Factuality and LLMs

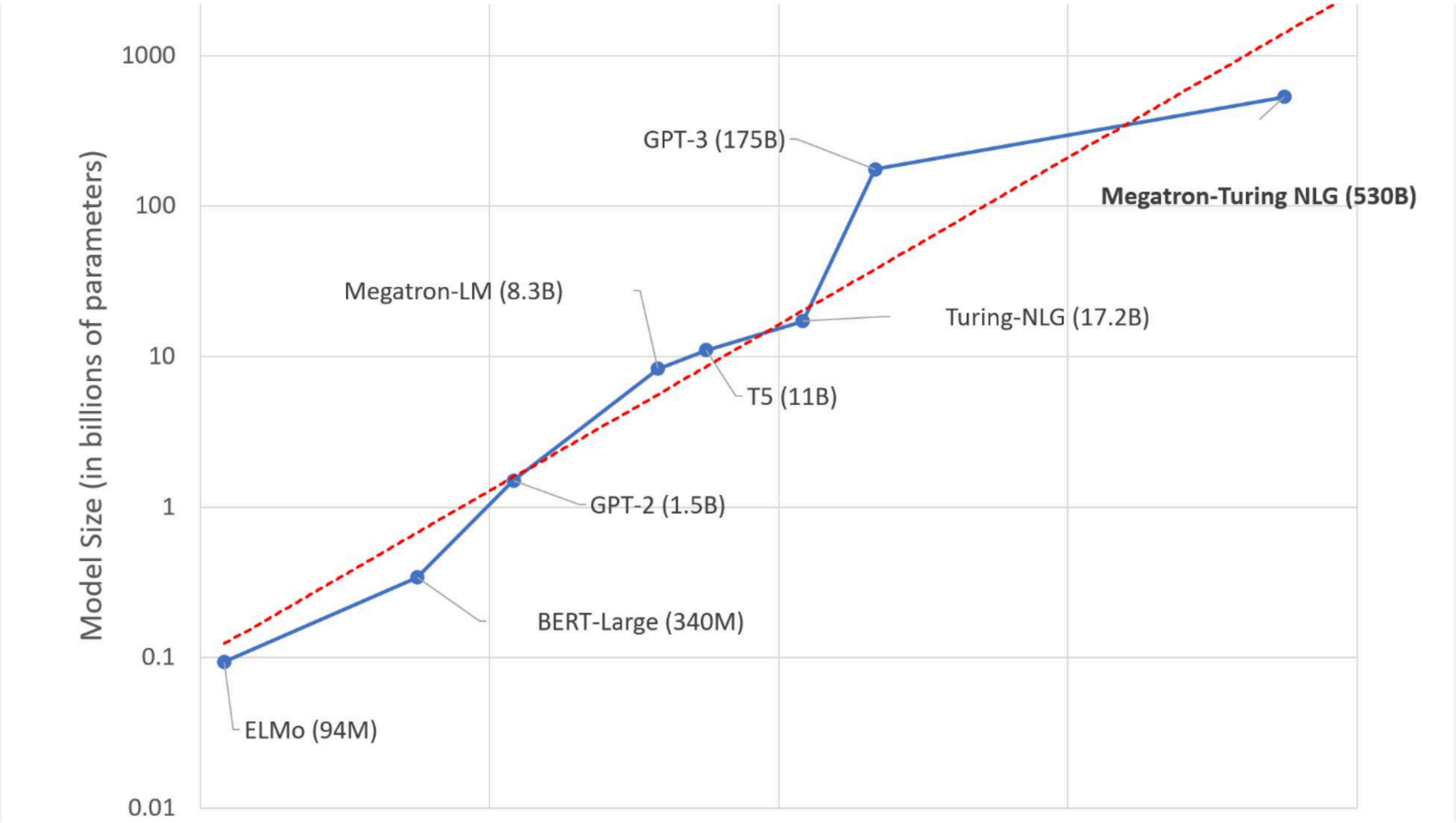
Agenda

- Introduction
  - LLMs intro
  - What is a hallucination
    - How the definition has changed — BART to LLMs
    - Two main categories of hallucinations
    - What we will focus on in this talk — context of summarization
  - Why should we care about hallucinations?
  - Factuality/LLMs and summarization
    - Summarization examples (zero-shot, before used to fine-tune models)
    - How summarization has changed in the era of GPT
    - How factuality has changed in summarization

  - Hallucination eval (humans) and limitations
    - Types : text/dialogue (include abridge work)
    - How humans Eval hallucinations
    - Exercise: Annotate 5 data points
  - Eval (automatic) and limitations
    - Rouge/BLEU
    - BERTscore
    - Models (QA/NLI)
    - Prompting
    - Models (MiniCheck, AlignScore)
    - Limitations : my recent work
  - Methods to address
  - What causes hallucinations
- Interactive Exercise
- Open Research Questions

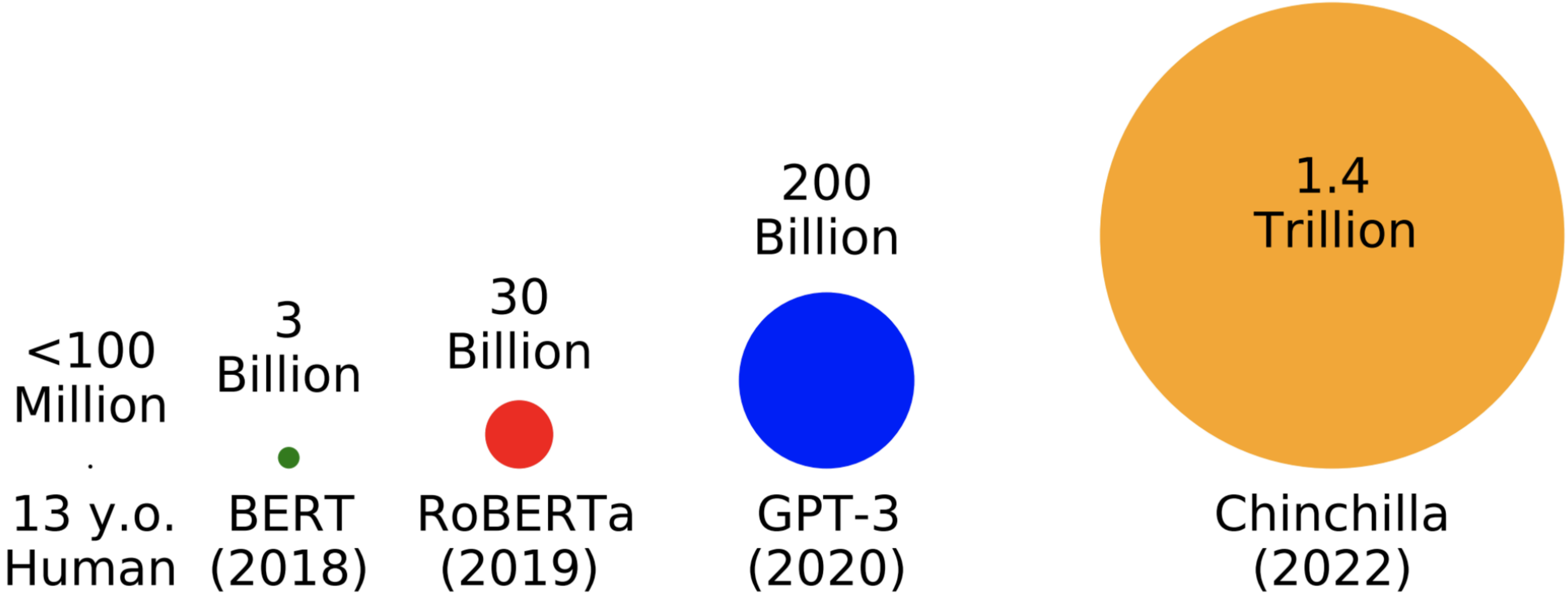
Large Language Models

- Trained on more and more parameters



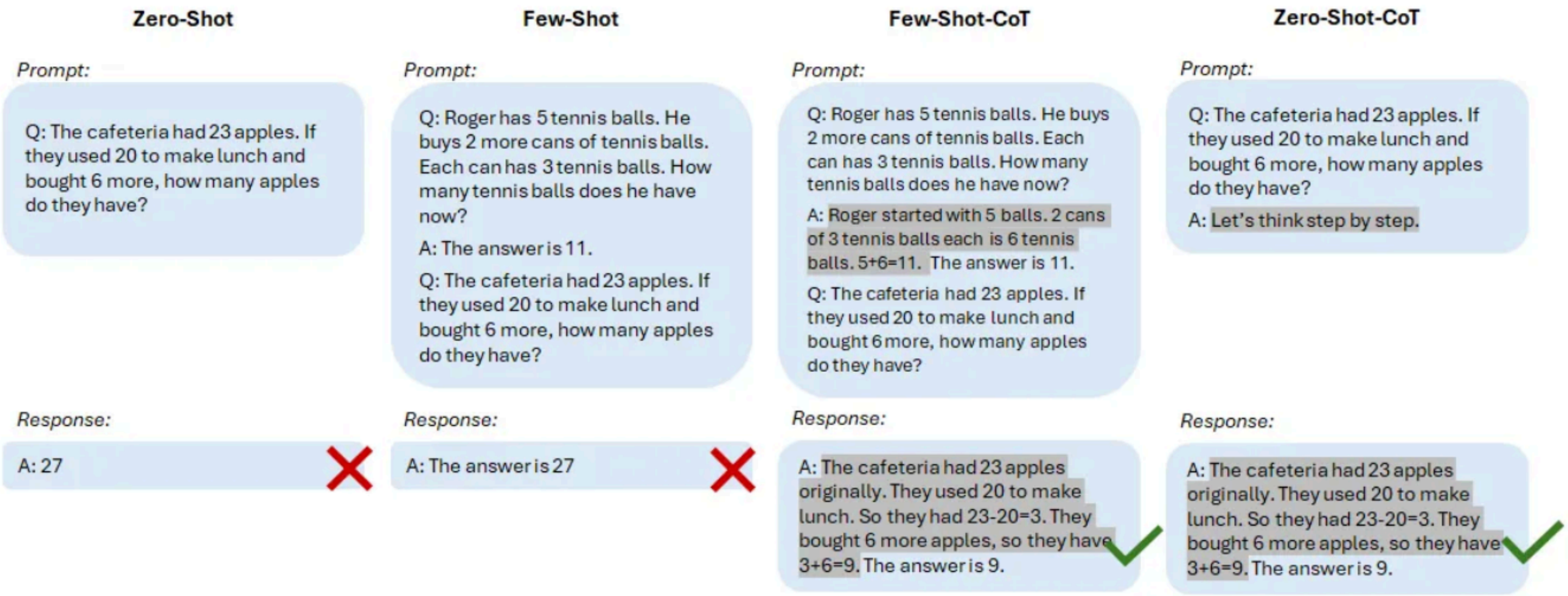
Source: https://huggingface.co/blog/large-language-models

- Trained on more and more data



tokens seen during training

- Scaling pre-trained language models has led to remarkable emergent abilities



Zero-Shot

Prompt:

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Response:

A: 27 ✗

Few-Shot

Prompt:

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Response:

A: The answer is 27 ✗

Few-Shot-CoT

Prompt:

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5+6=11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Response:

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23-20=3. They bought 6 more apples, so they have 3+6=9. The answer is 9. ✓

Zero-Shot-CoT

Prompt:

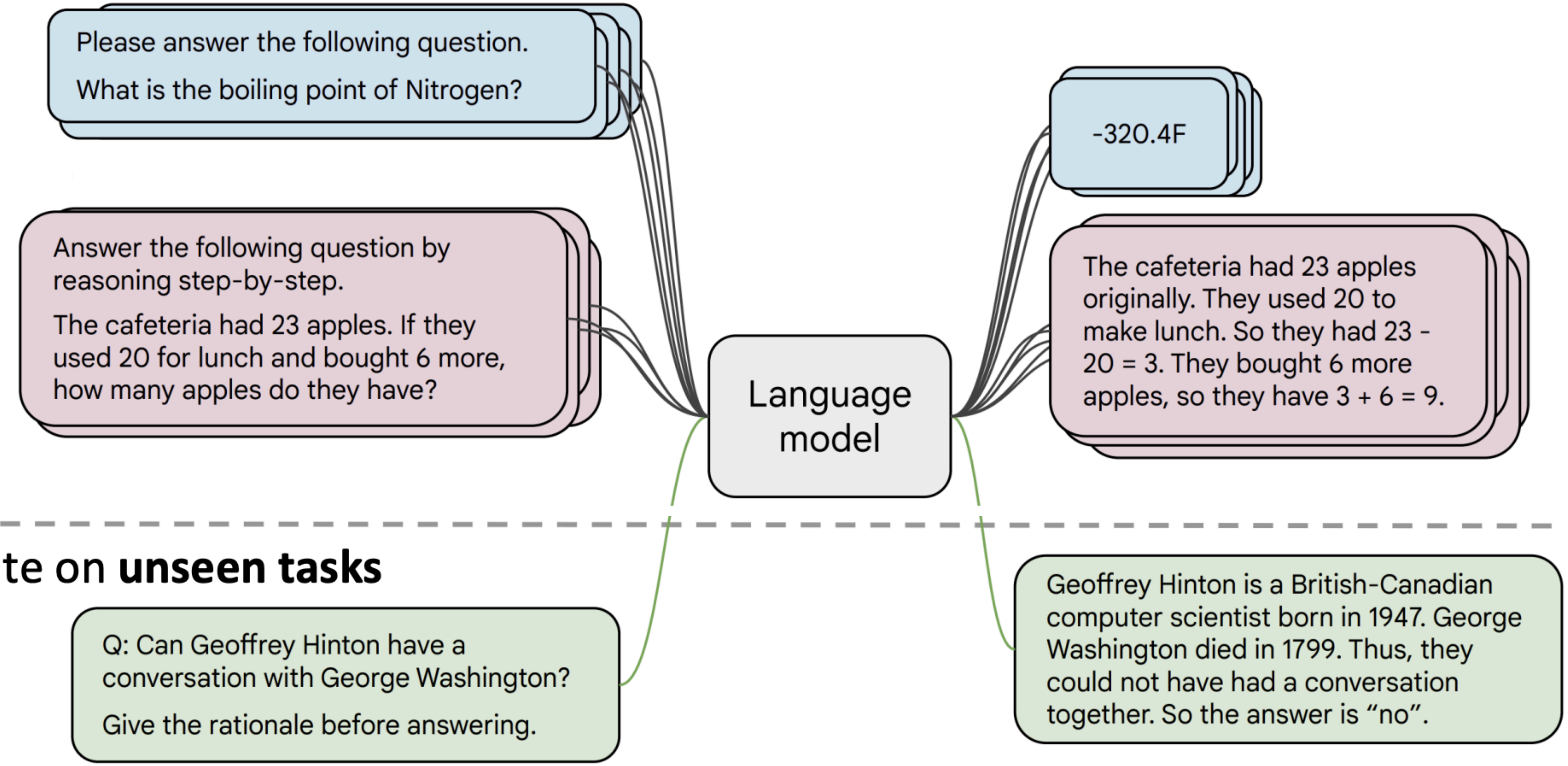
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: Let's think step by step.

