

# Post-Hoc Answer Attribution for Grounded and Trustworthy Long Document Comprehension: Task, Insights, and Challenges

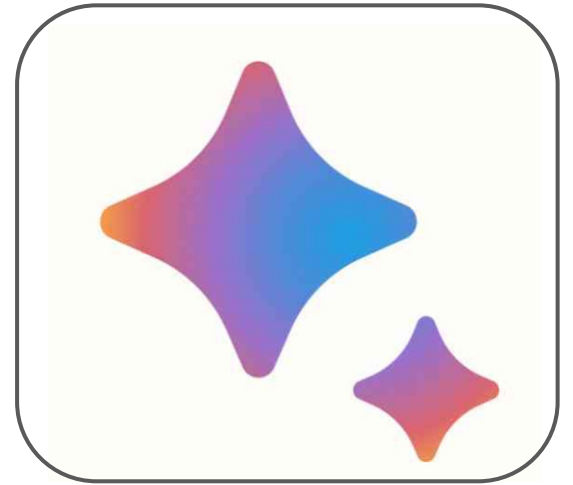
---

**Abhilasha Sancheti**<sup>1</sup>, Koustava Goswami<sup>2</sup>, Balaji Vasanth Srinivasan<sup>2</sup>

<sup>1</sup>University of Maryland, College Park

<sup>2</sup>Adobe Research

# Day-to-day use of digital assistants for seeking information



# Problem of hallucination and risk of spreading misinformation



**You**

Who was the sole survivor of Titanic?



**ChatGPT**

The sole survivor of the Titanic disaster was Charles Lightoller. He was the most senior officer to survive and managed to find a place on an overturned lifeboat (Collapsible B). Lightoller's survival was remarkable, considering the immense tragedy that claimed over 1,500 lives.

# Problem of hallucination and risk of spreading misinformation



You

Can you give me a bibliography of sources that prove that he was the sole survivor of Titanic?

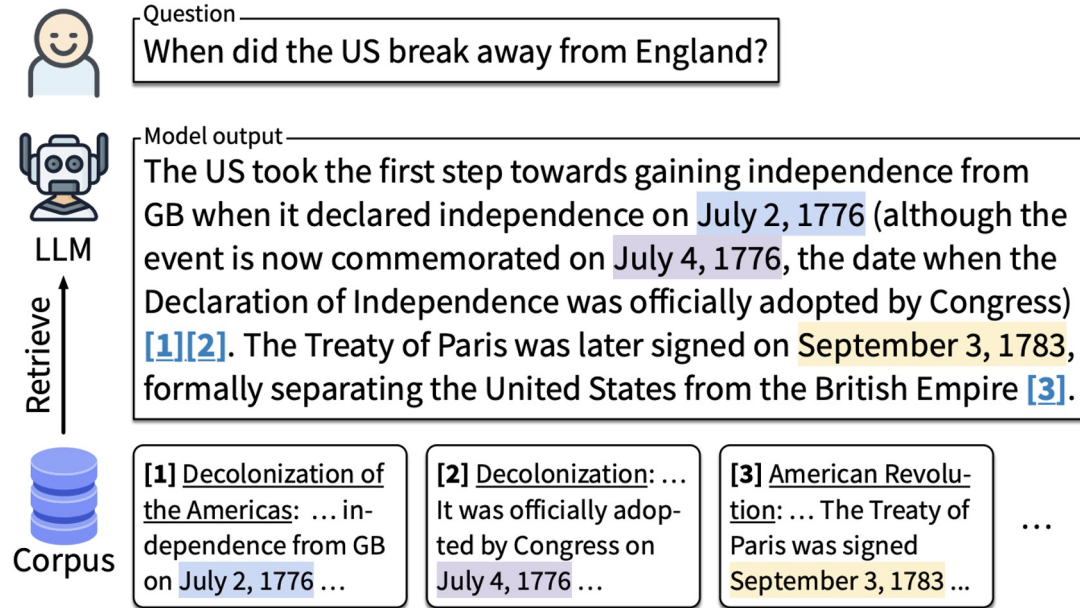


ChatGPT

Certainly! Charles Lightoller was not the sole survivor of the Titanic disaster. There were actually several survivors, both passengers and crew. Here are some reputable sources where you can verify information about survivors of the Titanic:

