutes for both drivers and LIRR customers

contributed to noise and air pollution

⋮

Chunks

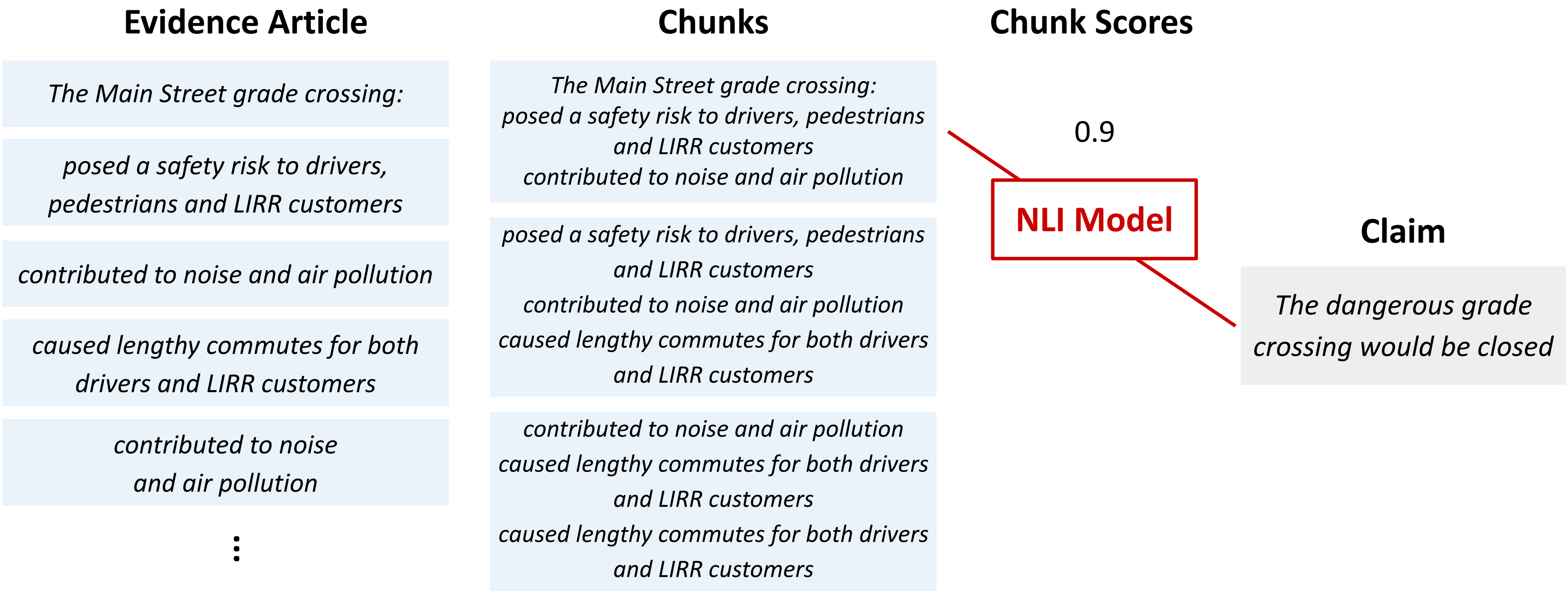
The Main Street grade crossing:
posed a safety risk to drivers, pedestrians and LIRR customers
contributed to noise and air pollution

posed a safety risk to drivers, pedestrians and LIRR customers
contributed to noise and air pollution
caused lengthy commutes for both drivers and LIRR customers

contributed to noise and air pollution
caused lengthy commutes for both drivers and LIRR customers
caused lengthy commutes for both drivers and LIRR customers

- Evidence in WiCE contains about **1600 tokens** in average
- Many NLI models cannot accommodate long evidence

How to Evaluate Long Evidence?



Evidence Article

The Main Street grade crossing:

posed a safety risk to drivers, pedestrians and LIRR customers

contributed to noise and air pollution

caused lengthy commutes for both drivers and LIRR customers

contributed to noise and air pollution

⋮

Chunks

The Main Street grade crossing:
posed a safety risk to drivers, pedestrians and LIRR customers
contributed to noise and air pollution

posed a safety risk to drivers, pedestrians and LIRR customers
contributed to noise and air pollution
caused lengthy commutes for both drivers and LIRR customers

contributed to noise and air pollution
caused lengthy commutes for both drivers and LIRR customers
caused lengthy commutes for both drivers and LIRR customers

