# Factuality and Hallucination

- Language models model distributions over text, not facts. There's no guarantee that what they generate is factual:
    - Language models are trained on the web. Widely-popularized falsehoods may be reproduced in language models
    - A language model may not be able to store all rare facts, and as a result moderate probability is assigned to several options
- RLHF improves this (particularly the calibration of when the model answers versus saying "I don't know) but doesn't eliminate it. How can we detect factual errors in order to evaluate our systems?

# Grounding LM Generations

- Suppose we have text generated from an LM. We want to check it against a source document. What techniques have we seen so far that can do this?

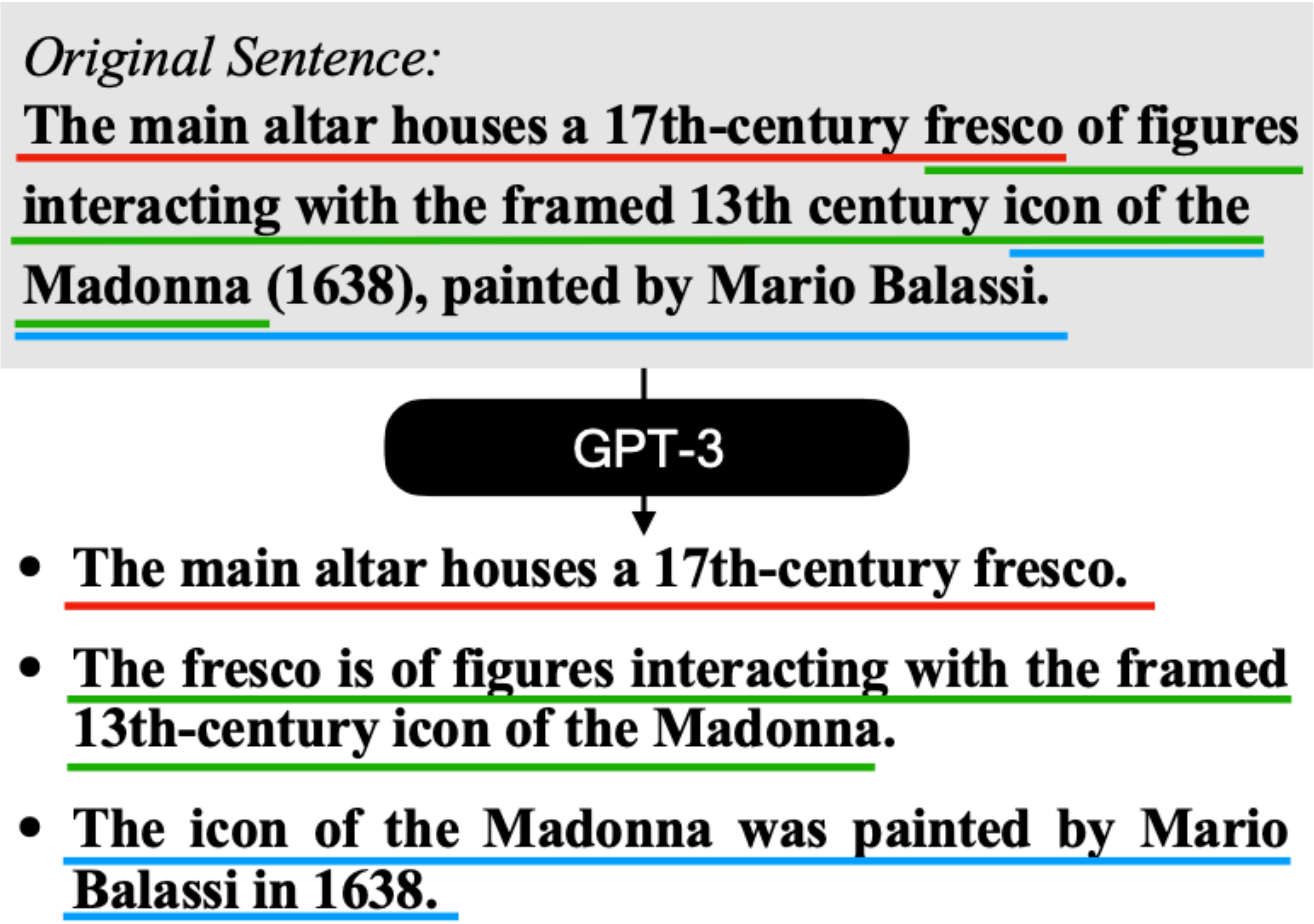
- What steps are involved?