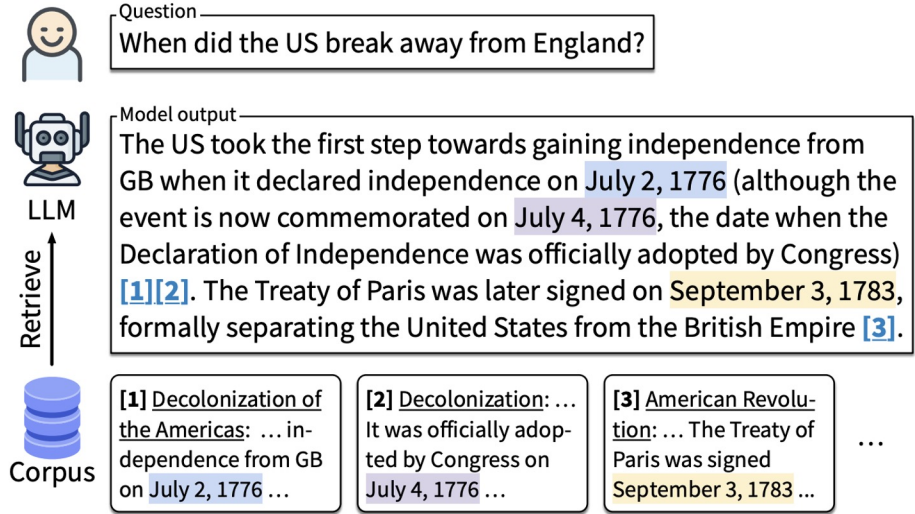
1. **Encyclopedia Titanica:** This website is dedicated to providing comprehensive information about the Titanic, including lists of survivors and victims. You can access it at: [Encyclopedia Titanica](#).

# Solution: Attributing Answers to their Source



Gao et al., 2023b

# Solution: Attributing Answers to their Source



Gao et al., 2023b



Trustworthy

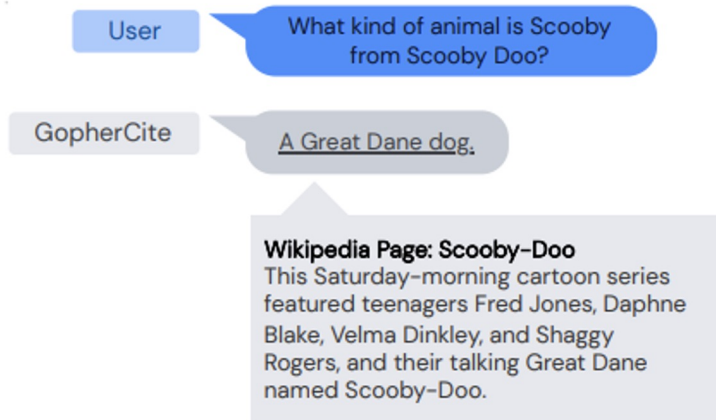


Accountable

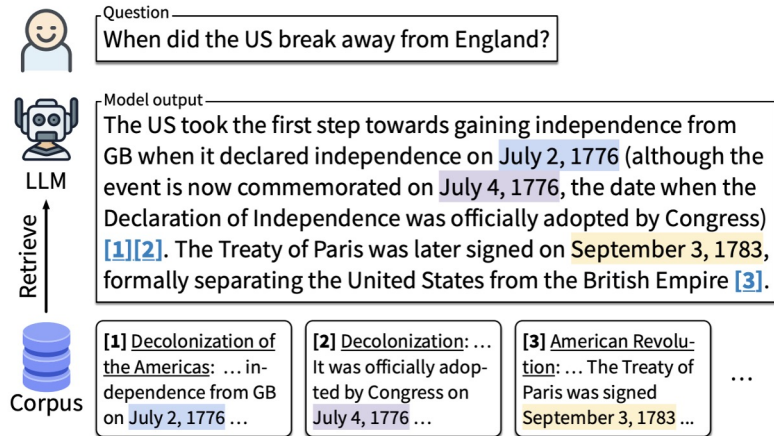


Reliable and Verifiable

# Answer Attribution so far ...



Menick et al., 2022



Gao et al., 2023b

## But ...

- Several scenarios when **user** wants to **ask queries grounded in a long document** (such as, financial, scientific, or medical reports)
- Existing datasets have **short and extractive answers** (Dasigi et al., 2021)
- Obtaining **attribution annotations is cognitively demanding, labor-intensive, and expensive**