Response:

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23-20=3. They bought 6 more apples, so they have 3+6=9. The answer is 9. ✓

- Scaling pre-trained language models has led to remarkable emergent abilities
  - **Collect examples** of (instruction, output) pairs across many tasks and finetune an LM



Please answer the following question.

What is the boiling point of Nitrogen?

-320.4F

Answer the following question by reasoning step-by-step.

The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?

The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9.

Language model

  - Evaluate on **unseen tasks**

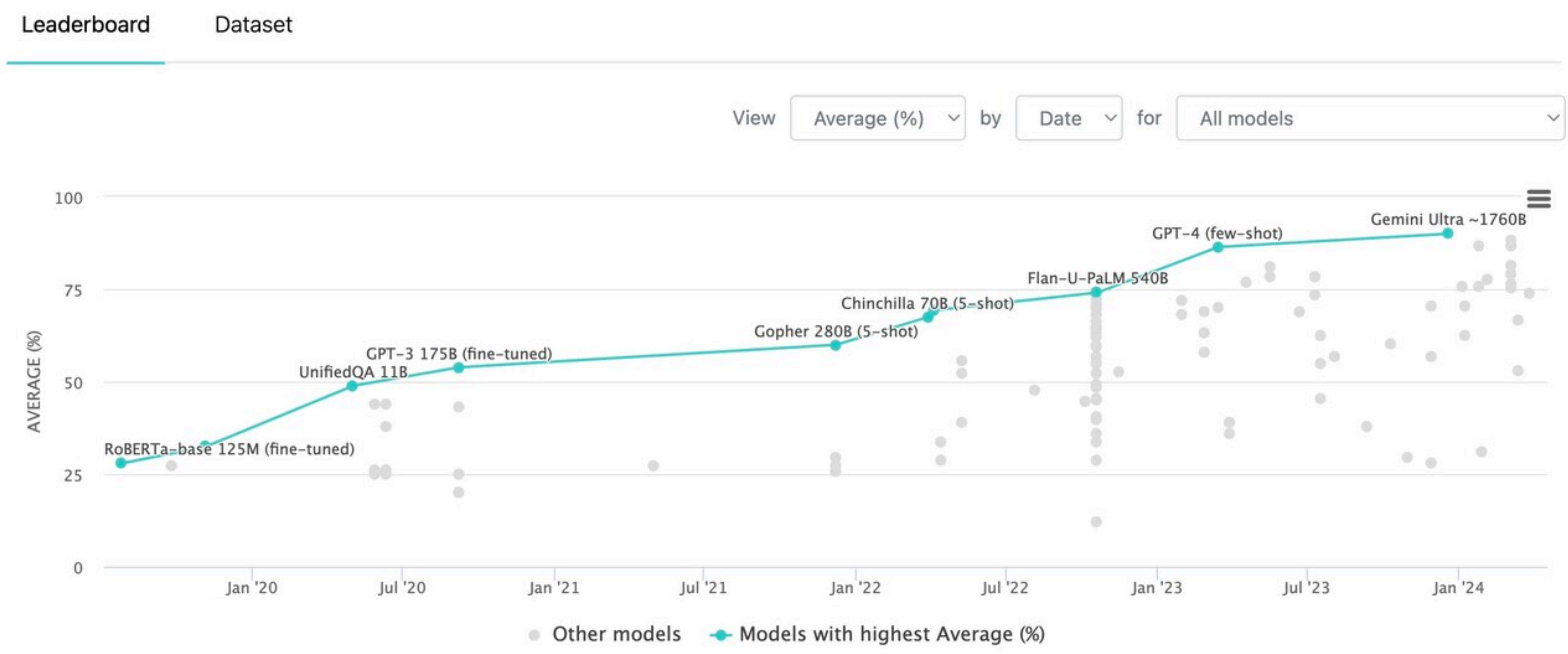
Q: Can Geoffrey Hinton have a conversation with George Washington?

Give the rationale before answering.

Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is “no”.

[FLAN-T5; Chung et al., 2022]

- LLMs have demonstrated substantial performance improvements across diverse NLP tasks (e.g., summarization, QA, translation).



- However hallucination has remained one of the more significant issues in practical applications

What is hallucination?

- In the field of pathology and psychology hallucination is defined as follows [1]:

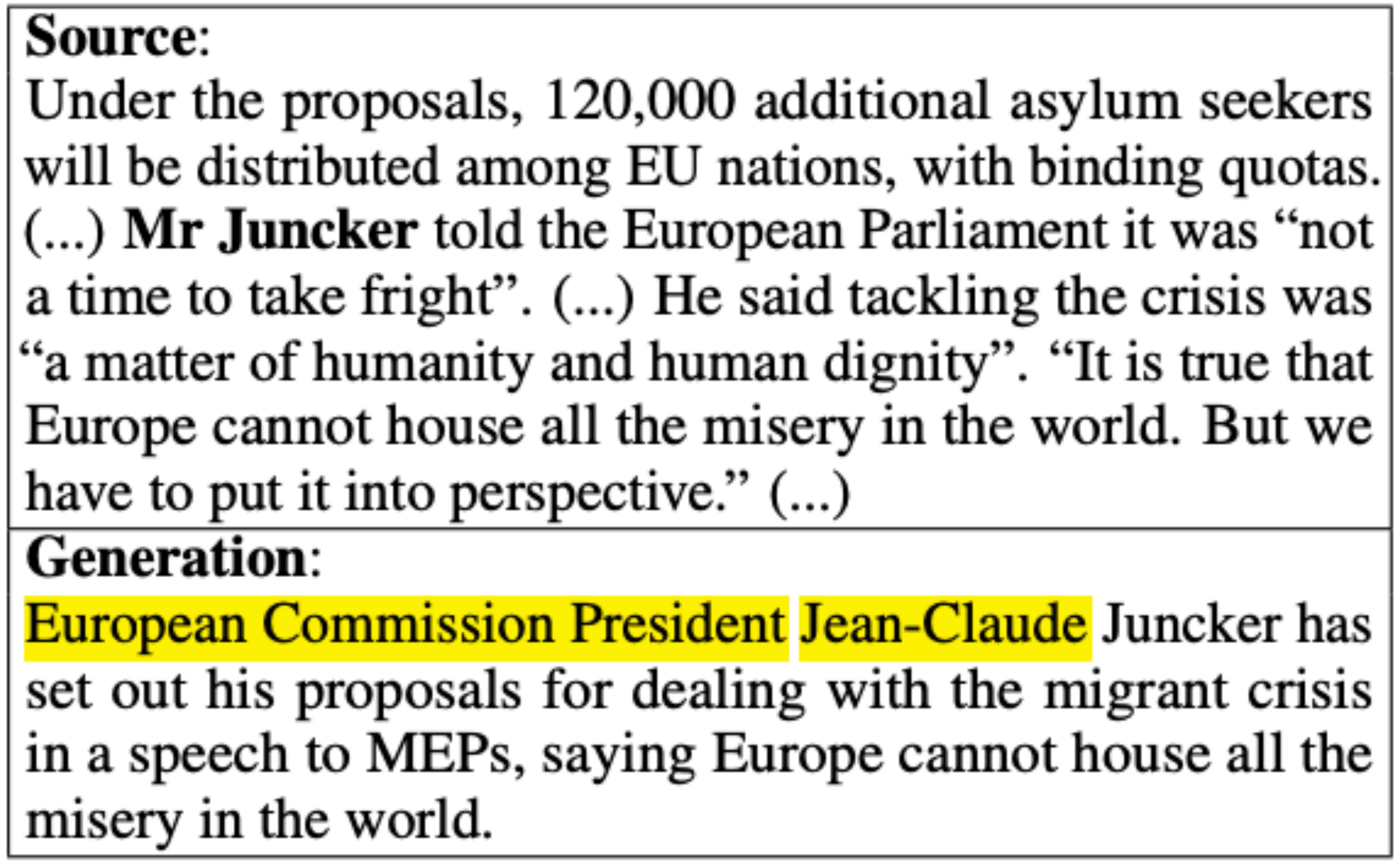
  “the perception of an entity or event that is absent in reality”

- In the field of NLP however this term is a little more conflated. There are many definitions of hallucinations

[1] Macpherson, F., & Platchias, D. (Eds.). (2013). *Hallucination: Philosophy and psychology*. MIT Press.

- Before LLMs a well accepted definition of hallucination was as follows:

  “generated content that is nonsensical or unfaithful to the provided source content”



Source:
Under the proposals, 120,000 additional asylum seekers will be distributed among EU nations, with binding quotas. (...) **Mr Juncker** told the European Parliament it was “not a time to take fright”. (...) He said tackling the crisis was “a matter of humanity and human dignity”. “It is true that Europe cannot house all the misery in the world. But we have to put it into perspective.” (...)

Generation:
European Commission President Jean-Claude Juncker has set out his proposals for dealing with the migrant crisis in a speech to MEPs, saying Europe cannot house all the misery in the world.

[Ji et al;2023]

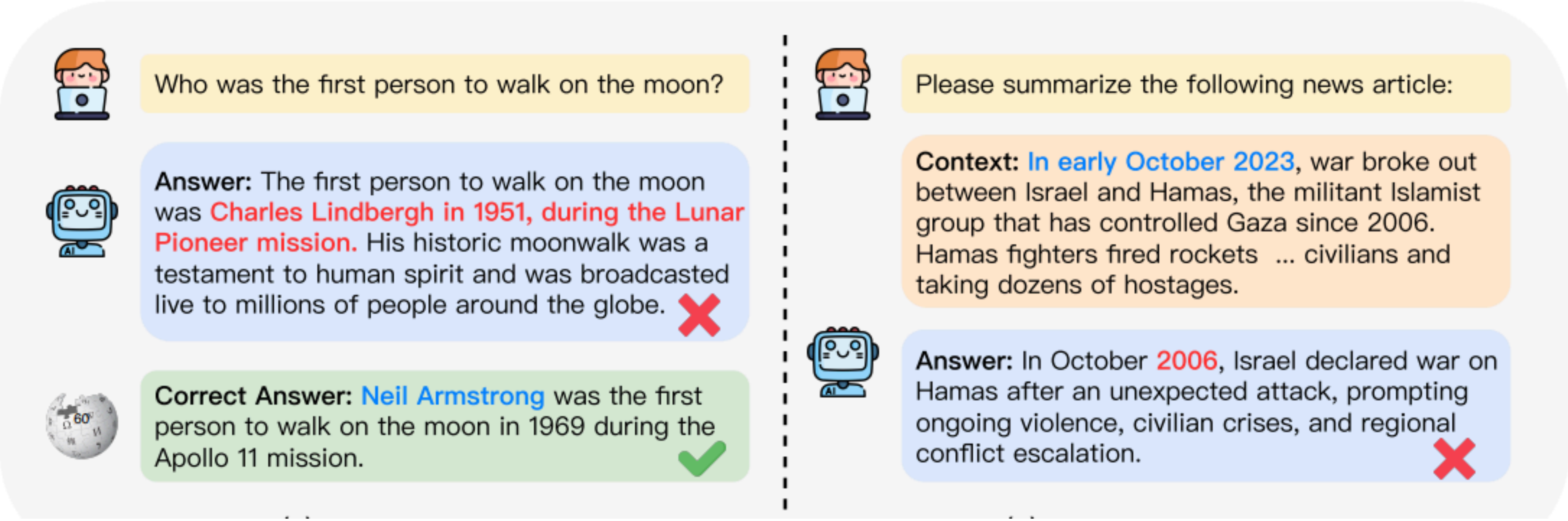
- LLMs can produce more sophisticated hallucinations — including completely fabricated facts

- The term hallucination encompasses:

  A. Factual Hallucinations :
  Discrepancy between generated content and real-world facts

  B. Faithful Hallucinations :
  Discrepancy between generated content and context

Example



Who was the first person to walk on the moon?

Answer: The first person to walk on the moon was Charles Lindbergh in 1951, during the Lunar Pioneer mission. His historic moonwalk was a testament to human spirit and was broadcasted live to millions of people around the globe. ✗

Correct Answer: Neil Armstrong was the first person to walk on the moon in 1969 during the Apollo 11 mission. ✓

Please summarize the following news article:

Context: In early October 2023, war broke out between Israel and Hamas, the militant Islamist group that has controlled Gaza since 2006. Hamas fighters fired rockets ... civilians and taking dozens of hostages.