  1. Decide what text you are grounding in (may involve retrieval)
  2. Decompose your text into pieces of meaning to ground
  3. Check each piece

- For now, we'll assume the reference text/documents are given to us and not focus on step 1

# Step 2: Fact Decompositions

- Simplest approach: each sentence needs to be grounded
- Can go deeper: think of sentences as expressing a collection of propositions
- Long history in frame semantics of defining these propositions. Many propositions anchor to verbs
- Recent work: extract propositions with LLMs:



*Original Sentence:*

**The main altar houses a 17th-century fresco of figures interacting with the framed 13th century icon of the Madonna (1638), painted by Mario Balassi.**

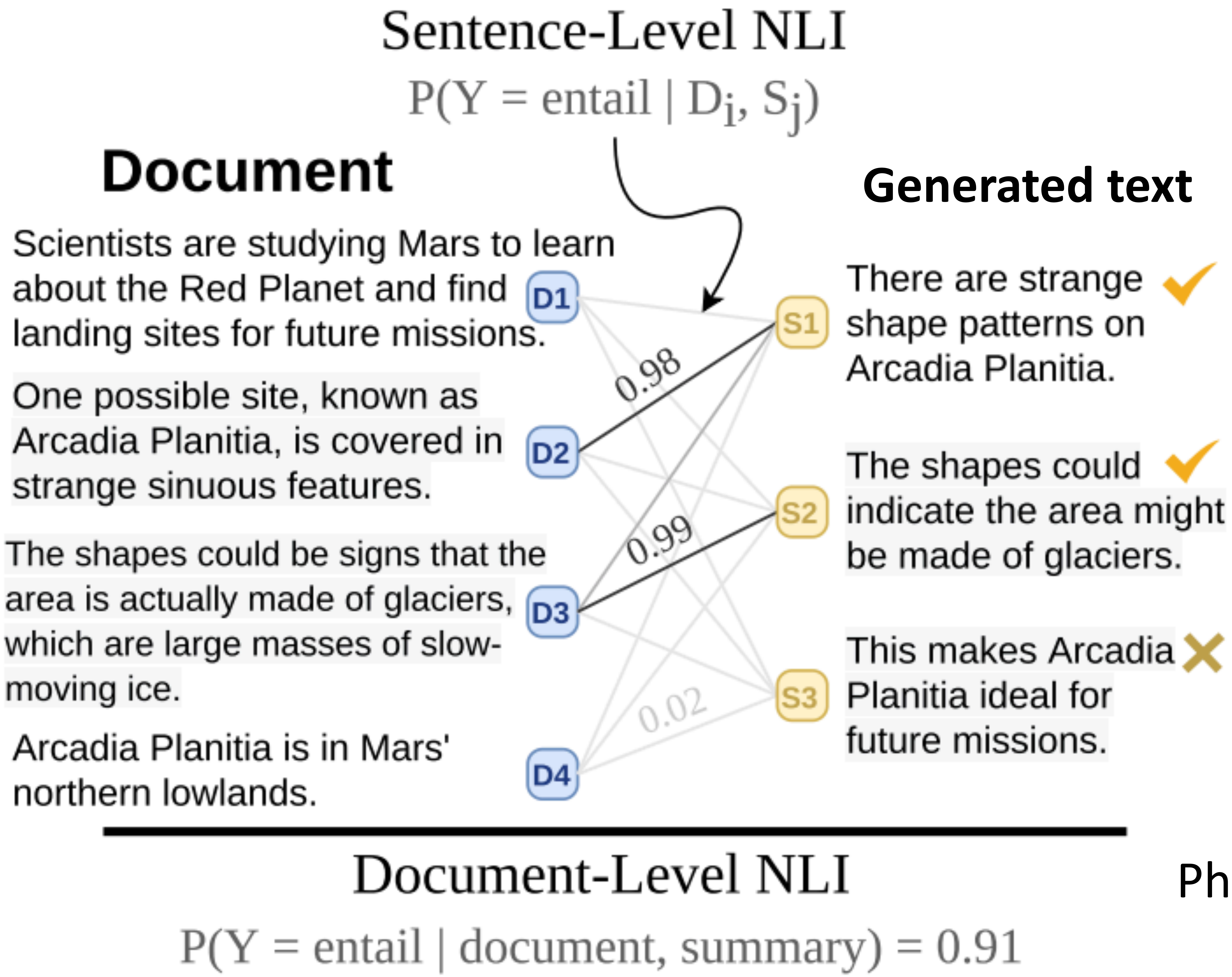
- **The main altar houses a 17th-century fresco.**
- **The fresco is of figures interacting with the framed 13th-century icon of the Madonna.**
- **The icon of the Madonna was painted by Mario Balassi in 1638.**

Yixin Liu et al. (2023)

Ryo Kamoi et al. (2023)

# Step 3: Checking

- One idea: use textual entailment to see if each piece to check is entailed by the source
- Simple version that originated in summarization: take the max entailment score over every document sentence



Sentence-Level NLI

$P(Y = \text{entail} \mid D_i, S_j)$

**Document**

Scientists are studying Mars to learn about the Red Planet and find landing sites for future missions.

