

Rich Knowledge Sources Bring Complex Knowledge Conflicts: Recalibrating Models to Reflect Conflicting Evidence



Hung-Ting Chen



Michael J.Q. Zhang



Eunsol Choi



Rich Knowledge Sources for QA Model

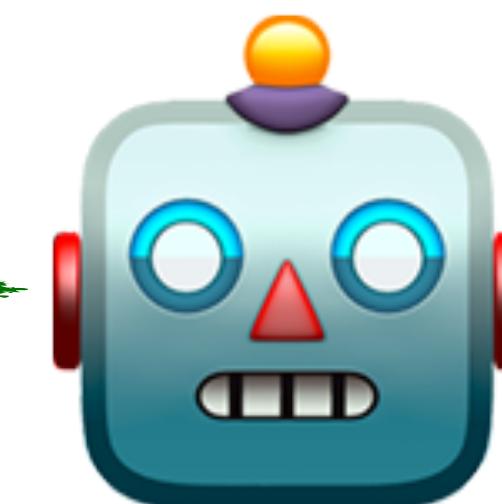


What is the third planet from the sun?

Facts memorized during training (parametric)

The third planet from the sun is Earth.

(Output from text-davinci-002)



Earth

Documents retrieved at inference time (non-parametric)

Passage 1

... From closest to farthest from the Sun, they are Mercury, Venus,

Passage 2

...Planet Earth is the third planet from the Sun...

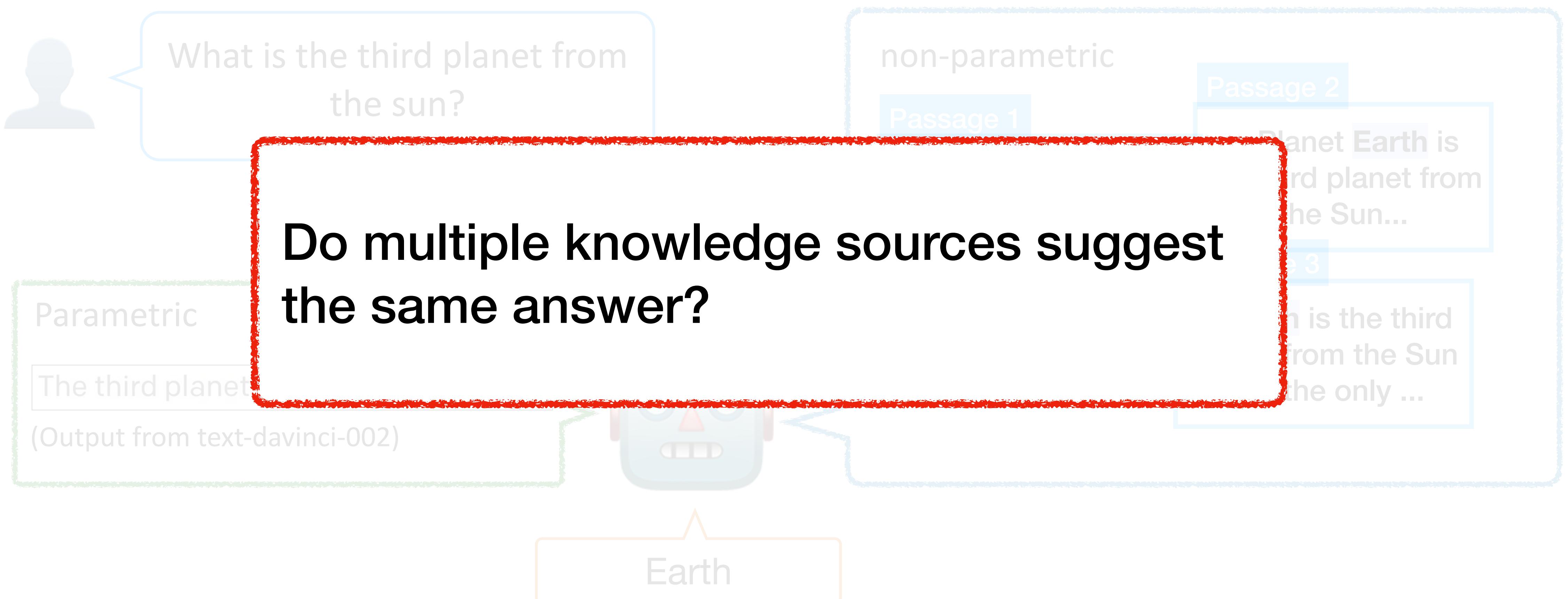
Passage 3

...Earth is the third planet from the Sun and the only ...



WIKIPEDIA
The Free Encyclopedia

Rich Knowledge Sources for QA Model



Multiple Sources Suggest the Same Answer

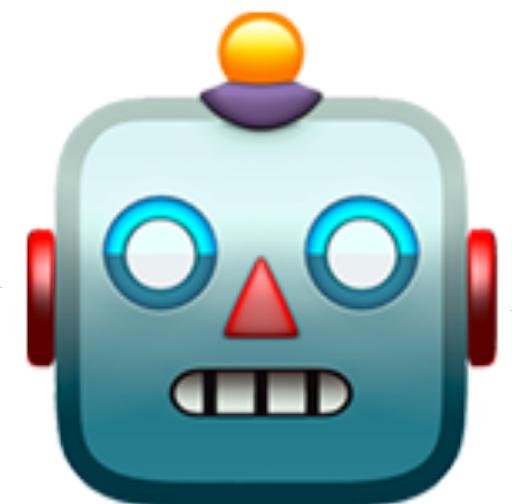


What is the third planet from the sun?

Parametric

The third planet from the sun is Earth.

(Output from text-davinci-002)



Earth

non-parametric

Passage 1

... From closest to farthest from the Sun, they are: Mercury, Venus, **Earth**, ...

Passage 2

...Planet **Earth** is the third planet from the Sun...

Passage 3

...**Earth** is the third planet from the Sun and the only ...

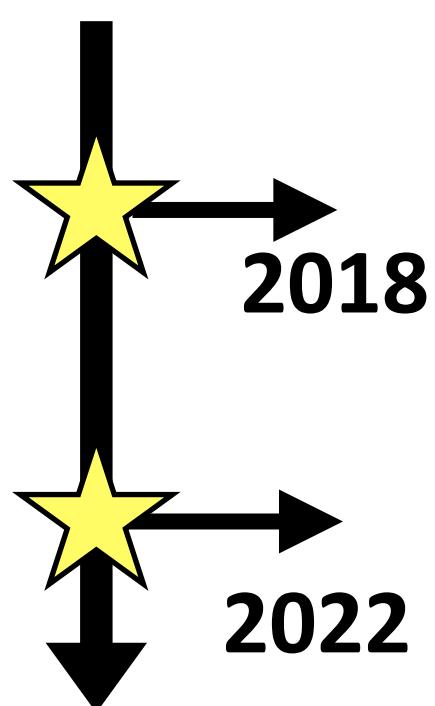
Real-world Knowledge Conflicts

- Temporal Conflicts

SituatedQA [Zhang and Choi, EMNLP 2021]



Where were the last Winter Olympic Games held?



Passage 1

The last winter Olympics were held in **Pyeongchang** in **South Korea**.

Passage 2

The last winter Olympic Games were held in **Beijing, China**.

- Ambiguous Questions

AmbigQA [Min et al, EMNLP 2020]



When did **harry potter and the sorcerer's stone** movie come out?

Passage 1

The film had its **world premiere** at the Odeon Leicester Square in London on **4 November 2001**.

Passage 2

The film was **released to cinemas** in the United Kingdom and the United States on **16 November 2001**.

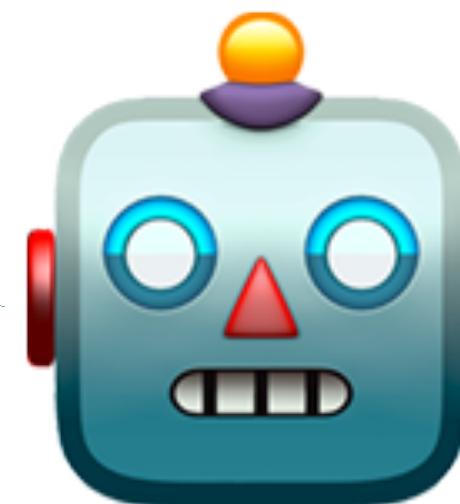
Rich Knowledge Sources for QA Model



Which country won the most medals in winter olympics?

Facts memorized during training (parametric)

Norway won the most medals in the winter olympics.



Documents retrieved at inference time (non-parametric)

Passage 1

Passage 2 of the

...Nor

re

n

Wi

39

Passage 3

...With 36 total
medals, **Germany**
set a record for
most total medals at
a Winter Olympics...



Types of Knowledge Conflicts

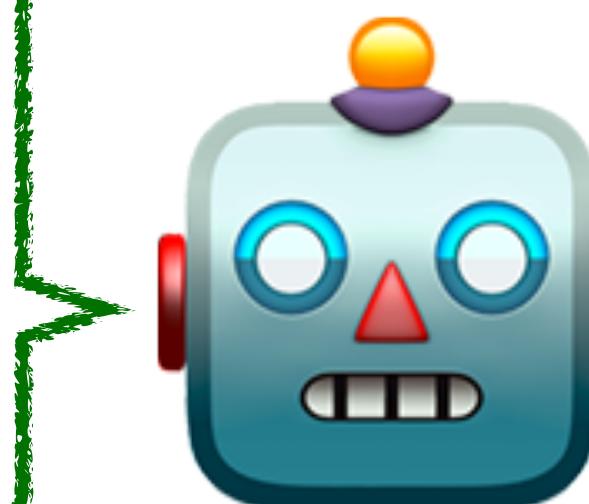
Parametric vs. non-parametric knowledge conflicts



Which country won the most medals in winter olympics?

Facts memorized during training (parametric)

Norway won the most medals in the winter olympics.



Documents retrieved at inference time (non-parametric)

Passage 1

...Norway for a single success with 39 medals.

Passage 2

...Norway successfully won 39 medals.

Passage 3

...With 36 total medals, Germany set a record for most total medals at a Winter Olympics...



Types of Knowledge Conflicts

Parametric vs. non-parametric knowledge conflicts



Which country won the most medals in winter olympics?

Facts memorized during training (parametric)



Norway won the most medals in the winter olympics.

Conflicts within non-parametric sources

Documents retrieved at inference time (non-parametric)



WIKIPEDIA
The Free Encyclopedia



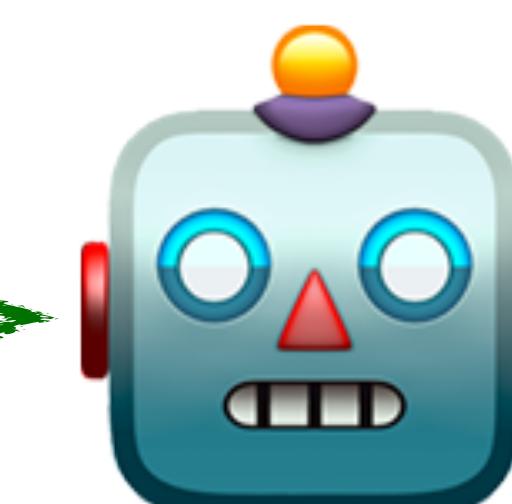
Passage 1

Passage 2

Passage 3

...Norway won the most medals in the winter olympics for a second consecutive time, with 36 total medals, setting a new record for most total medals at a Winter Olympics...

...With 36 total medals, Germany set a record for most total medals at a Winter Olympics...



This Talk

Analysis

RQ1. Conflicts between memorized vs. retrieved knowledge

- Do QA models rely on parametric or non-parametric knowledge?

RQ2. Conflicts within retrieved knowledge

- What models do when retrieved passages suggest conflicting answers?

Solution

RQ3. Recalibration given knowledge conflicts

- Can we teach models to abstain from answering when knowledge source points multiple answers?

Experimental Setup

Dataset

- NQ-Open (Open-retrieval split of NQ), TriviaQA

Models

- Fusion-in-Decoder (FiD): [Izacard and Grave, EACL 2021]
 - Retrieve up to 100 passages
- Retrieval-Augmented Generation (RAG): [Lewis et al., NeurIPS 2020]
 - Retrieves up to with 5 passages
 - Retriever trained together with the reader

This Talk

Analysis

RQ1. Conflicts between memorized vs. retrieved knowledge

- Do QA models rely on parametric or non-parametric knowledge?

RQ2. Conflicts within retrieved knowledge

- How models select one answer from a set of potential answers?