Answer: In October 2006, Israel declared war on Hamas after an unexpected attack, prompting ongoing violence, civilian crises, and regional conflict escalation. ✗

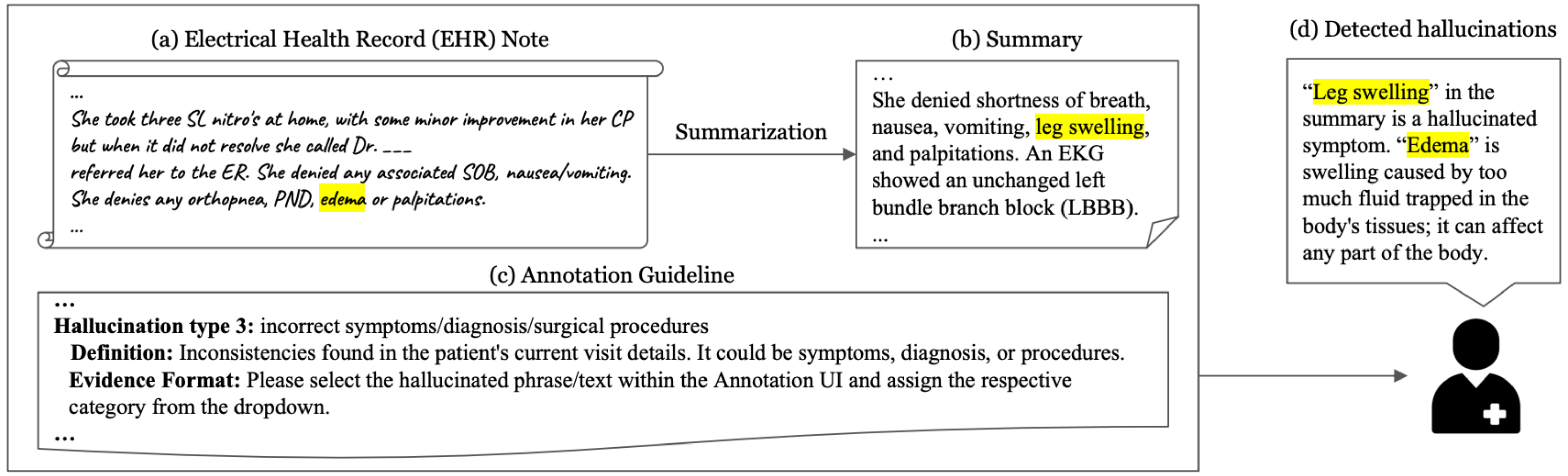
The example on the left illustrates fabricated content, where the model incorrectly provides an answer to the question of who was the first person to walk on the moon. The example on the right demonstrates unsupported content, where the model inaccurately generated the date of the Israel-Hamas conflict, shifting it from October 2023 to October 2006.

[Huang et al;2023]

Note: For this talk we will focus on faithful hallucinations in summarization

Why care about hallucinations?

- Hinders performance and raises safety concerns in real-world applications
  - Misinformation
  - Risk in consequential domains (law, medicine)
  - Potential privacy violations

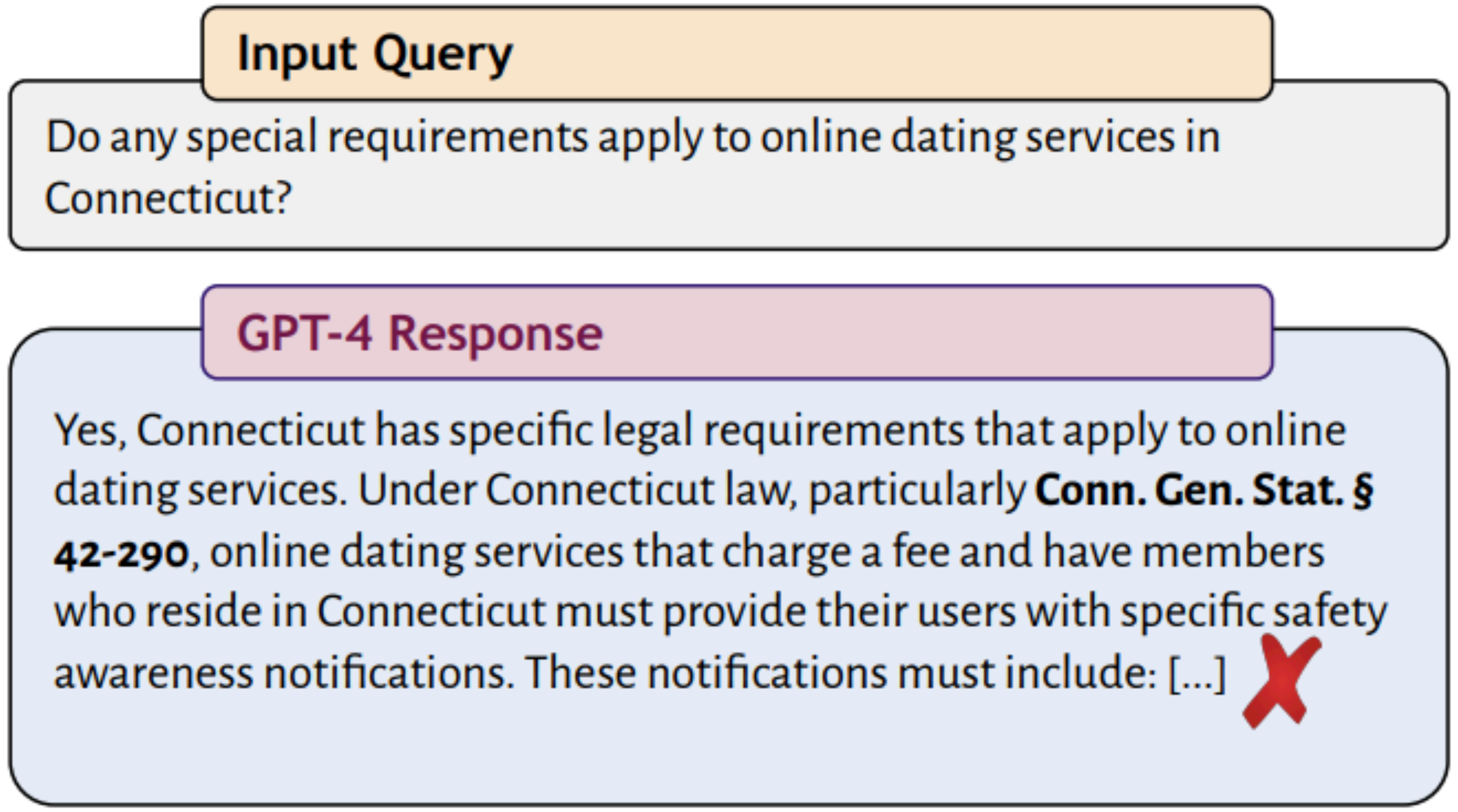

(a) Electrical Health Record (EHR) Note
Summarization
(b) Summary
(c) Annotation Guideline
(d) Detected hallucinations

...
She took three SL nitro's at home, with some minor improvement in her CP but when it did not resolve she called Dr. ___ referred her to the ER. She denied any associated SOB, nausea/vomiting. She denies any orthopnea, PND, edema or palpitations.
...

...
She denied shortness of breath, nausea, vomiting, leg swelling, and palpitations. An EKG showed an unchanged left bundle branch block (LBBB).
...

...
Hallucination type 3: incorrect symptoms/diagnosis/surgical procedures
Definition: Inconsistencies found in the patient's current visit details. It could be symptoms, diagnosis, or procedures.
Evidence Format: Please select the hallucinated phrase/text within the Annotation UI and assign the respective category from the dropdown.
...

"Leg swelling" in the summary is a hallucinated symptom. "Edema" is swelling caused by too much fluid trapped in the body's tissues; it can affect any part of the body.



Input Query

Do any special requirements apply to online dating services in Connecticut?

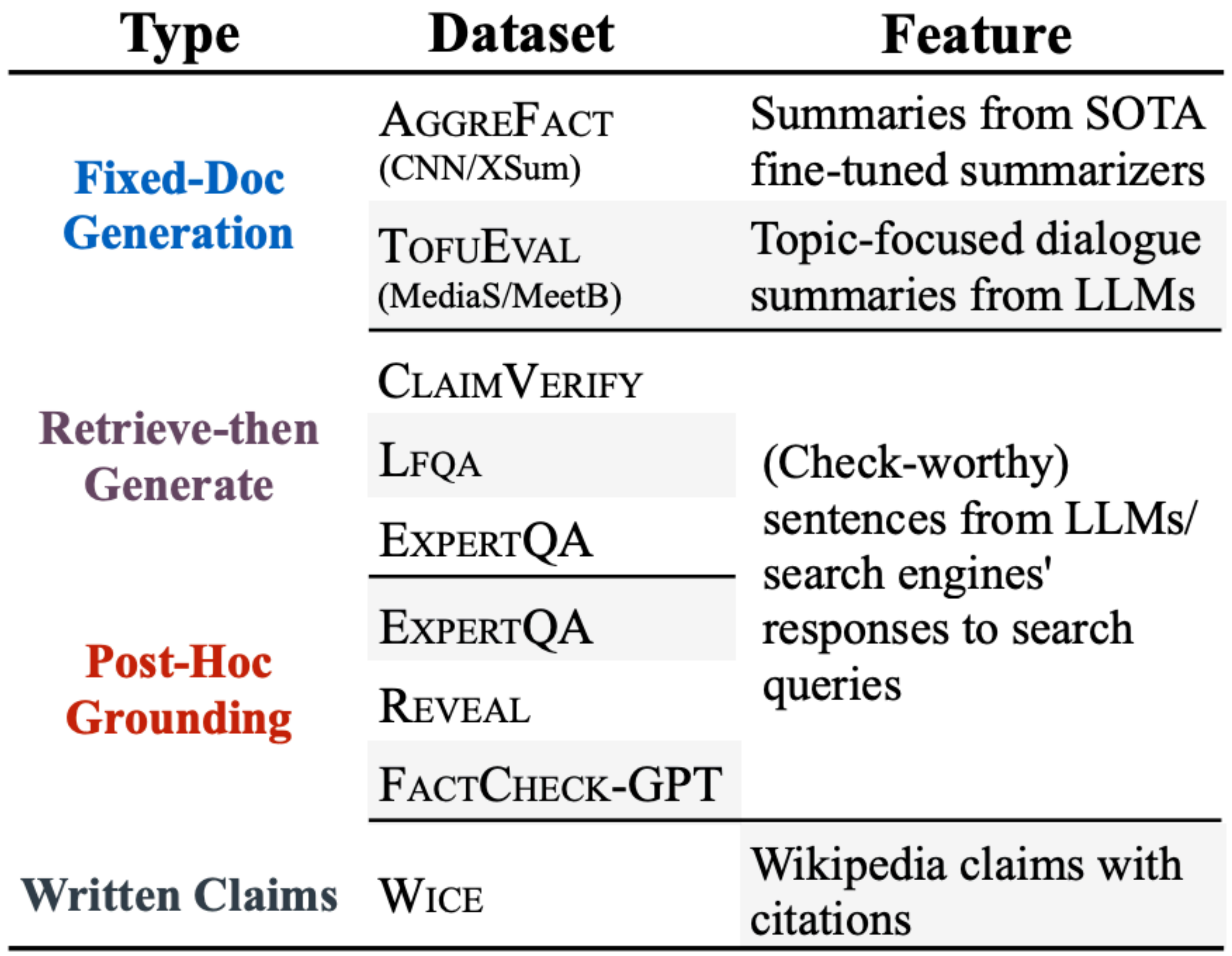
GPT-4 Response

Yes, Connecticut has specific legal requirements that apply to online dating services. Under Connecticut law, particularly **Conn. Gen. Stat. § 42-290**, online dating services that charge a fee and have members who reside in Connecticut must provide their users with specific safety awareness notifications. These notifications must include: [...] ✗

Factuality and Summarization

Experimental Setups

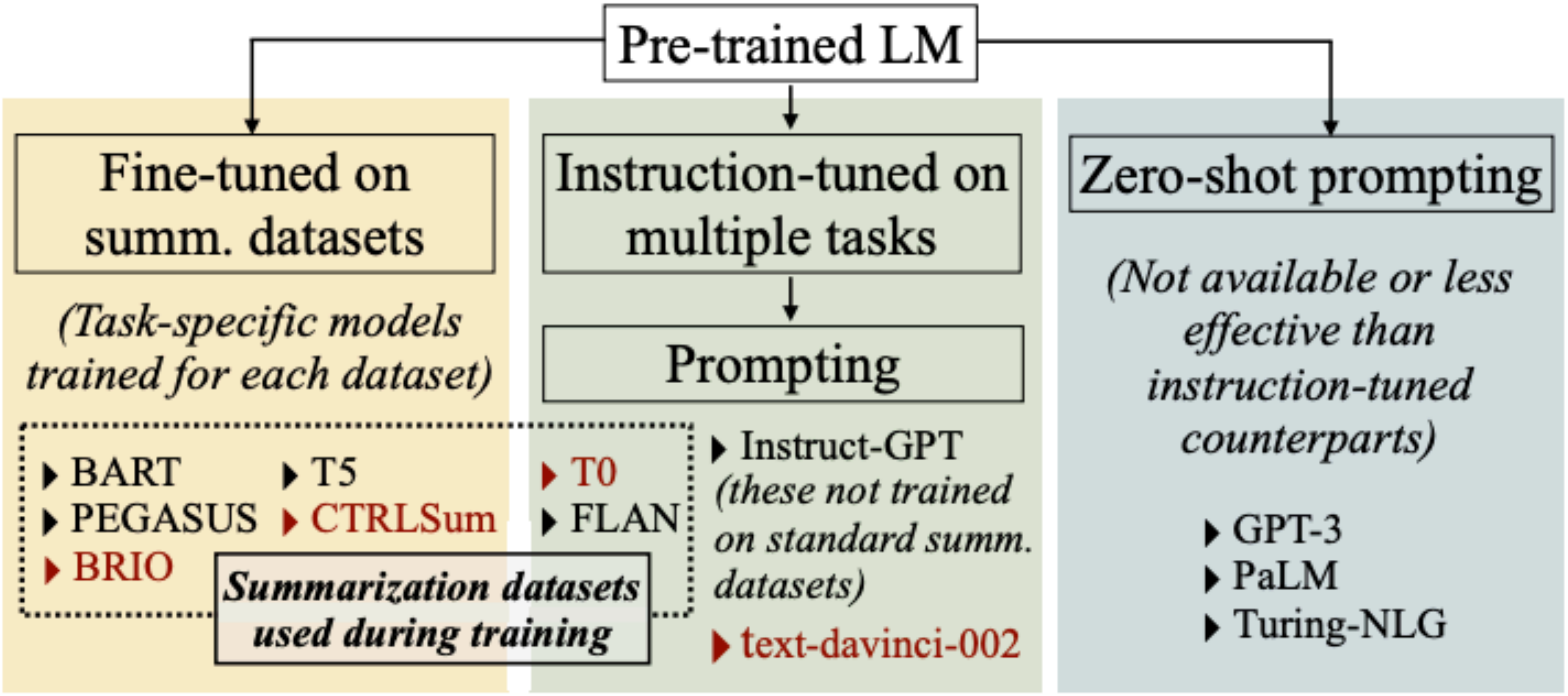
- Summarization datasets typically include



Type	Dataset	Feature
Fixed-Doc Generation	AggreFact (CNN/XSum)	Summaries from SOTA fine-tuned summarizers
	TofuEval (MediaS/MeetB)	Topic-focused dialogue summaries from LLMs
Retrieve-then Generate	ClaimVerify	(Check-worthy) sentences from LLMs/ search engines' responses to search queries
	LFQA	
	ExpertQA	
Post-Hoc Grounding	ExpertQA	
	Reveal	
	FactCheck-GPT	
Written Claims	Wice	Wikipedia claims with citations

- Most popular among all to evaluate on has been news (cnn/xsum/bbc)

- Broad categorization of summarization systems is as follows


Pre-trained LM
Fine-tuned on summ. datasets
(Task-specific models trained for each dataset)
BART
PEGASUS
BRIO
T5
CTRLSum
Summarization datasets used during training
Instruction-tuned on multiple tasks
Prompting
T0
FLAN
Instruct-GPT
(these not trained on standard summ. datasets)
text-davinci-002
Zero-shot prompting
(Not available or less effective than instruction-tuned counterparts)
GPT-3
PaLM
Turing-NLG

News Summarization and Evaluation in the Era of GPT-3

Tanya Goyal[1] **Junyi Jessy Li**[2] **Greg Durrett**[1]
[1] Department of Computer Science [2] Department of Linguistics
The University of Texas at Austin
`tanyagoyal@utexas.edu`

CNN Article: *(CNN) Mortgage rates fell slightly this week, marking the third consecutive week of declines. But with rates hovering above 5% and home prices well above where they were this time last year, prospective buyers are finding it increasingly difficult to afford a home. The 30-year, fixed-rate mortgage averaged 5.09% in the week ending June 2, down from 5.10% the week before, according to Freddie Mac. It is still well above the 2.99% average from this time last year. […] Rising prices had already been pushing many prospective buyers to the sidelines. […] The Fed has been seeking to tame inflation by raising interest rates over the past couple of months. And the central bank has signaled there are more rate hikes to come. […]*

BRIO The 30-year, fixed-rate mortgage averaged 5.09% in the week ending June 2. It's the third consecutive week of declines. But rates are still above 5% and home prices are well above where they were this time last year. Rising prices have pushed some prospective buyers to the sidelines.