Chunk Scores

0.9

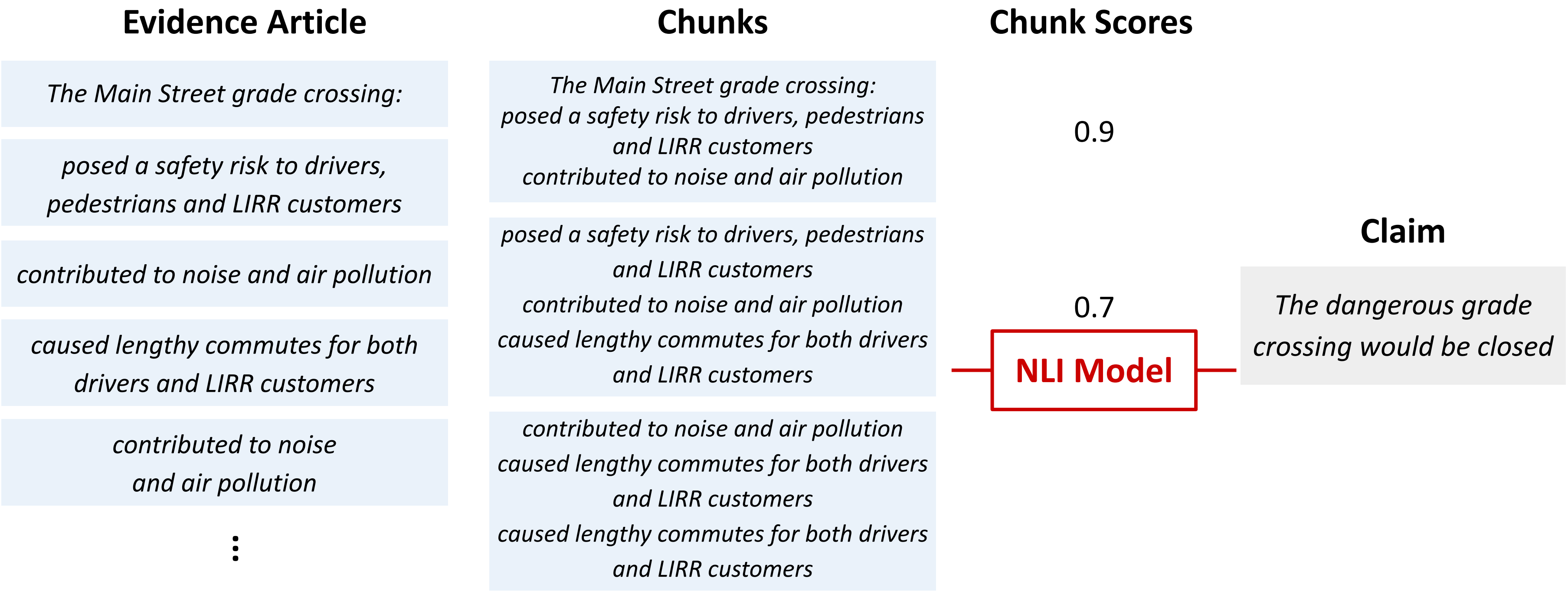
NLI Model

Claim

The dangerous grade crossing would be closed

- Evidence in WiCE contains about **1600 tokens** in average
- Many NLI models cannot accommodate long evidence

How to Evaluate Long Evidence?



Evidence Article

The Main Street grade crossing:

posed a safety risk to drivers, pedestrians and LIRR customers

contributed to noise and air pollution

caused lengthy commutes for both drivers and LIRR customers

contributed to noise and air pollution

⋮

Chunks

The Main Street grade crossing: posed a safety risk to drivers, pedestrians and LIRR customers contributed to noise and air pollution

posed a safety risk to drivers, pedestrians and LIRR customers contributed to noise and air pollution caused lengthy commutes for both drivers and LIRR customers

contributed to noise and air pollution caused lengthy commutes for both drivers and LIRR customers caused lengthy commutes for both drivers and LIRR customers

