ACL 2023 Matching Workshop

July 13, 2023

Editing Models

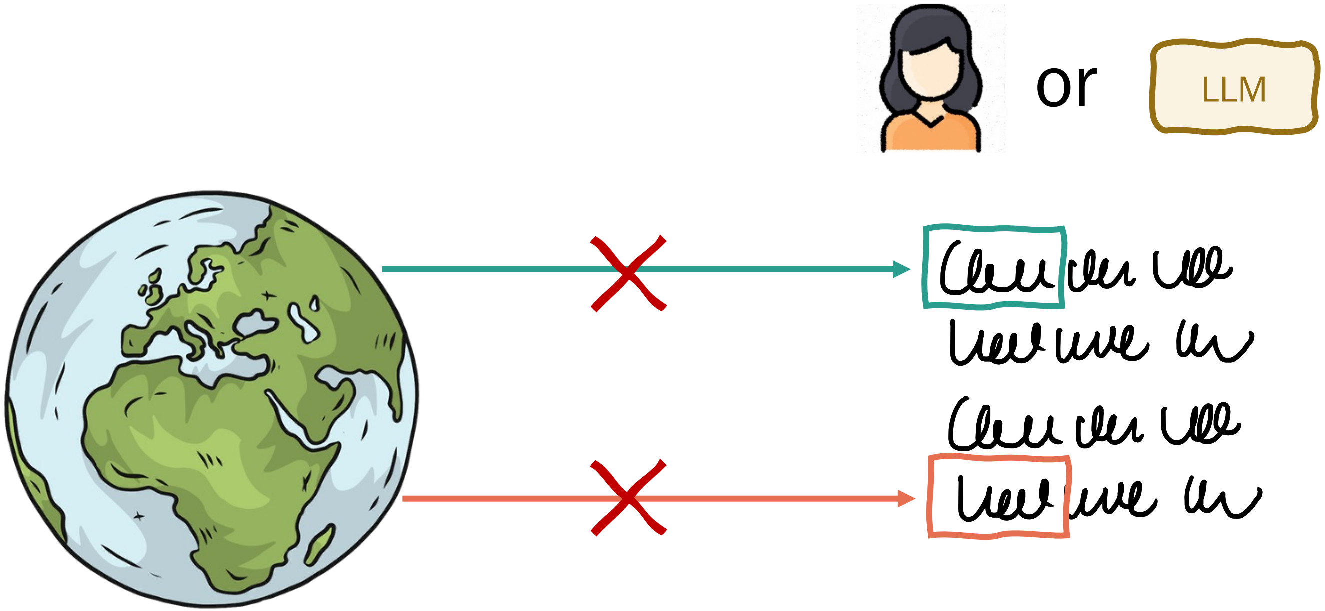
Updating Text with Up-to-date Information

UCI nlp

Sameer Singh

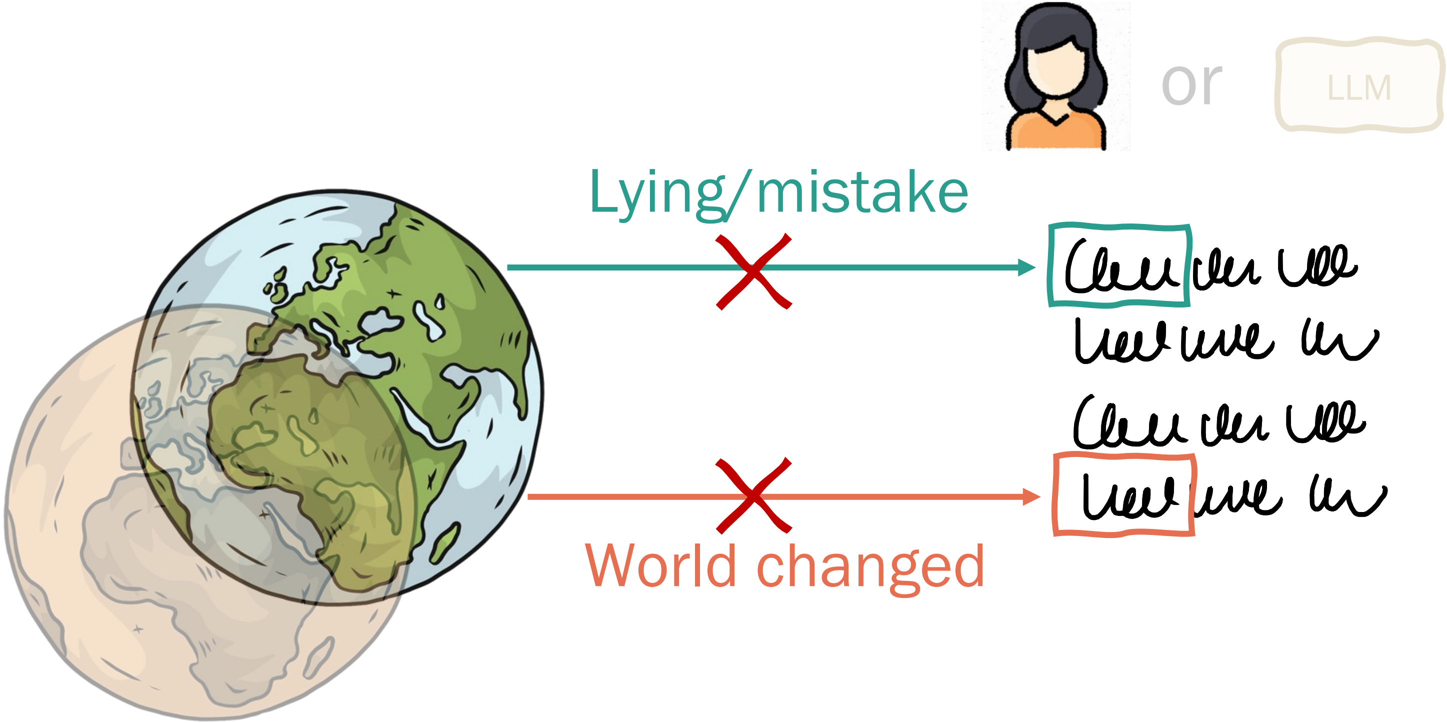
sameersingh.org

Facts in text should match the world



But it often does not!