T0 30-year, fixed-rate mortgage averaged 5.09% in the week ending June 2. It is still well above the 2.99% average from this time last year. Rising prices had already been pushing many prospective buyers to the sidelines.

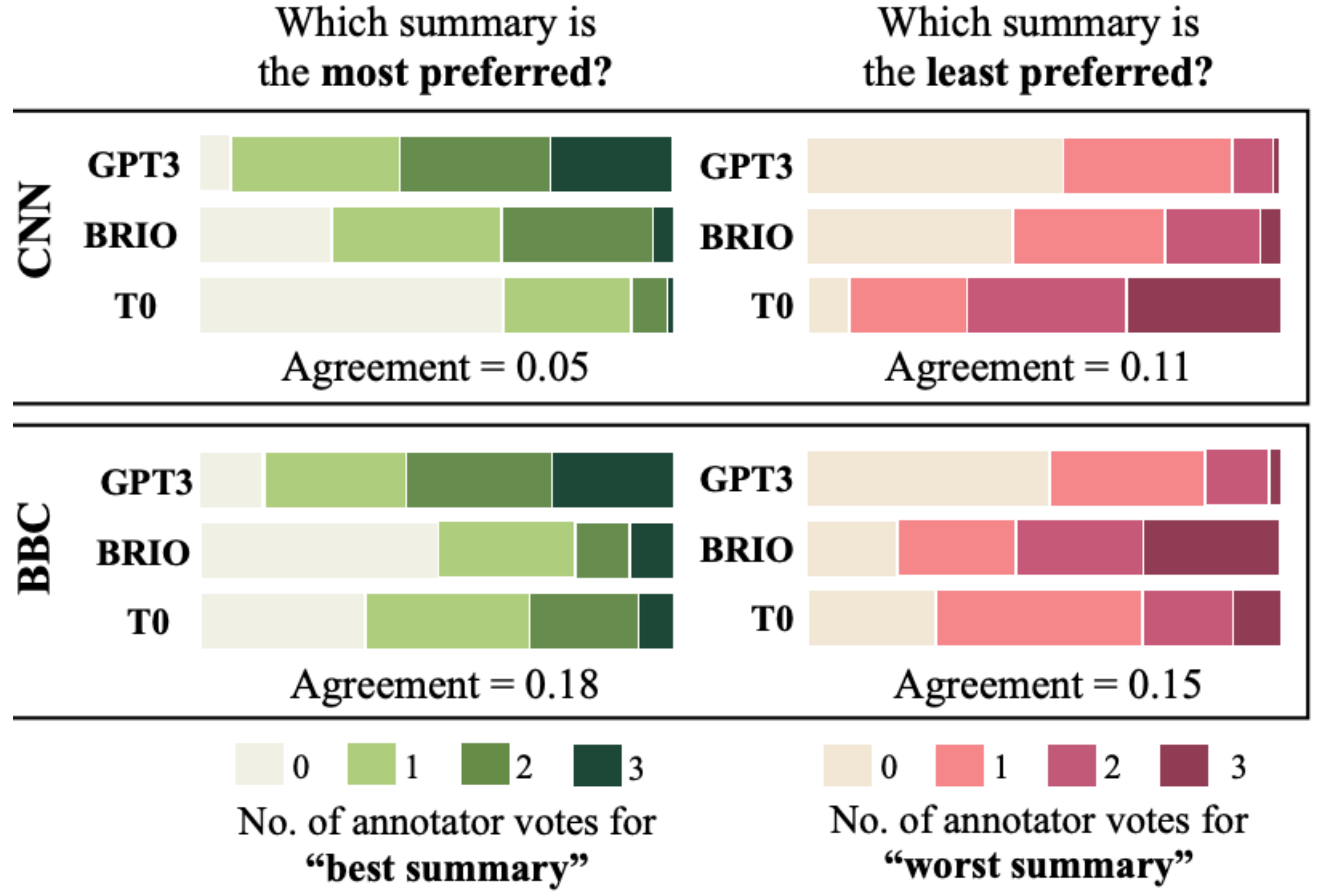
GPT3 -D2 Mortgage rates have fallen slightly in the past few weeks, but they are still well above last year's levels. This is making it difficult for many prospective buyers to afford a home. The Fed has been raising interest rates in an effort to control inflation, and more rate hikes are expected.

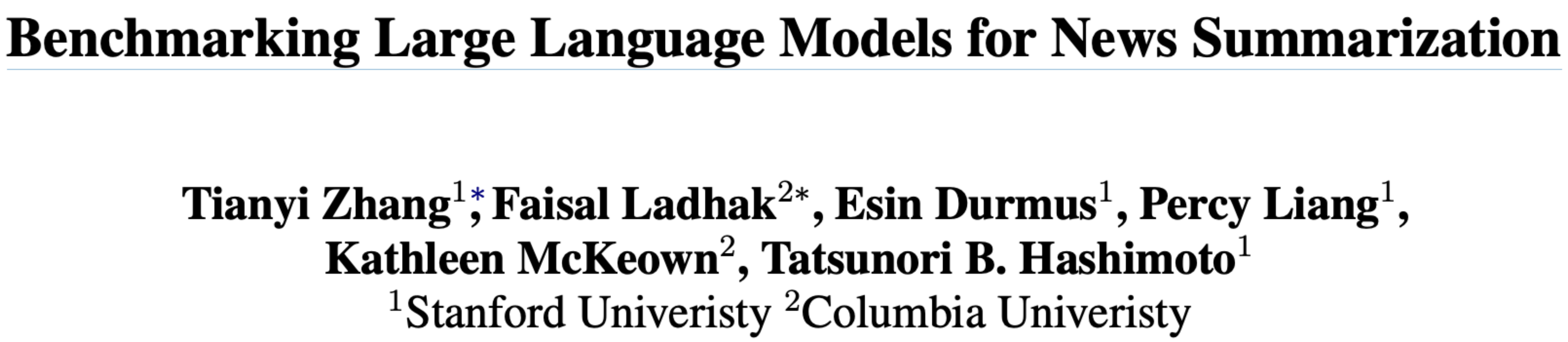
BBC Article: *The full opening of a landmark shared education campus in Omagh is facing another delay. Education Minister Michelle McIlveen has now said the Strule Shared Education Campus is planned to open in September 2026. The minister clarified the new date in response to an assembly question from the SDLP MLA Daniel McCrossan. The campus had originally been due to open in 2020, but has been delayed a number of times.[…]*

BRIO The full opening of the Strule shared education campus in Omagh, County Tyrone, has been delayed to September 2026.

T0 Strule, Northern Ireland's biggest ever school-building project, is now scheduled to open in 2026, the Education Minister has said.

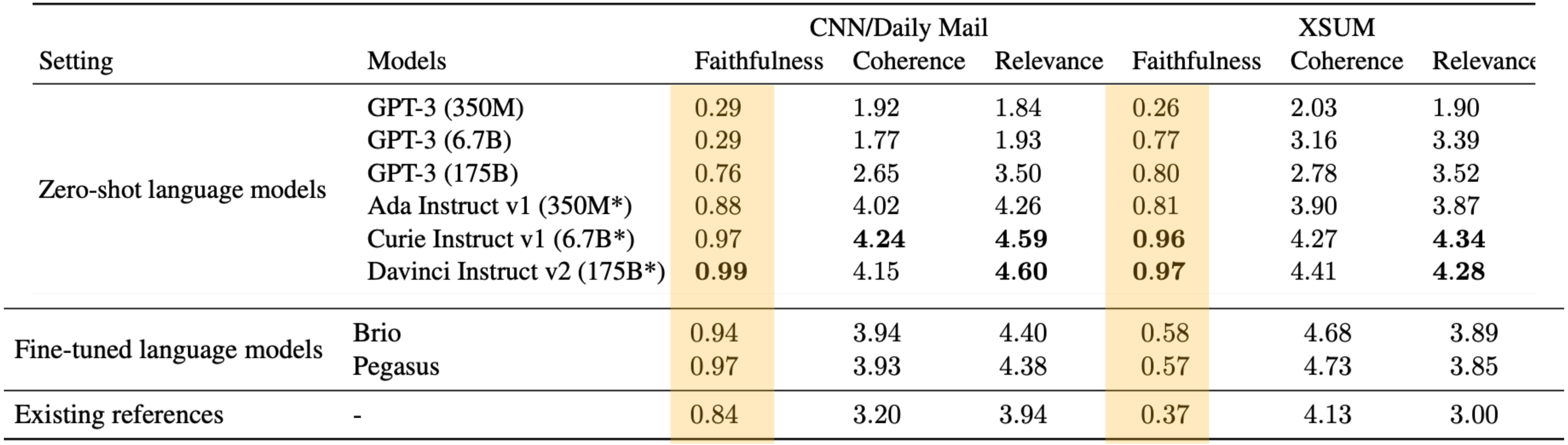
GPT3 -D2 The Strule Shared Education Campus is facing another delay, and is now planned to open in September 2026.



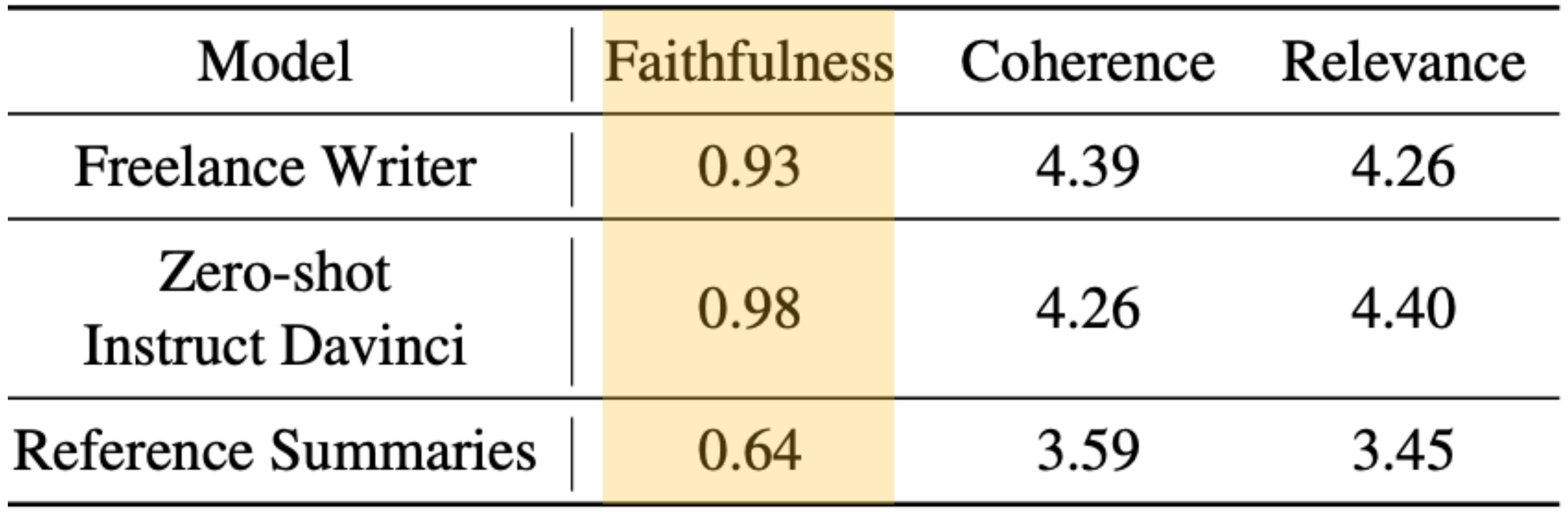


Benchmarking Large Language Models for News Summarization

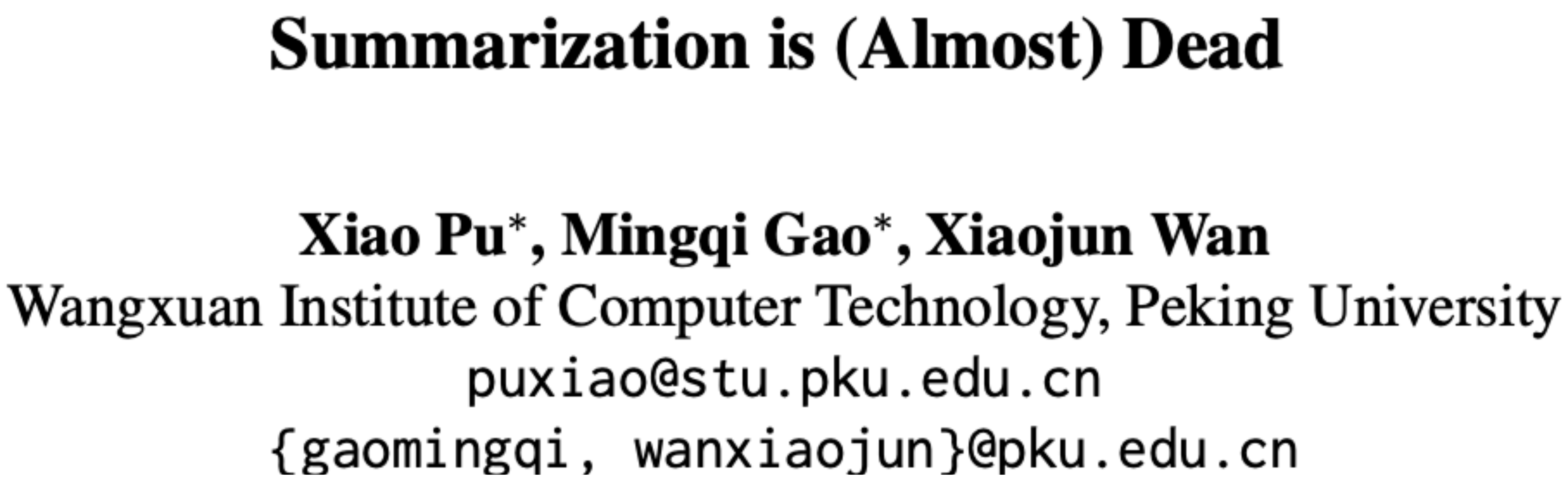
Tianyi Zhang[1]*, Faisal Ladhak[2*], Esin Durmus[1], Percy Liang[1],
Kathleen McKeown[2], Tatsunori B. Hashimoto[1]
[1]Stanford Univeristy [2]Columbia Univeristy