Chunk Scores

0.9

0.7

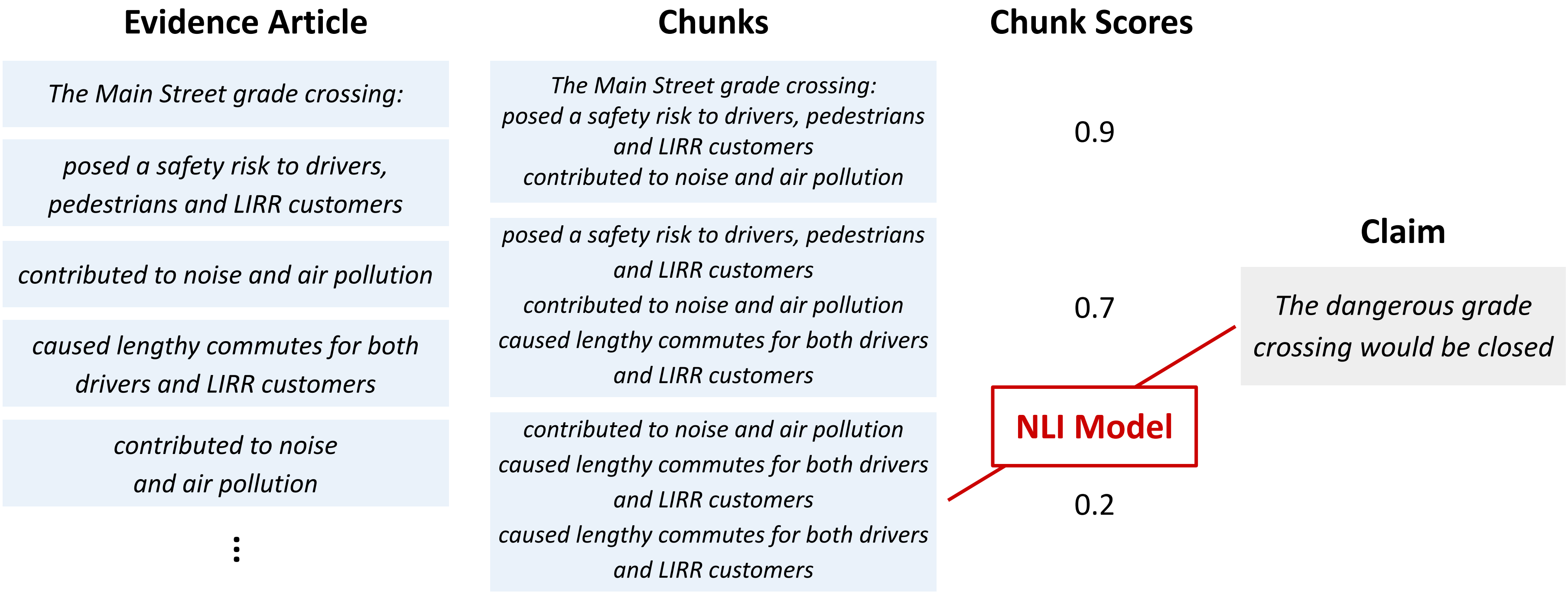
NLI Model

Claim

The dangerous grade crossing would be closed

- Evidence in WiCE contains about **1600 tokens** in average
- Many NLI models cannot accommodate long evidence

How to Evaluate Long Evidence?



Evidence Article

The Main Street grade crossing:

posed a safety risk to drivers, pedestrians and LIRR customers

contributed to noise and air pollution

caused lengthy commutes for both drivers and LIRR customers

contributed to noise and air pollution

⋮

Chunks

The Main Street grade crossing: posed a safety risk to drivers, pedestrians and LIRR customers contributed to noise and air pollution

posed a safety risk to drivers, pedestrians and LIRR customers contributed to noise and air pollution caused lengthy commutes for both drivers and LIRR customers

contributed to noise and air pollution caused lengthy commutes for both drivers and LIRR customers caused lengthy commutes for both drivers and LIRR customers

Chunk Scores

0.9

0.7

0.2

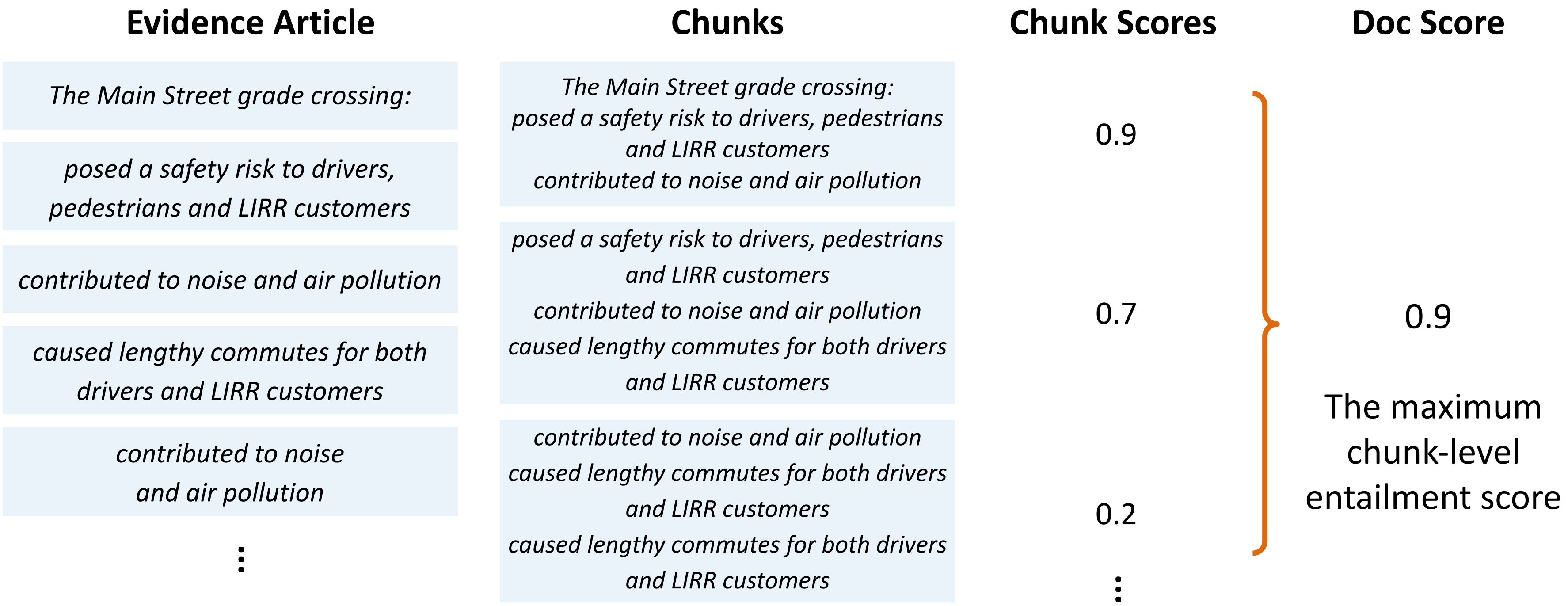
Claim

The dangerous grade crossing would be closed

NLI Model

- Evidence in WiCE contains about **1600 tokens** in average
- Many NLI models cannot accommodate long evidence

How to Evaluate Long Evidence?