Solution

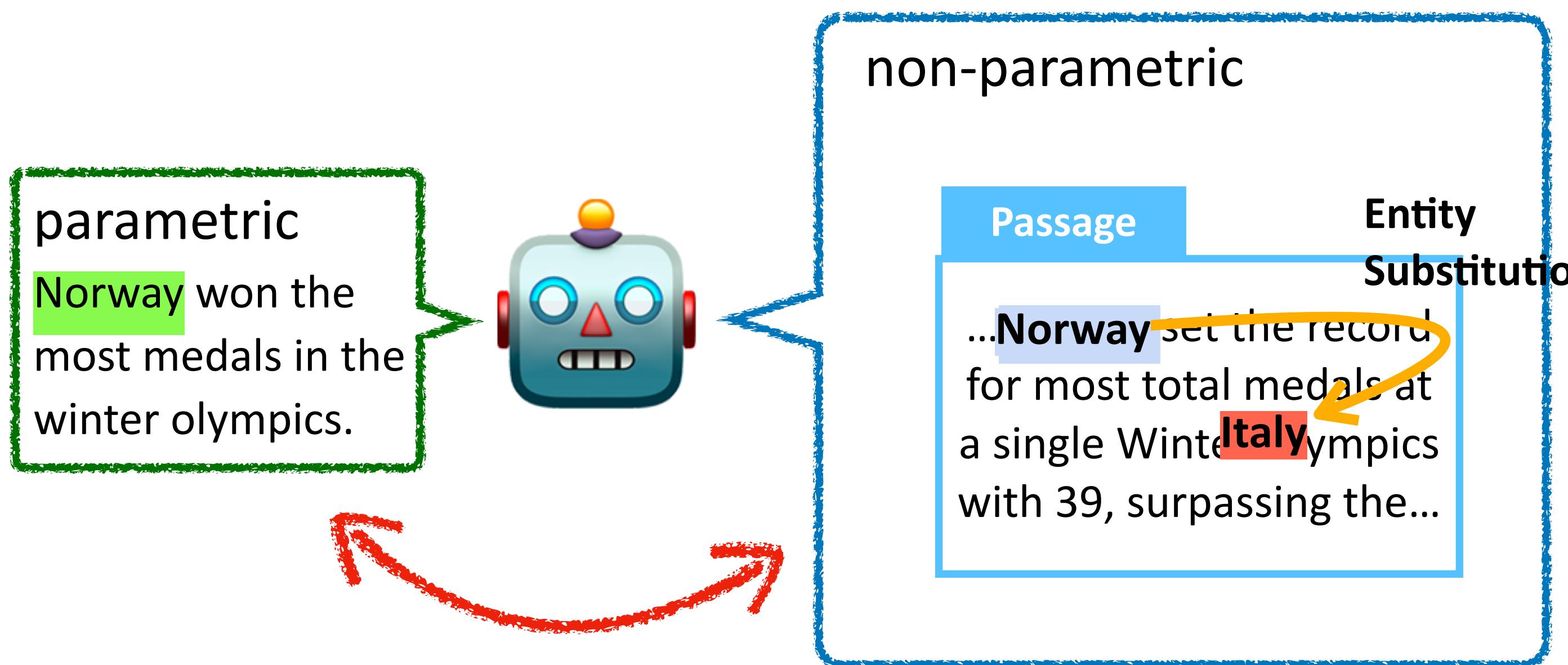
RQ3. Recalibration provided knowledge conflicts

- Can we teach models to abstain from answering when knowledge source points multiple answers?

Prior Work: Simulated parametric vs. non-parametric knowledge conflict



Which country won the most medals in winter olympics?



$$M_R = \frac{p_o}{p_o + p_s}$$

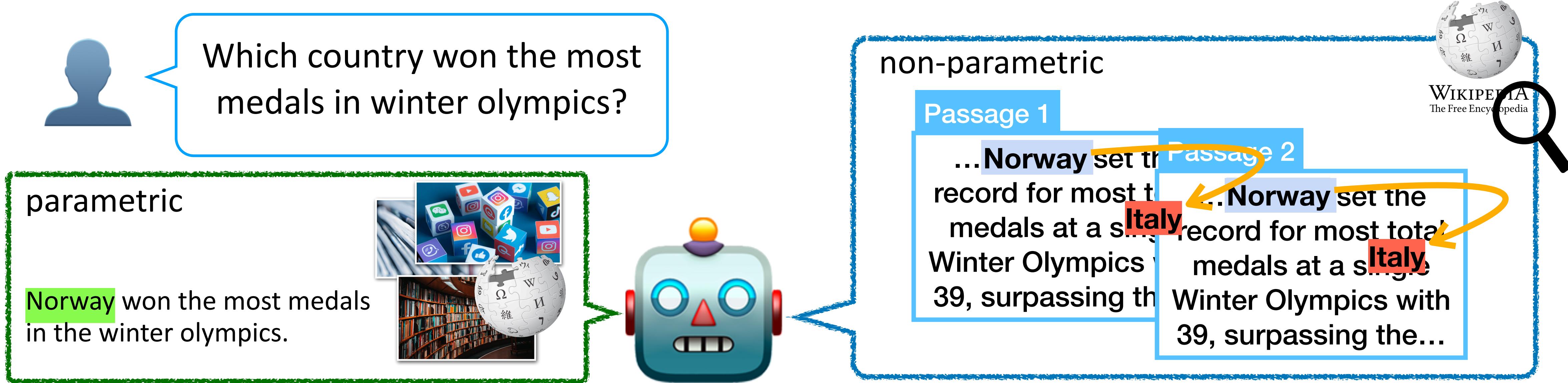
= 27% in NQ

examples predicting original answers
examples predicting substitute answers

Only perturb examples where the model predicts **correctly**

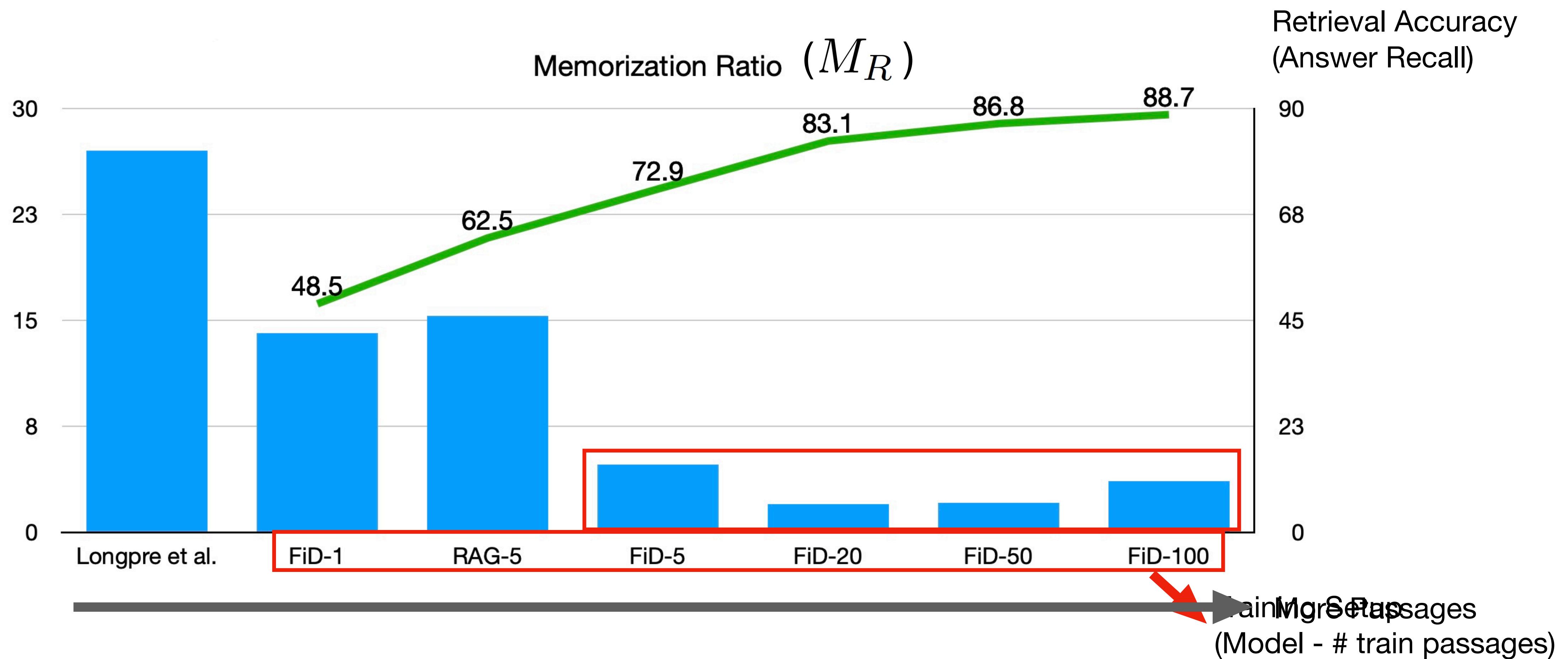
[Longpre et al, EMNLP 2021]

Extension to Multiple Documents Setting

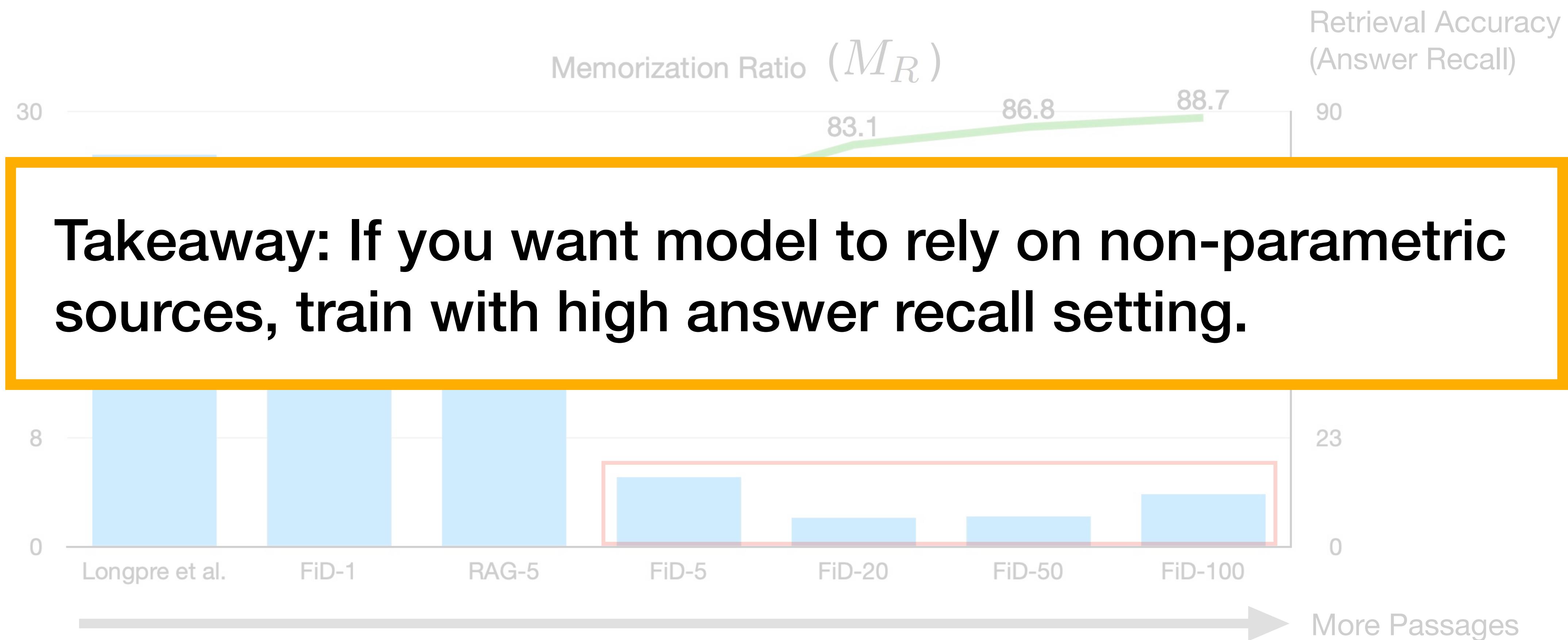


- We study entity substitution in more realistic, 100 evidence document setting, and substituting all answer entity mentions.