Setting	Models	CNN/Daily Mail Faithfulness	CNN/Daily Mail Coherence	CNN/Daily Mail Relevance	XSUM Faithfulness	XSUM Coherence	XSUM Relevance
Zero-shot language models	GPT-3 (350M)	0.29	1.92	1.84	0.26	2.03	1.90
	GPT-3 (6.7B)	0.29	1.77	1.93	0.77	3.16	3.39
	GPT-3 (175B)	0.76	2.65	3.50	0.80	2.78	3.52
	Ada Instruct v1 (350M*)	0.88	4.02	4.26	0.81	3.90	3.87
	Curie Instruct v1 (6.7B*)	0.97	**4.24**	**4.59**	**0.96**	4.27	**4.34**
	Davinci Instruct v2 (175B*)	**0.99**	4.15	**4.60**	**0.97**	4.41	**4.28**
Fine-tuned language models	Brio	0.94	3.94	4.40	0.58	4.68	3.89
	Pegasus	0.97	3.93	4.38	0.57	4.73	3.85
Existing references	-	0.84	3.20	3.94	0.37	4.13	3.00

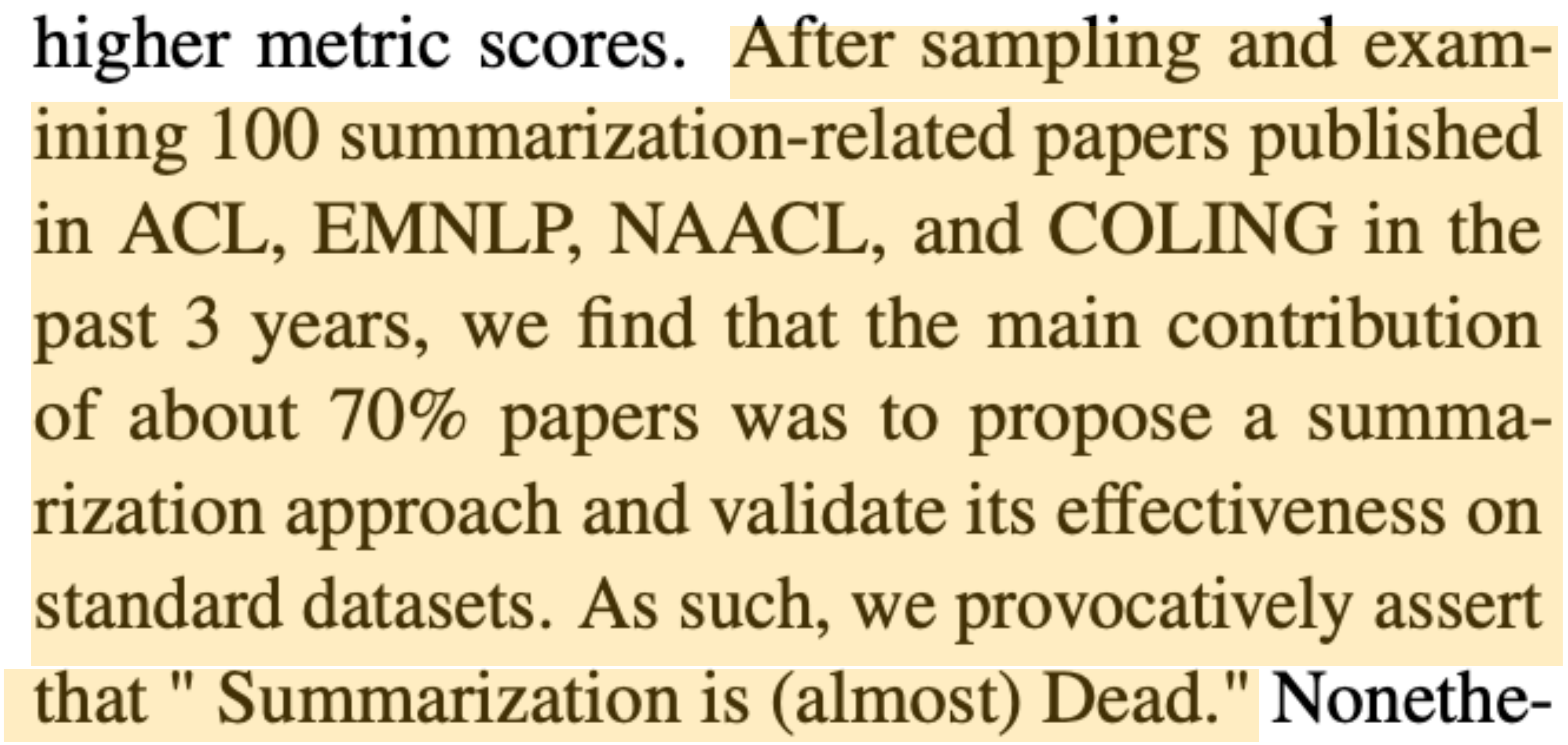


Model	Faithfulness	Coherence	Relevance
Freelance Writer	0.93	4.39	4.26
Zero-shot Instruct Davinci	0.98	4.26	4.40
Reference Summaries	0.64	3.59	3.45

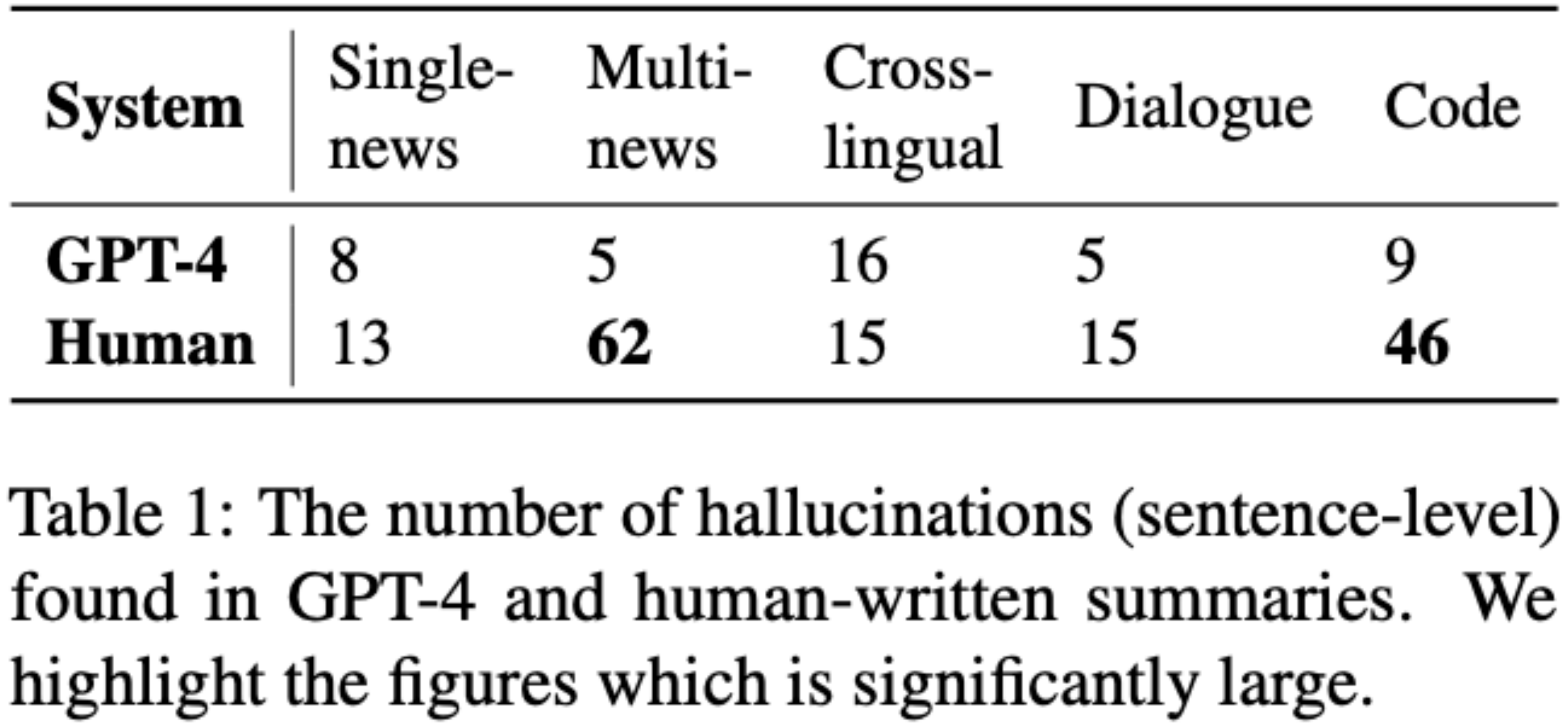


Summarization is (Almost) Dead

Xiao Pu[*], Mingqi Gao[*], Xiaojun Wan
Wangxuan Institute of Computer Technology, Peking University
`puxiao@stu.pku.edu.cn`
`{gaomingqi, wanxiaojun}@pku.edu.cn`

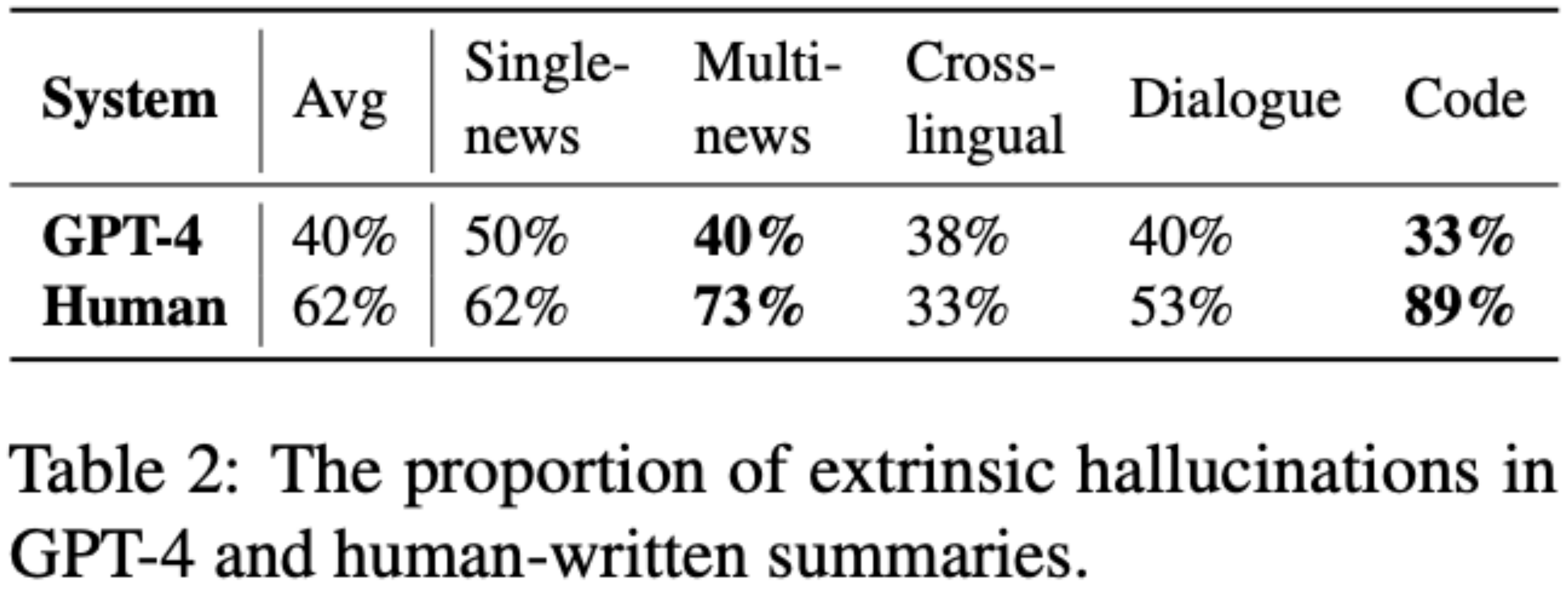


higher metric scores. After sampling and examining 100 summarization-related papers published in ACL, EMNLP, NAACL, and COLING in the past 3 years, we find that the main contribution of about 70% papers was to propose a summarization approach and validate its effectiveness on standard datasets. As such, we provocatively assert that " Summarization is (almost) Dead." Nonethe-



System	Single-news	Multi-news	Cross-lingual	Dialogue	Code
GPT-4	8	5	16	5	9
Human	13	**62**	15	15	**46**

Table 1: The number of hallucinations (sentence-level) found in GPT-4 and human-written summaries. We highlight the figures which is significantly large.