Human Written Text



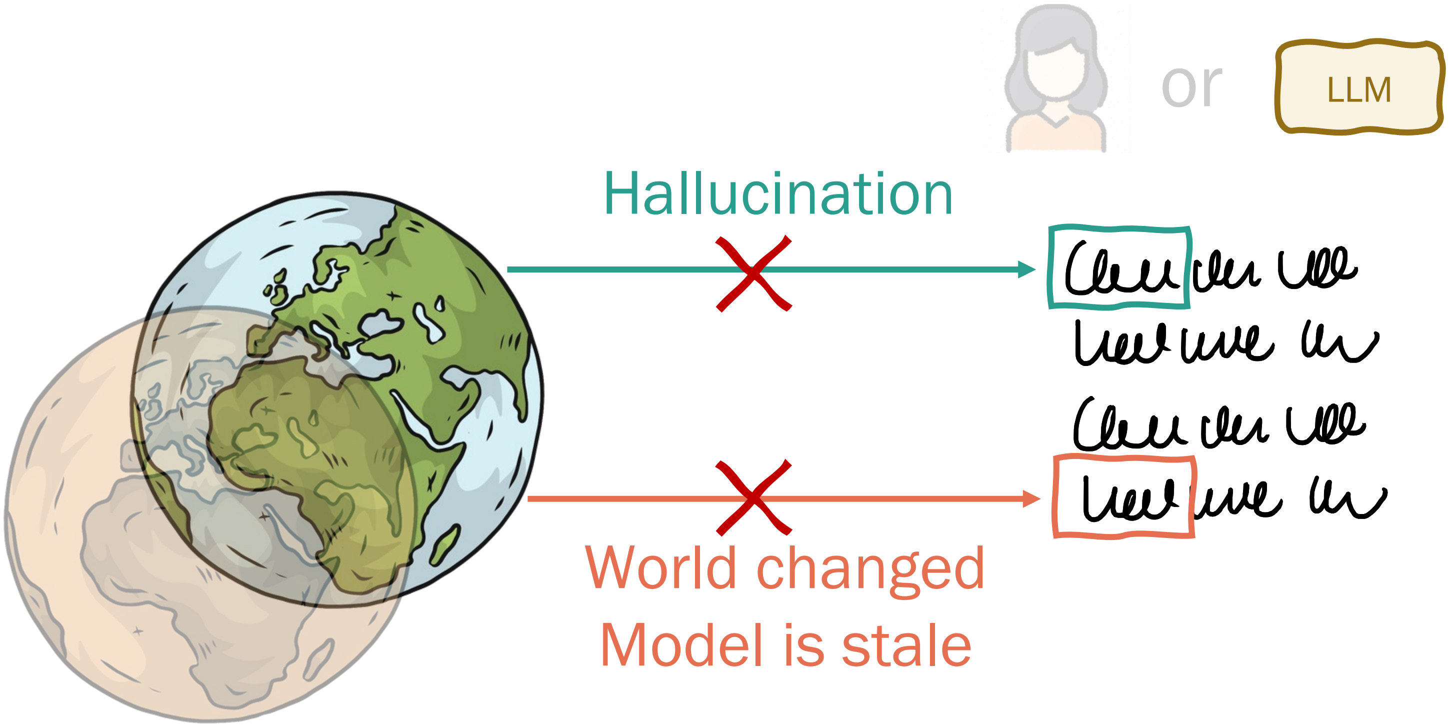
or

LLM

Lying/mistake

World changed

Language Model Generated Text



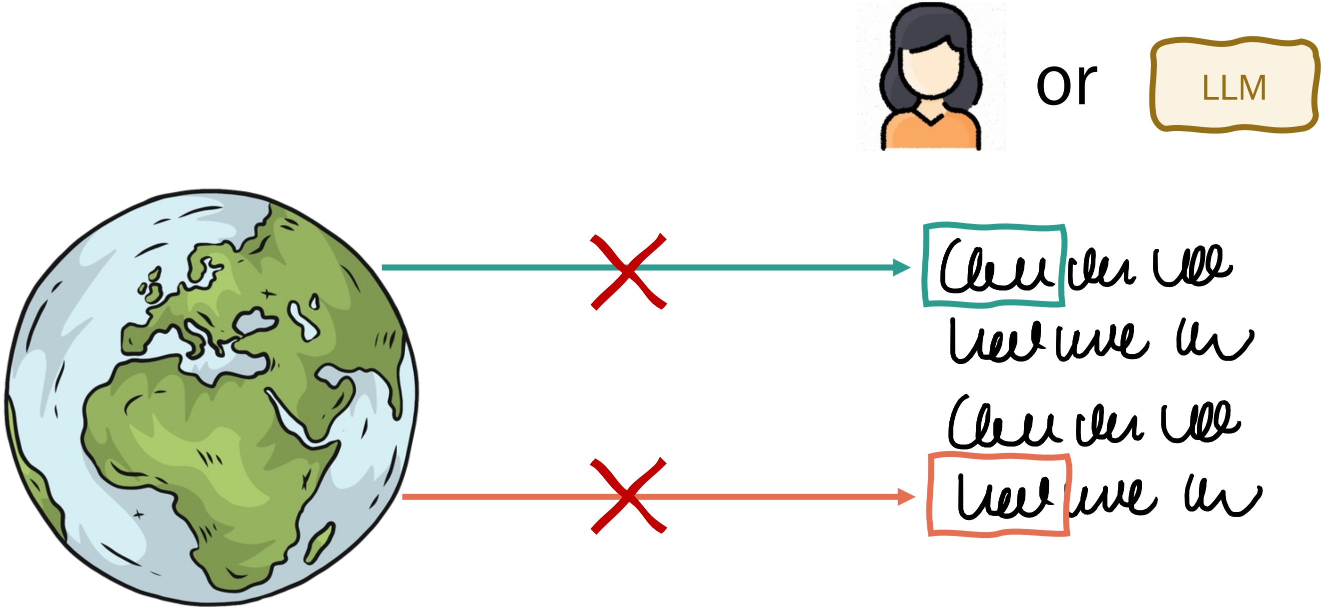
or

LLM

Hallucination

World changed
Model is stale

Facts in text should match the world

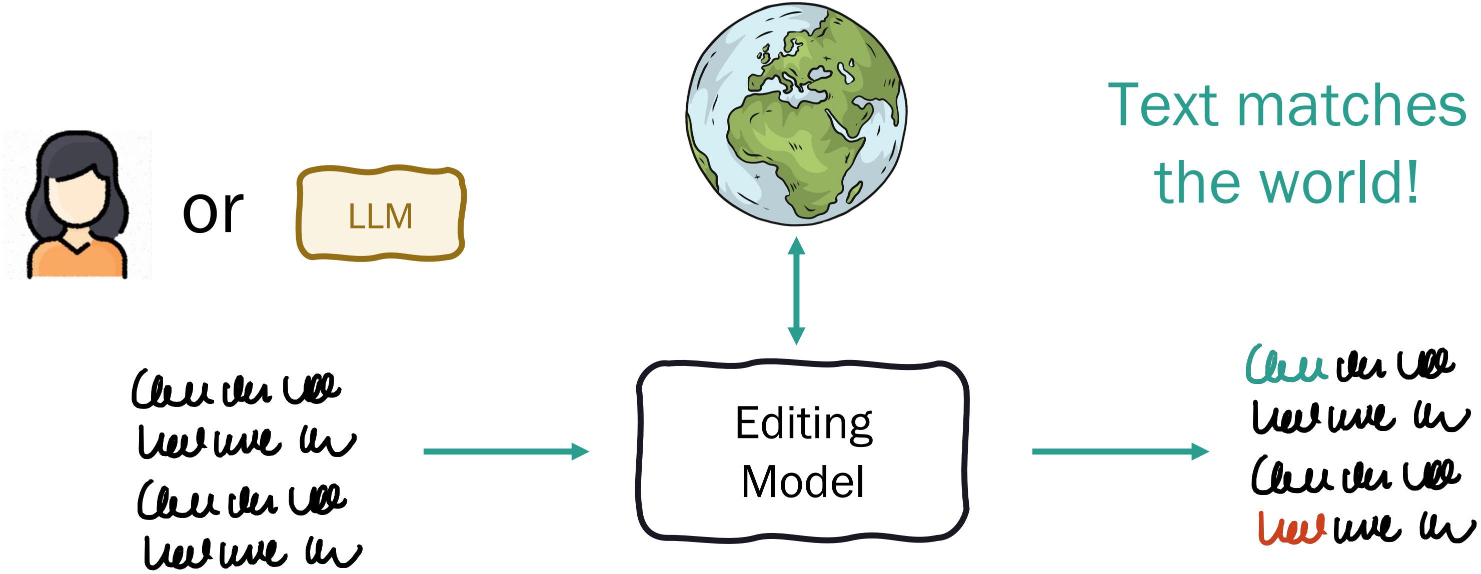


or

LLM

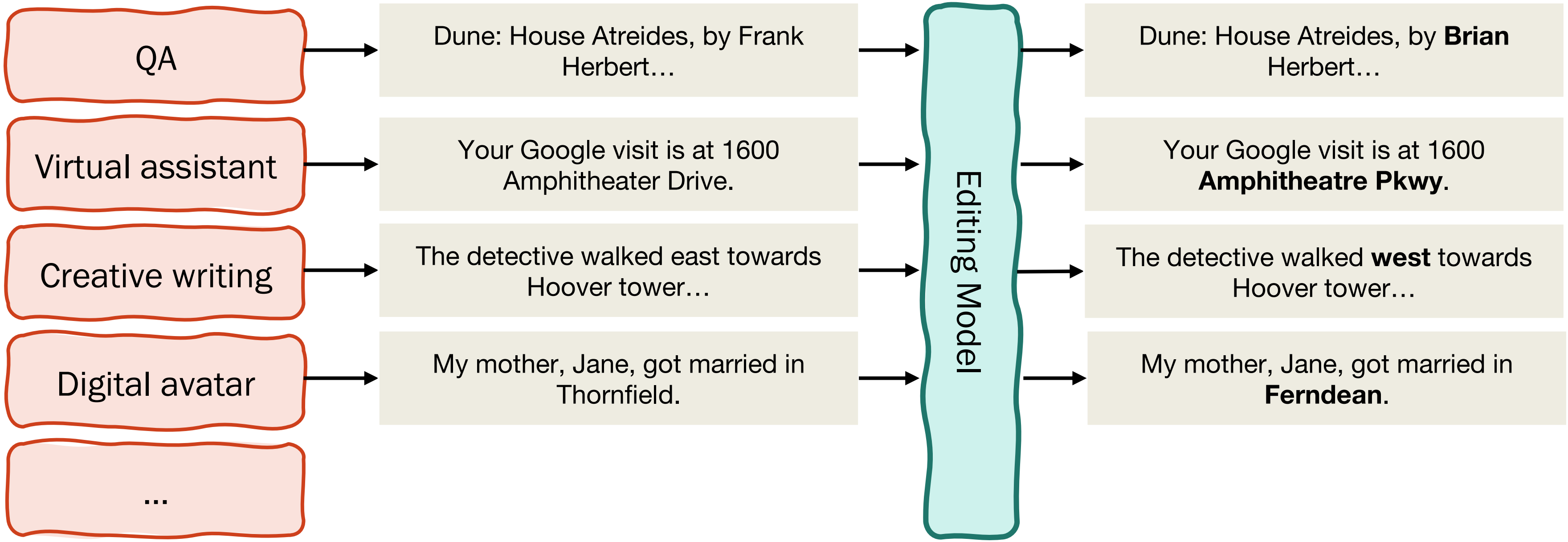
Update text to match the information

Task: Make text match information



Text matches the world!

Applications for LLM generated text



QA

Virtual assistant

Creative writing

Digital avatar

…

Dune: House Atreides, by Frank Herbert…

Your Google visit is at 1600 Amphitheater Drive.

The detective walked east towards Hoover tower…

My mother, Jane, got married in Thornfield.

Editing Model

Dune: House Atreides, by **Brian** Herbert…

Your Google visit is at 1600 **Amphitheatre Pkwy.**

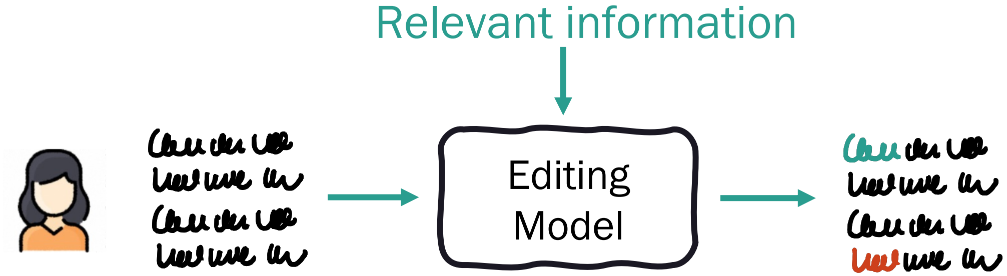
The detective walked **west** towards Hoover tower…

My mother, Jane, got married in **Ferndean**.

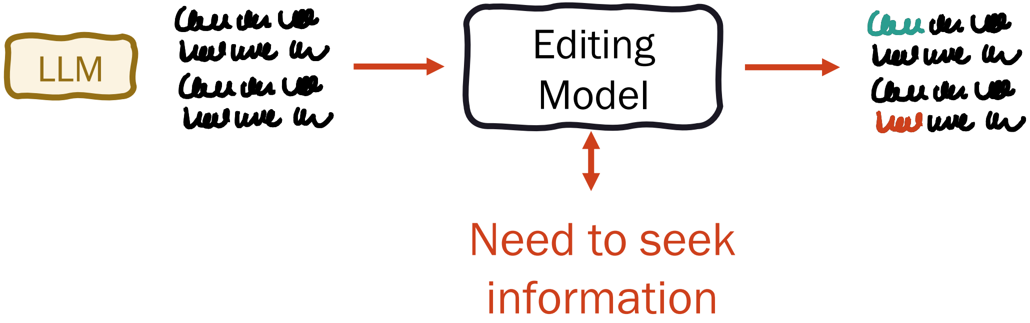
Keep this small, efficient

Outline

Part 1

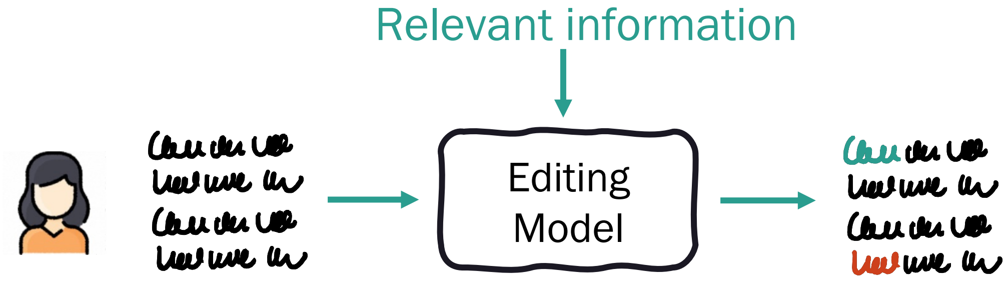


Part 2



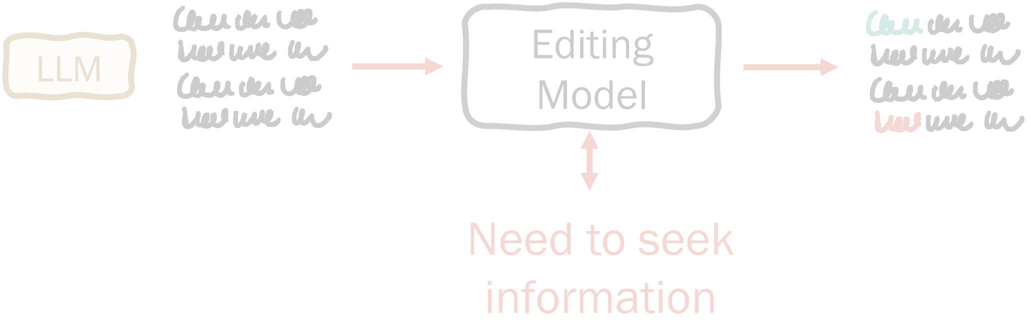
Outline

Part 1

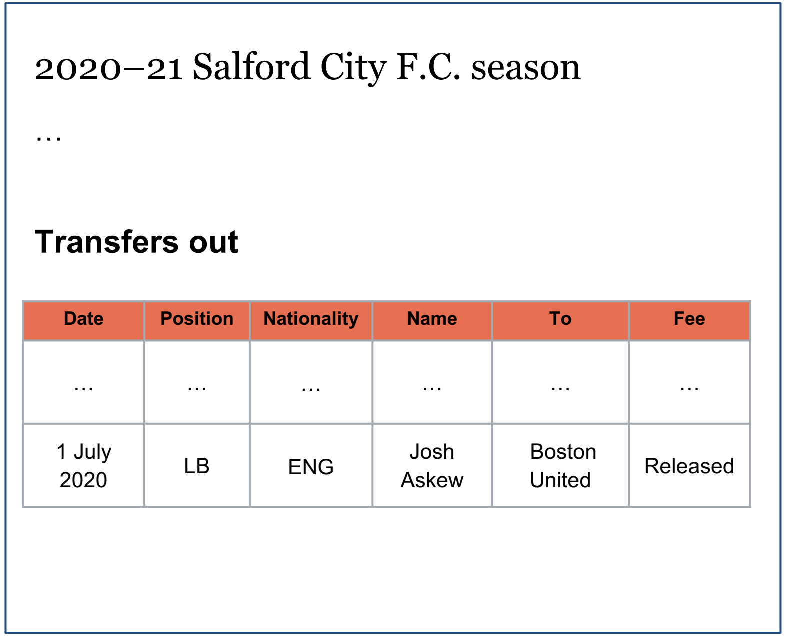


FRUIT: Faithfully Reflecting Updated Information in Text
Robert L. Logan IV, Alexandre Passos, Sameer Singh, Ming-Wei Chang
NAACL 2022 **BEST TASK PAPER AWARD**

Part 2



Motivation



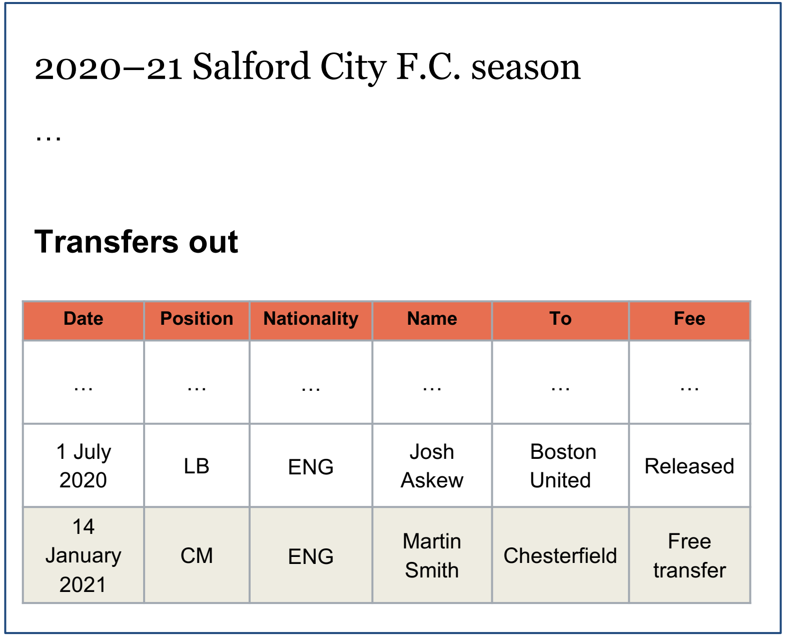
2020–21 Salford City F.C. season

…

Transfers out

Date	Position	Nationality	Name	To	Fee
…	…	…	…	…	…
1 July 2020	LB	ENG	Josh Askew	Boston United	Released

Motivation



2020–21 Salford City F.C. season

…

Transfers out

Date	Position	Nationality	Name	To	Fee
…	…	…	…	…	…
1 July 2020	LB	ENG	Josh Askew	Boston United	Released
14 January 2021	CM	ENG	Martin Smith	Chesterfield	Free transfer

Motivation

2020–21 Salford City F.C. season

…

Transfers out

Date	Position	Nationality	Name	To	Fee
…	…	…	…	…	…
1 July 2020	LB	ENG	Josh Askew	Boston United	Released
14 January 2021	CM	ENG	Martin Smith	Chesterfield	Free transfer

Martin Smith (footballer, born 1995)

Martin Smith (born 2 October 1995) is an English footballer who plays for club Salford City.

Motivation

2020–21 Salford City F.C. season

…

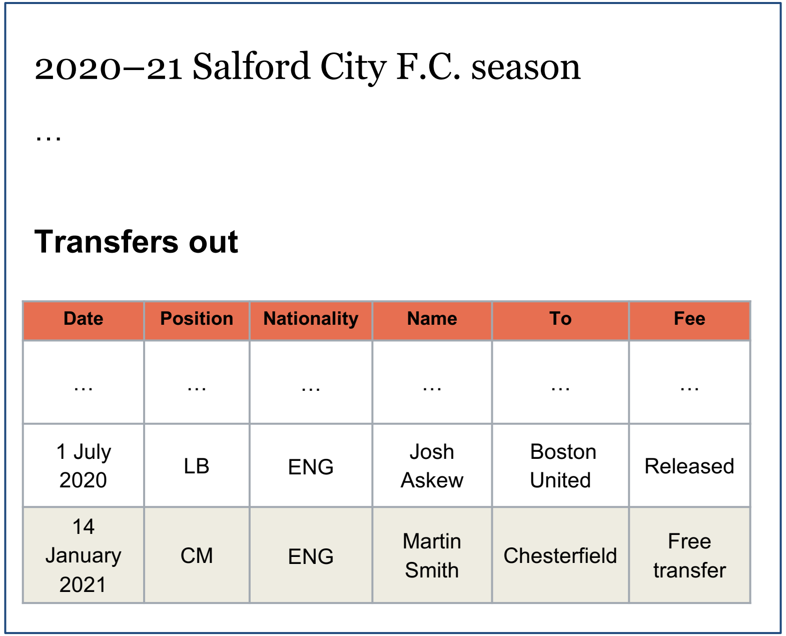
Transfers out

Date	Position	Nationality	Name	To	Fee
…	…	…	…	…	…
1 July 2020	LB	ENG	Josh Askew	Boston United	Released
14 January 2021	CM	ENG	Martin Smith	Chesterfield	Free transfer

Martin Smith (footballer, born 1995)

Martin Smith (born 2 October 1995) is an English footballer who plays ~~for club Salford City~~.