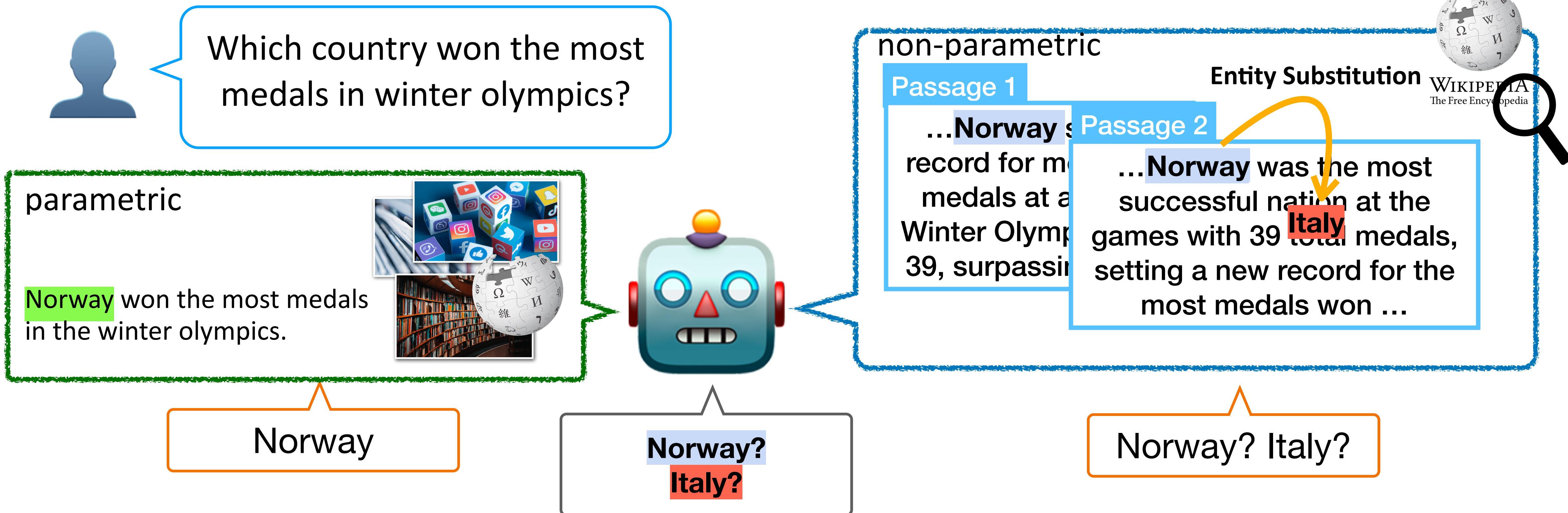
High-Recall Retriever Leads to Low Memorization



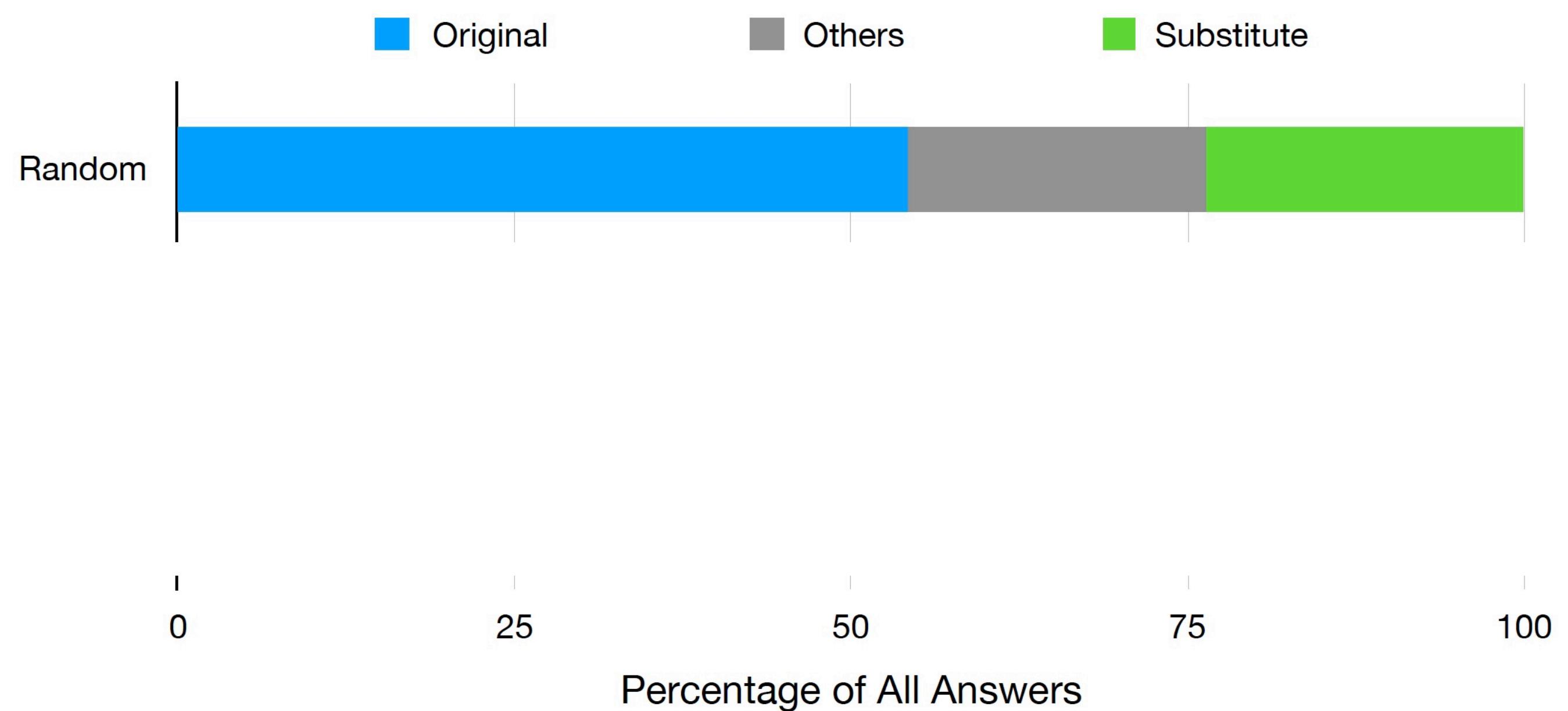
High-Recall Retriever Leads to Low Memorization



Simulating Mixed Bag of Evidence Passages

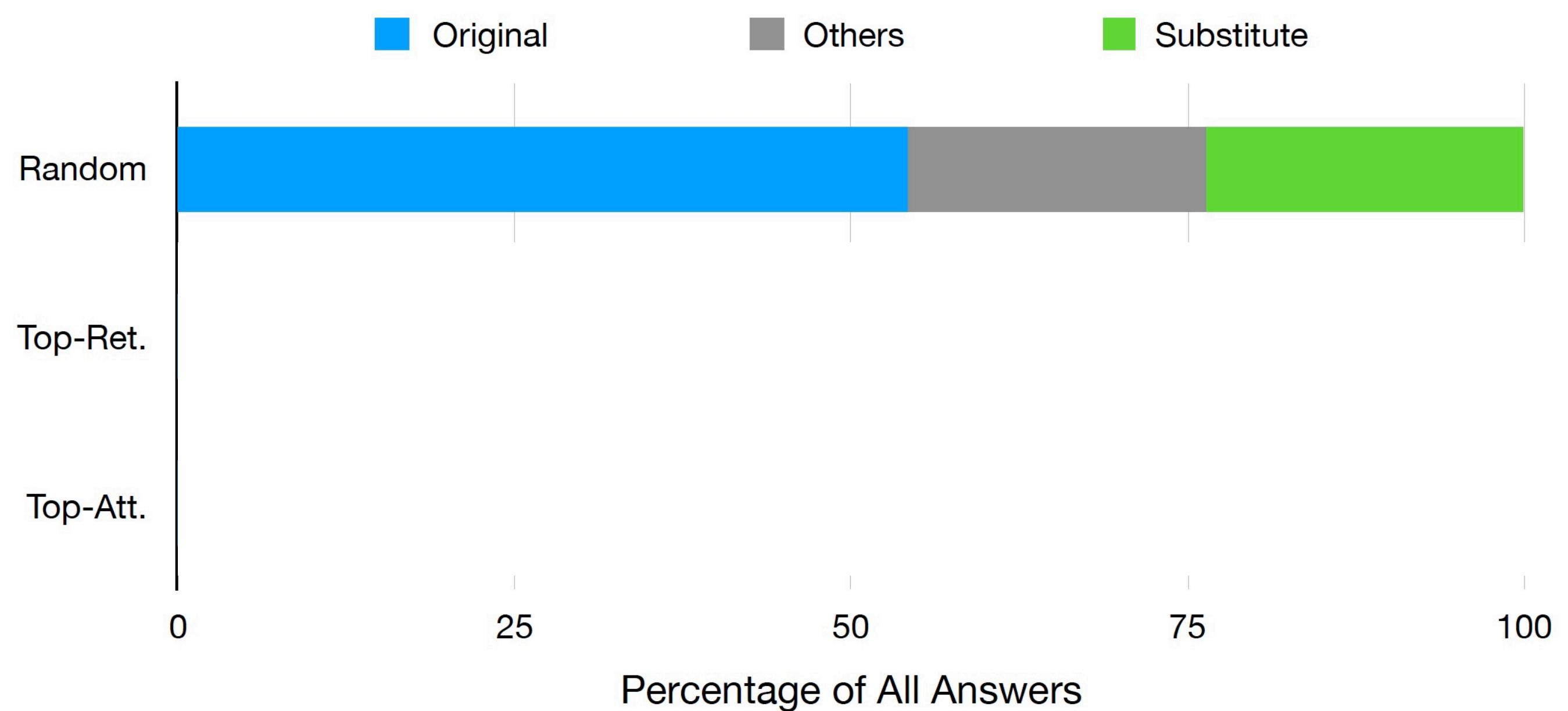


Results: Mixed Bag of Evidence



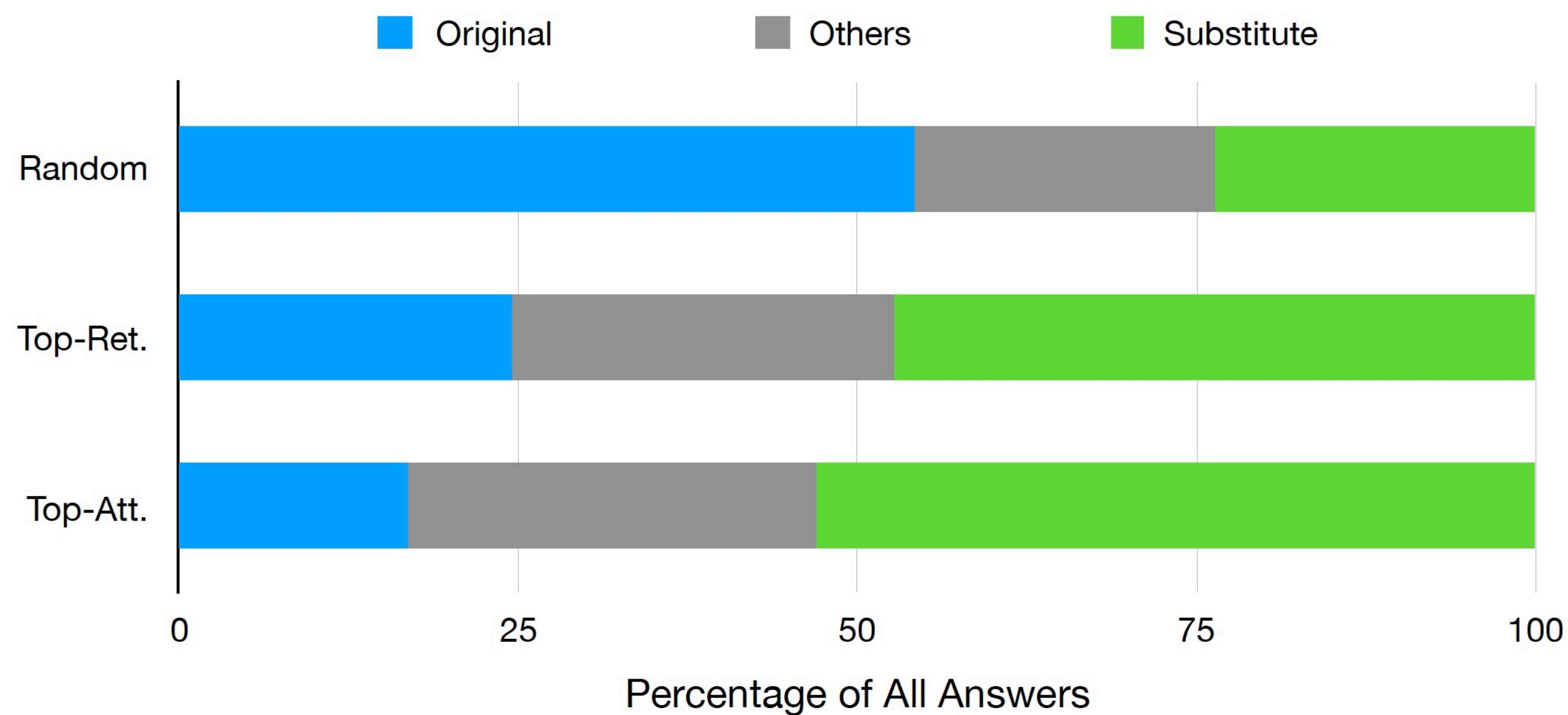
- When randomly perturbing 50% of passages, model favors the original answer.
 - ➡ Model still uses parametric knowledge.