# Task: Post-Hoc Answer Attribution for LDC

---

## Input

**Question:** When does the next assassins creed come out?

**Document:** [1] Ubisoft has announced that its next Assassin's Creed game will be revealed in September 2022.

[2] Ubisoft shared the first trailer for the game on Saturday.

[3] Assassin's Creed Mirage, the next entry in Ubisoft's long-running action-adventure series, will arrive in 2023.

[4] The publisher announced the release date today during its Ubisoft Forward event. ...

**Answer:** The next Assassin's Creed game, Assassin's Creed Mirage, will arrive in 2023 according to Ubisoft's announcement during its Ubisoft Forward event. It will be released for Xbox ... The game will be revealed in September 2022.

---

## Output

**Attributed answer:** The next Assassin's Creed game, Assassin's Creed Mirage, ... Ubisoft's announcement during its Ubisoft Forward event [3,4] ... The game will be revealed in September 2022 [1].

---

← Unattributed  
answer

← Answer sentences attributed to  
the sentences in the source  
document

# Dataset Challenge: Reformulating Existing Datasets



What are the latest discoveries from the James Webb Space Telescope?



The James Webb Space Telescope is designed to peer into the dusty clouds of gas where stars and planetary systems are born. Webb has captured the first direct image of an exoplanet, and the Pillars of Creation in the Eagle Nebula[1][2]. Additionally, the telescope will be used to study the next interstellar interloper[3].

(\*Some generated statements may *not be fully supported* by citations, while others are fully supported.)

## Cited Webpages

- [1]: nasa.gov (✗ citation does not support its associated statement)  
[NASA's Webb Confirms Its First Exoplanet](#)  
... Researchers confirmed an exoplanet, a planet that orbits another star, using NASA's James Webb Space Telescope for the first time. ...
- [2]: cnn.com (⚠ citation partially supports its associated statement)  
[Pillars of Creation: James Webb Space Telescope ...](#)  
... The Pillars of Creation, in the Eagle Nebula, is a star-forming region captured in a new image (right) by the James Webb Space Telescope that reveals more detail than a 2014 image (left) by Hubble ...
- [3]: nasa.gov (✔ citation fully supports its associated statement)  
[Studying the Next Interstellar Interloper with Webb](#)  
...Scientists have had only limited ability to study these objects once discovered, but all of that is about to change with NASA's James Webb Space Telescope...The team will use Webb's spectroscopic capabilities in both the near-infrared and mid-infrared bands to study two different aspects of the interstellar object.

**Task:** Verify the citations in the LLM generated answers

**Original Input (Open-ended):** Question + LLM generated (Bing Chat, NeevaAI, perplexity.ai, and YouChat) in-line cited answers + cited webpages

**Annotations:** Is the answer fully/partially/ not supported by cited webpages? + supporting sentences in the cited webpages

**Task:** Attribute answer sentences to sentences in source document

**Reformulated Input (Closed-domain):** Document (webpages corresponding to fully supported citations) + Question + Answer (without citations)

**Reformulated Annotations:** sentences in the webpages which fully support the answer sentences

# Dataset Challenge: Reformulating Existing Datasets

---

## Question

What was Octavia E. Butler's first novel?

---

## Quotes

[1] Survivor is a science fiction novel by American writer Octavia E. Butler. First published in 1978 as part of Butler's "Patternist series"...

[2] Butler's first work published was "Crossover" in the 1971 Clarion Workshop anthology... Starting in 1974, Butler worked on a series of novels that would later be collected as the Patternist series... The first novel, "Patternmaster" (1976), eventually became the last installment in the series' internal chronology...