Evidence Article

The Main Street grade crossing:

posed a safety risk to drivers, pedestrians and LIRR customers

contributed to noise and air pollution

caused lengthy commutes for both drivers and LIRR customers

contributed to noise and air pollution

⋮

Chunks

The Main Street grade crossing:
posed a safety risk to drivers, pedestrians and LIRR customers
contributed to noise and air pollution

posed a safety risk to drivers, pedestrians and LIRR customers
contributed to noise and air pollution
caused lengthy commutes for both drivers and LIRR customers

contributed to noise and air pollution
caused lengthy commutes for both drivers and LIRR customers
caused lengthy commutes for both drivers and LIRR customers

Chunk Scores

0.9

0.7

0.2

⋮

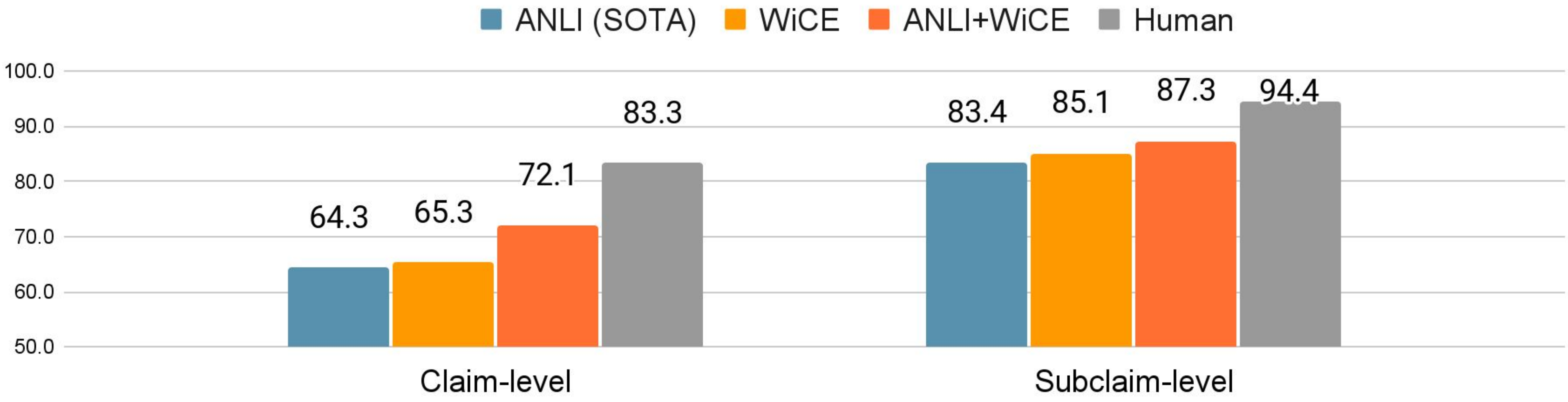
Doc Score

0.9

The maximum chunk-level entailment score

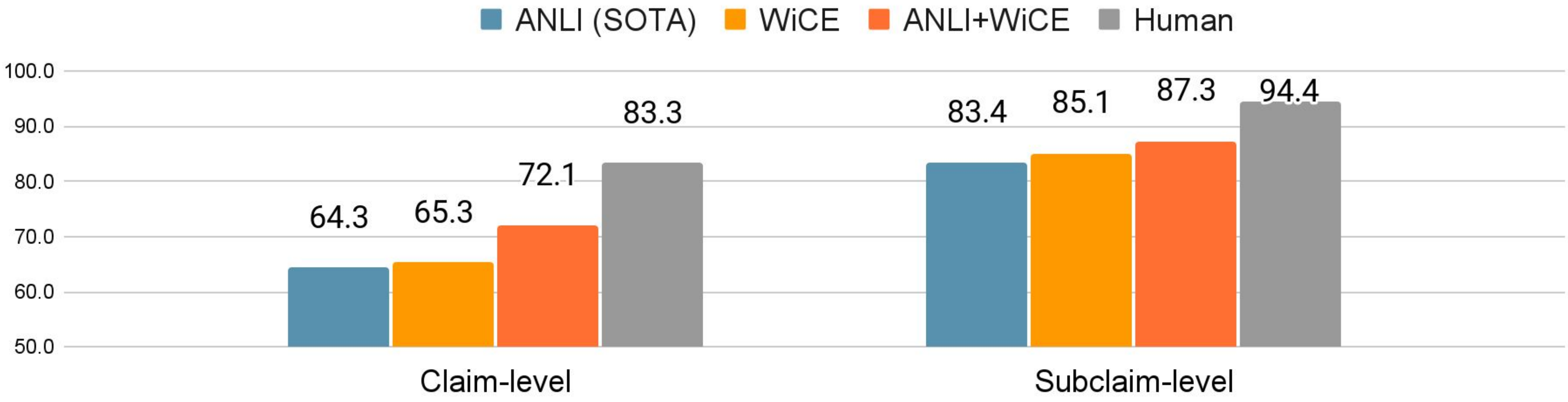
- Evidence in WiCE contains about **1600 tokens** in average
- Many NLI models cannot accommodate long evidence

Results: Maximum Chunk-level Scores of T5-3B



- We evaluate binary classification
  (Supported vs. Partially- or Not-Supported)