Results: Mixed Bag of Evidence



- When randomly perturbing 50% of passages, model favors the original answer.
 - ➡ Model still uses parametric knowledge.

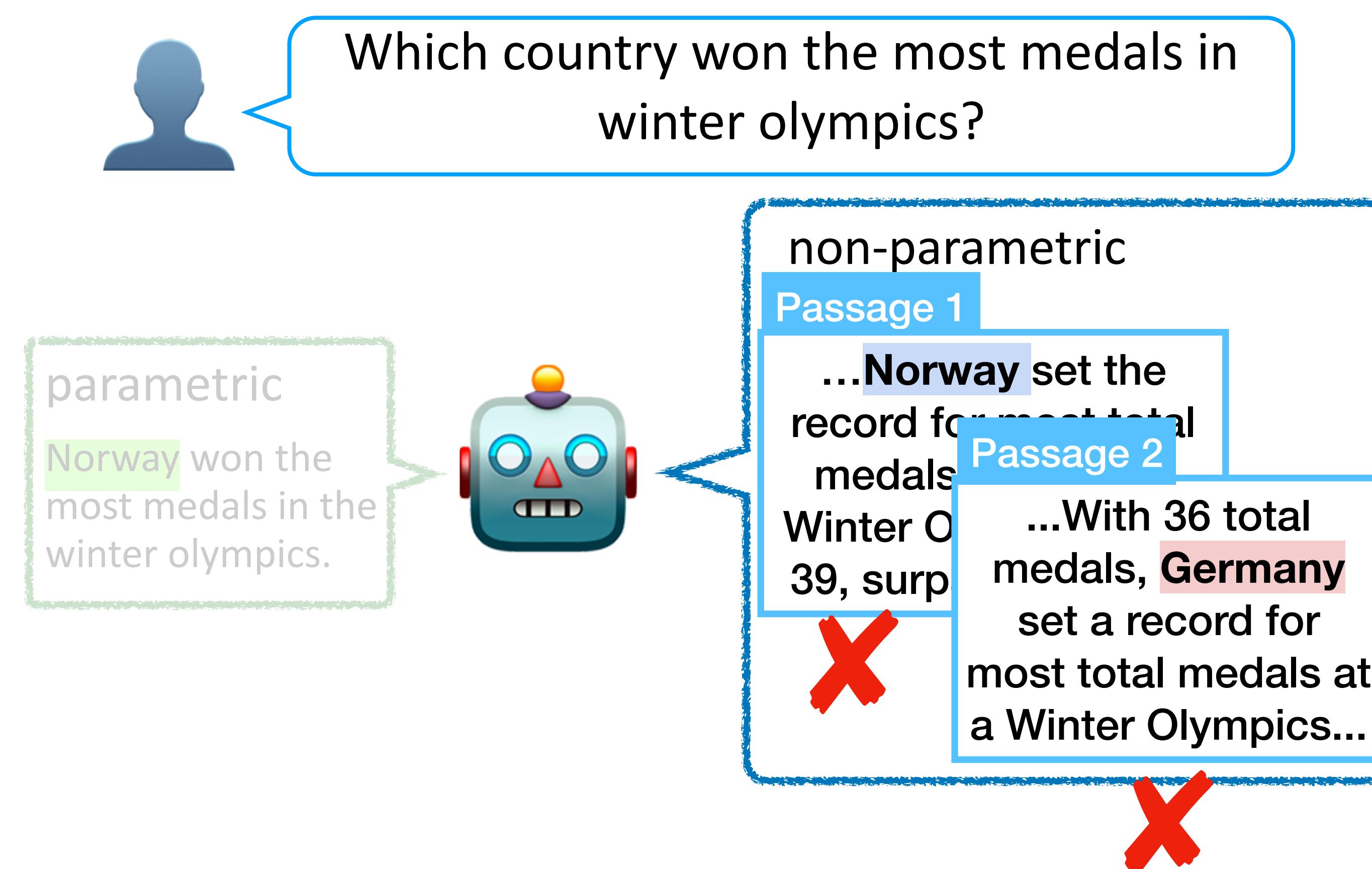
Results: Mixed Bag of Evidence



- When randomly perturbing 50% of passages, model favors the original answer.
 - ➡ Model still uses parametric knowledge.
- If we perturb top passages (by attention or retriever scores), model outputs substitute answers more.
 - ➡ Model might be using only the top few passages.

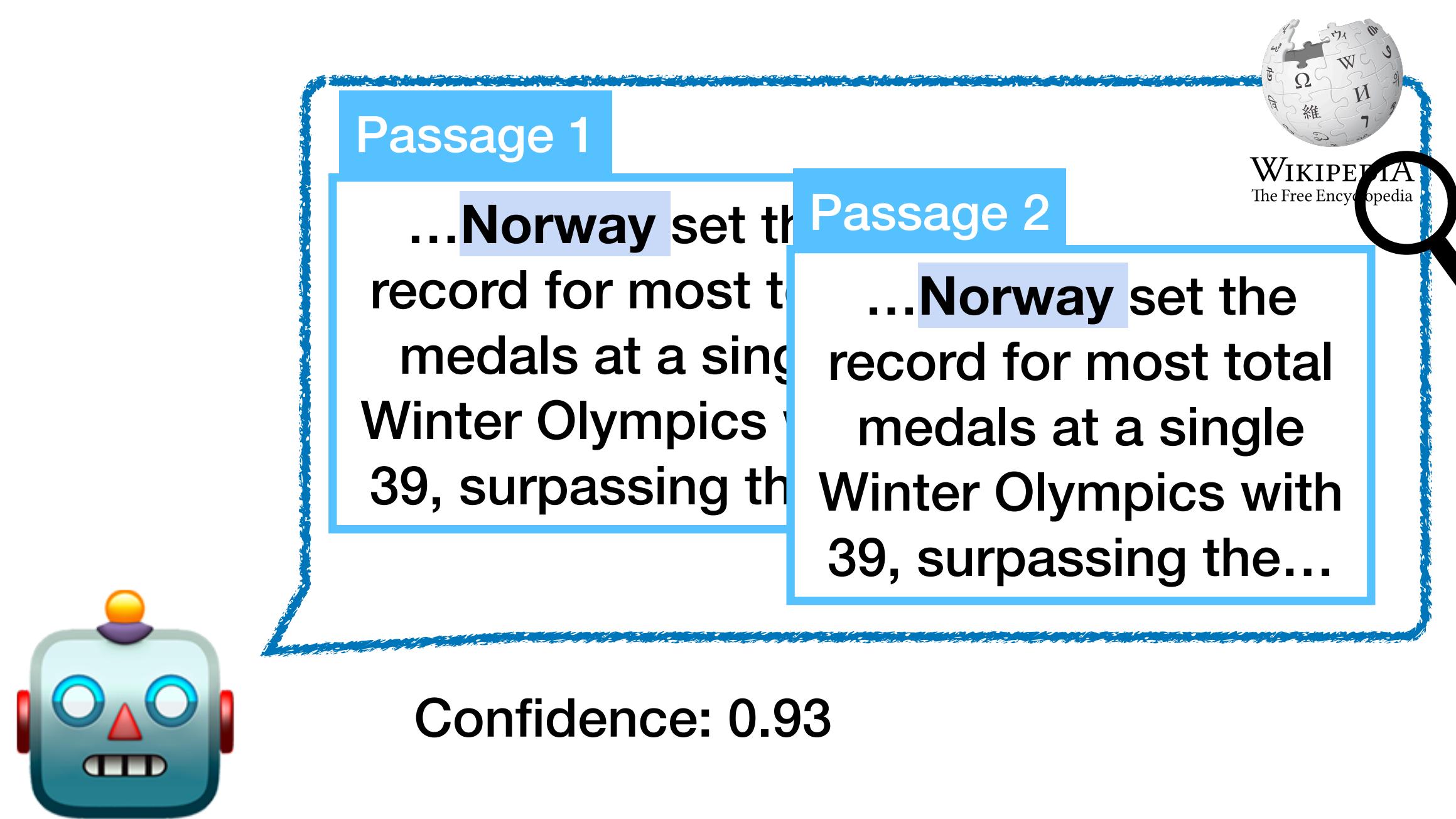
What To Do Under Knowledge Conflict?

- Solution: Abstain from answering

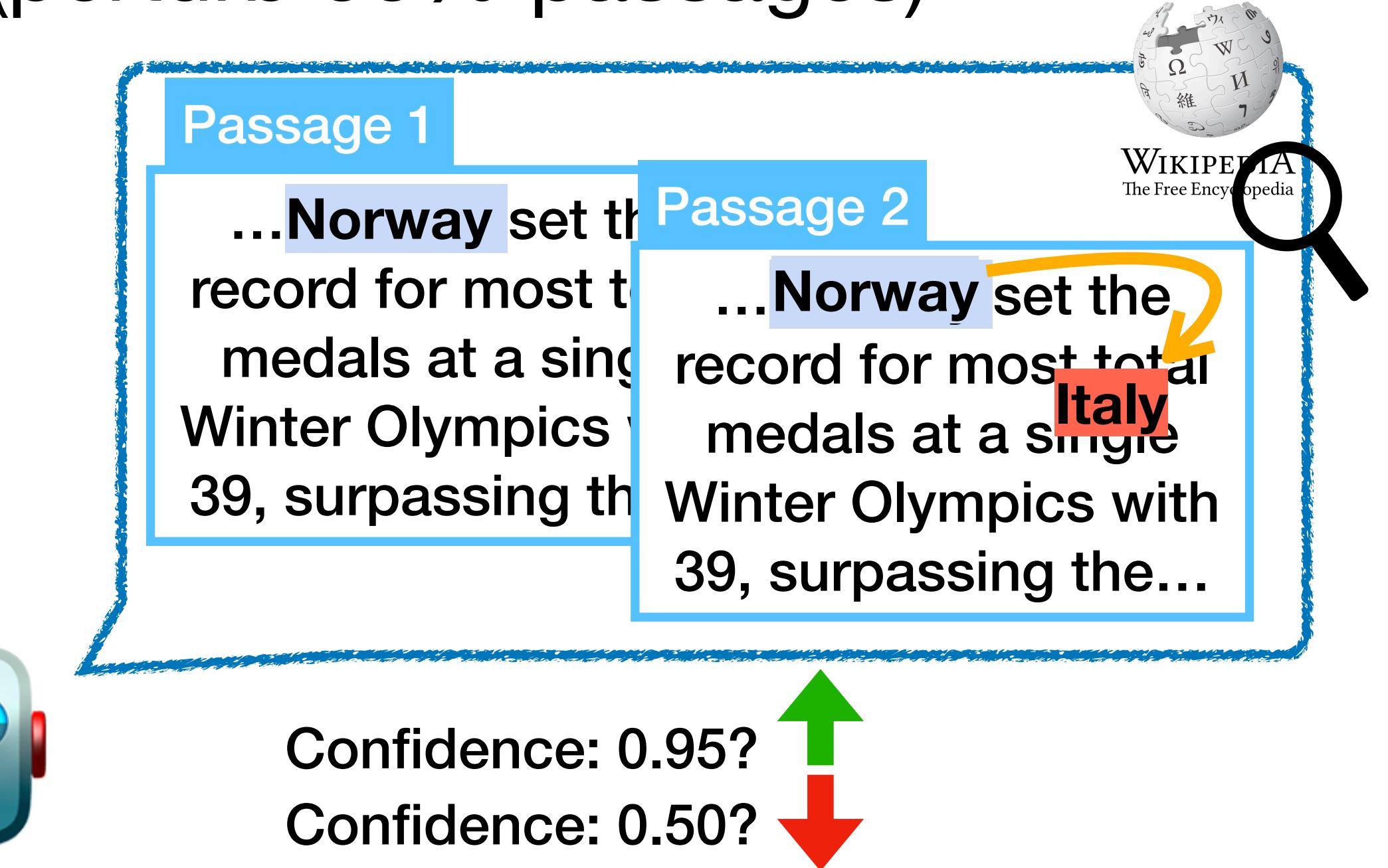


Is model less confident under knowledge conflict?