Motivation

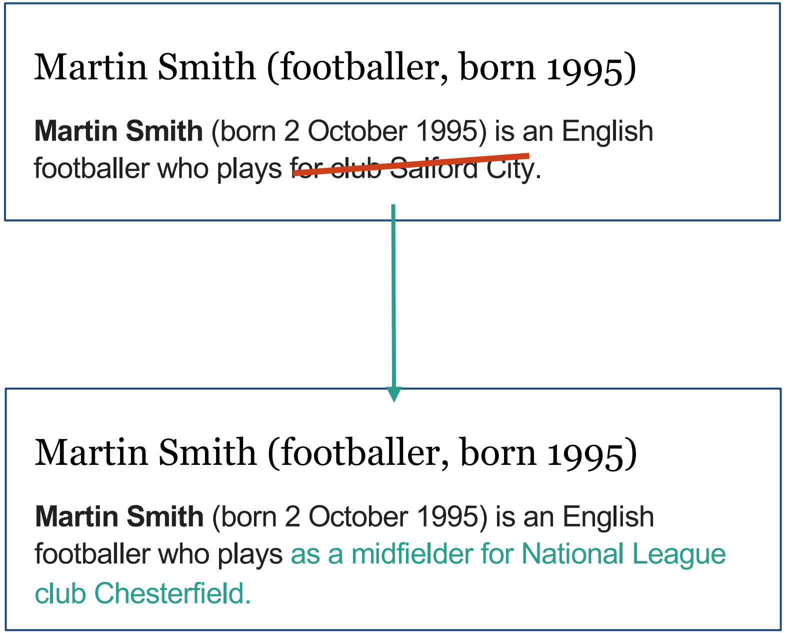


2020–21 Salford City F.C. season

…

Transfers out

Date	Position	Nationality	Name	To	Fee
…	…	…	…	…	…
1 July 2020	LB	ENG	Josh Askew	Boston United	Released
14 January 2021	CM	ENG	Martin Smith	Chesterfield	Free transfer



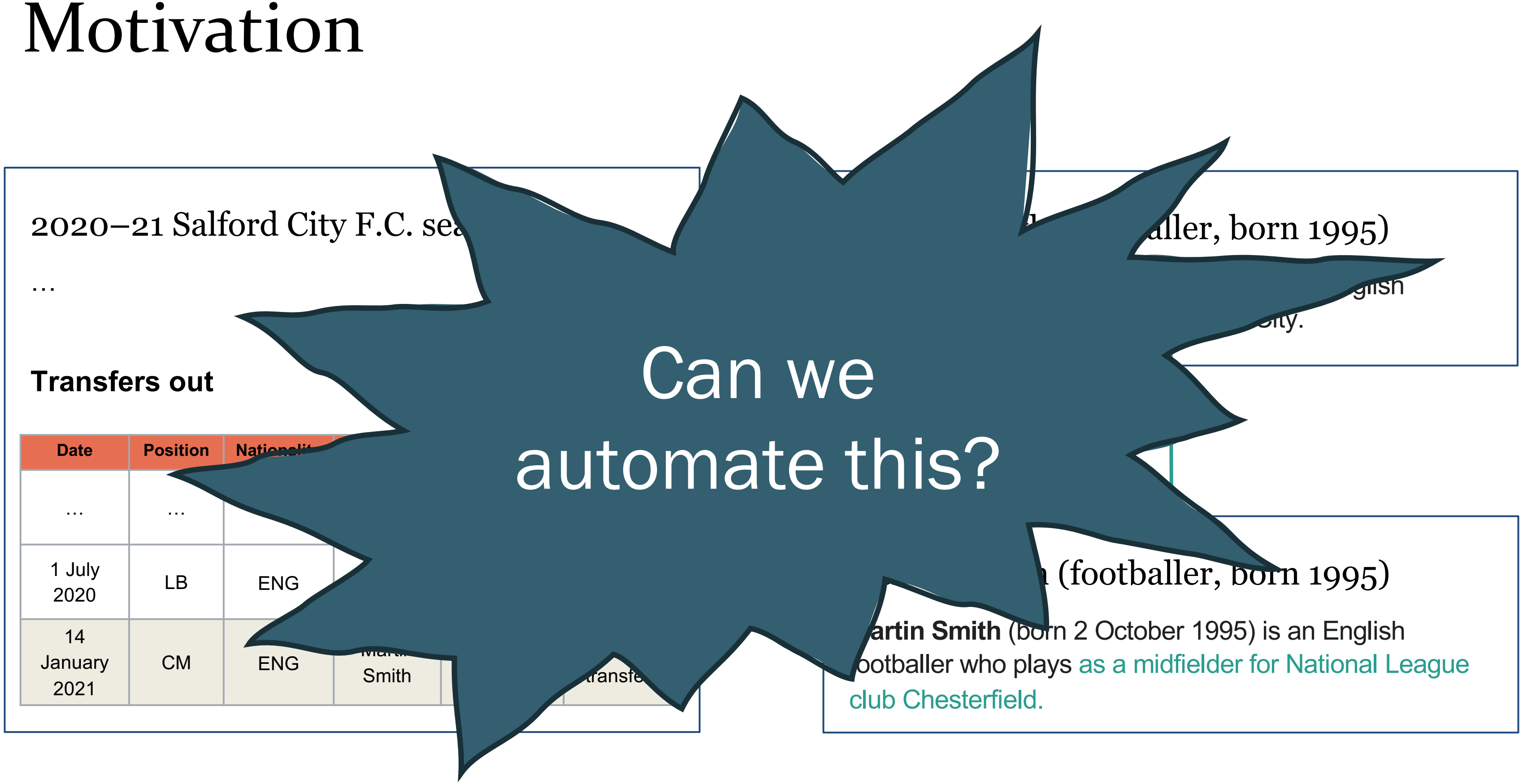
Martin Smith (footballer, born 1995)

Martin Smith (born 2 October 1995) is an English footballer who plays ~~for club Salford City~~.

Martin Smith (footballer, born 1995)

Martin Smith (born 2 October 1995) is an English footballer who plays as a midfielder for National League club Chesterfield.

Motivation



2020–21 Salford City F.C. sea

…

Transfers out

Date	Position	Nationalit
…	…	
1 July 2020	LB	ENG
14 January 2021	CM	ENG

Smith

ransfe

aller, born 1995)

glish

City.

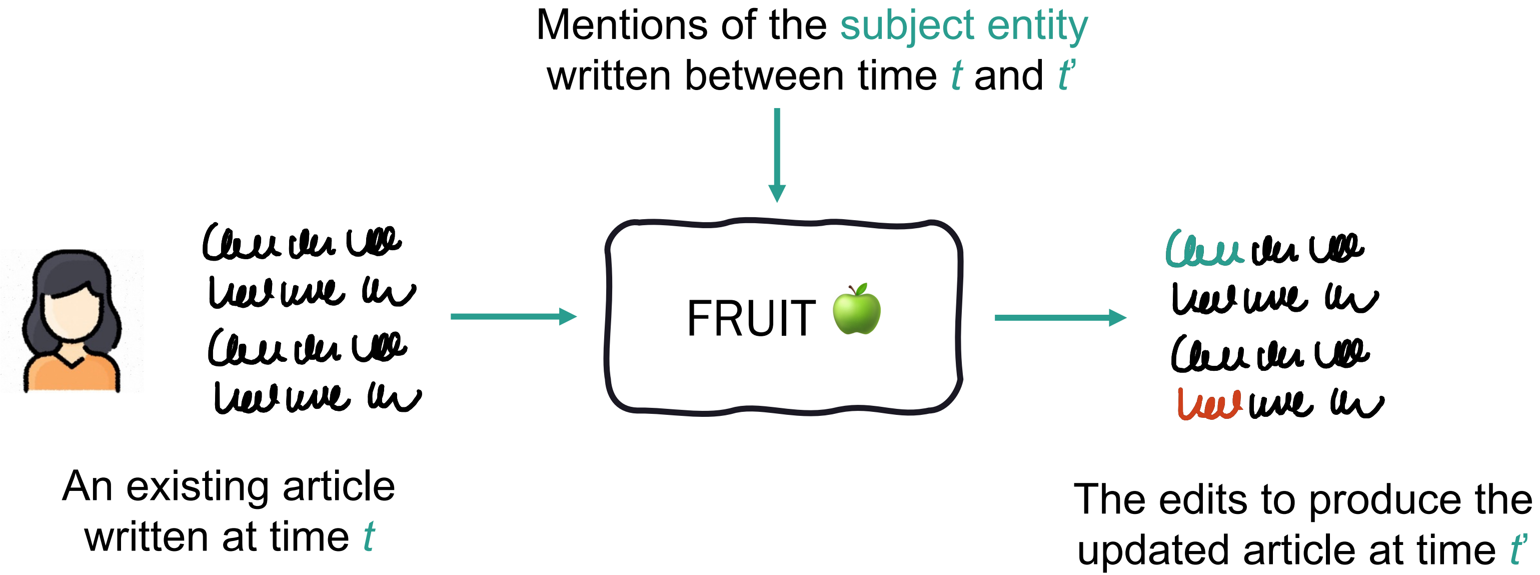
(footballer, born 1995)

artin Smith (born 2 October 1995) is an English ootballer who plays as a midfielder for National League club Chesterfield.

Can we automate this?

FRUIT🍏

Faithfully reflecting updated information in text (FRUIT).



Mentions of the subject entity written between time *t* and *t’*

FRUIT 🍏

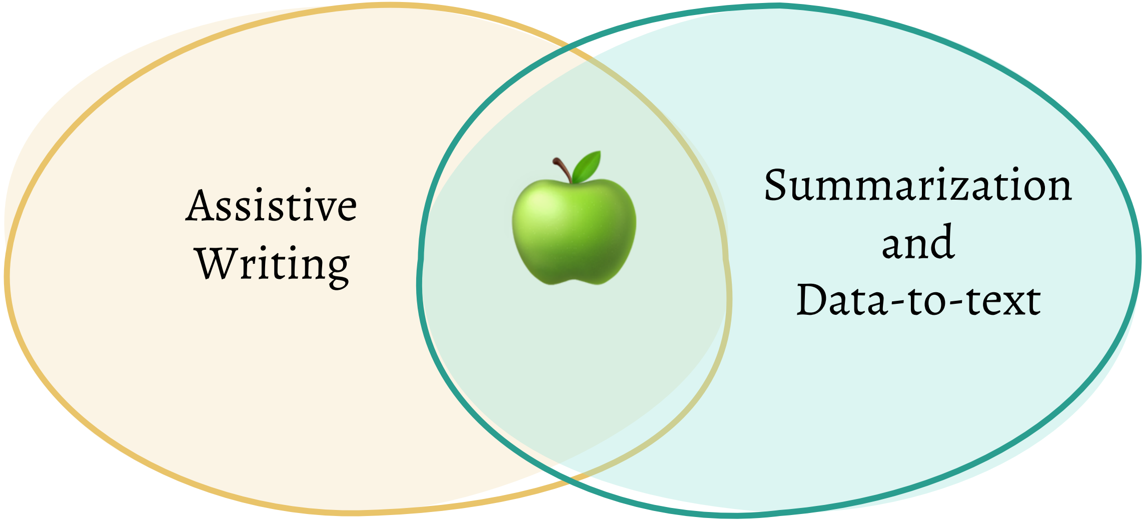
An existing article written at time *t*

The edits to produce the updated article at time *t’*

Why This is Challenging

Writing assistants focus on grammar/autocomplete capabilities instead of making grounded edits.

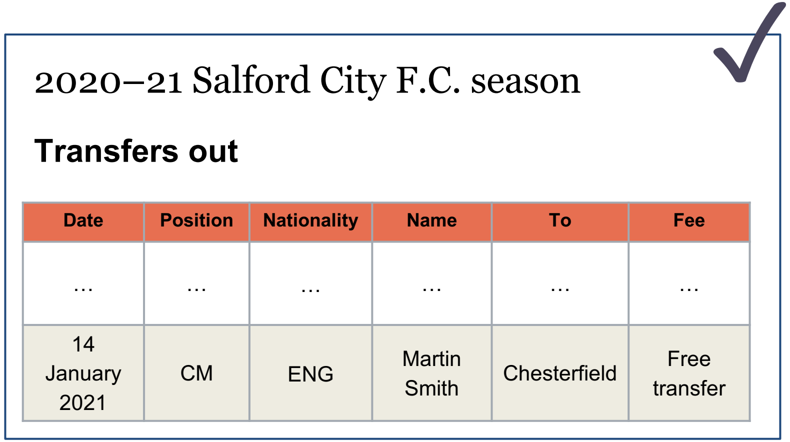
Data-to-text and summarization do not focus on updating existing text.



Why This is Challenging

Generated text needs to be faithful to original text and new evidence, preferring evidence when there is conflict.

Requires contrastive fact verification!

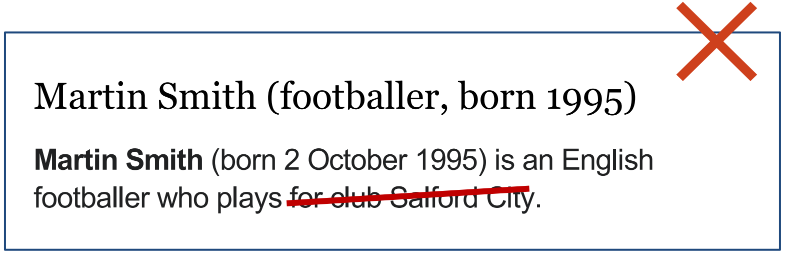


✓

2020–21 Salford City F.C. season

Transfers out

Date	Position	Nationality	Name	To	Fee
…	…	…	…	…	…
14 January 2021	CM	ENG	Martin Smith	Chesterfield	Free transfer



✗

Martin Smith (footballer, born 1995)

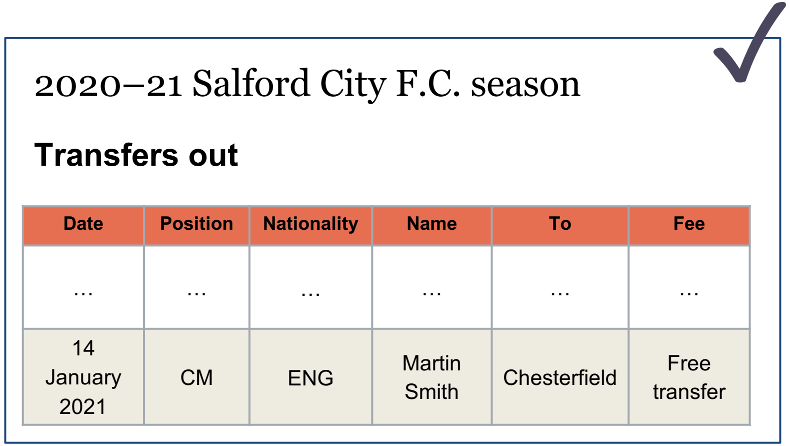
Martin Smith (born 2 October 1995) is an English footballer who plays ~~for club Salford City~~.

[1] Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence (Schuster et al., NAACL 2021)

Why This is Challenging

Editing requires using context instead of parametric knowledge.

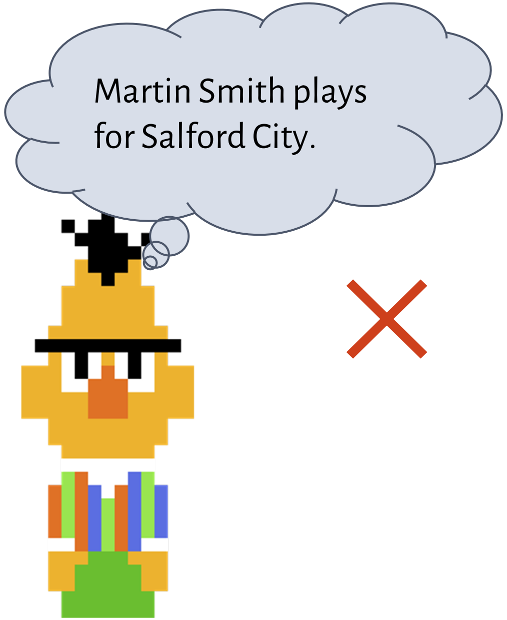
Pretrained models find this difficult.



2020–21 Salford City F.C. season

Transfers out

Date	Position	Nationality	Name	To	Fee
…	…	…	…	…	…
14 January 2021	CM	ENG	Martin Smith	Chesterfield	Free transfer



[1] Hurdles to Progress in Long-form Question Answering (Krishna et al., NAACL 2021)
[2]Entity-Based Knowledge Conflicts in Question Answering (Longpre et al., EMNLP 2021)

Contributions

- “Silver” Training Data: Weak supervision from Wikipedia.
- Gold Evaluation Data: Human annotated evaluation dataset.
- Models: Strong T5-based baselines
- Analysis: Error analysis.

Data Collection

2020

WIKIPEDIA
The Free Encyclopedia

-

2019

WIKIPEDIA
The Free Encyclopedia

=

Updated Article Intros & Mentions

Joe Biden (2019)
Joseph Robinette Biden Jr. (; born November 20, 1942) is an American politician who served as the 47th vice president of the United States from 2009 to 2017.