Results: Maximum Chunk-level Scores of T5-3B



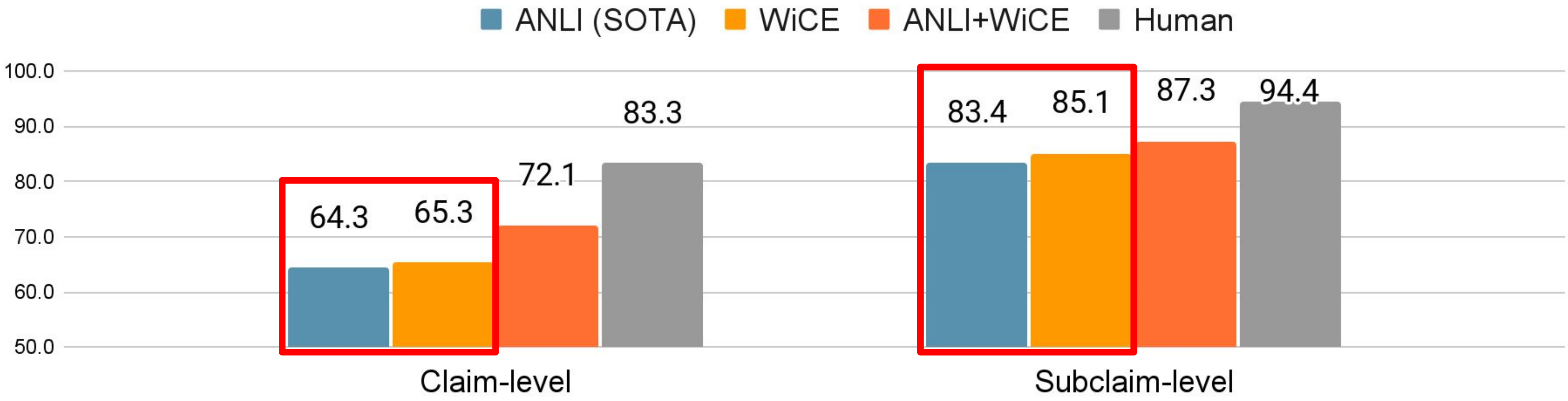
- We evaluate binary classification

  (Supported vs. Partially- or Not-Supported)

- Subclaim-level is much easier

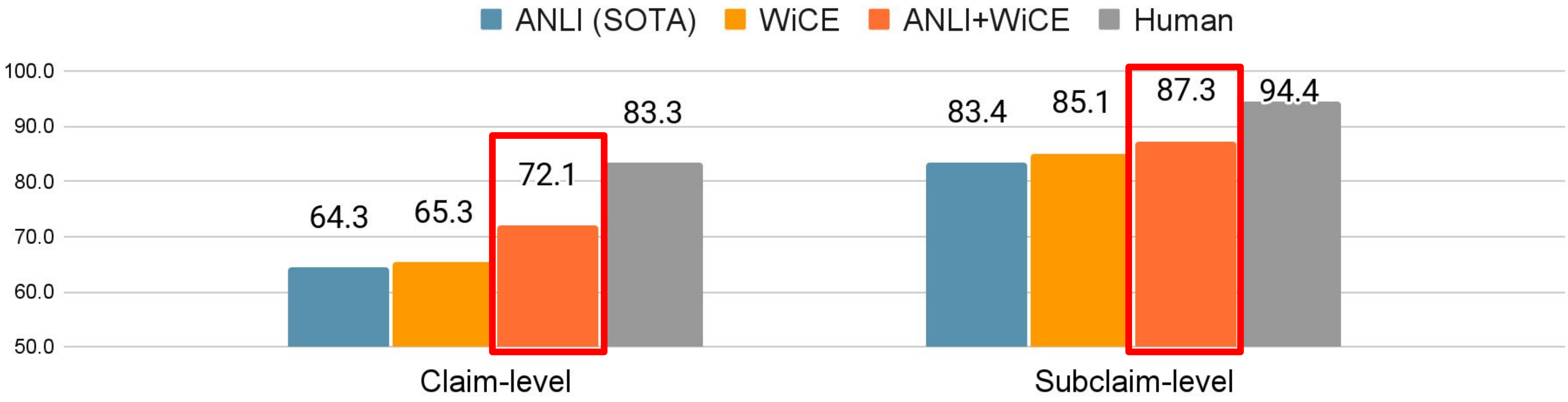
  (Claim-Split is useful for improving inference)

Results: Maximum Chunk-level Scores of T5-3B



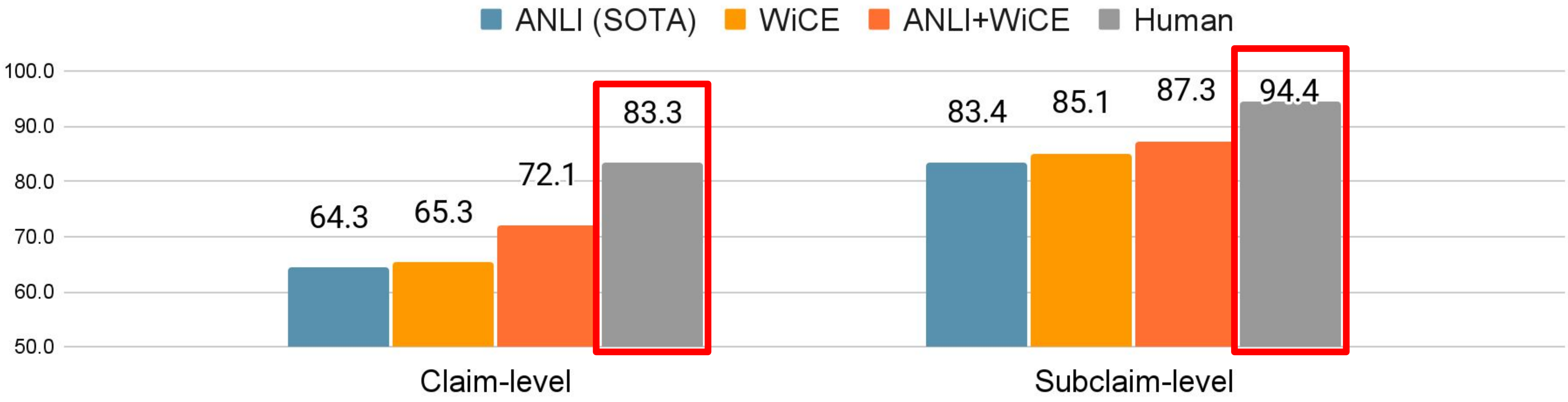
- We evaluate binary classification
  (Supported vs. Partially- or Not-Supported)
- Subclaim-level is much easier
  (Claim-Split is useful for improving inference)
- Fine-tuning on WiCE achieves the best single-dataset performance