One possible site, known as Arcadia Planitia, is covered in strange sinuous features.

The shapes could be signs that the area is actually made of glaciers, which are large masses of slow-moving ice.

Arcadia Planitia is in Mars' northern lowlands.

**Generated text**

There are strange shape patterns on Arcadia Planitia. ✓

The shapes could indicate the area might be made of glaciers. ✓

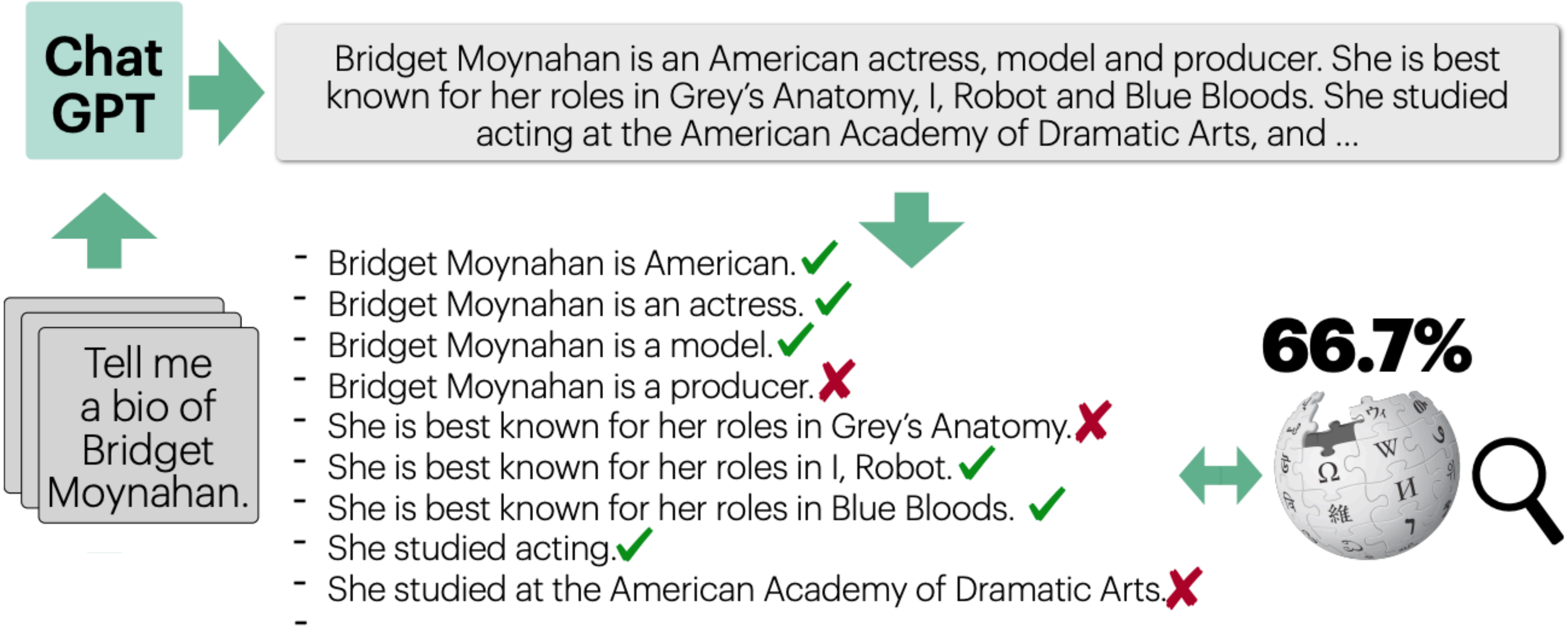
This makes Arcadia Planitia ideal for future missions. ✗

Document-Level NLI

$P(Y = \text{entail} \mid \text{document}, \text{summary}) = 0.91$

Philippe Laban et al. (2022)

# FActScore



Chat GPT

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

Tell me a bio of Bridget Moynahan.