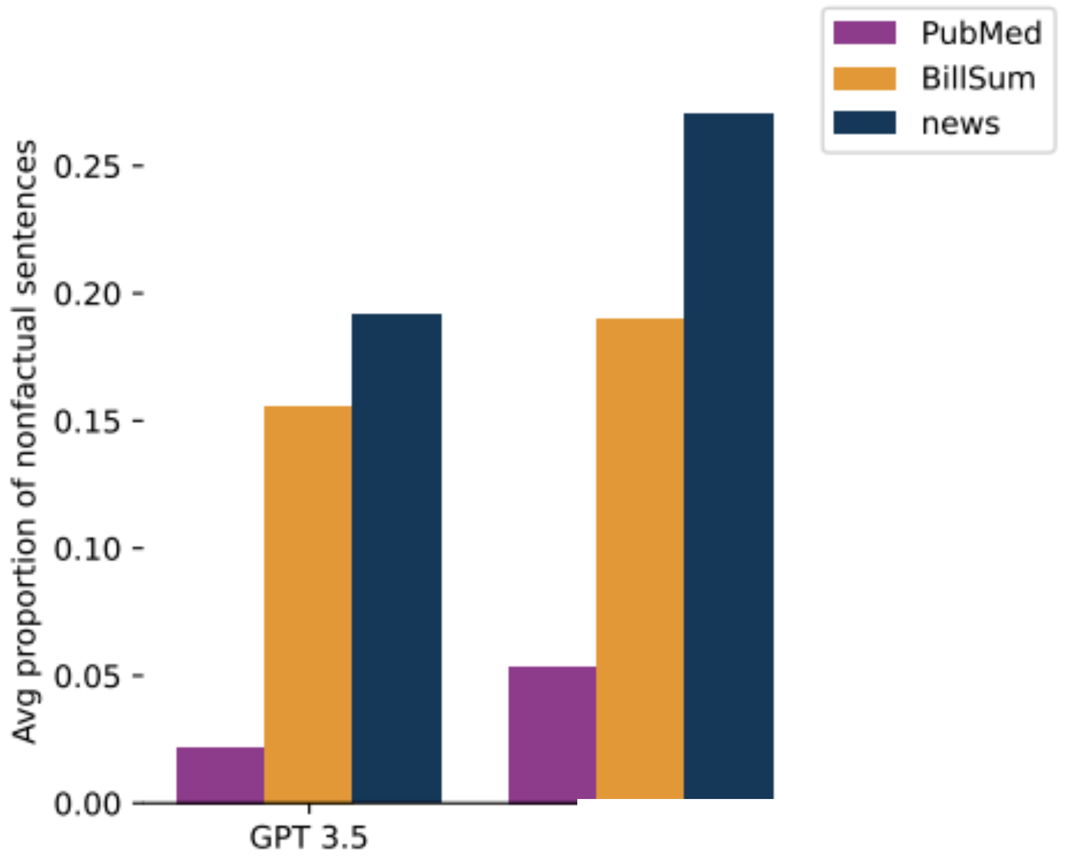


System	Avg	Single-news	Multi-news	Cross-lingual	Dialogue	Code
GPT-4	40%	50%	**40%**	38%	40%	**33%**
Human	62%	62%	**73%**	33%	53%	**89%**

Table 2: The proportion of extrinsic hallucinations in GPT-4 and human-written summaries.

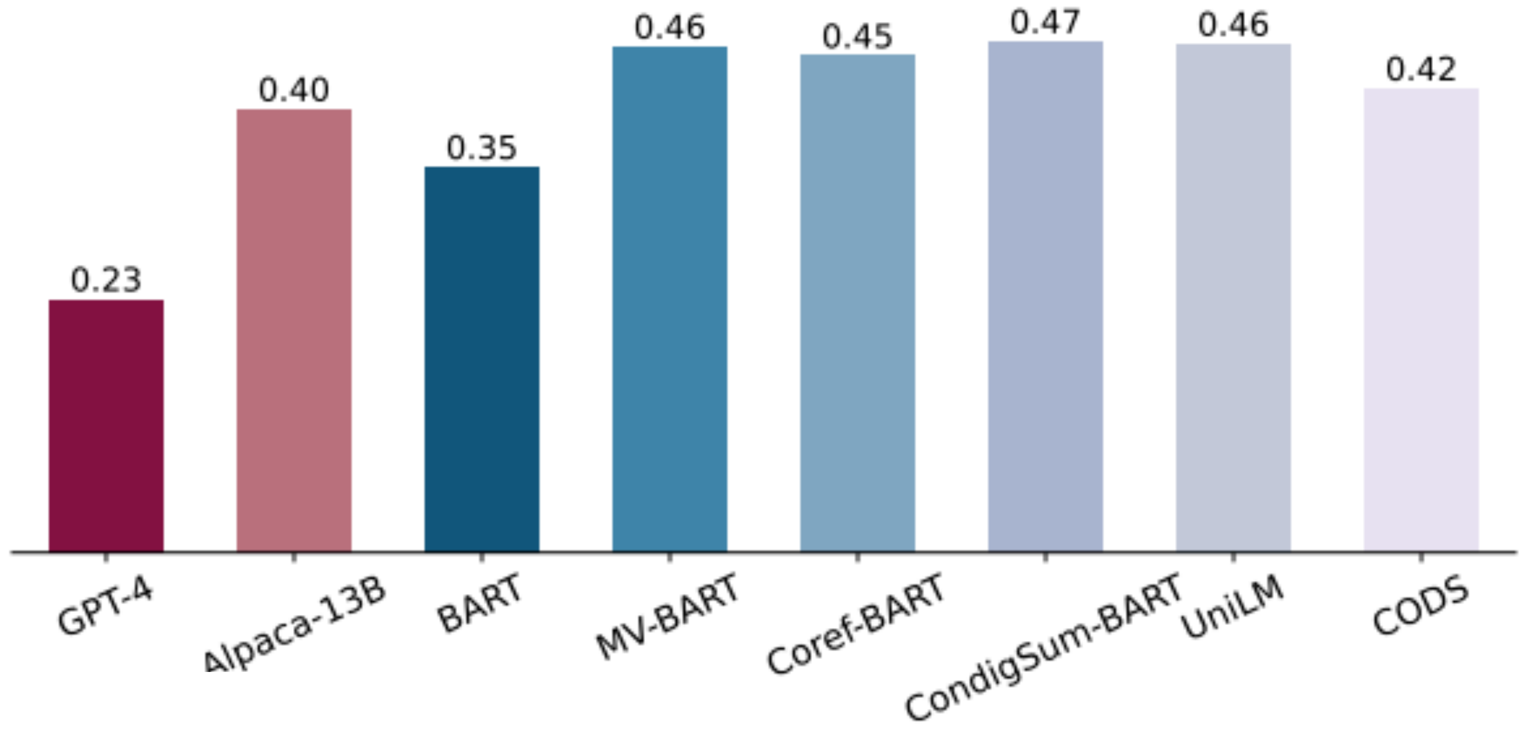
Evaluating the Factuality of Zero-shot Summarizers Across Varied Domains

Sanjana Ramprasad[♢] **Kundan Krishna**[♣] **Zachary C. Lipton**[♣] **Byron C. Wallace**
[♢]Northeastern University
[♣] Carnegie Mellon University



Analyzing LLM Behavior in Dialogue Summarization: Unveiling Circumstantial Hallucination Trends

Sanjana Ramprasad[♢] **Elisa Ferracane**[♣] **Zachary C. Lipton**[♣]
[♢]Northeastern University
[♣] Abridge AI

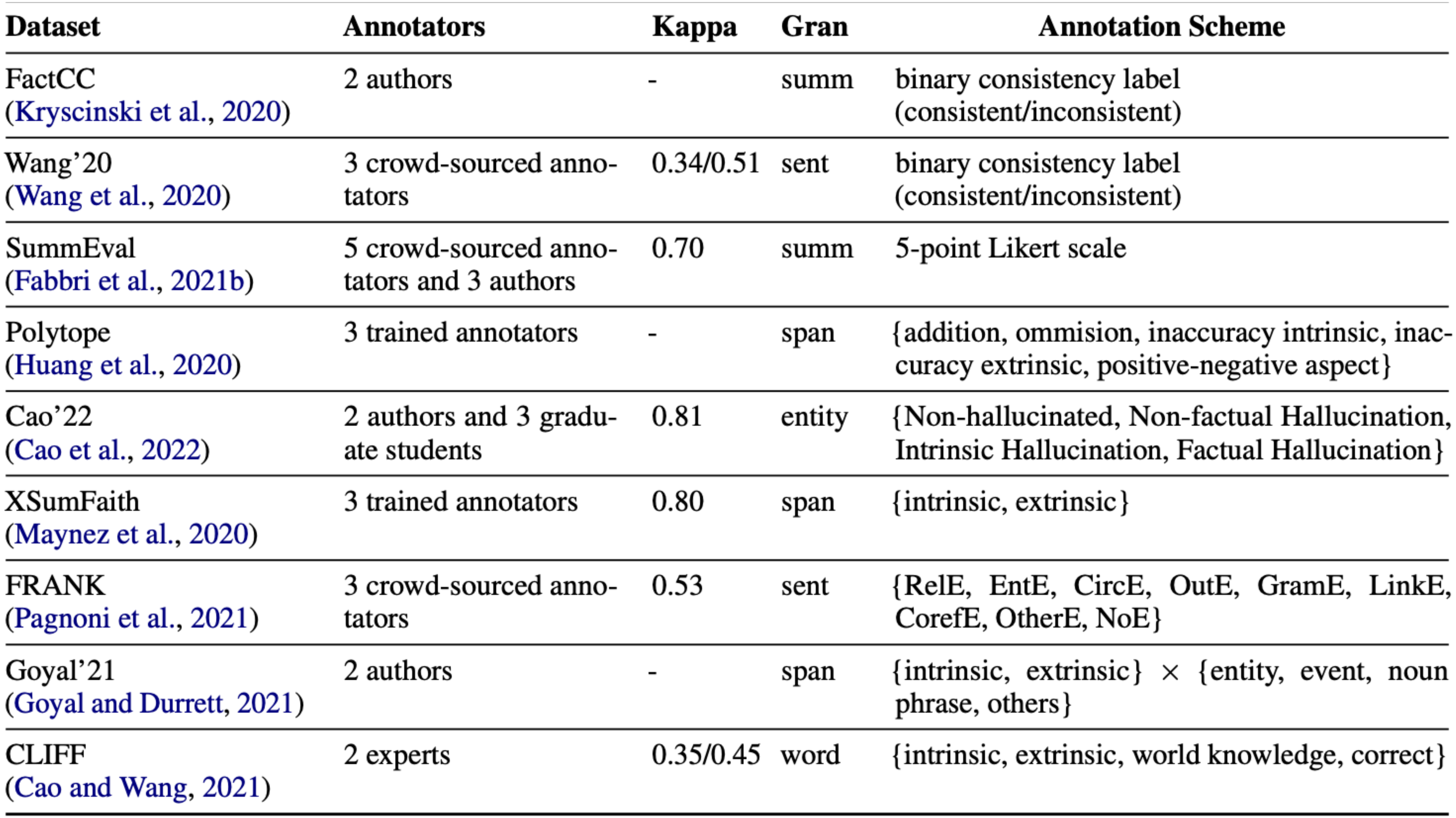


TofuEval: Evaluating Hallucinations of LLMs on Topic-Focused Dialogue Summarization

Liyan Tang[♢†], **Igor Shalyminov**[♠], **Amy Wing-mei Wong**[♠], **Jon Burnsky**[♠], **Jake W. Vincent**[♠]
Yu'an Yang[♠], **Siffi Singh**[♠], **Song Feng**[♠], **Hwanjun Song**[♡‡], **Hang Su**[♠], **Lijia Sun**[♠],
Yi Zhang[♠], **Saab Mansour**[♠], **Kathleen McKeown**[♠]

Summ. Model	**Sentence-Level (% Error)**				**Summary-Level (% Error)**			
	MediaSum		**Meetingbank**		**MediaSum**		**Meetingbank**	
	Main	**Marginal**	**Main**	**Marginal**	**Main**	**Marginal**	**Main**	**Marginal**
Vicuna-7B	19.6	35.8	17.6	36.8	42.7	55.4	33.0	58.0
WizardLM-7B	29.1	36.4	21.3	42.4	49.6	54.8	35.6	49.0
WizardLM-13B	17.4	27.2	15.8	25.4	35.9	44.4	41.3	46.8
WizardLM-30B	14.6	27.2	13.7	26.2	35.9	48.2	31.5	44.8
GPT-3.5-Turbo	8.8	13.6	4.4	9.4	22.2	27.2	10.9	19.8
Average	17.5	27.8	14.4	27.8	37.2	46.0	30.4	43.6