Results: Maximum Chunk-level Scores of T5-3B



- We evaluate binary classification (Supported vs. Partially- or Not-Supported)
- Subclaim-level is much easier (Claim-Split is useful for improving inference)
- Fine-tuning on WiCE achieves the best single-dataset performance
- Using multiple datasets is beneficial

Results: Maximum Chunk-level Scores of T5-3B



- We evaluate binary classification

  (Supported vs. Partially- or Not-Supported)

- Subclaim-level is much easier

  (Claim-Split is useful for improving inference)

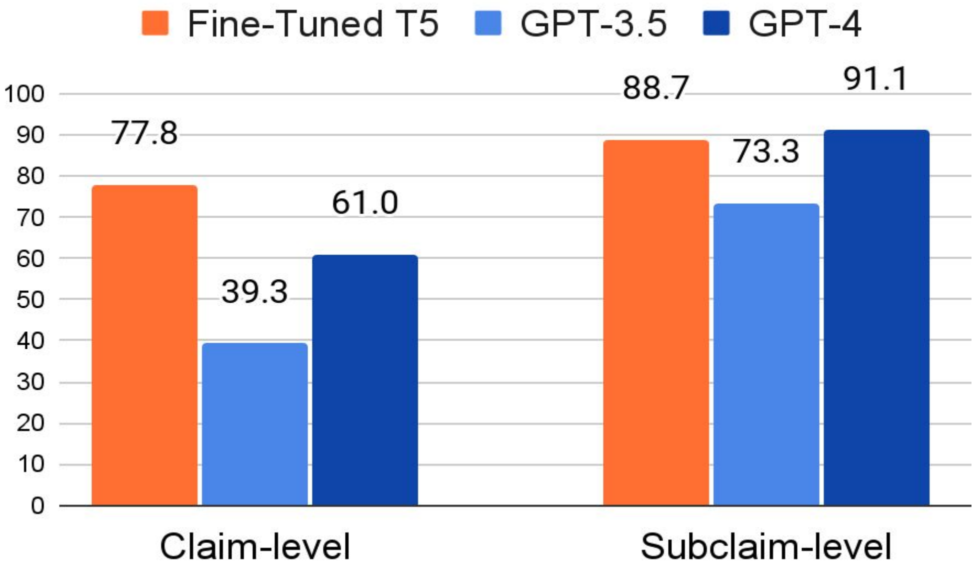
- Fine-tuning on WiCE achieves the best single-dataset performance

- Using multiple datasets is beneficial

- Human Performance is much better

Results: Few-Shot GPT-4

Performance (F1) with **Oracle Retrieval**



- Few-shot GPT-4 works well at subclaim-level but **not very well at claim-level**
- Much smaller fine-tuned model (T5-3B) is better than GPT-4 at claim-level

WiCE is challenging for GPT-4 even with oracle retrieval (perfect evidence retrieval)

Takeaways

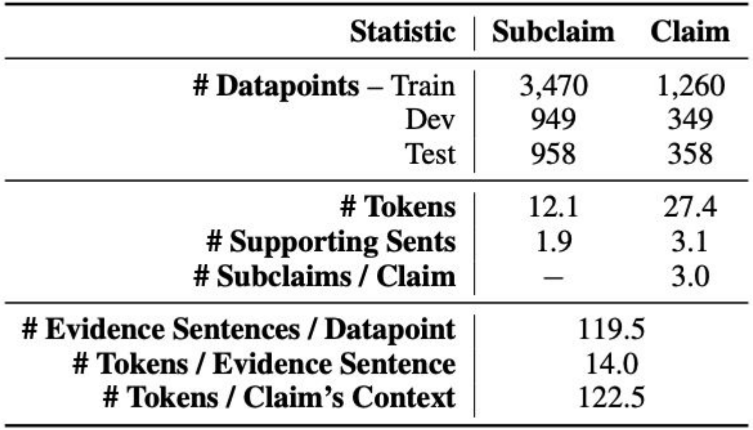
- WiCE is suitable for training and analyzing realistic fact-verification systems
  - Long premises
  - Natural negative cases
  - Fine-grained annotation
- WiCE is challenging for fine-tuned T5 and few-shot GPT-4

Thank you!