Joe Biden (2020)
Joseph Robinette Biden Jr. (; born November 20, 1942) is an American politician who is the 46th and current president of the United States.

Filtering Heuristic

Substance vs. Style

Substantive edits can be identified by the addition of new entity mentions

Filtering Heuristic

Substance vs. Style

Substantive edits can be identified by the addition of new entity mentions

Martin Smith (born 2 October 1995) is an English footballer who plays for club Salford City.

→

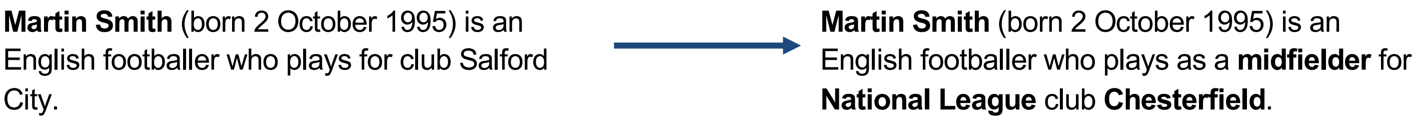
Martin Smith (born 2 October 1995) is an English footballer. He currently plays for club Salford City.

No Added Entities = Stylistic Edit

Filtering Heuristic

Substance vs. Style

Substantive edits can be identified by the addition of new entity mentions



Martin Smith (born 2 October 1995) is an English footballer who plays for club Salford City.

Martin Smith (born 2 October 1995) is an English footballer who plays as a **midfielder** for **National League** club **Chesterfield**.

Added Entities = Substantive Edit

Filtering Heuristic

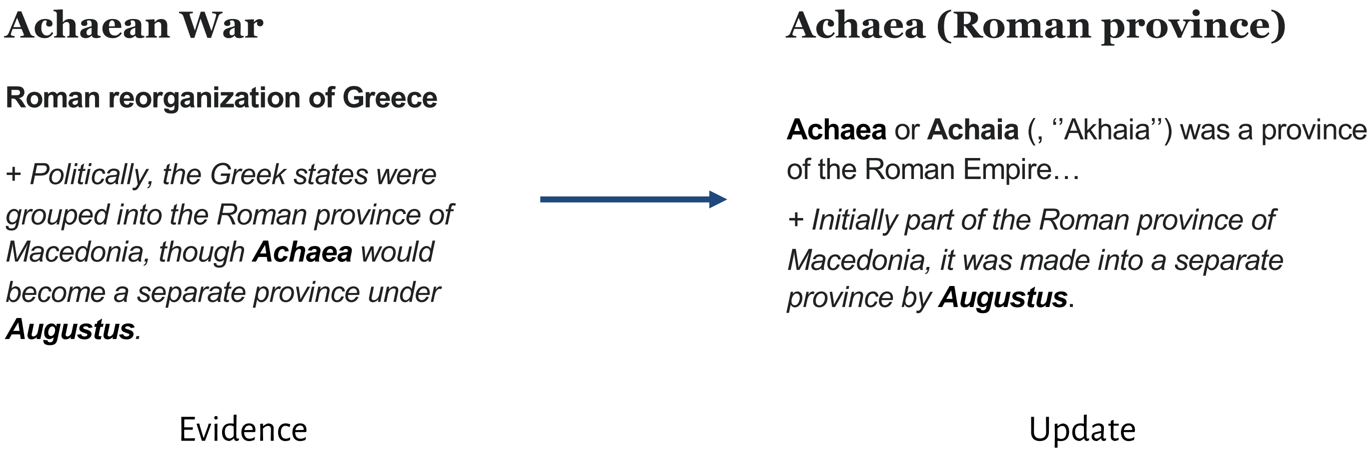
Support

A piece of evidence provides support for an update to an article if the evidence mentions both the article subject, and one of the added entities.

Filtering Heuristic

Support

A piece of evidence provides support for an update to an article if the evidence mentions both the article subject, and one of the added entities.



Achaean War

Roman reorganization of Greece

+ *Politically, the Greek states were grouped into the Roman province of Macedonia, though* ***Achaea*** *would become a separate province under* ***Augustus****.*

Evidence

Achaea (Roman province)

Achaea or **Achaia** (, ''Akhaia'') was a province of the Roman Empire…

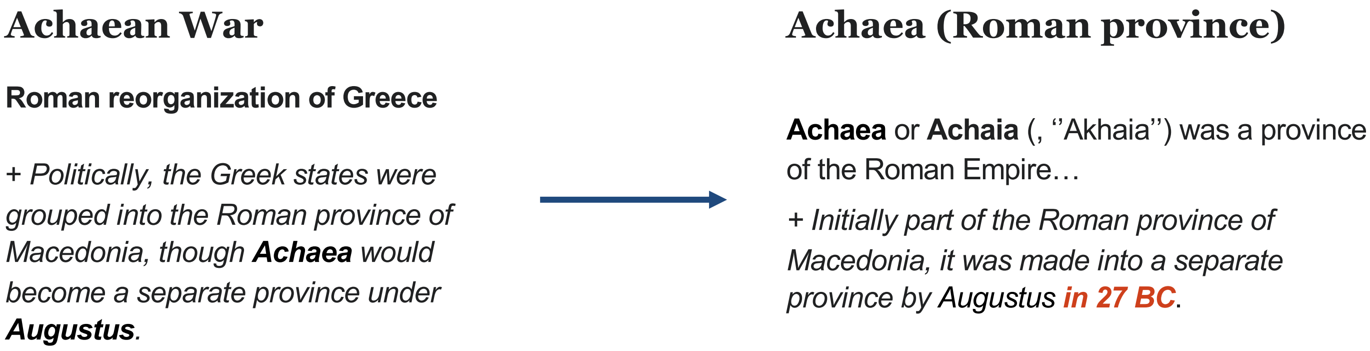
+ *Initially part of the Roman province of Macedonia, it was made into a separate province by* ***Augustus***.

Update

Gold Data Collection

Eliminate Unsupported Edits

Not all edits in the target will be supported. We hire annotators to remove them.



Achaean War

Roman reorganization of Greece

+ *Politically, the Greek states were grouped into the Roman province of Macedonia, though* ***Achaea*** *would become a separate province under* ***Augustus****.*

Achaea (Roman province)

Achaea or **Achaia** (, ‘’Akhaia’’) was a province of the Roman Empire…

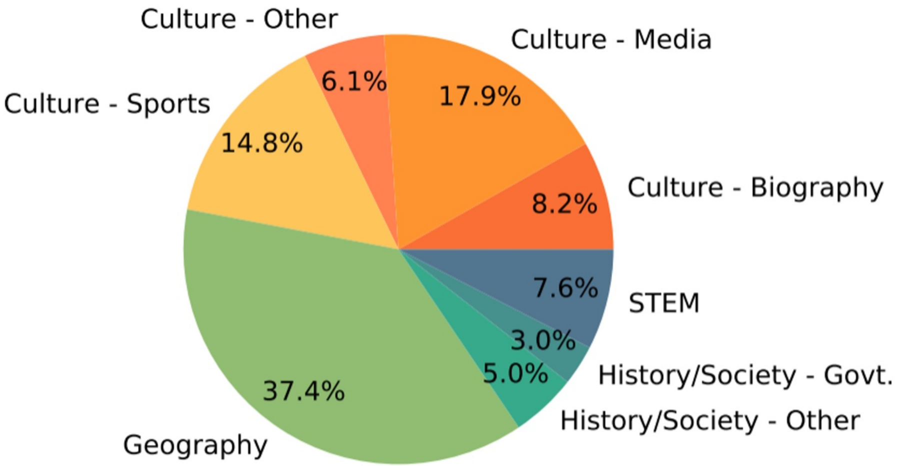
+ *Initially part of the Roman province of Macedonia, it was made into a separate province by Augustus* ***in 27 BC****.*

FRUIT-Wiki Dataset

>100 thousand training article pairs from 2019-2020.

~900 annotated evaluation pairs from 2020-2021.

High agreement between automatic and human annotations.



Model - T5 Baseline

Input: Concatenate source article and evidence, delimited by sentinel tokens.

```
[0] Martin Smith (born 2 October 1995) is an English footballer who plays for club Salford City.
[CONTEXT]
(0) 2020–21 Salford City F.C. season - Transfers out
[HEADER] Date [COL] Position [COL] Nationality [COL] Name [COL] To [COL] Fee [COL]
[ROW]  14 January 2021 [COL] CM [COL] ENG [COL] Martin Smith [COL] Chesterfield [COL] Free transfer
(1) List of Christian Brothers school alumni - Sport - Football (soccer)
* Martin Smith - current Kilmarnock footballer - St Aidan’s Catholic Academy, Sunderland
```

Model - T5 Baseline

Input: Concatenate source article and evidence, delimited by sentinel tokens.

```
[0] Martin Smith (born 2 October 1995) is an English footballer who plays for club Salford City.
[CONTEXT]
(0) 2020–21 Salford City F.C. season - Transfers out
[HEADER] Date [COL] Position [COL] Nationality [COL] Name [COL] To [COL] Fee [COL]
[ROW]  14 January 2021 [COL] CM [COL] ENG [COL] Martin Smith [COL] Chesterfield [COL] Free transfer
(1) List of Christian Brothers school alumni - Sport - Football (soccer)
* Martin Smith - current Kilmarnock footballer - St Aidan’s Catholic Academy, Sunderland
```

Model - T5 Baseline

Input: Concatenate source article and evidence, delimited by sentinel tokens.

```
[0] Martin Smith (born 2 October 1995) is an English footballer who plays for club Salford City.
[CONTEXT]
(0) 2020–21 Salford City F.C. season - Transfers out
[HEADER] Date [COL] Position [COL] Nationality [COL] Name [COL] To [COL] Fee [COL]
[ROW]  14 January 2021 [COL] CM [COL] ENG [COL] Martin Smith [COL] Chesterfield [COL] Free transfer
(1) List of Christian Brothers school alumni - Sport - Football (soccer)
* Martin Smith - current Kilmarnock footballer - St Aidan’s Catholic Academy, Sunderland
```

Model - T5 Baseline

Input: Concatenate source article and evidence, delimited by sentinel tokens.

```
[0] Martin Smith (born 2 October 1995) is an English footballer who plays for club Salford City.
[CONTEXT]
(0) 2020–21 Salford City F.C. season - Transfers out
[HEADER] Date [COL] Position [COL] Nationality [COL] Name [COL] To [COL] Fee [COL]
[ROW]  14 January 2021 [COL] CM [COL] ENG [COL] Martin Smith [COL] Chesterfield [COL] Free transfer
(1) List of Christian Brothers school alumni - Sport - Football (soccer)
* Martin Smith - current Kilmarnock footballer - St Aidan’s Catholic Academy, Sunderland
```

Model - T5 Baseline

Output: Generate full text of article.

```
Tom Krister Kristensson (born 30 April 1991) is a Swedish rally driver, who drives in the ADAC Opel
Rallye Cup. He Is originally from Lund, but now lives in Hörby In 2016 and 2017, Kristensson Competed
in the ADAC Opel Rallye Cup, finishing second in 2016 and winning the 2017 championship. In the 2019
season of JWRC, Tom finished second behind Jan Solans. The next season he went  on to become the 2020
Junior World Rally Champion.
```

Model - EdiT5

Output: Generate Diff + References

```
(2) Tom Krister Kristensson (born 30 April 1991) is a Swedish rally driver, who drives in the Junior
World Championship. [1] [2] (1) In the 2019 season of JWRC, Tom finished second behind Jan Solans.
(2) The next season he went on to become the 2020 Junior World Rally champion.
```

Model - EdiT5

Output: Generate **Diff** + References

`(2) Tom Krister Kristensson (born 30 April 1991) is a Swedish rally driver, who drives in the Junior World Championship. [1] [2] (1) In the 2019 season of JWRC, Tom finished second behind Jan Solans. (2) The next season he went on to become the 2020 Junior World Rally champion.`

Copy Sentences

Model - EdiT5

Output: Generate Diff + **References**

(2) Tom Krister Kristensson (born 30 April 1991) is a Swedish rally driver, who drives in the Junior World Championship. [1] [2] (1) In the 2019 season of JWRC, Tom finished second behind Jan Solans. (2) The next season he went on to become the 2020 Junior World Rally champion.

Reference Tokens

Quantitative Evaluation

UpdateROUGE

- ROUGE score measured only on updated sentences.
- Prevents high scores from being assigned to no-ops.

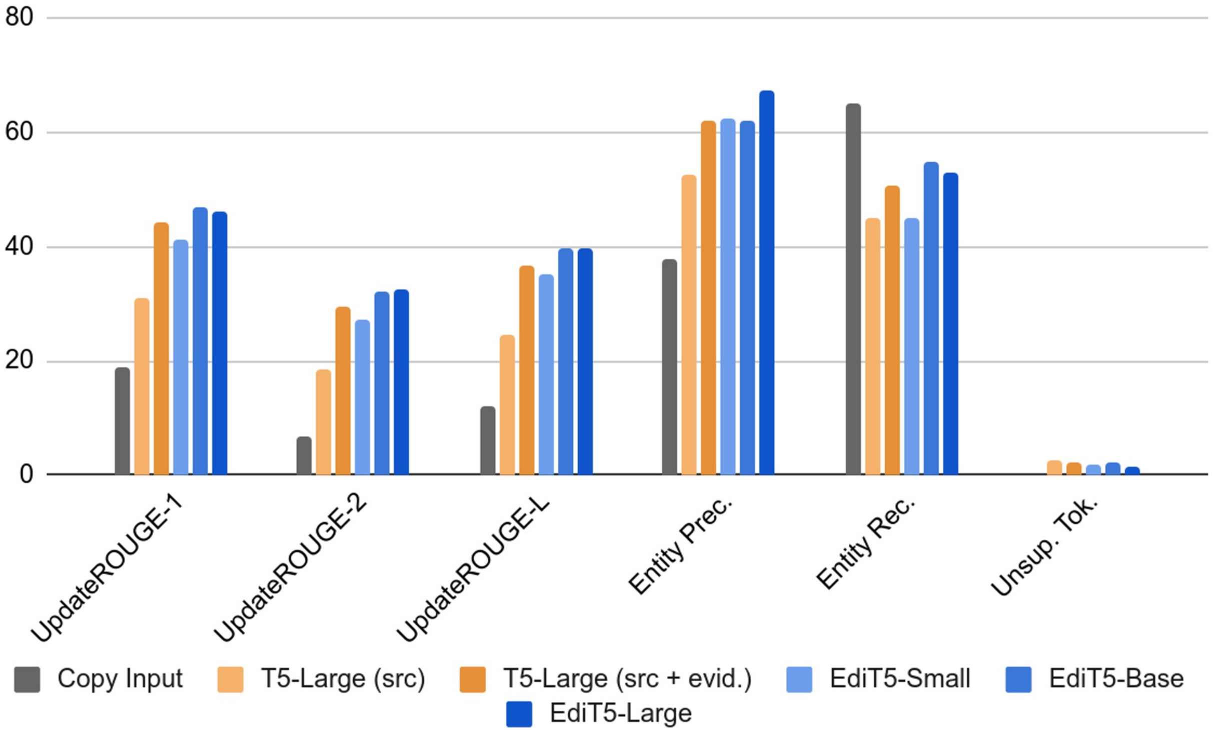
Entity Precision/Recall w.r.t. Target

- Uses NER to detect entities in target vs. generated updates.
- Affected by content selection.