Human Evaluation

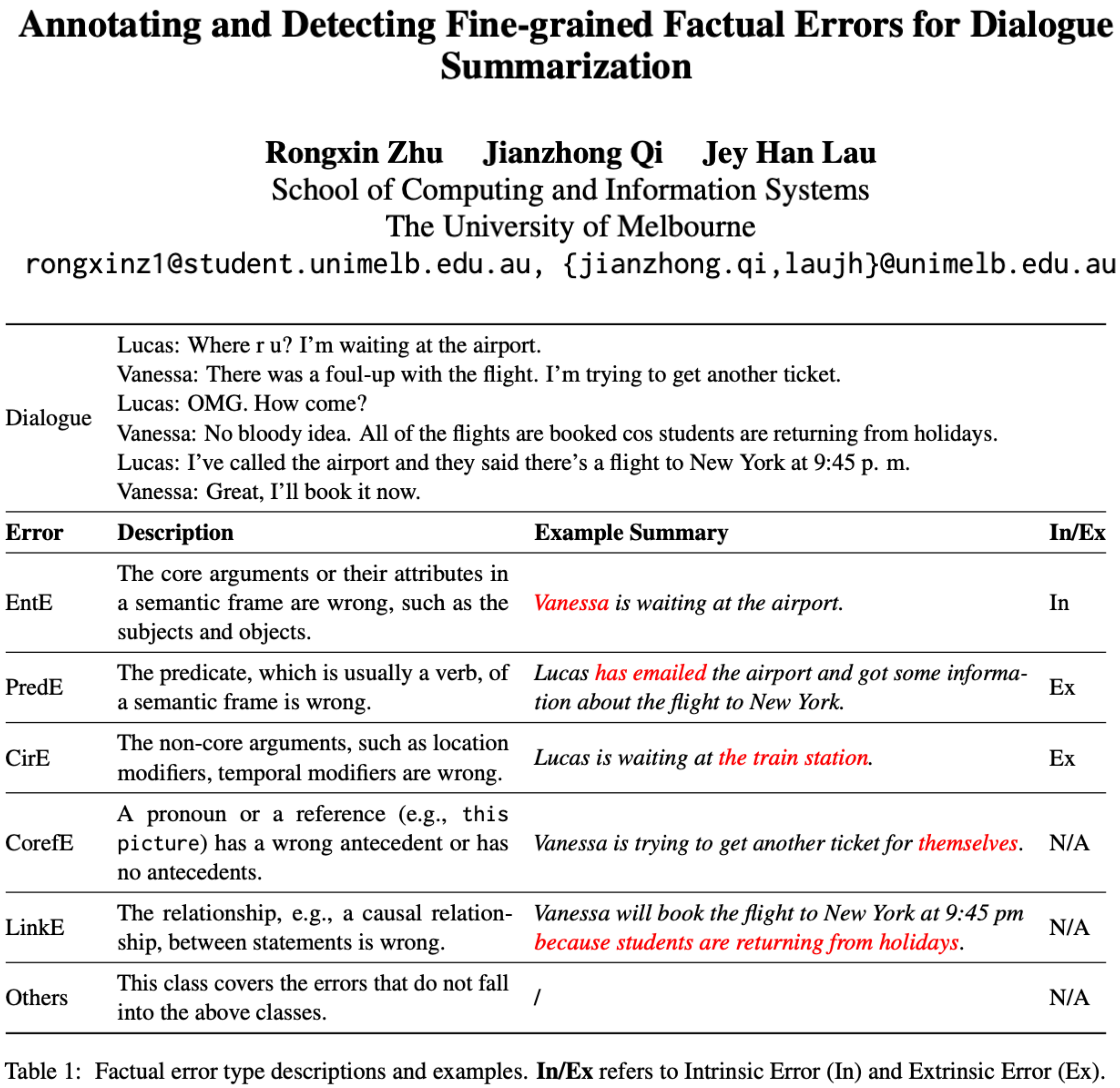


Dataset	Annotators	Kappa	Gran	Annotation Scheme
FactCC (Kryscinski et al., 2020)	2 authors	-	summ	binary consistency label (consistent/inconsistent)
Wang'20 (Wang et al., 2020)	3 crowd-sourced annotators	0.34/0.51	sent	binary consistency label (consistent/inconsistent)
SummEval (Fabbri et al., 2021b)	5 crowd-sourced annotators and 3 authors	0.70	summ	5-point Likert scale
Polytope (Huang et al., 2020)	3 trained annotators	-	span	{addition, ommision, inaccuracy intrinsic, inaccuracy extrinsic, positive-negative aspect}
Cao'22 (Cao et al., 2022)	2 authors and 3 graduate students	0.81	entity	{Non-hallucinated, Non-factual Hallucination, Intrinsic Hallucination, Factual Hallucination}
XSumFaith (Maynez et al., 2020)	3 trained annotators	0.80	span	{intrinsic, extrinsic}
FRANK (Pagnoni et al., 2021)	3 crowd-sourced annotators	0.53	sent	{RelE, EntE, CircE, OutE, GramE, LinkE, CorefE, OtherE, NoE}
Goyal'21 (Goyal and Durrett, 2021)	2 authors	-	span	{intrinsic, extrinsic} × {entity, event, noun phrase, others}
CLIFF (Cao and Wang, 2021)	2 experts	0.35/0.45	word	{intrinsic, extrinsic, world knowledge, correct}

Proposed benchmarks for summarization

Some annotation instructions

- Is the summary faithful to the source?
- Is the summary factually correct based on the article provided?
- Is the claim correct (supported by the news article) or incorrect (unsupported)? A claim may be incorrect because it mis-states information in the article or introduces new information



Annotating and Detecting Fine-grained Factual Errors for Dialogue Summarization

Rongxin Zhu **Jianzhong Qi** **Jey Han Lau**
School of Computing and Information Systems
The University of Melbourne
`rongxinz1@student.unimelb.edu.au, {jianzhong.qi,laujh}@unimelb.edu.au`

Dialogue	Lucas: Where r u? I'm waiting at the airport. Vanessa: There was a foul-up with the flight. I'm trying to get another ticket. Lucas: OMG. How come? Vanessa: No bloody idea. All of the flights are booked cos students are returning from holidays. Lucas: I've called the airport and they said there's a flight to New York at 9:45 p. m. Vanessa: Great, I'll book it now.		
Error	**Description**	**Example Summary**	**In/Ex**
EntE	The core arguments or their attributes in a semantic frame are wrong, such as the subjects and objects.	*Vanessa is waiting at the airport.*	In
PredE	The predicate, which is usually a verb, of a semantic frame is wrong.	*Lucas has emailed the airport and got some information about the flight to New York.*	Ex
CirE	The non-core arguments, such as location modifiers, temporal modifiers are wrong.	*Lucas is waiting at the train station.*	Ex
CorefE	A pronoun or a reference (e.g., `this picture`) has a wrong antecedent or has no antecedents.	*Vanessa is trying to get another ticket for themselves.*	N/A
LinkE	The relationship, e.g., a causal relationship, between statements is wrong.	*Vanessa will book the flight to New York at 9:45 pm because students are returning from holidays.*	N/A
Others	This class covers the errors that do not fall into the above classes.	/	N/A

Table 1: Factual error type descriptions and examples. **In/Ex** refers to Intrinsic Error (In) and Extrinsic Error (Ex).

Understanding Factual Errors in Summarization: Errors, Summarizers, Datasets, Error Detectors

Liyan Tang[♢], Tanya Goyal[♢], Alexander R. Fabbri[♠], Philippe Laban[♠],
Jiacheng Xu[♢,♠], Semih Yavuz[♠], Wojciech Kryściński[♠], Justin F. Rousseau[♢], Greg Durrett[♢]

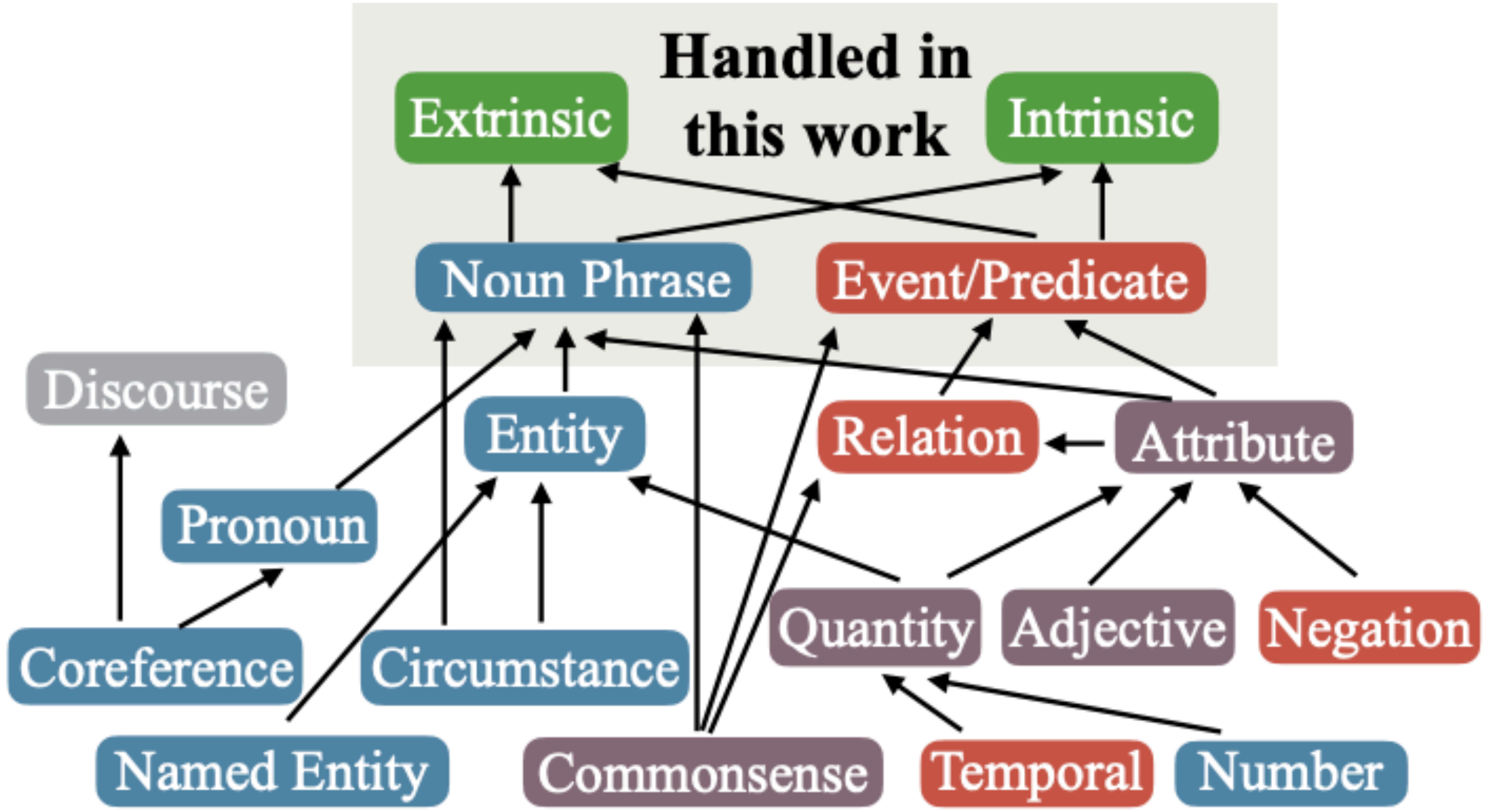


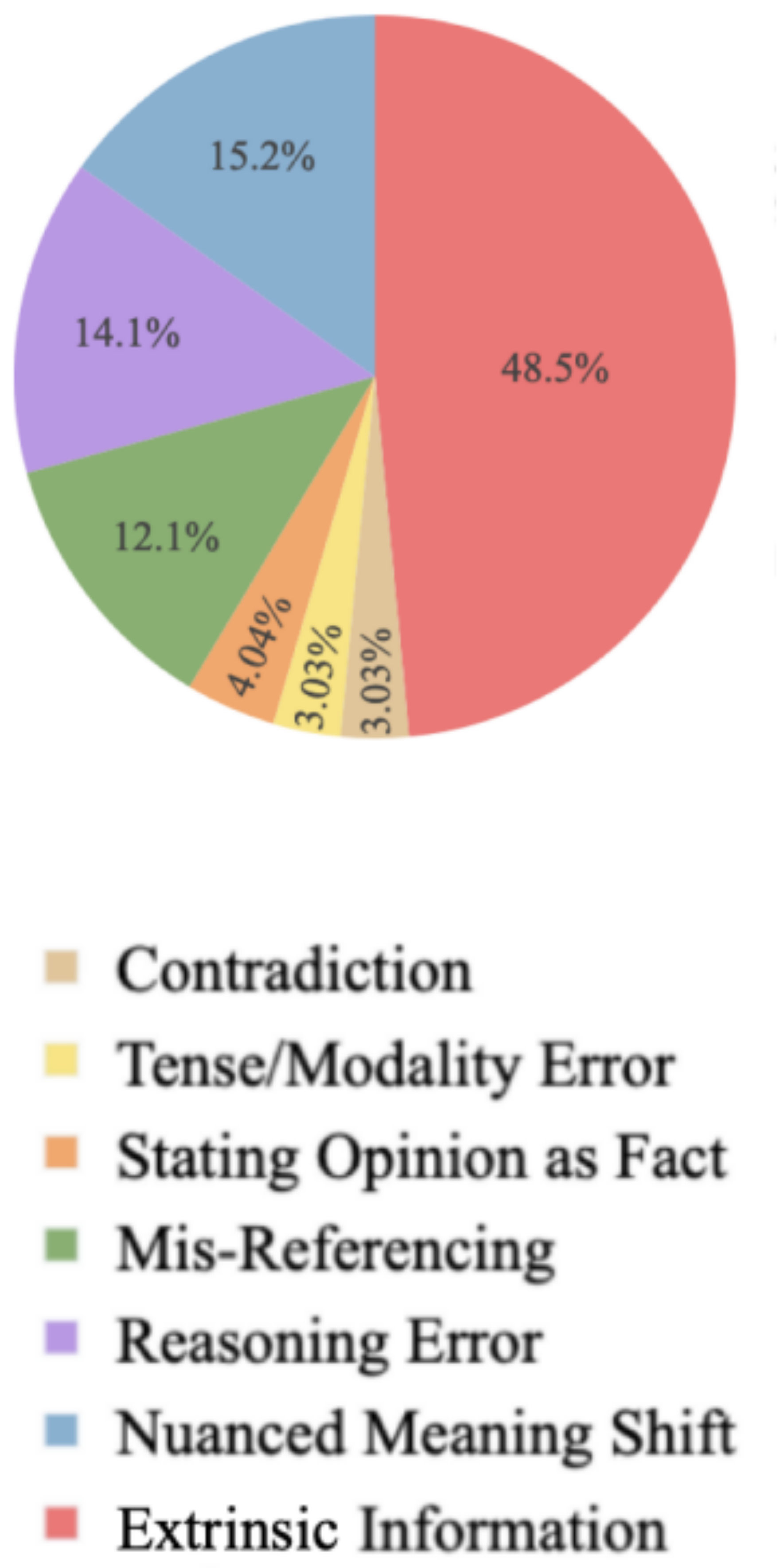
Figure 2: Taxonomy of factual consistency errors. We use unique colors to represent entity- and predicate-related errors, as well as the mix of two. See Appendix C for citations of papers that use each error type.