Please reach out us for any questions

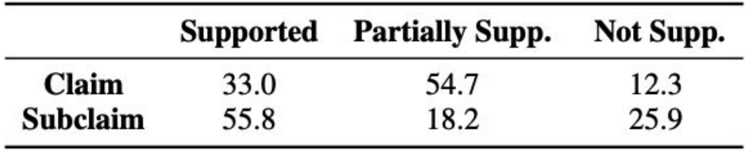
ryokamoi@utexas.edu

Dataset Statistics



Statistic	Subclaim	Claim
# Datapoints – Train	3,470	1,260
Dev	949	349
Test	958	358
# Tokens	12.1	27.4
# Supporting Sents	1.9	3.1
# Subclaims / Claim	–	3.0
# Evidence Sentences / Datapoint	119.5	
# Tokens / Evidence Sentence	14.0	
# Tokens / Claim's Context	122.5	

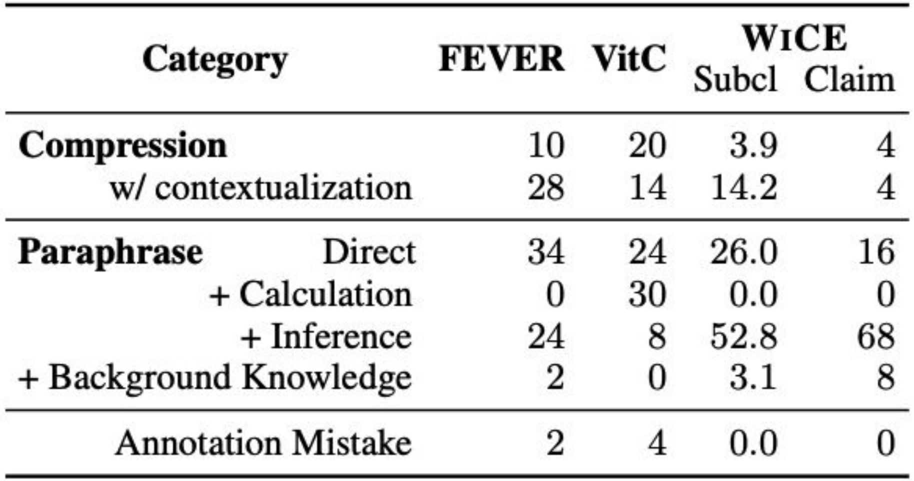
Table 1: Statistics of the WICE dataset.



	Supported	Partially Supp.	Not Supp.
Claim	33.0	54.7	12.3
Subclaim	55.8	18.2	25.9

Table 2: Entailment label distribution (%) of claims and subclaims in the development set of WICE.

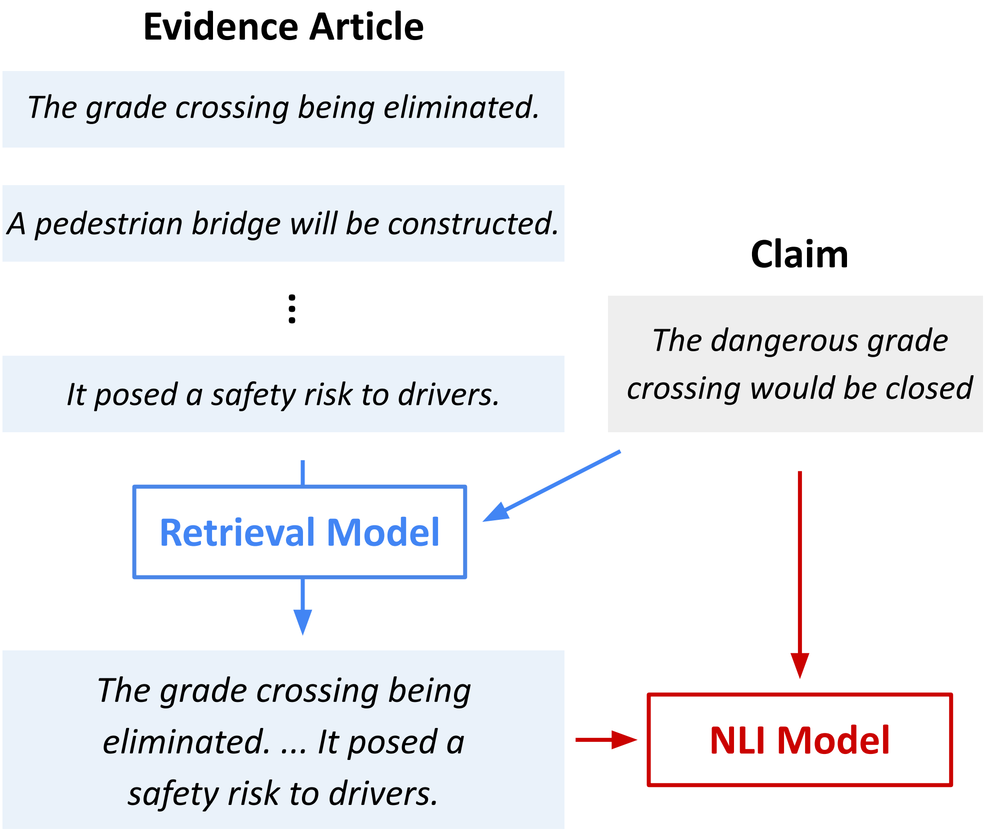
Verification Categories



Category		FEVER	VitC	WiCE	
				Subcl	Claim
Compression		10	20	3.9	4
	w/ contextualization	28	14	14.2	4
Paraphrase	Direct	34	24	26.0	16
	+ Calculation	0	30	0.0	0
	+ Inference	24	8	52.8	68
	+ Background Knowledge	2	0	3.1	8
	Annotation Mistake	2	4	0.0	0

Table 3: Distribution (%) of verification problems estimated from annotation for 50 claims in each dataset (127 subclaims in WiCE). We evaluated claims labeled as entailed in FEVER and VitaminC, and claims labeled as supported or partially supported in WiCE.