- No Knowledge Conflict



- Simulated Knowledge Conflict (perturb 50% passages)

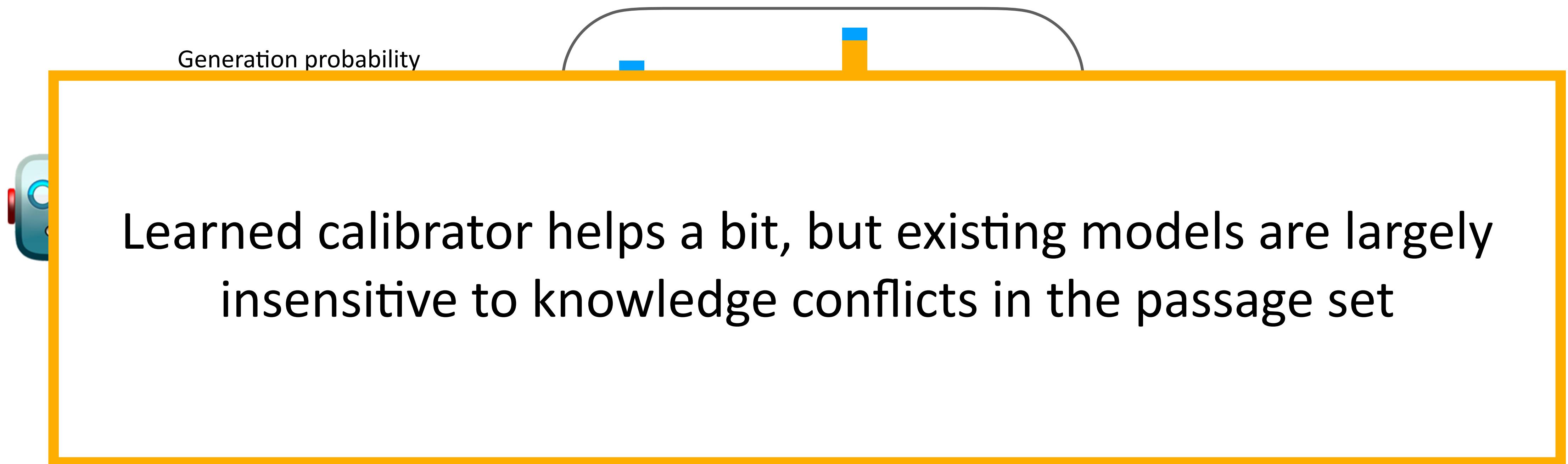


- How frequently model confidence **decrease** under simulated knowledge conflict?

50% in NQ, 70% in TriviaQA

Can Learned Calibrators Help?

- Train a simple classifier feature-correctness pairs from the target domain



6%/10% increase in confidence drop in NQ/TriviaQA

Summary: Analyzing Model Behaviors Under Knowledge Conflict

- Whether model rely on parametric vs. non parametric knowledge depends on the answer recall during its training
- Models tend to focus on top few passages, and use parametric knowledge when passages suggest multiple answers
- Model confidence score does not decrease significantly at knowledge conflicts

This Talk

Analysis

RQ1. Conflicts between memorized vs. retrieved knowledge

- Do QA models rely on parametric or non-parametric knowledge?

RQ2. Conflicts within retrieved knowledge

- What models do when retrieved passages suggest conflicting answers?

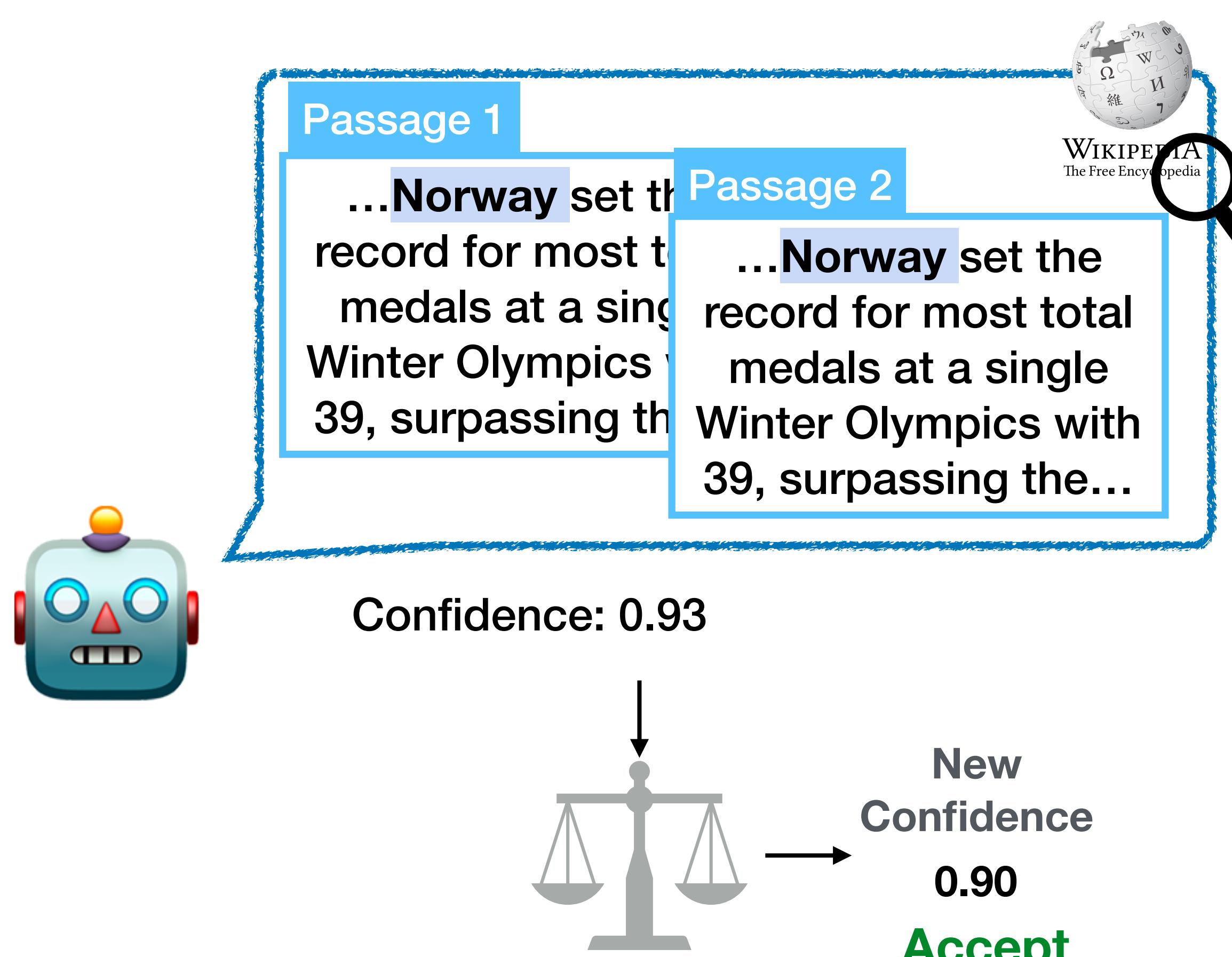
Solution

RQ3. Recalibration provided knowledge conflicts

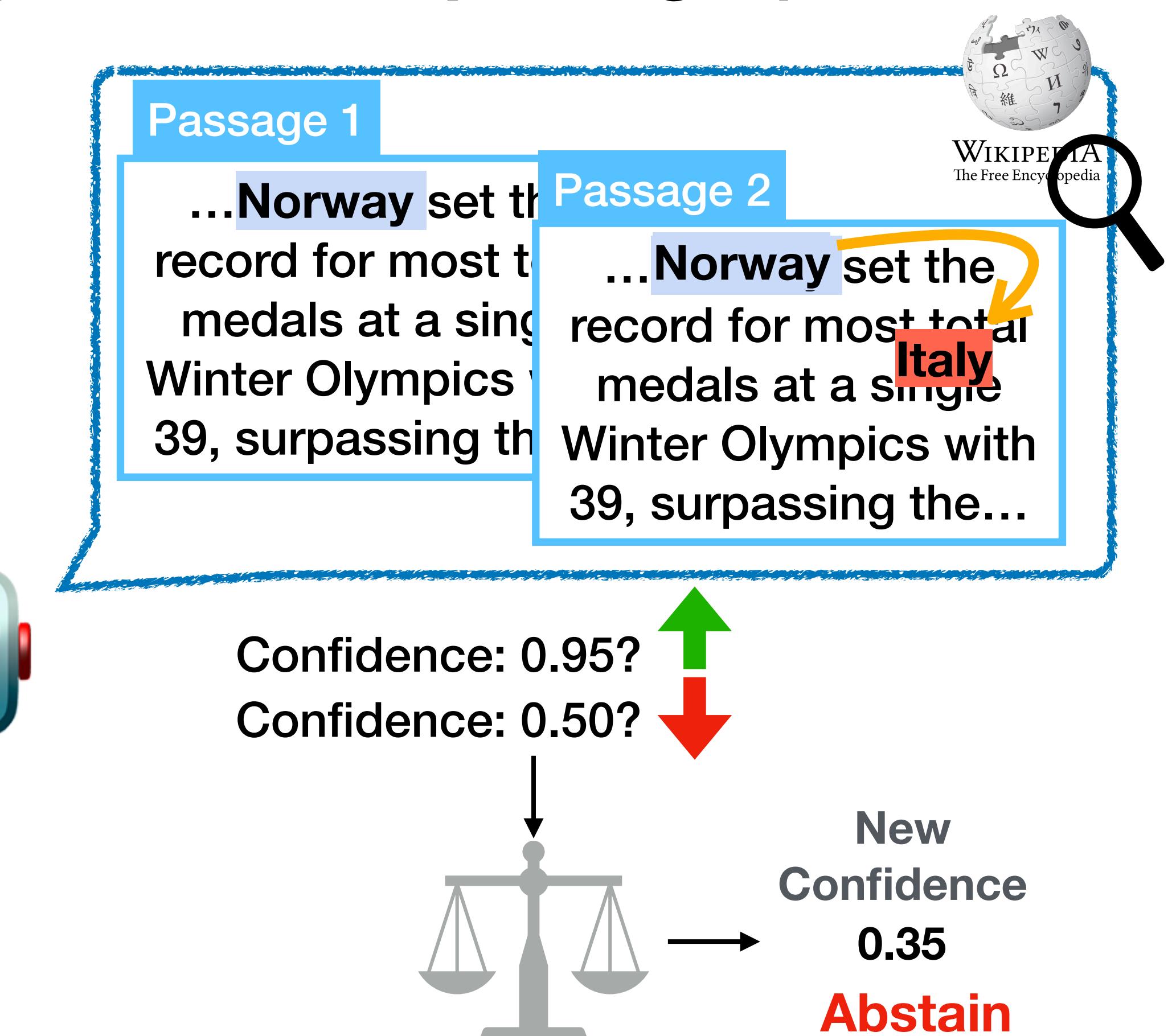
- Can we teach models to abstain from answering when knowledge source points multiple answers?