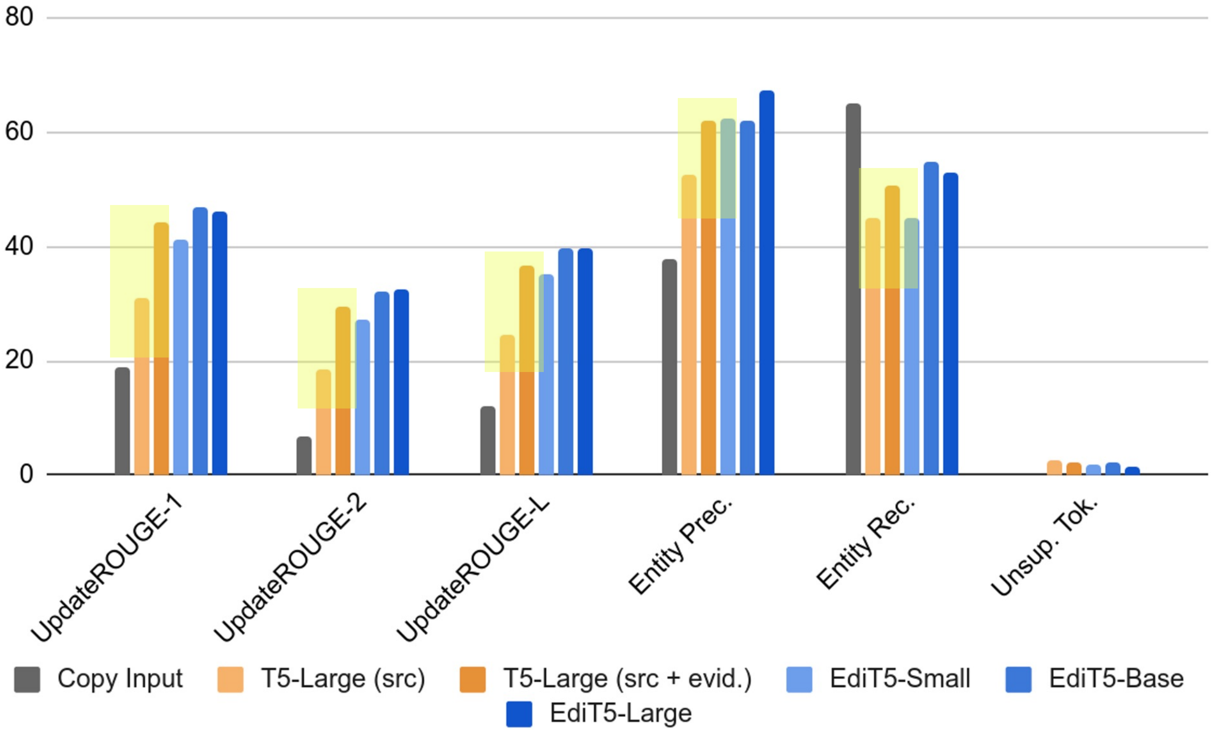
Unsupported Entity Tokens w.r.t. Evidence

- Average number of tokens that the model is “hallucinating”.

Quantitative Evaluation

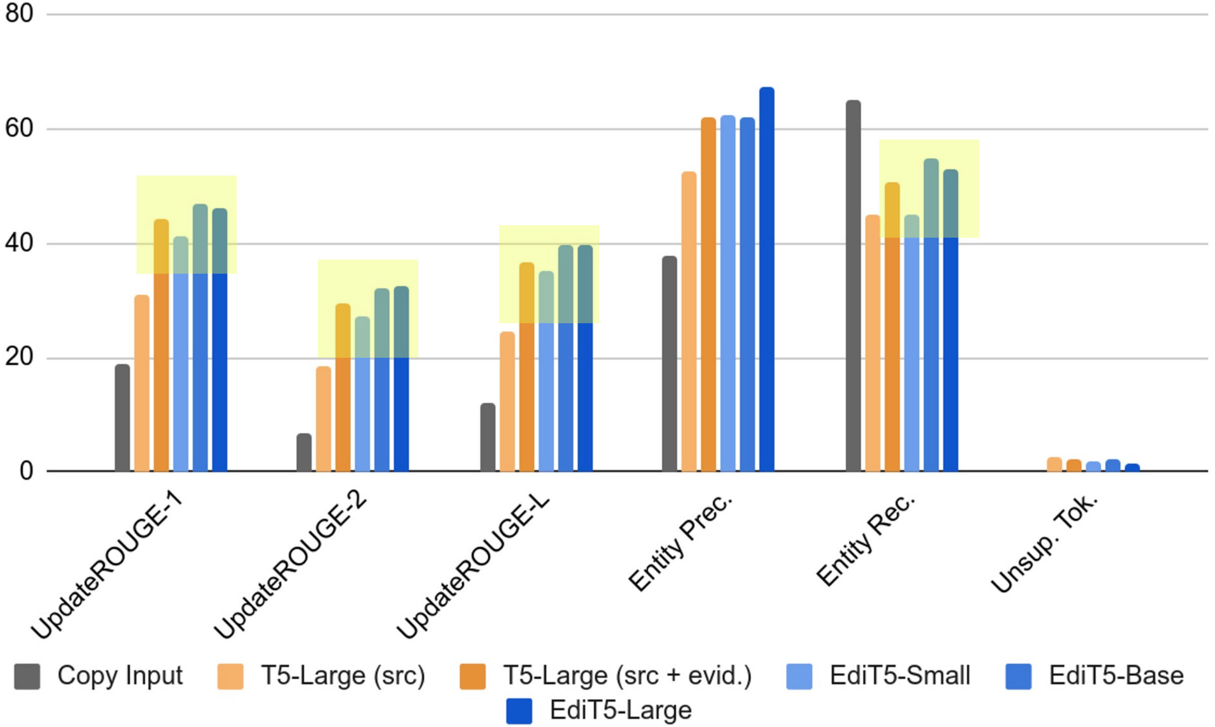


Quantitative Evaluation



Evidence is needed to obtain good performance!

Quantitative Evaluation



EdiT5 further improves performance!

Error Analysis

Three Common Types of Error:

1. Incorrect numbers and dates

   e.g., the model outputs 2020 instead of 2021.

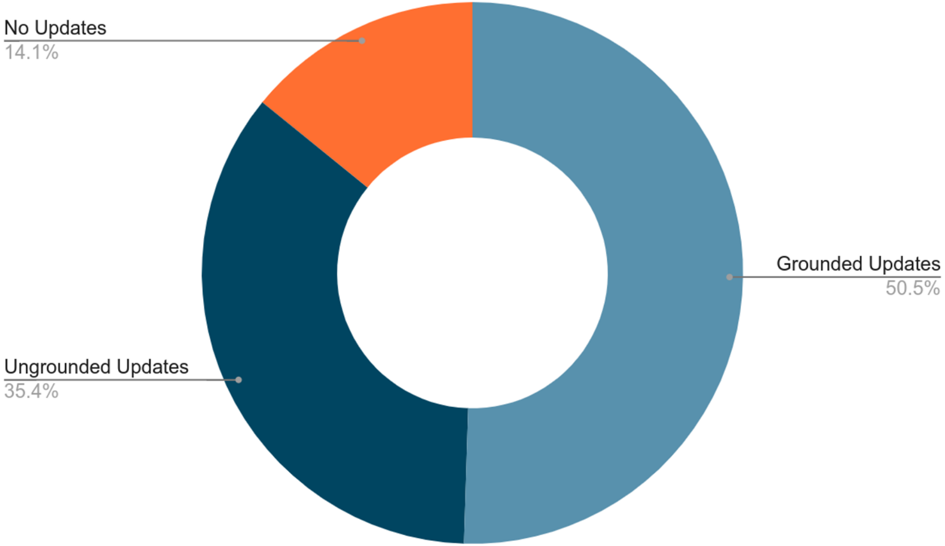
2. Distorted evidence

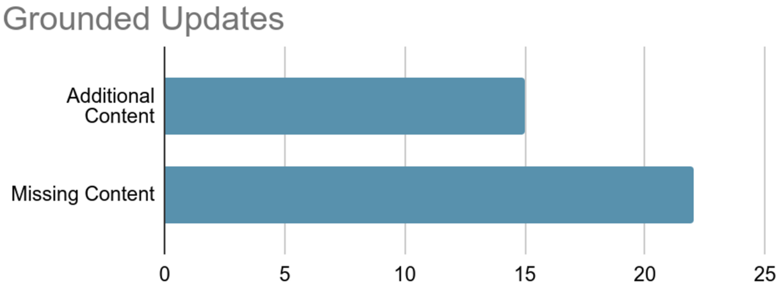
   e.g., model conflates columns in a table and makes an incorrect claim.

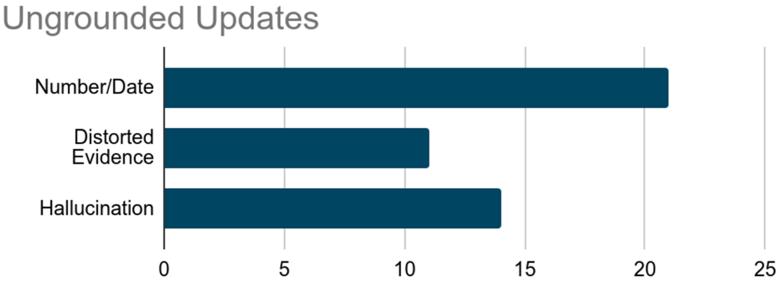
3. Hallucinations

   e.g., model output includes new text completely unrelated to the evidence.

Error Analysis




Grounded Updates


Ungrounded Updates

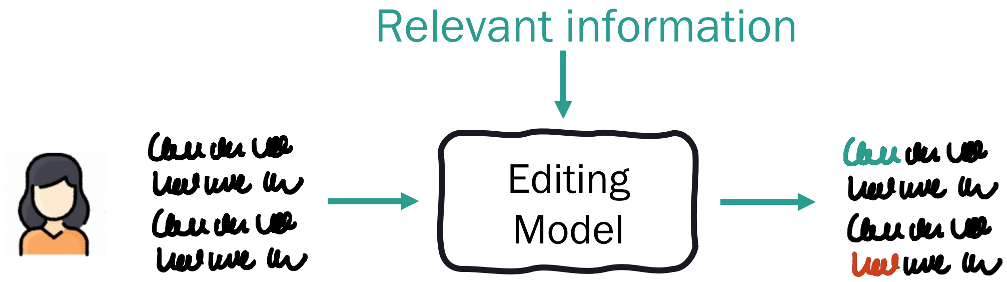
Conclusion

- Introduce a new task and dataset to explore collaborative methods for improving consistency in knowledge bases.
- Results show that you can use small language models for this task.

- Limitations
  - Need to provide relevant information
  - Limited to Wikipedia
  - Training signal is noisy

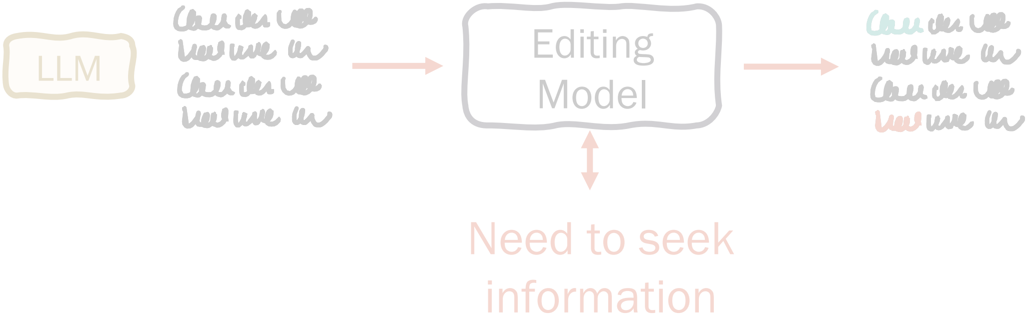
Outline

Part 1



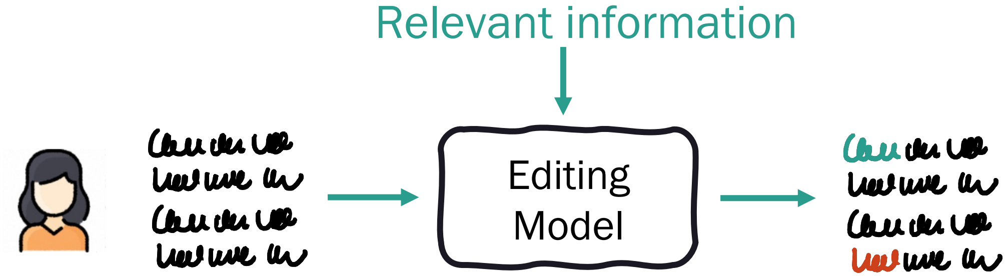
FRUIT: Faithfully Reflecting Updated Information in Text
Robert L. Logan IV, Alexandre Passos, Sameer Singh, Ming-Wei Chang
NAACL 2022 **BEST TASK PAPER AWARD**