Retrieve-then-Predict: Claim Verification by NLI



Evidence Article

The grade crossing being eliminated.

A pedestrian bridge will be constructed.

⋮

It posed a safety risk to drivers.

Claim

The dangerous grade crossing would be closed

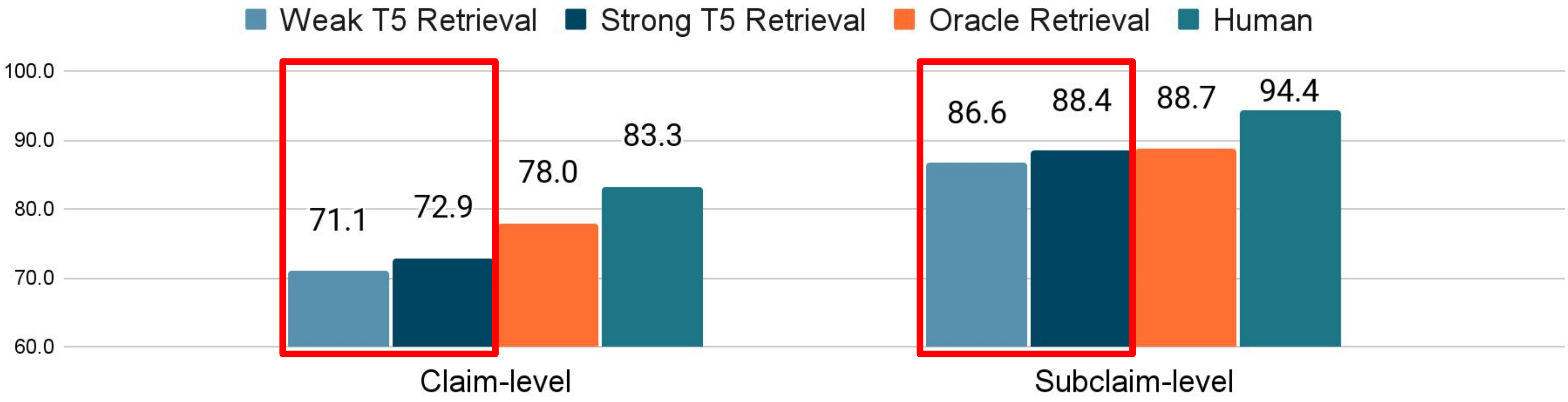
The grade crossing being eliminated. ... It posed a safety risk to drivers.

- Evidence in WiCE contains about **1600 tokens** in average
- Many NLI models cannot accommodate long evidence
- We **retrieve useful evidence sentences** and feed the retrieved sentences to NLI models

(Nie et al., AAAI 2019)

Results: Retrieve-then-Predict

Entailment Classifier is **T5-3B** fine-tuned on ANLI + WiCE

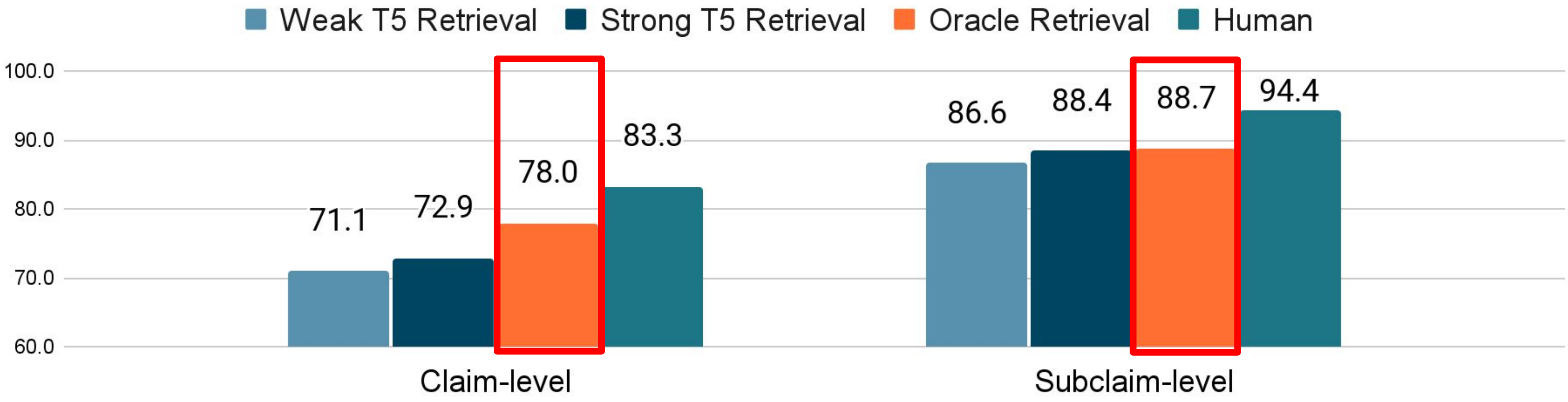


T5-based Retrievers

- Better retrieval model improves the entailment classification performance

Results: Retrieve-then-Predict

Entailment Classifier is **T5-3B** fine-tuned on ANLI + WiCE



T5-based Retrievers

- Better retrieval model improves the entailment classification performance

Oracle Retrieval

- Provide gold supporting sentences
- The improvement suggests that the retrieval is important