---

## Answer

Octavia E. Butler's first novel was "Patternmaster" which was published in 1976 and was also the first installment in her "Patternist series" [2].

---

**Informative?** Yes

**Attributable?** Yes

---

**Task:** Generate an attributed answer given a query and some quotes

**Original Input:** Question + Quotes

**Annotations:** Is the generated answer correctly attributed?

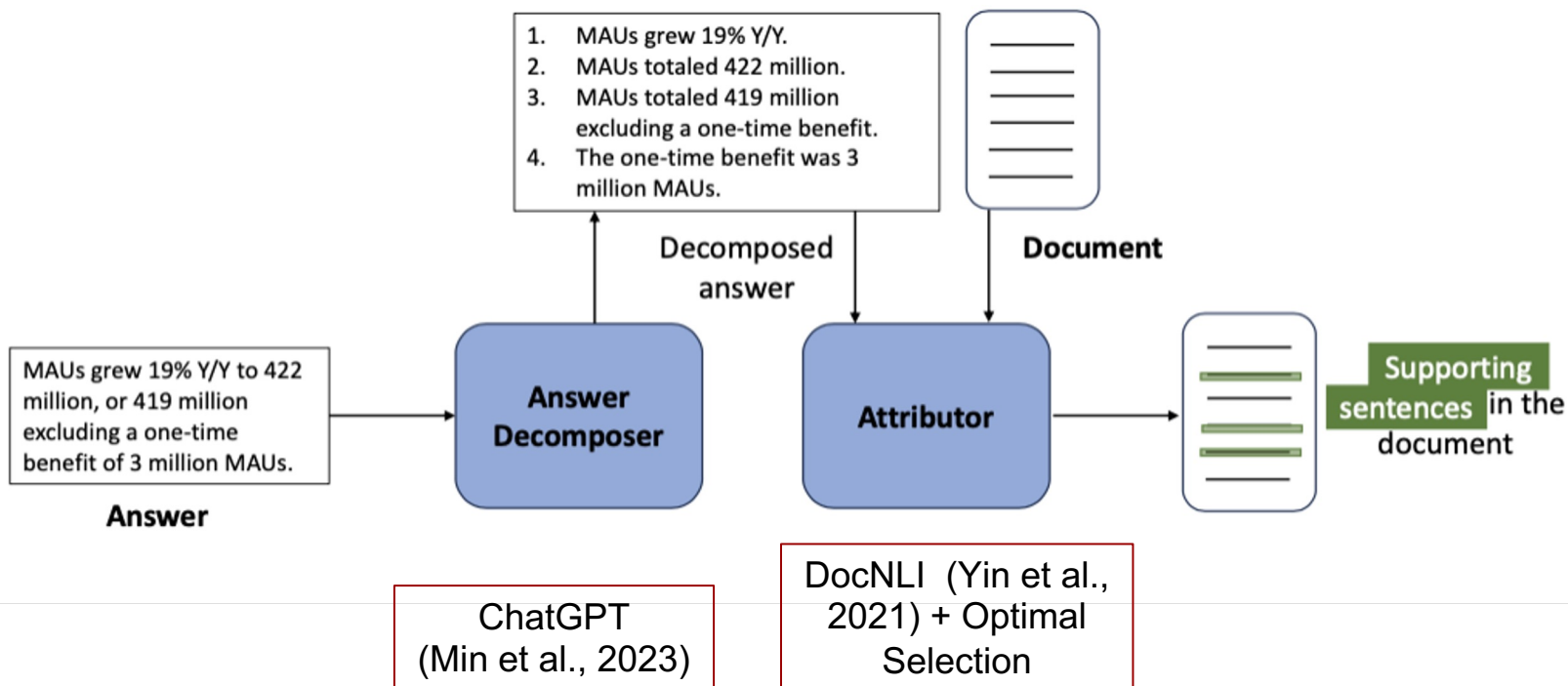
**Task:** Attribute answer sentences to sentences in source document