Part 2

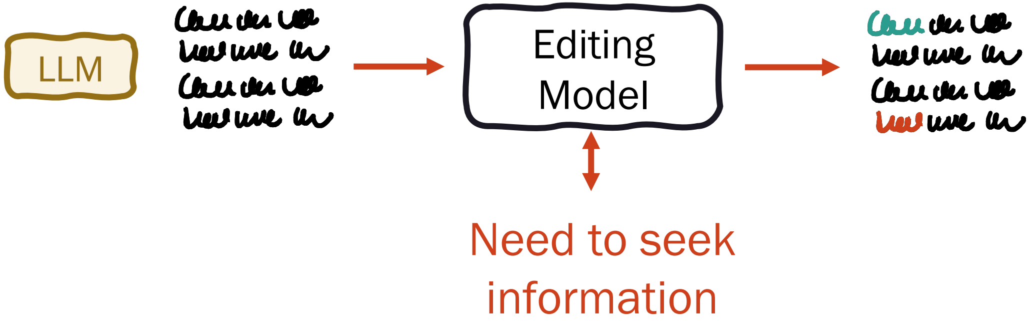


Outline

Part 1

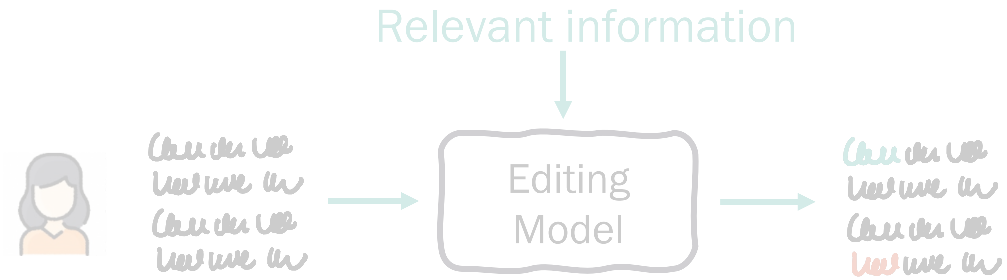


Part 2



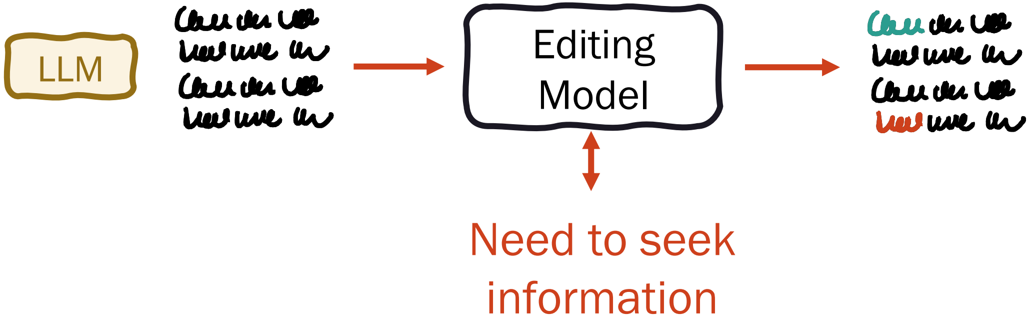
Outline

Part 1

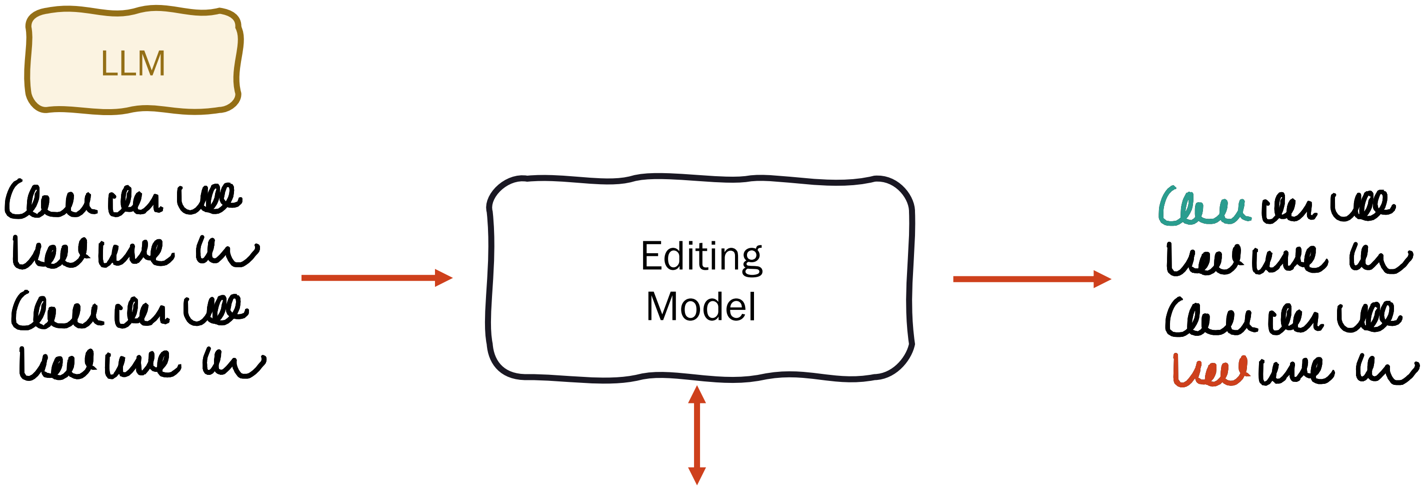


PURR: Efficiently Editing Language Model Hallucinations by Denoising Language Model Corruptions
Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, Kelvin Guu
ArXiv 2023

Part 2

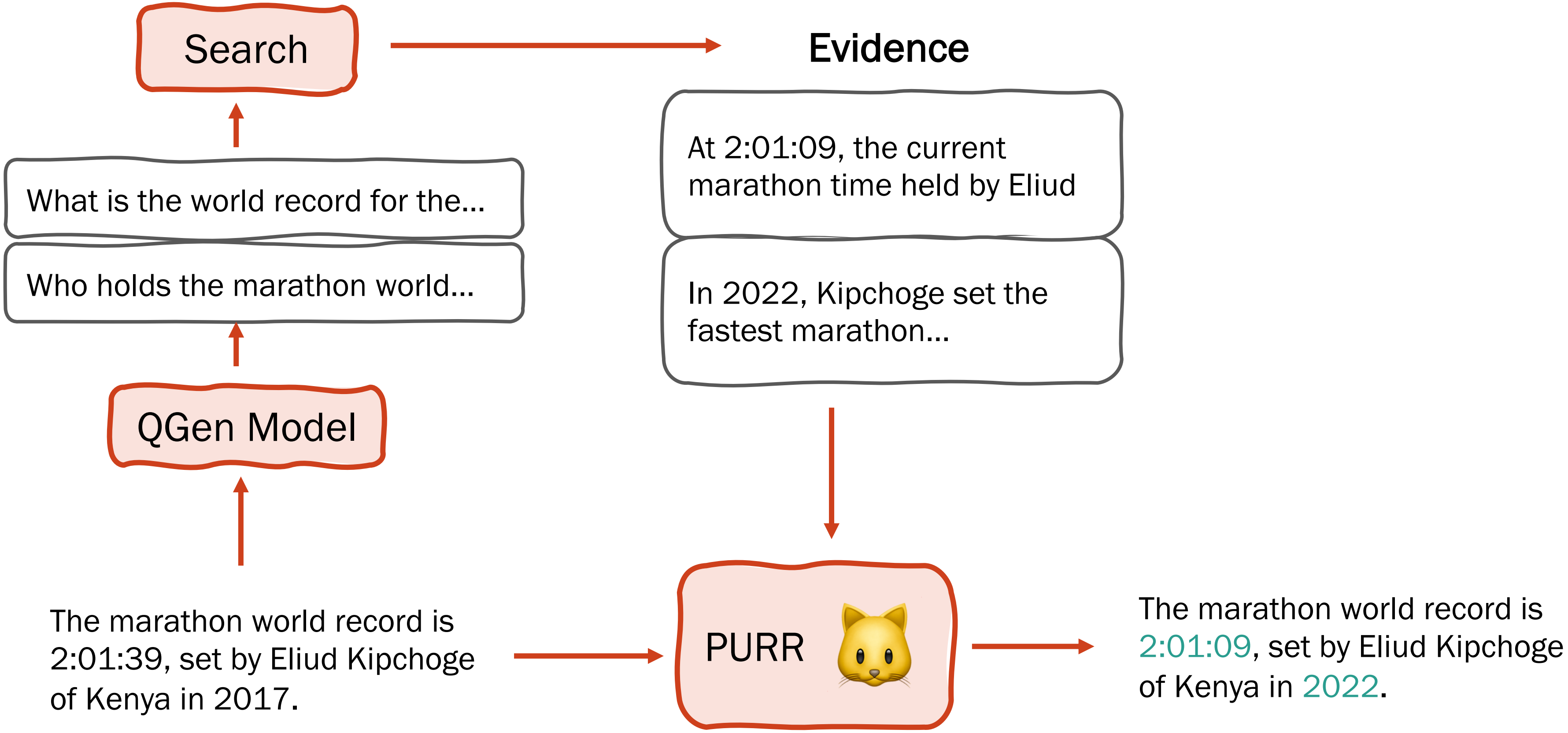


Task: Get Information and Revise



Need to seek information

PURR Overview



Search

Evidence

What is the world record for the...

Who holds the marathon world...

At 2:01:09, the current marathon time held by Eliud

In 2022, Kipchoge set the fastest marathon...

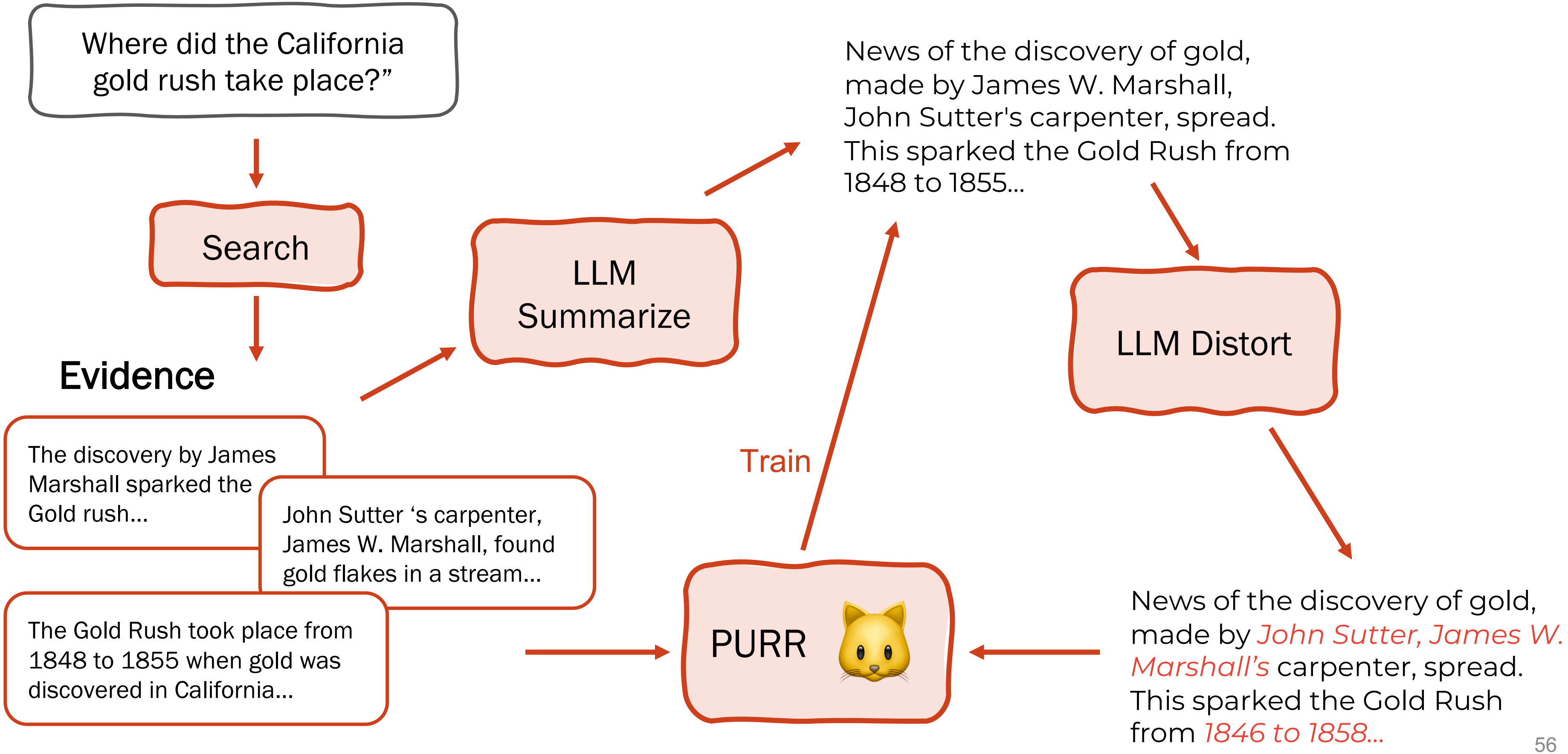
QGen Model

PURR 🐱

The marathon world record is 2:01:39, set by Eliud Kipchoge of Kenya in 2017.

The marathon world record is 2:01:09, set by Eliud Kipchoge of Kenya in 2022.

Automated Training Pipeline



Where did the California gold rush take place?”

Search

Evidence

The discovery by James Marshall sparked the Gold rush...

John Sutter ‘s carpenter, James W. Marshall, found gold flakes in a stream...

The Gold Rush took place from 1848 to 1855 when gold was discovered in California...

LLM Summarize

News of the discovery of gold, made by James W. Marshall, John Sutter's carpenter, spread. This sparked the Gold Rush from 1848 to 1855...

LLM Distort

News of the discovery of gold, made by *John Sutter, James W. Marshall’s* carpenter, spread. This sparked the Gold Rush from *1846 to 1858...*

PURR 🐱

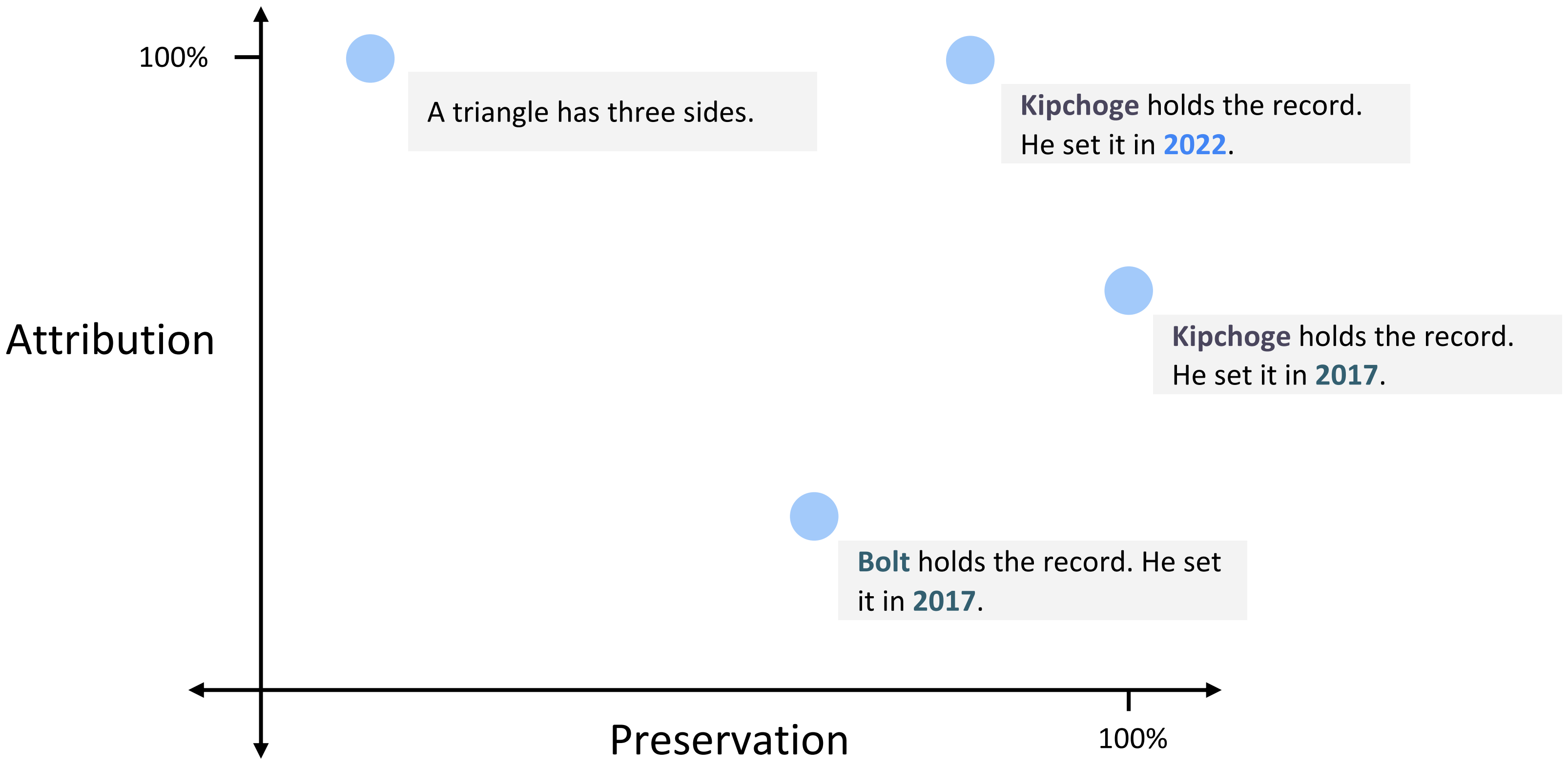
Train

Evaluation

Given any LM output, revise it so that all claims are **factually consistent** to trusted sources…

… while **preserving** the original text as much as possible.