Re-calibrate models to be sensitive to knowledge conflict

- No Knowledge Conflict

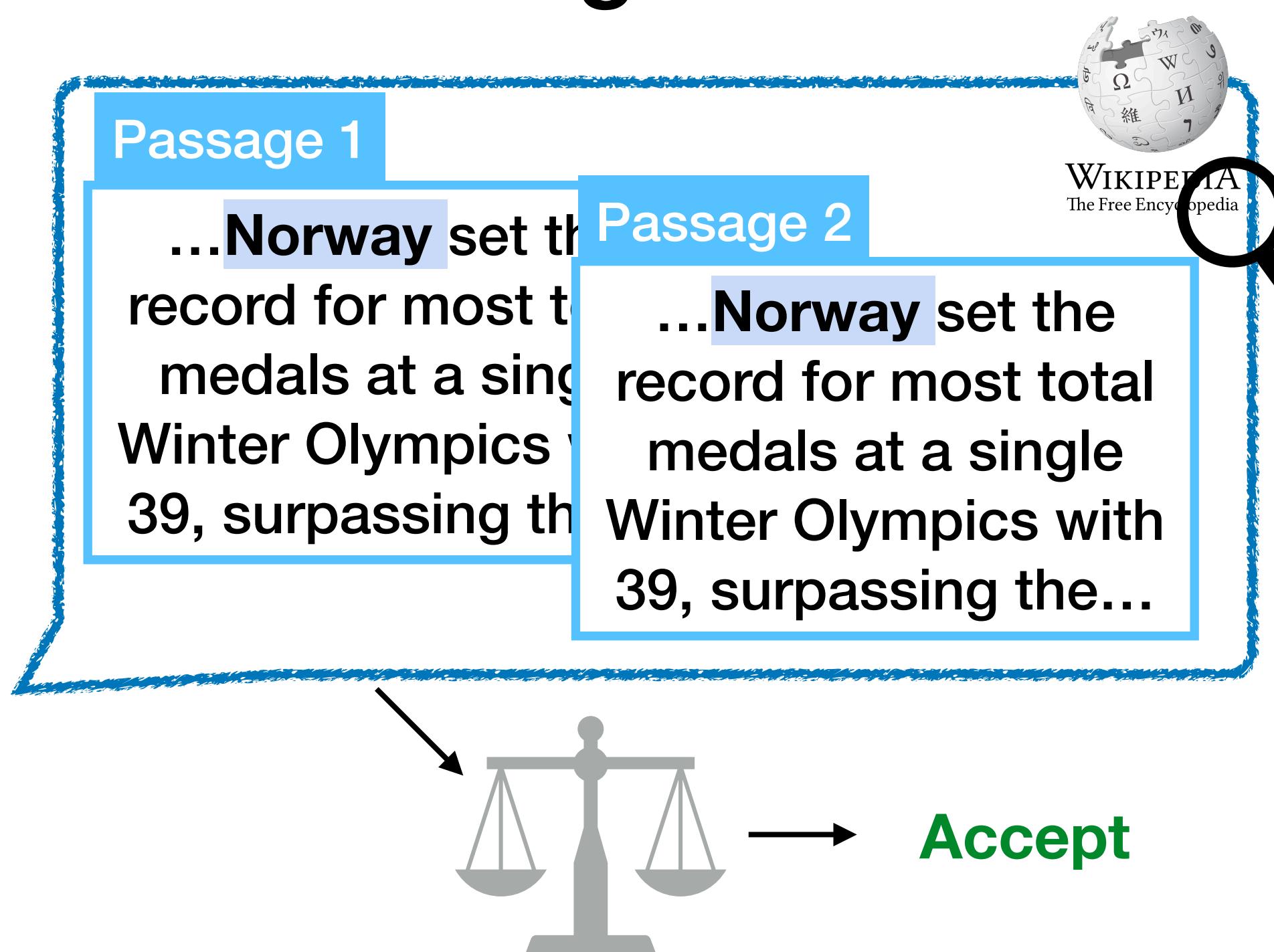


- Simulated Knowledge Conflict (perturb 50% passages)

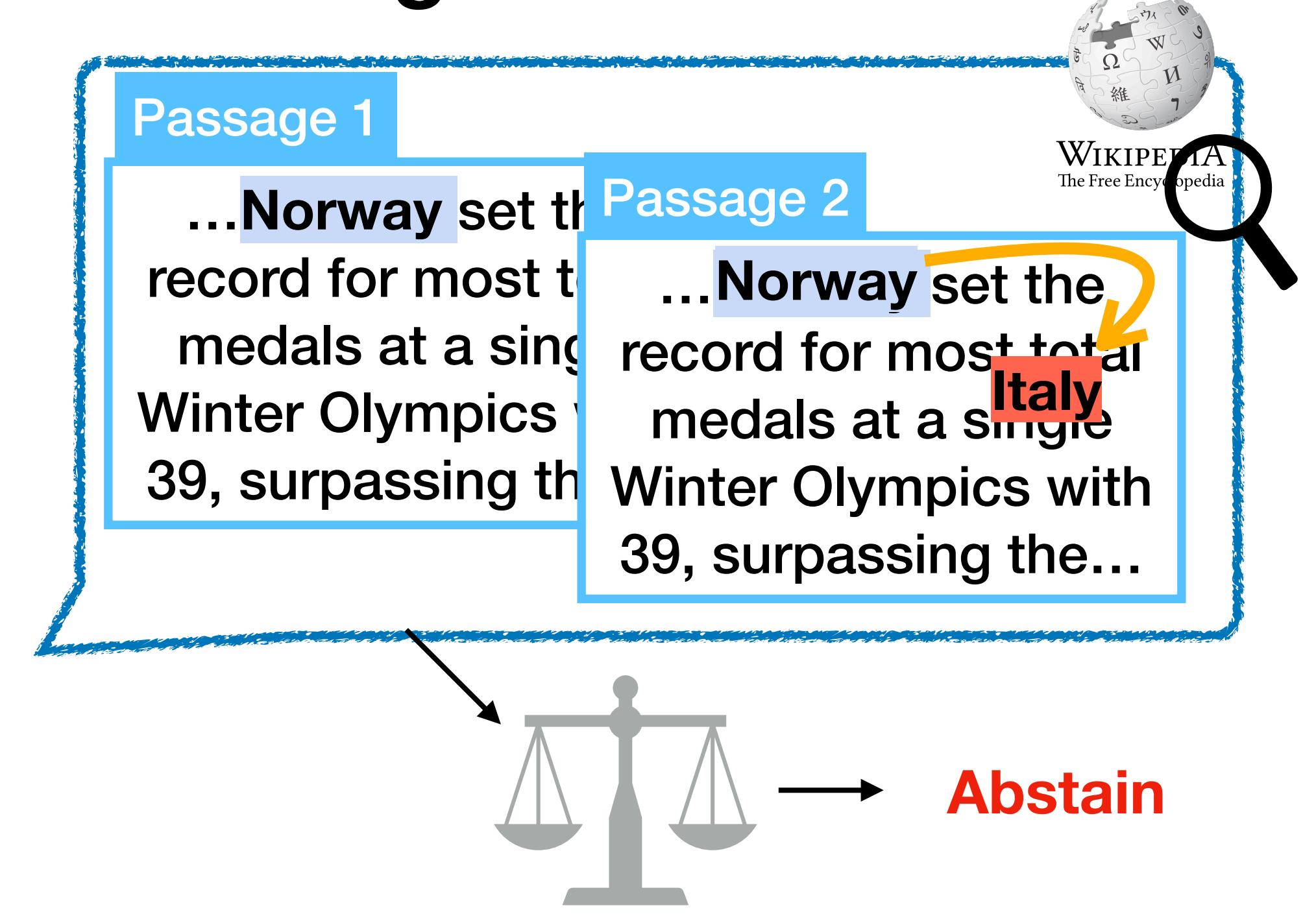


Evaluation Setting

- No Knowledge Conflict



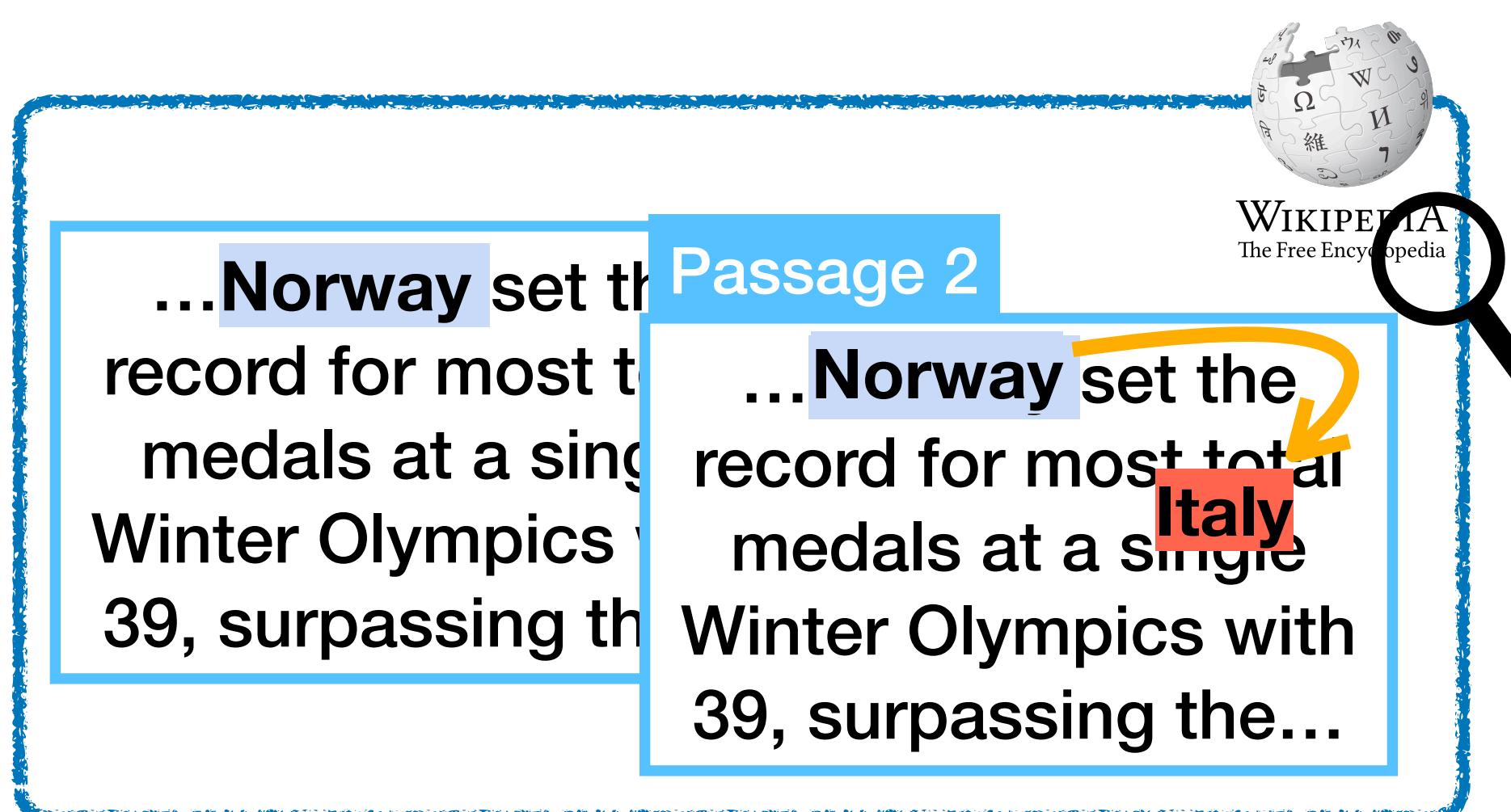
- Knowledge Conflict



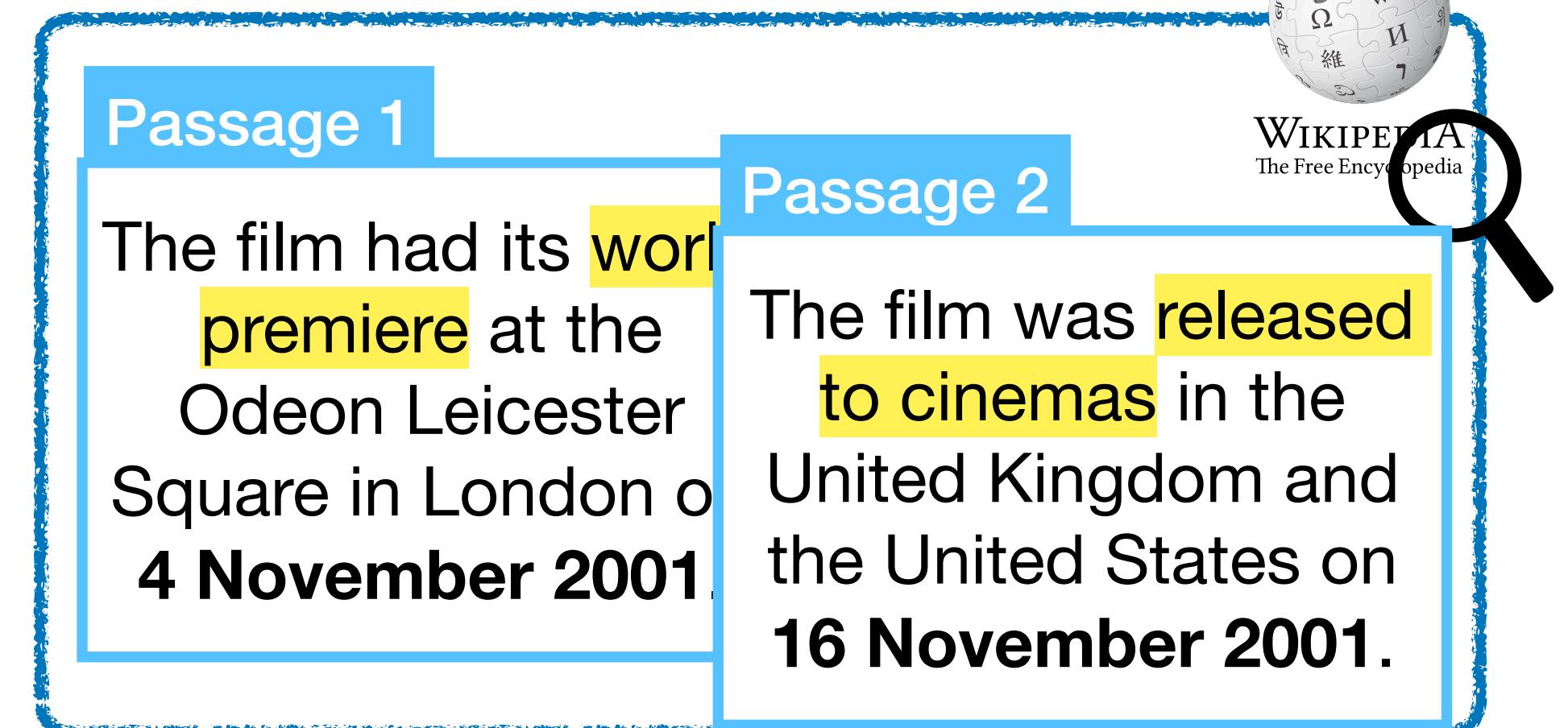
- To create evaluation dataset, we sample 50% of examples without knowledge conflict, 50% with knowledge conflict and report binary classification accuracy.