**Reformulated Input:** Document (Quotes) + Question + Answer  
(without citations/attributions)

**Reformulated Annotations:** answer sentences which are attributable to one or more quotes

Hagrid (Kamalloo et al., 2023)

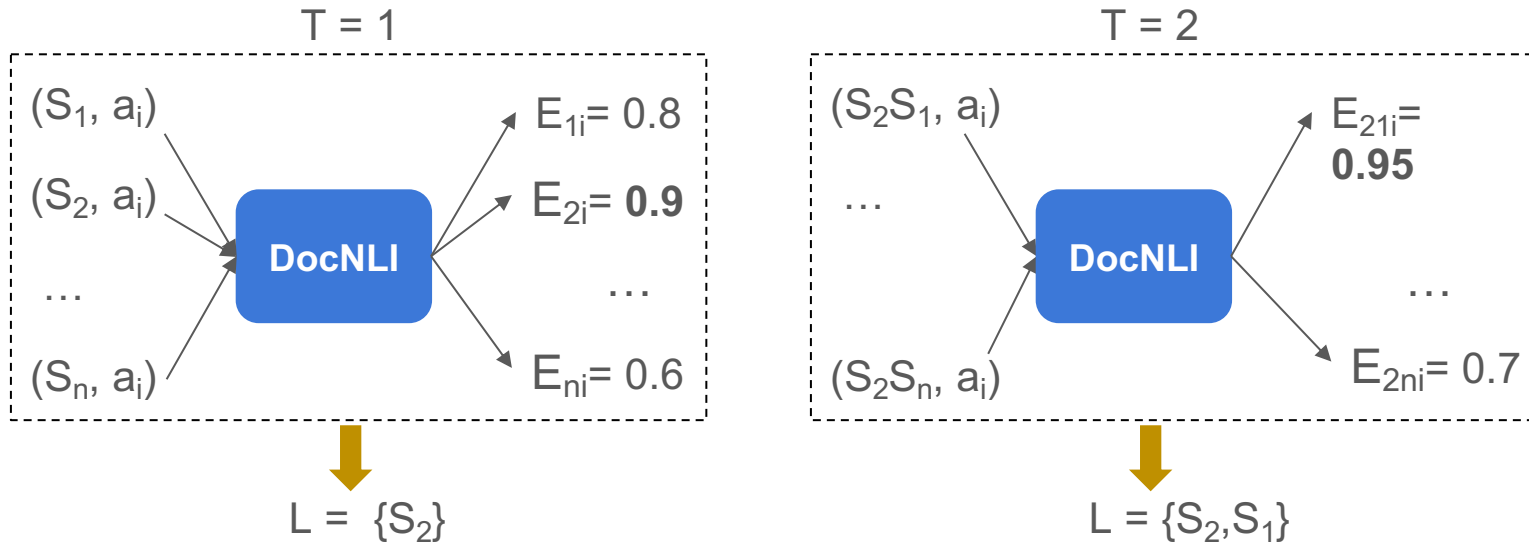
# ADiOSAA: Answer Decomposition and Optimal Selection for Answer Attribution



# Optimal Selection: An iterative and greedy algorithm

**Input:**  $D = (S_1, S_2, \dots, S_n)$   $A = (a_1, a_2, \dots, a_m)$   $\text{DocNLI}(p, h) = E$  (probability of  $p \rightarrow h$ )