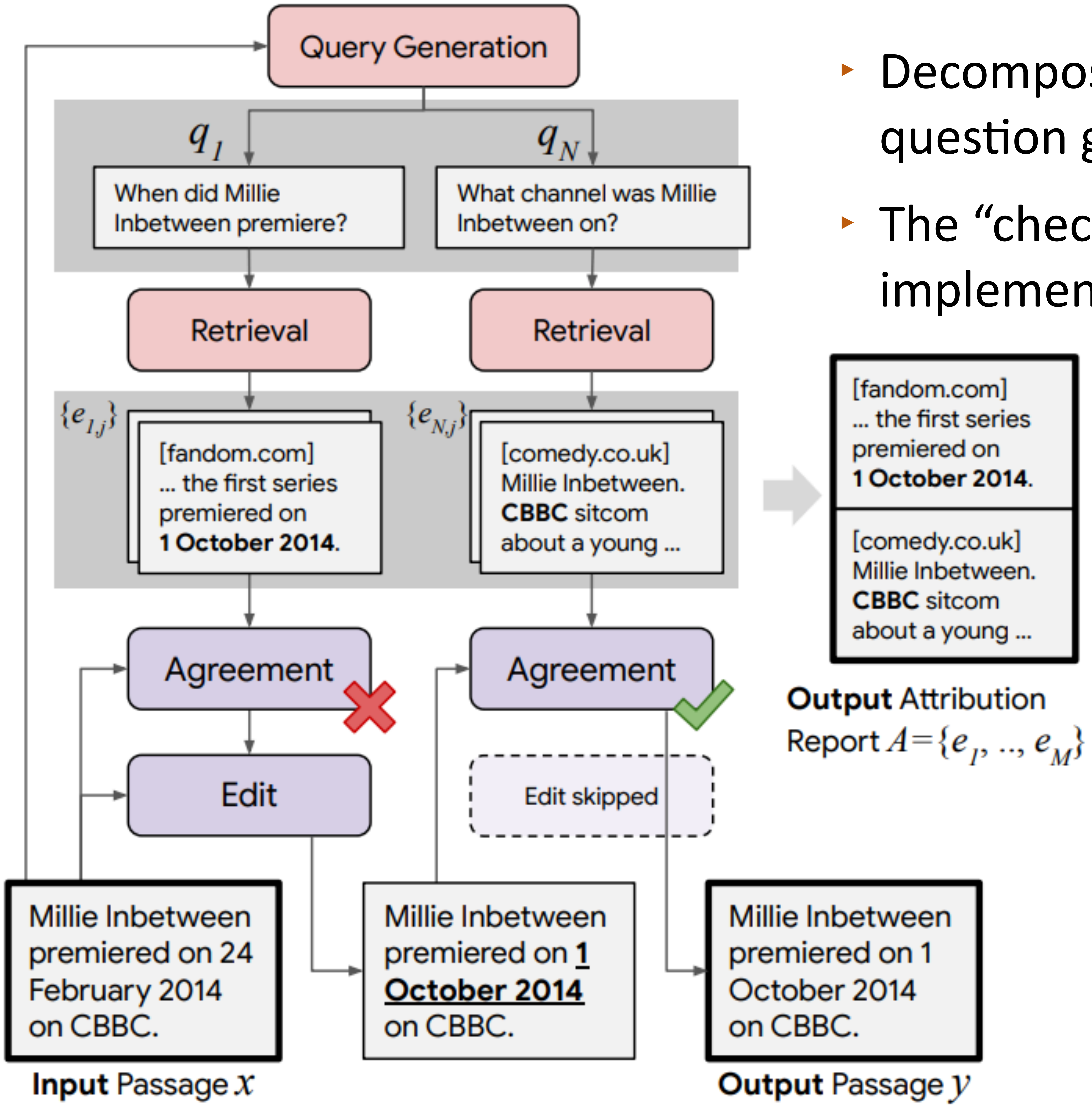
- Bridget Moynahan is American. ✓
- Bridget Moynahan is an actress. ✓
- Bridget Moynahan is a model. ✓
- Bridget Moynahan is a producer. ✗
- She is best known for her roles in Grey's Anatomy. ✗
- She is best known for her roles in I, Robot. ✓
- She is best known for her roles in Blue Bloods. ✓
- She studied acting. ✓
- She studied at the American Academy of Dramatic Arts. ✗
-

66.7%

- Dataset: ChatGPT-generated biographies of people. May contain errors, particularly when dealing with obscure peole!
- Uses LLMs both for decomposition and for checking

Sewon Min and Kalpesh Krishna et al. (2023)

# Pipeline: RARR



Query Generation
$q_1$
$q_N$
When did Millie Inbetween premiere?
What channel was Millie Inbetween on?
Retrieval
Retrieval
$\{e_{1,j}\}$
$\{e_{N,j}\}$
[fandom.com] ... the first series premiered on **1 October 2014**.
[comedy.co.uk] Millie Inbetween. **CBBC** sitcom about a young ...
Agreement
Agreement
Edit
Edit skipped
[fandom.com] ... the first series premiered on **1 October 2014**.
[comedy.co.uk] Millie Inbetween. **CBBC** sitcom about a young ...
**Output** Attribution Report $A=\{e_1, .., e_M\}$
Millie Inbetween premiered on 24 February 2014 on CBBC.
**Input** Passage $x$
Millie Inbetween premiered on **1 October 2014** on CBBC.
Millie Inbetween premiered on 1 October 2014 on CBBC.
**Output** Passage $y$

- Full pipeline including retrieval
- Decomposition is framed as question generation
- The “checking” stage is also implemented with LLMs here

Luyu Gao et al. (2022)