Kipchoge holds the record. He set it in **2017**.



A triangle has three sides.

Kipchoge holds the record. He set it in **2022**.

Kipchoge holds the record. He set it in **2017**.

Bolt holds the record. He set it in **2017**.

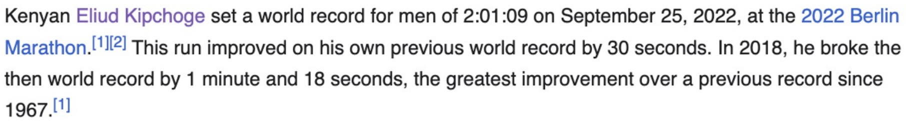
Measuring Attribution

Percentage of sentences that are **fully entailed** by at least one evidence.

Eliud Kipchoge set a marathon world record in 2022.

He trains in Eldoret.

entailed



Kenyan Eliud Kipchoge set a world record for men of 2:01:09 on September 25, 2022, at the 2022 Berlin Marathon.[1][2] This run improved on his own previous world record by 30 seconds. In 2018, he broke the then world record by 1 minute and 18 seconds, the greatest improvement over a previous record since 1967.[1]

https://en.wikipedia.org/wiki/Marathon_world_record_progression

attribution = 50%
since 1 out of 2
sentences are entailed.

NOTE: *your system needs to be good at retrieving evidence.*
You only get credit if you can find support for each sentence (edited or not).

Evaluation Set: LLM Outputs on Variety of Tasks

Natural Questions (NQ)
Long-form question answering

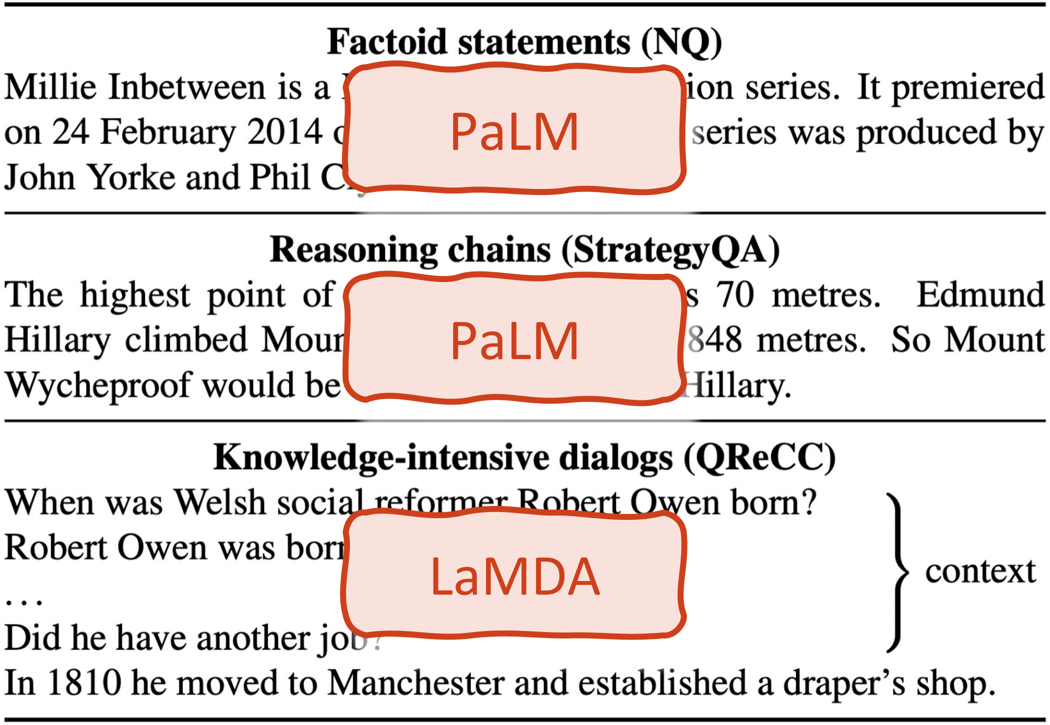
"When did Millie Inbetween premiere?"

StrategyQA
Multi-step reasoning

"Would it be hard for Edmund Hillary to climb Mount Wycheproof?"

QReCC
Knowledge-intensive dialog

"When was Robert Owen born? … What was his job? … "



Factoid statements (NQ)

Millie Inbetween is a ion series. It premiered
on 24 February 2014 series was produced by
John Yorke and Phil C

PaLM

Reasoning chains (StrategyQA)

The highest point of s 70 metres. Edmund
Hillary climbed Mour 848 metres. So Mount
Wycheproof would be Hillary.

PaLM

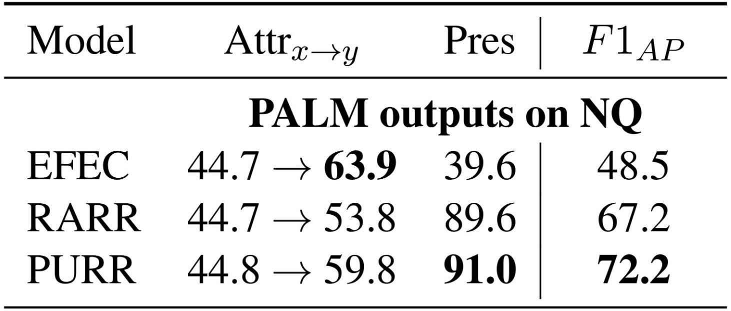
Knowledge-intensive dialogs (QReCC)

When was Welsh social reformer Robert Owen born?
Robert Owen was borr
. . .
Did he have another job?
In 1810 he moved to Manchester and established a draper's shop.

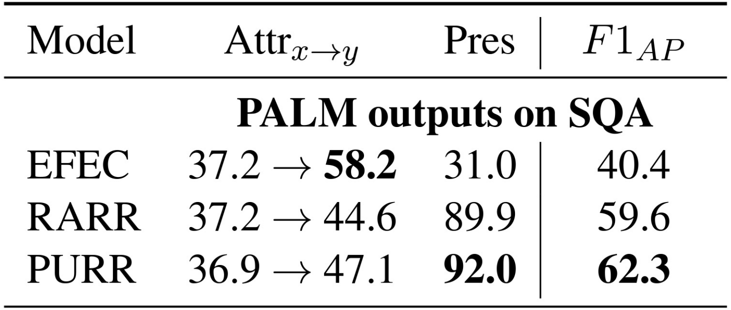
LaMDA

context

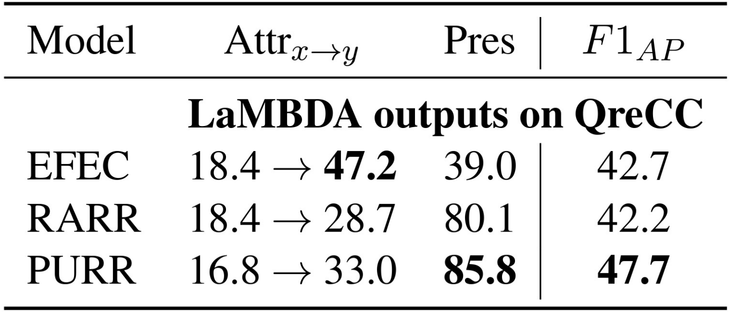
Results



Model	Attr$_{x \to y}$	Pres	$F1_{AP}$
	PALM outputs on NQ		
EFEC	44.7 → **63.9**	39.6	48.5
RARR	44.7 → 53.8	89.6	67.2
PURR	44.8 → 59.8	**91.0**	**72.2**



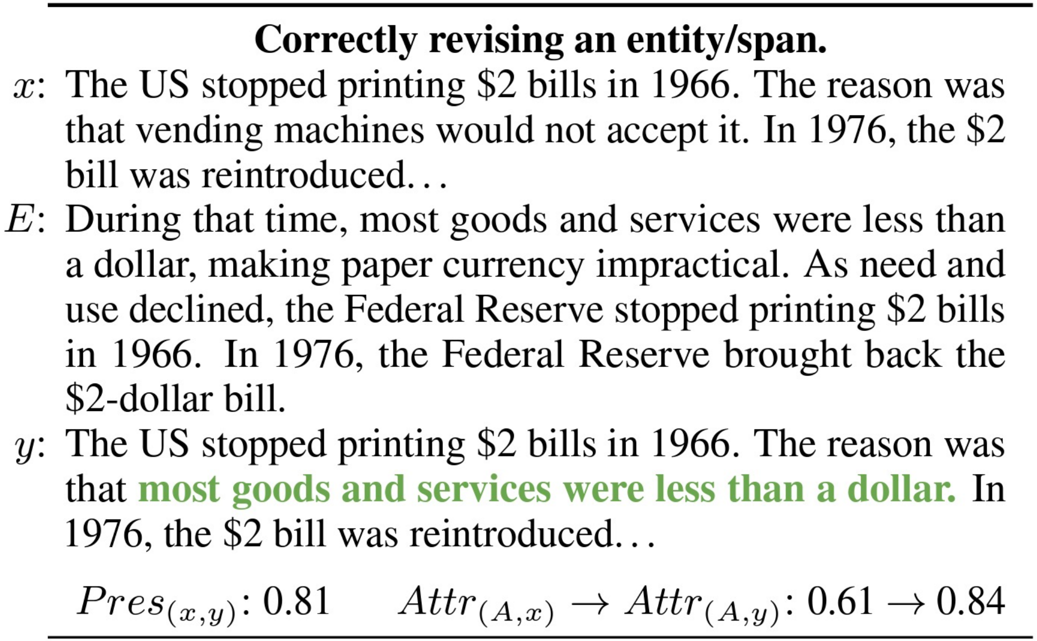
Model	Attr$_{x \to y}$	Pres	$F1_{AP}$
	PALM outputs on SQA		
EFEC	37.2 → **58.2**	31.0	40.4
RARR	37.2 → 44.6	89.9	59.6
PURR	36.9 → 47.1	**92.0**	**62.3**



Model	Attr$_{x \to y}$	Pres	$F1_{AP}$
	LaMBDA outputs on QreCC		
EFEC	18.4 → **47.2**	39.0	42.7
RARR	18.4 → 28.7	80.1	42.2
PURR	16.8 → 33.0	**85.8**	**47.7**

PURR is more than 100x faster than RARR

Examples



Correctly revising an entity/span.

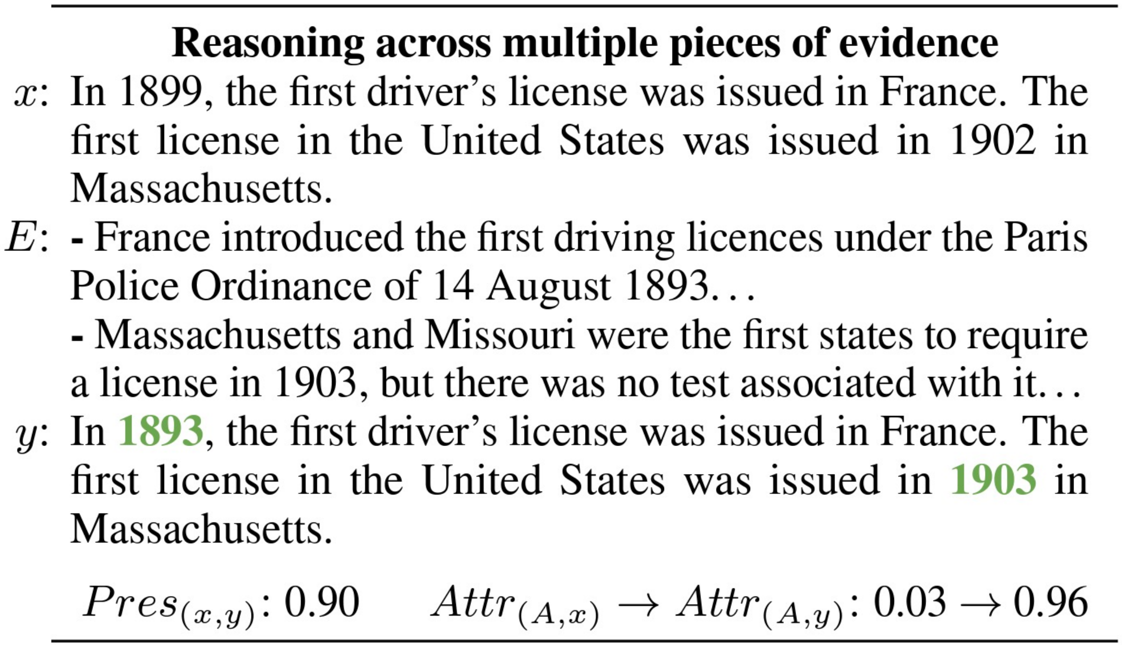
x: The US stopped printing \$2 bills in 1966. The reason was that vending machines would not accept it. In 1976, the \$2 bill was reintroduced…

E: During that time, most goods and services were less than a dollar, making paper currency impractical. As need and use declined, the Federal Reserve stopped printing \$2 bills in 1966. In 1976, the Federal Reserve brought back the \$2-dollar bill.

y: The US stopped printing \$2 bills in 1966. The reason was that **most goods and services were less than a dollar.** In 1976, the \$2 bill was reintroduced…

$Pres_{(x,y)}$: 0.81 $Attr_{(A,x)} \rightarrow Attr_{(A,y)}$: 0.61 → 0.84

Examples



Reasoning across multiple pieces of evidence

x: In 1899, the first driver's license was issued in France. The first license in the United States was issued in 1902 in Massachusetts.

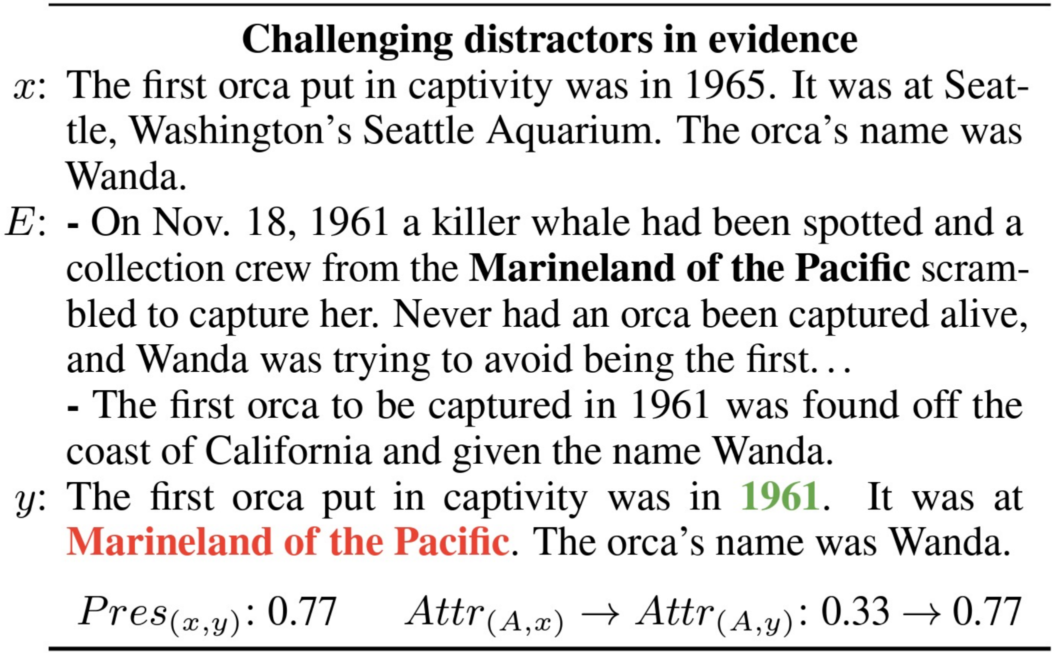
E: - France introduced the first driving licences under the Paris Police Ordinance of 14 August 1893…

- Massachusetts and Missouri were the first states to require a license in 1903, but there was no test associated with it…

y: In **1893**, the first driver's license was issued in France. The first license in the United States was issued in **1903** in Massachusetts.