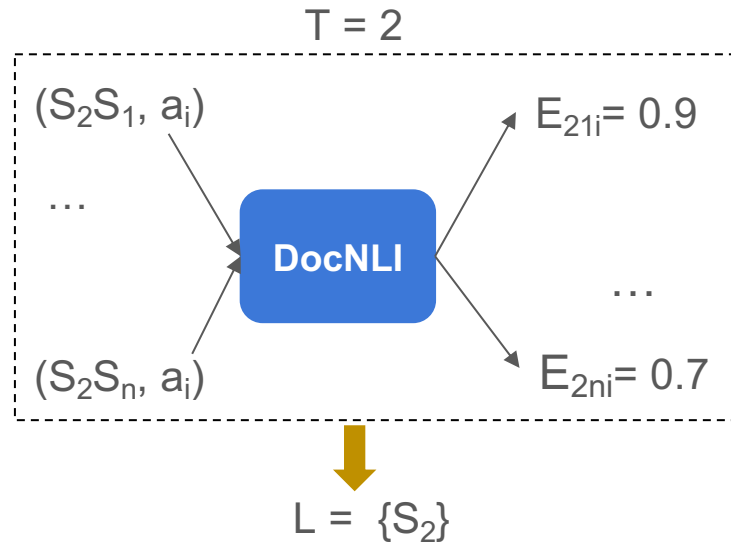
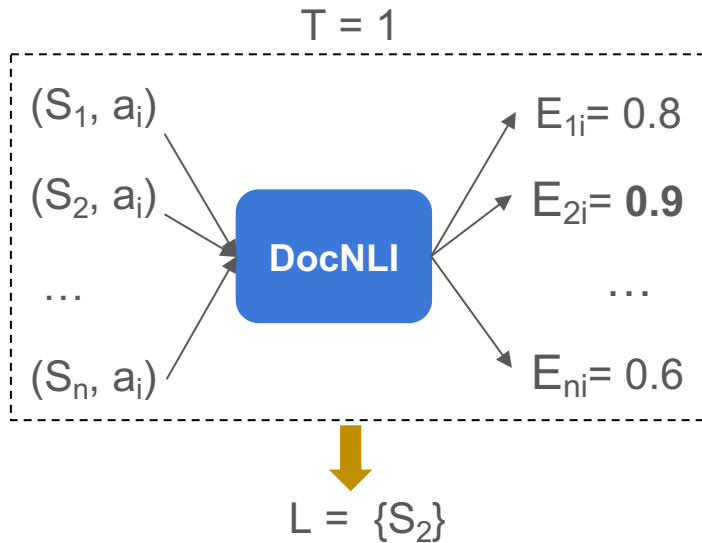
**Output:**  $L =$  List of sentences in  $D$  that can be attributed to  $a_i$  (for all  $i$  in  $1:m$ )



# Optimal Selection

**Input:**  $D = (S_1, S_2, \dots, S_n)$   $A = (a_1, a_2, \dots, a_n)$   $\text{DocNLI}(p, h) = E$  (probability of  $p \rightarrow h$ )

**Output:**  $L =$  List of sentences in  $D$  that can be attributed to  $a_i$



# Evaluation

- **Retrieval-based systems**
  - BM25, GTR, MonoT5
- **ADiOSAA ablations**
  - ADiOSAA - D (Decomposer)
  - ADiOSAA - OS (Optimal Selection)
  - ADiOSAA - D - OS
- **Metrics**
  - Precision@k, Recall@k, F1@k

# Findings

Model	Verifiability		
	(P/R/F1)@1	(P/R/F1)@2	(P/R/F1)@4
BM25	0.669/0.529/0.567	0.443/0.648/0.499	0.270/0.722/0.369
GTR	0.656/0.511/0.550	0.432/0.623/0.483	0.270/0.723/0.371
MonoT5	<b>0.698/0.552/0.593</b>	0.466/ <b>0.675/0.522</b>	0.284/ <b>0.757/0.389</b>
ADiOSAA	0.545/0.428/0.459	<b>0.484</b> /0.546/0.487	<b>0.476</b> /0.604/ <b>0.499</b>
ADiOSAA - D	0.473/0.388/0.412	0.445/0.418/0.412	0.442/0.418/0.411
ADiOSAA - OS	0.375/0.295/0.317	0.280/0.333/0.284	0.256/0.360/0.276
ADiOSAA - D - OS	0.269/0.234/0.243	0.269/0.234/0.243	0.269/0.234/0.243