Types of knowledge conflicts

- Simulated Conflicts



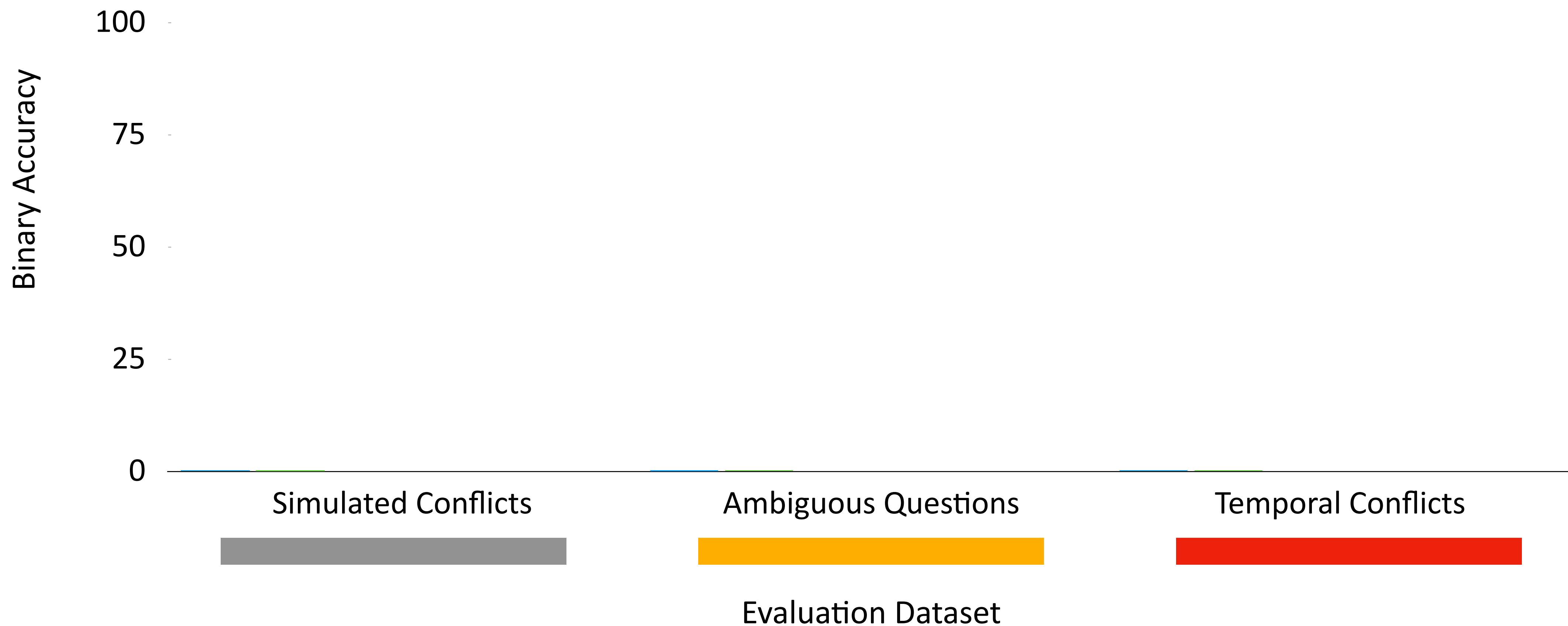
- Ambiguous Questions



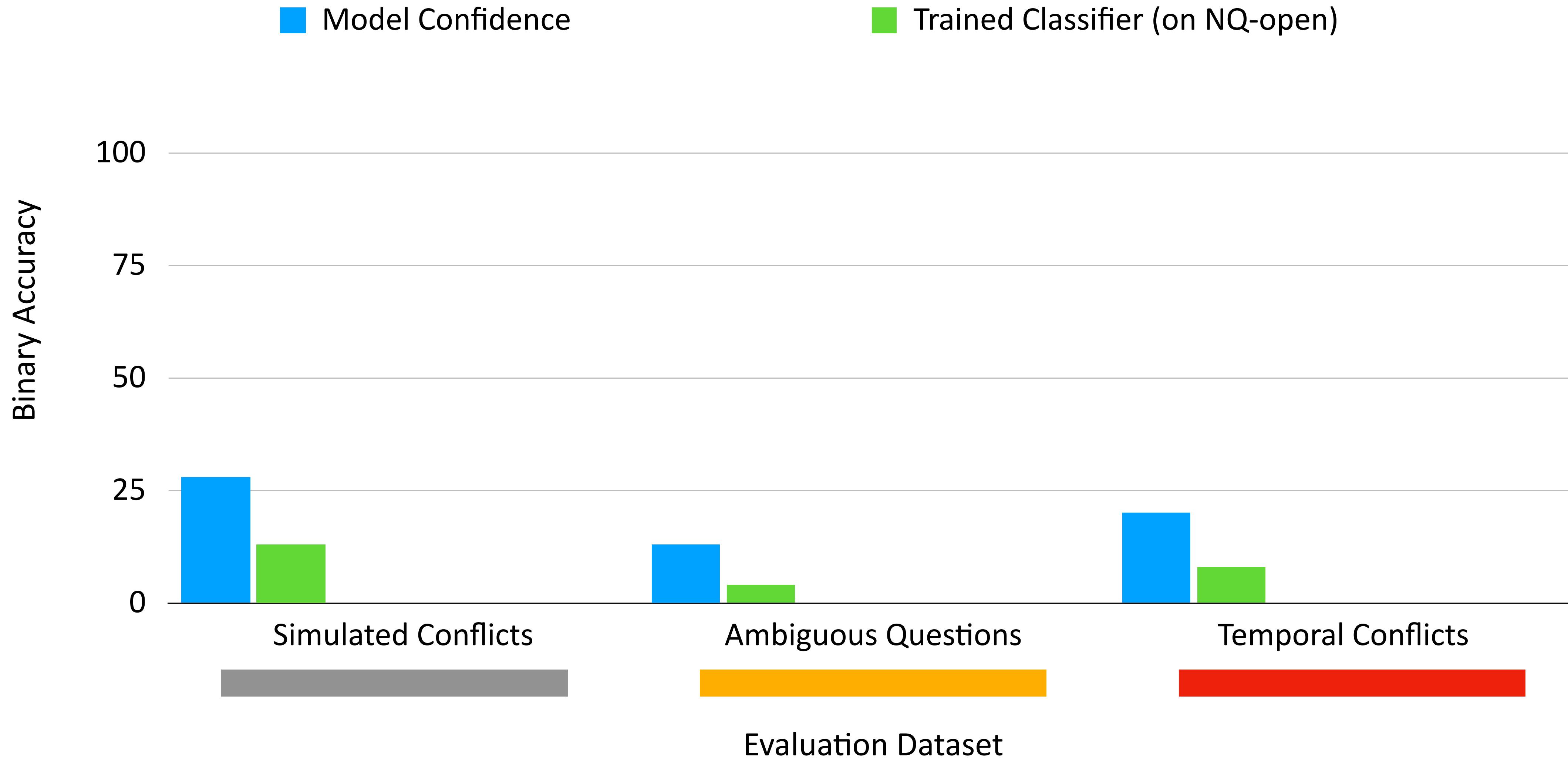
- Temporal Conflicts



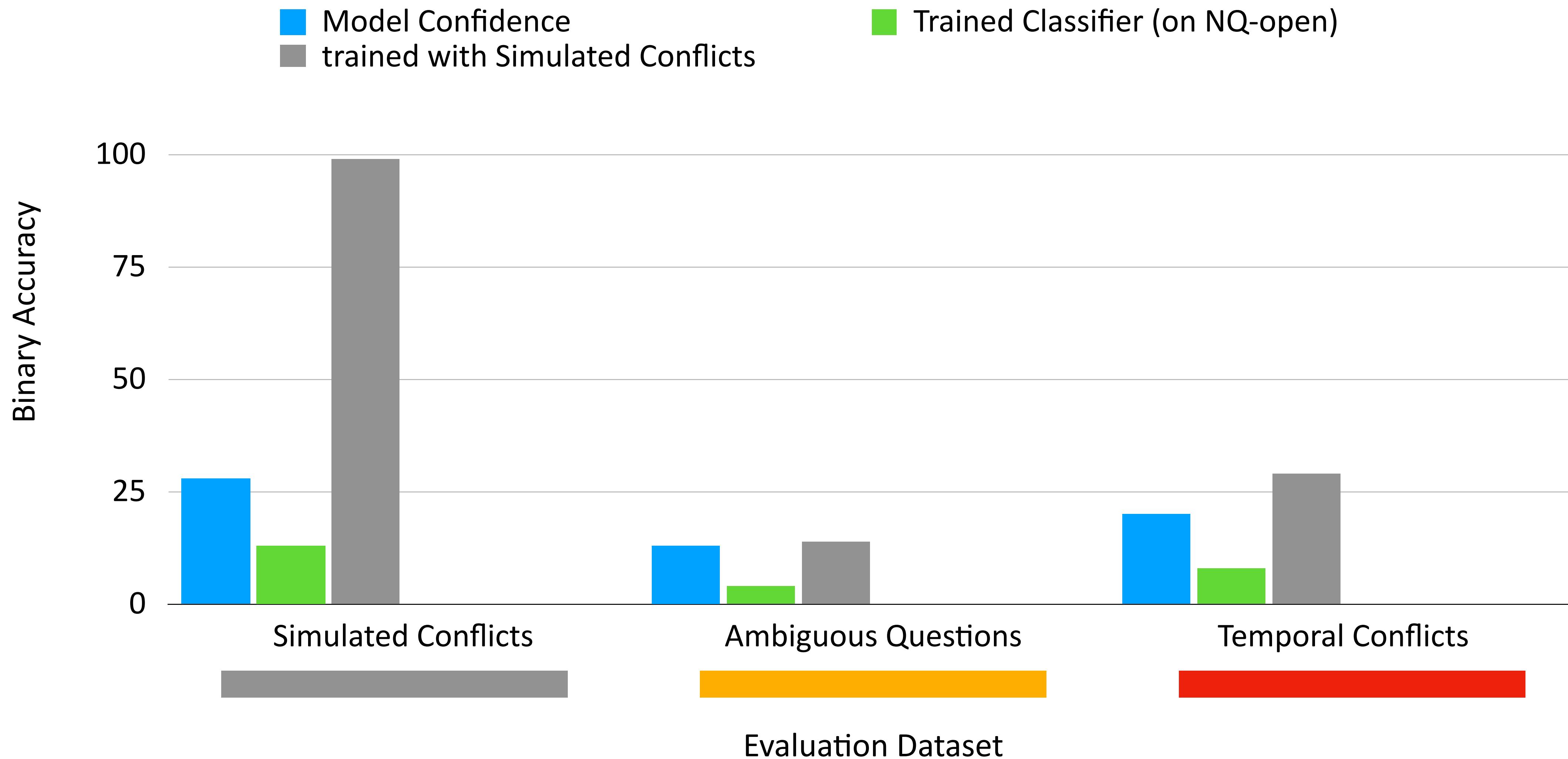
Result: Knowledge Conflict Detection



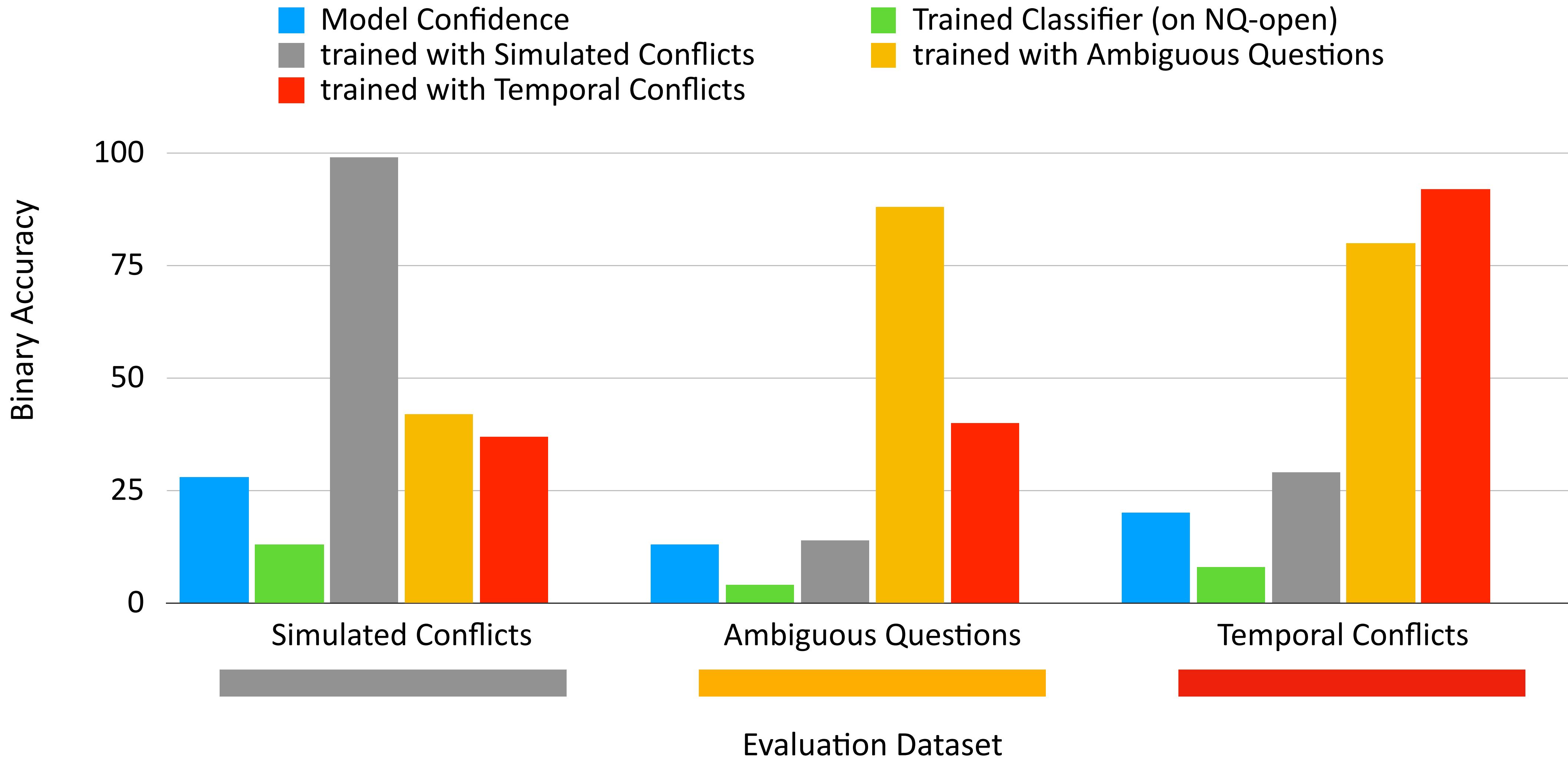
Result: Knowledge Conflict Detection



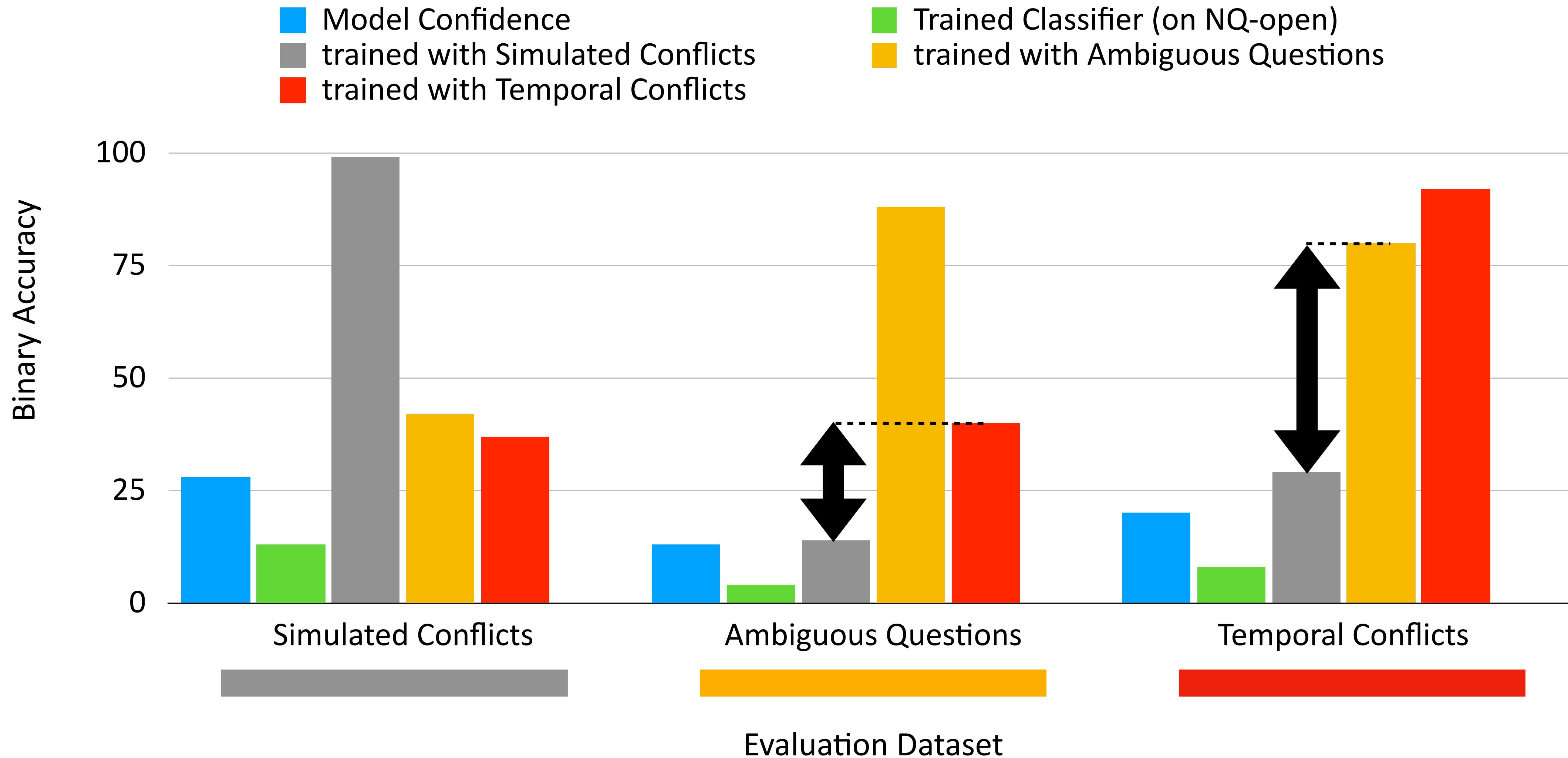
Result: Knowledge Conflict Detection



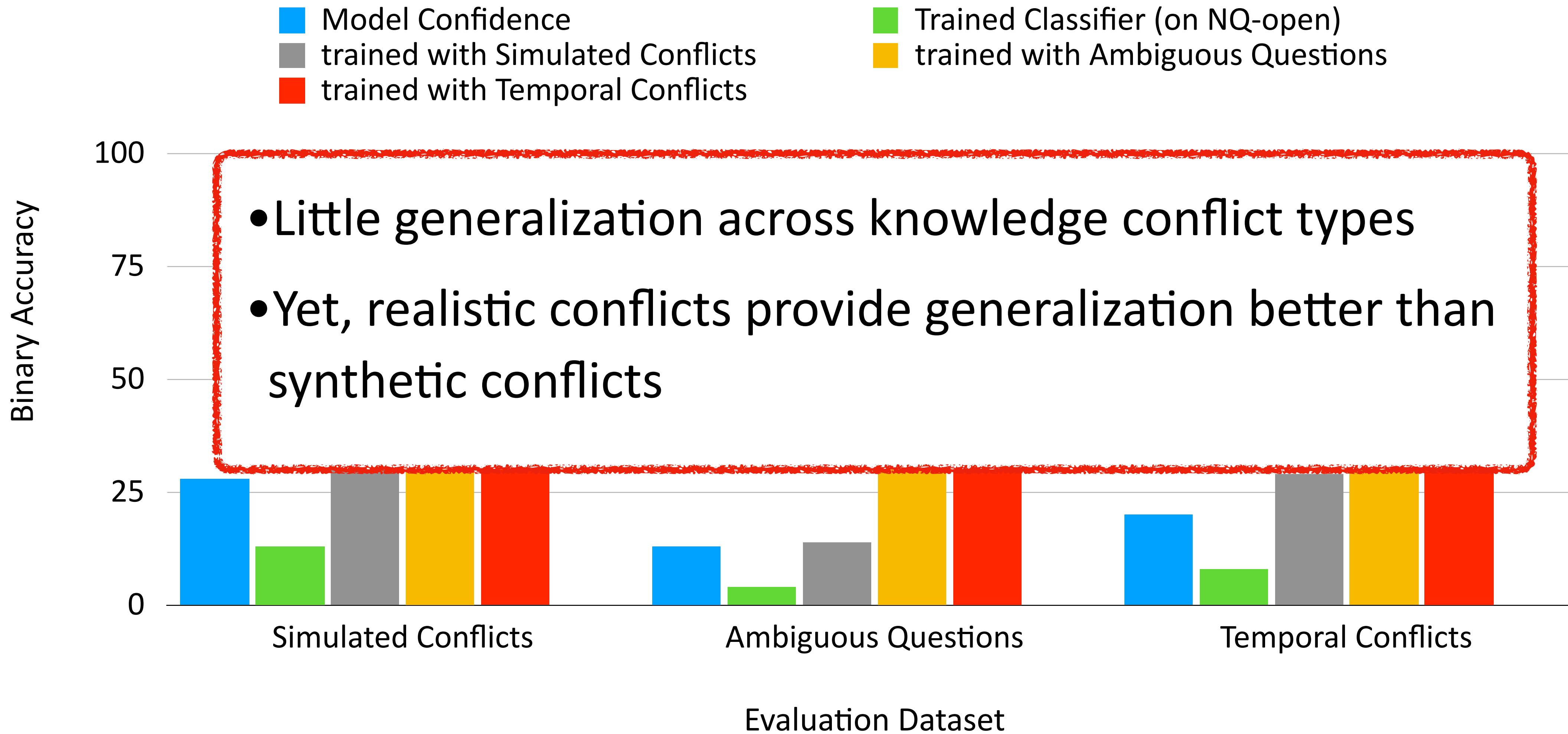
Result: Knowledge Conflict Detection



Result: Knowledge Conflict Detection



Result: Knowledge Conflict Detection



Conclusion

- We study realistic knowledge conflicts in open domain QA
- Models mostly rely on **retrieved passages**, when provided with a high-quality retriever
- Models rely on **most relevant** passages and use the **parametric knowledge** to break ties
- Model confidence is **not sensitive** to knowledge conflicts
- We can train a **separate calibrator** which detect knowledge conflicts, but with limited performance

Future Work

- Calibrators that could generalize to different types of knowledge conflicts
- Build a model capable of composing an answer consist of all the possible answers under knowledge conflicts
- Datasets containing real-world knowledge conflicts
- Studying knowledge conflicts with more complex QA settings
 - e.g. long-form QA (Fan et al., 2019), conditional QA (Sun et al., 2022)