$Pres_{(x,y)}$: 0.90 $Attr_{(A,x)} \rightarrow Attr_{(A,y)}$: 0.03 → 0.96

Examples



Challenging distractors in evidence

x: The first orca put in captivity was in 1965. It was at Seattle, Washington's Seattle Aquarium. The orca's name was Wanda.

E: **-** On Nov. 18, 1961 a killer whale had been spotted and a collection crew from the **Marineland of the Pacific** scrambled to capture her. Never had an orca been captured alive, and Wanda was trying to avoid being the first…

- The first orca to be captured in 1961 was found off the coast of California and given the name Wanda.

y: The first orca put in captivity was in **1961**. It was at **Marineland of the Pacific**. The orca's name was Wanda.

$Pres_{(x,y)}$: 0.77 $Attr_{(A,x)} \rightarrow Attr_{(A,y)}$: 0.33 → 0.77

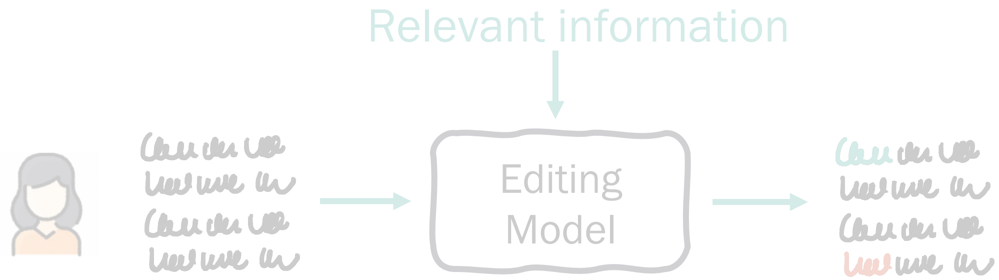
Conclusion

- Introduce training data creation pipeline
  - Automated, combination of real and generated data
  - Find real text, but use summarization and corruptions as instances
- Train extremely efficient editors to denoise the text

- Limitations
  - Tries to find justification for *everything*
  - Assumes corpus is trusted and doesn’t contradict itself
  - Cannot *guarantee* factuality or consistency

Outline

Part 1

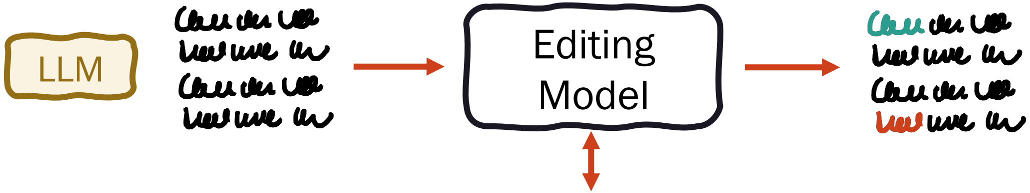


Relevant information

Editing Model

PURR: Efficiently Editing Language Model Hallucinations by Denoising Language Model Corruptions
Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, Kelvin Guu
ArXiv 2023

Part 2



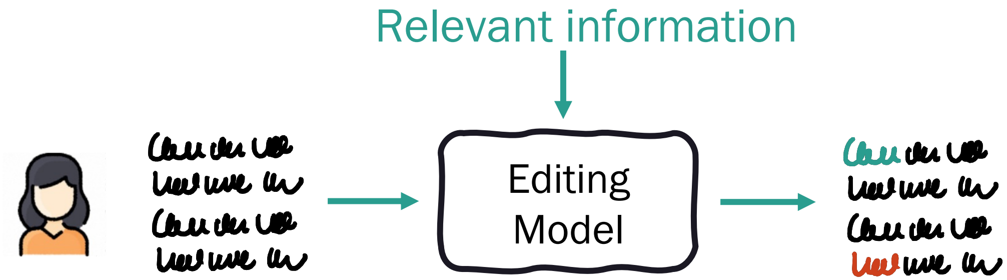
LLM

Editing Model

Need to seek information

Outline

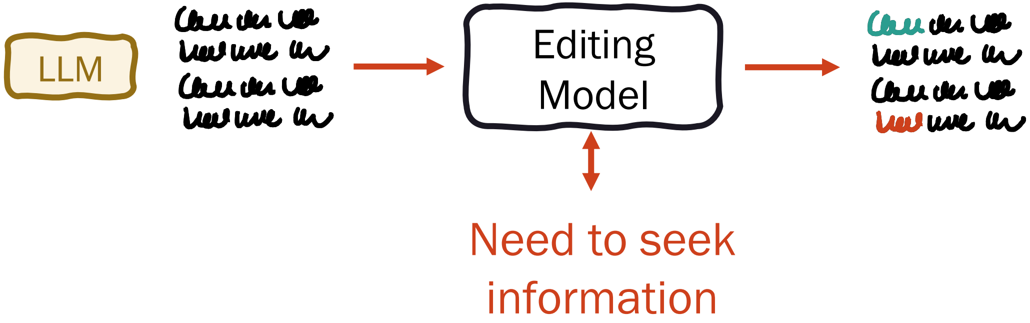
Part 1



Relevant information

Editing Model

Part 2

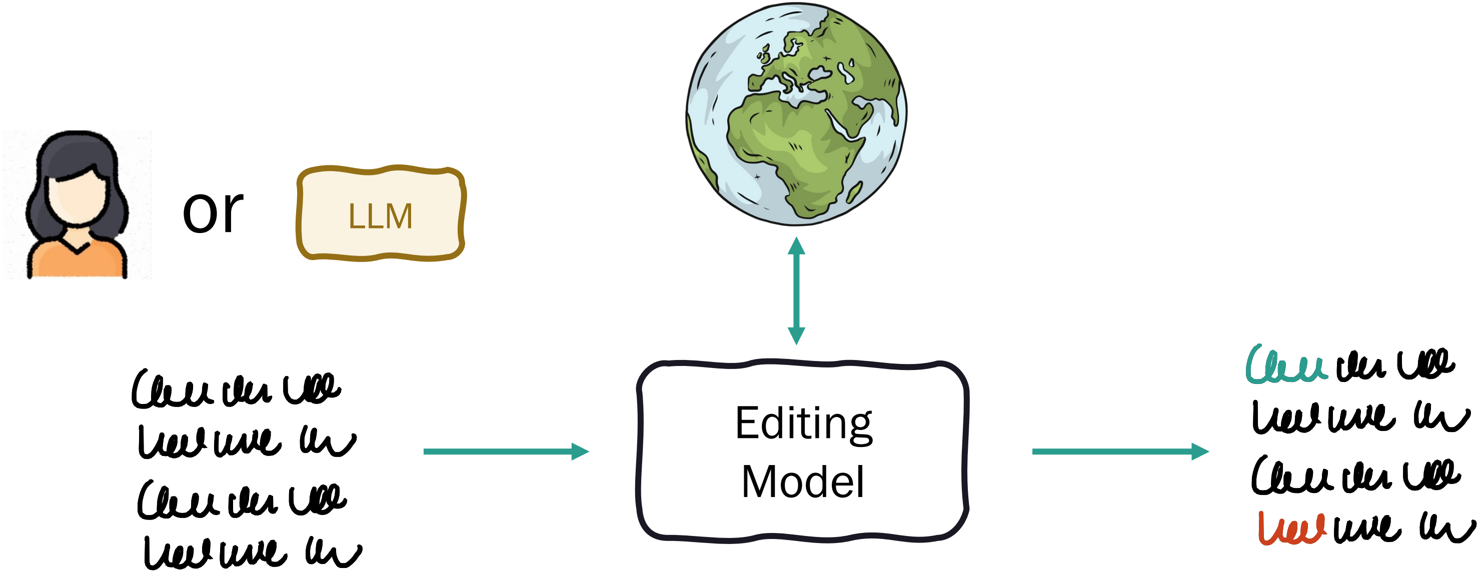


LLM

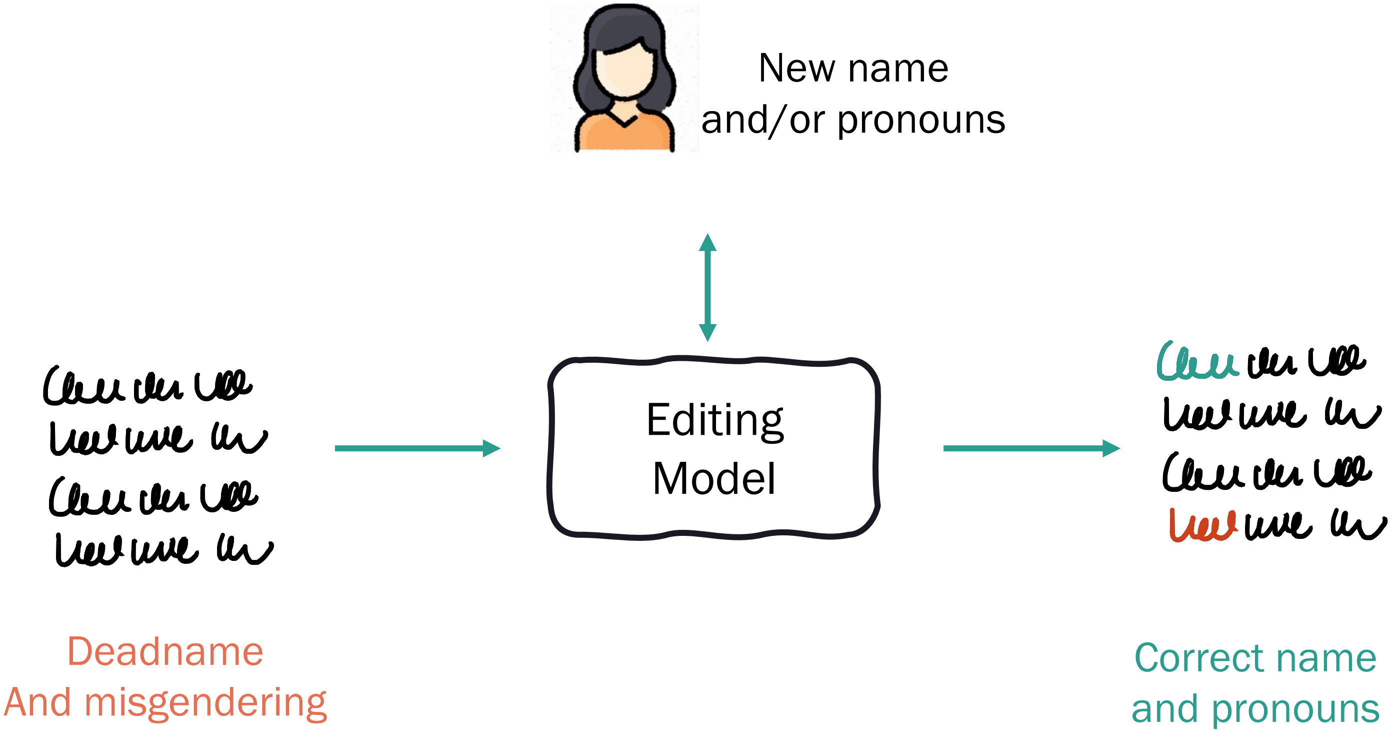
Editing Model

Need to seek information

Task: Make text match information



Application: Misgendering/Deadnaming



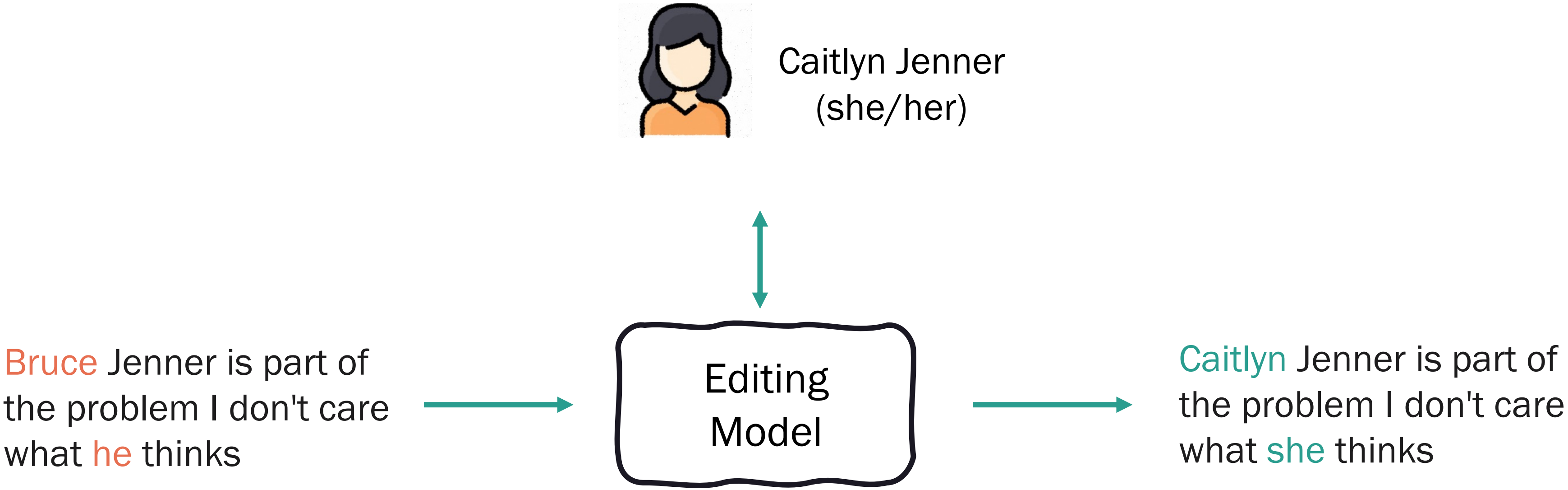
New name
and/or pronouns

Editing
Model

Deadname
And misgendering

Correct name
and pronouns

Application: Misgendering/Deadnaming



Caitlyn Jenner
(she/her)

Bruce Jenner is part of the problem I don't care what he thinks

Editing Model

Caitlyn Jenner is part of the problem I don't care what she thinks

Deadname
And misgendering

Correct name
and pronouns

Thank you!

@sameer_

sameer@uci.edu

sameersingh.org