- MonoT5 outperforms others on all the top-1 measures



# Findings

Model	Verifiability		
	(P/R/F1)@1	(P/R/F1)@2	(P/R/F1)@4
BM25	0.669/0.529/0.567	0.443/0.648/0.499	0.270/0.722/0.369
GTR	0.656/0.511/0.550	0.432/0.623/0.483	0.270/0.723/0.371
MonoT5	<b>0.698/0.552/0.593</b>	0.466/ <b>0.675/0.522</b>	0.284/ <b>0.757/0.389</b>
ADiOSAA	0.545/0.428/0.459	<b>0.484</b> /0.546/0.487	<b>0.476</b> /0.604/ <b>0.499</b>
ADiOSAA - D	0.473/0.388/0.412	0.445/0.418/0.412	0.442/0.418/0.411
ADiOSAA - OS	0.375/0.295/0.317	0.280/0.333/0.284	0.256/0.360/0.276
ADiOSAA - D - OS	0.269/0.234/0.243	0.269/0.234/0.243	0.269/0.234/0.243

- MonoT5 outperforms others on top-1 measures
- ADiOSAA attains the highest top-2/4 precision

# Findings

Model	Verifiability		
	(P/R/F1)@1	(P/R/F1)@2	(P/R/F1)@4
BM25	0.669/0.529/0.567	0.443/0.648/0.499	0.270/0.722/0.369
GTR	0.656/0.511/0.550	0.432/0.623/0.483	0.270/0.723/0.371
MonoT5	<b>0.698/0.552/0.593</b>	0.466/ <b>0.675/0.522</b>	0.284/ <b>0.757/0.389</b>
ADiOSAA	0.545/0.428/0.459	<b>0.484/0.546/0.487</b>	<b>0.476/0.604/0.499</b>
ADiOSAA - D	0.473/0.388/0.412	0.445/0.418/0.412	0.442/0.418/0.411
ADiOSAA - OS	0.375/0.295/0.317	0.280/0.333/0.284	0.256/0.360/0.276
ADiOSAA - D - OS	0.269/0.234/0.243	0.269/0.234/0.243	0.269/0.234/0.243

- MonoT5 outperforms others on top-1 measures
- ADiOSAA attains the highest top-2/4 precision
- ADiOSAA variants retain high precision with increase in k
  - Retrieval-based systems are good in correctly identifying one attribution but fail in case of more than one attribution

# Findings

Model	Verifiability		
	(P/R/F1)@1	(P/R/F1)@2	(P/R/F1)@4
ADiOSAA	0.545/0.428/0.459	<b>0.484</b> /0.546/0.487	<b>0.476</b> /0.604/ <b>0.499</b>
ADiOSAA - D	0.473/0.388/0.412	0.445/0.418/0.412	0.442/0.418/0.411
ADiOSAA - OS	0.375/0.295/0.317	0.280/0.333/0.284	0.256/0.360/0.276
ADiOSAA - D - OS	0.269/0.234/0.243	0.269/0.234/0.243	0.269/0.234/0.243

- Optimal selection results in a significant improvement (ADiOSAA > ADiOSAA - OS and ADiOSAA - D > ADiOSAA - D - OS)

# Findings

Model	Verifiability		
	(P/R/F1)@1	(P/R/F1)@2	(P/R/F1)@4
ADiOSAA	0.545/0.428/0.459	<b>0.484</b> /0.546/0.487	<b>0.476</b> /0.604/ <b>0.499</b>
ADiOSAA - D	0.473/0.388/0.412	0.445/0.418/0.412	0.442/0.418/0.411
ADiOSAA - OS	0.375/0.295/0.317	0.280/0.333/0.284	0.256/0.360/0.276
ADiOSAA - D - OS	0.269/0.234/0.243	0.269/0.234/0.243	0.269/0.234/0.243

- Optimal selection results in a significant improvement (ADiOSAA > ADiOSAA - OS and ADiOSAA - D > ADiOSAA - D - OS)
- Decomposition results in improved scores (ADiOSAA > ADiOSAA - D and ADiOSAA - OS > ADiOSAA - OS - D)

# Takeaway

- **Good performance of retrieval-based systems** indicate that the **existing datasets are less abstractive** for long-form comprehension
  - **need** for more **challenging datasets** to foster development of trustworthy and reliable systems
- **Decomposition** of answer sentences into information units and **optimal selection** of source sentences **results in improved performance**

# Thanks!

sancheti@umd.edu



Scan me!