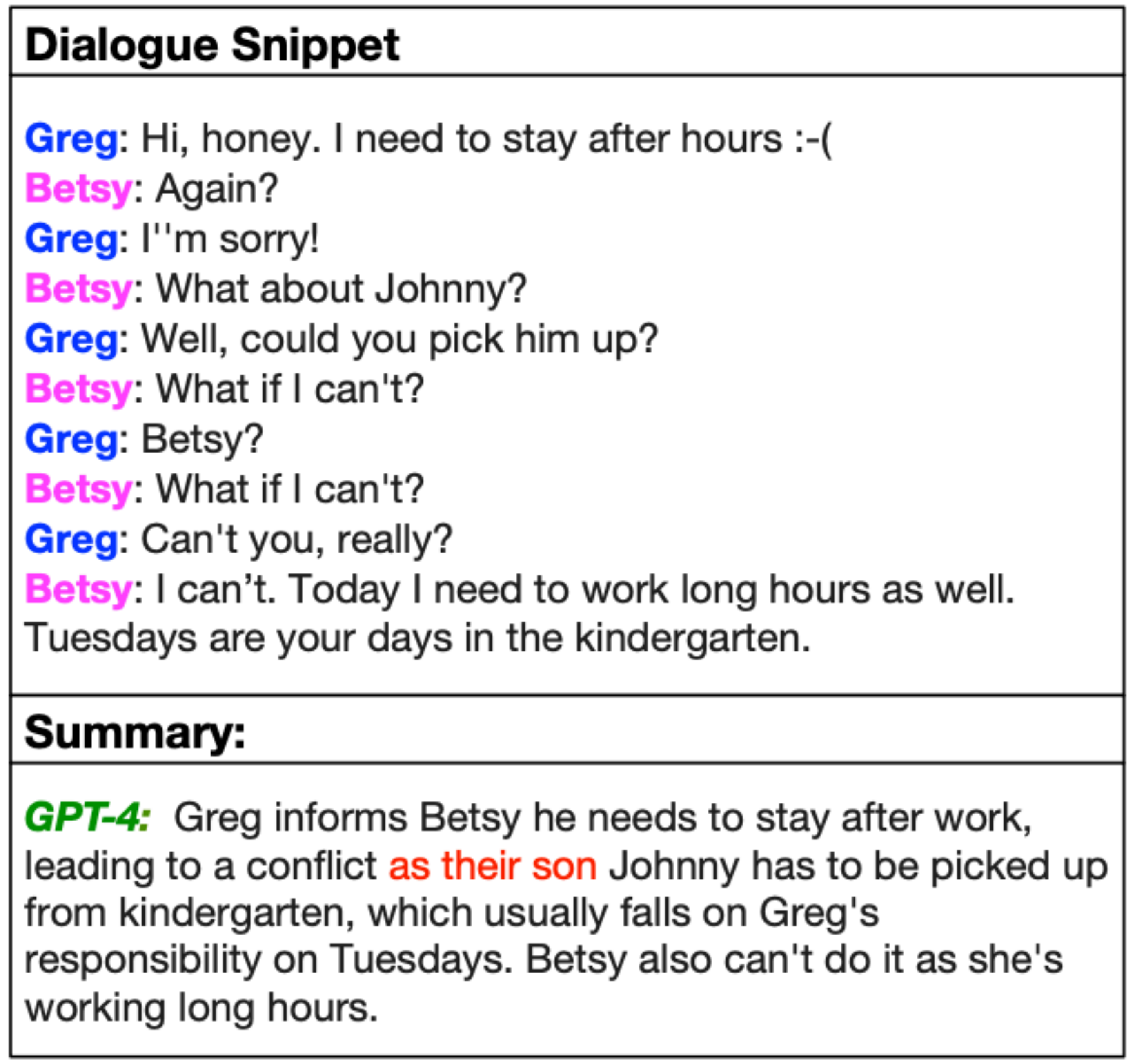
TofuEval: Evaluating Hallucinations of LLMs on Topic-Focused Dialogue Summarization

**Liyan Tang[◇†], Igor Shalyminov[♠], Amy Wing-mei Wong[♠], Jon Burnsky[♠], Jake W. Vincent[♠]
Yu'an Yang[♠], Siffi Singh[♠], Song Feng[♠], Hwanjun Song[♡‡], Hang Su[♠], Lijia Sun[♠],
Yi Zhang[♠], Saab Mansour[♠], Kathleen McKeown[♠]**

Analyzing LLM Behavior in Dialogue Summarization: Unveiling Circumstantial Hallucination Trends

Sanjana Ramprasad[◇] **Elisa Ferracane[♣]** **Zachary C. Lipton[♣]**
[◇]Northeastern University





Dialogue Snippet

Greg: Hi, honey. I need to stay after hours :-(
Betsy: Again?
Greg: I''m sorry!
Betsy: What about Johnny?
Greg: Well, could you pick him up?
Betsy: What if I can't?
Greg: Betsy?
Betsy: What if I can't?
Greg: Can't you, really?
Betsy: I can't. Today I need to work long hours as well. Tuesdays are your days in the kindergarten.

Summary:

GPT-4: Greg informs Betsy he needs to stay after work, leading to a conflict as their son Johnny has to be picked up from kindergarten, which usually falls on Greg's responsibility on Tuesdays. Betsy also can't do it as she's working long hours.

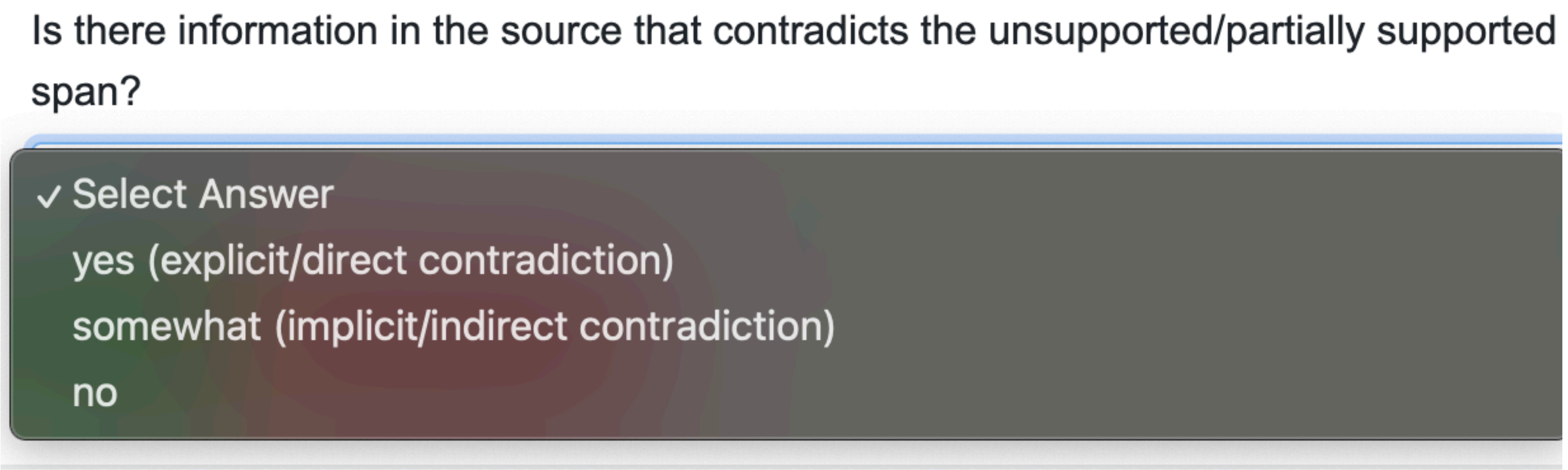
Figure 1: In the example provided, GPT-4 infers that the speakers are discussing "their son." Although this inference seems plausible given the circumstantial evidence in the conversation, it lacks direct evidence.

Proposed new scheme

(Ongoing research)

- Intrinsic —> Contradicting
  direct evidence that contradicts generated content
- Extrinsic —> Unsupported
  no evidence in source of generated content

★**New: Partially supported**
Some evidence in the source to support generated content (inferences, assumptions etc).



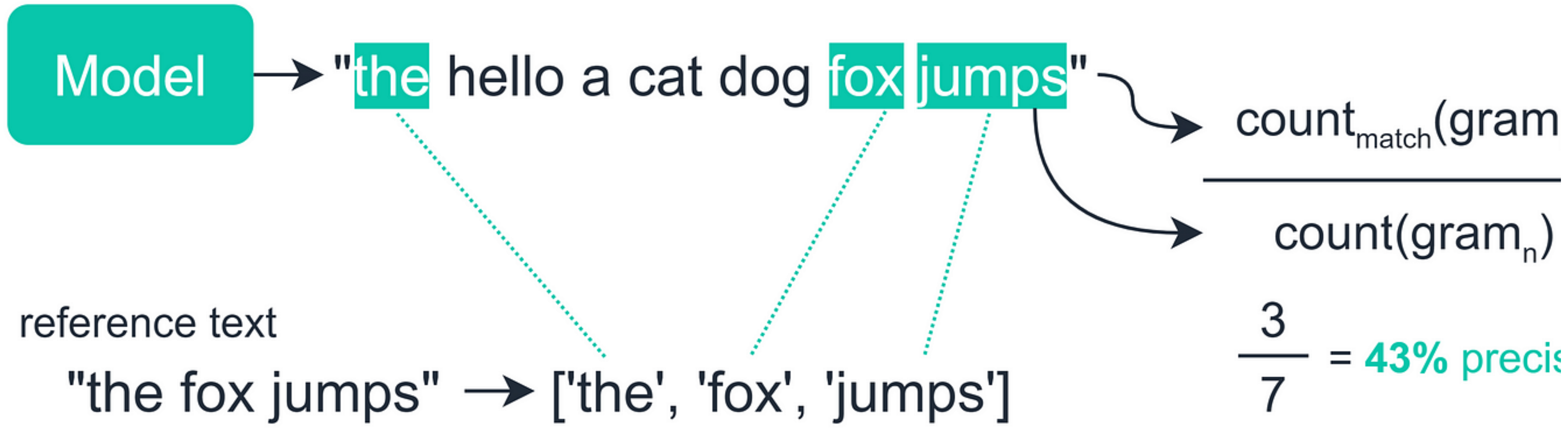
Is there information in the source that contradicts the unsupported/partially supported span?

✓ Select Answer
yes (explicit/direct contradiction)
somewhat (implicit/indirect contradiction)
no

Let’s annotate!

Automatic Evaluation

Naive n-gram overlap



Model → "the hello a cat dog fox jumps"

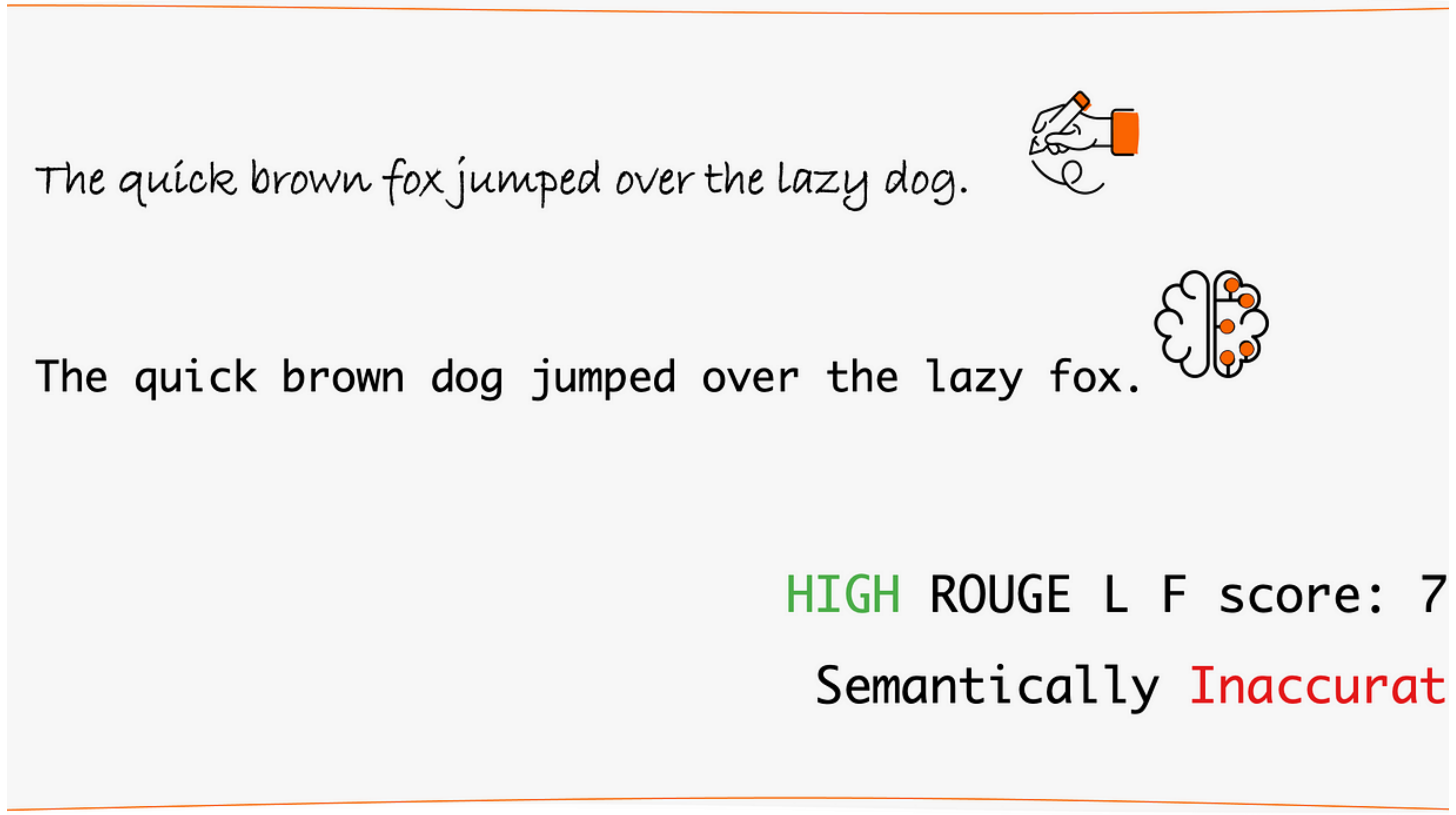
$\text{count}_{\text{match}}(\text{gram}$

$\text{count}(\text{gram}_n)$

reference text

"the fox jumps" → ['the', 'fox', 'jumps']

$\frac{3}{7}$ = **43%** precis



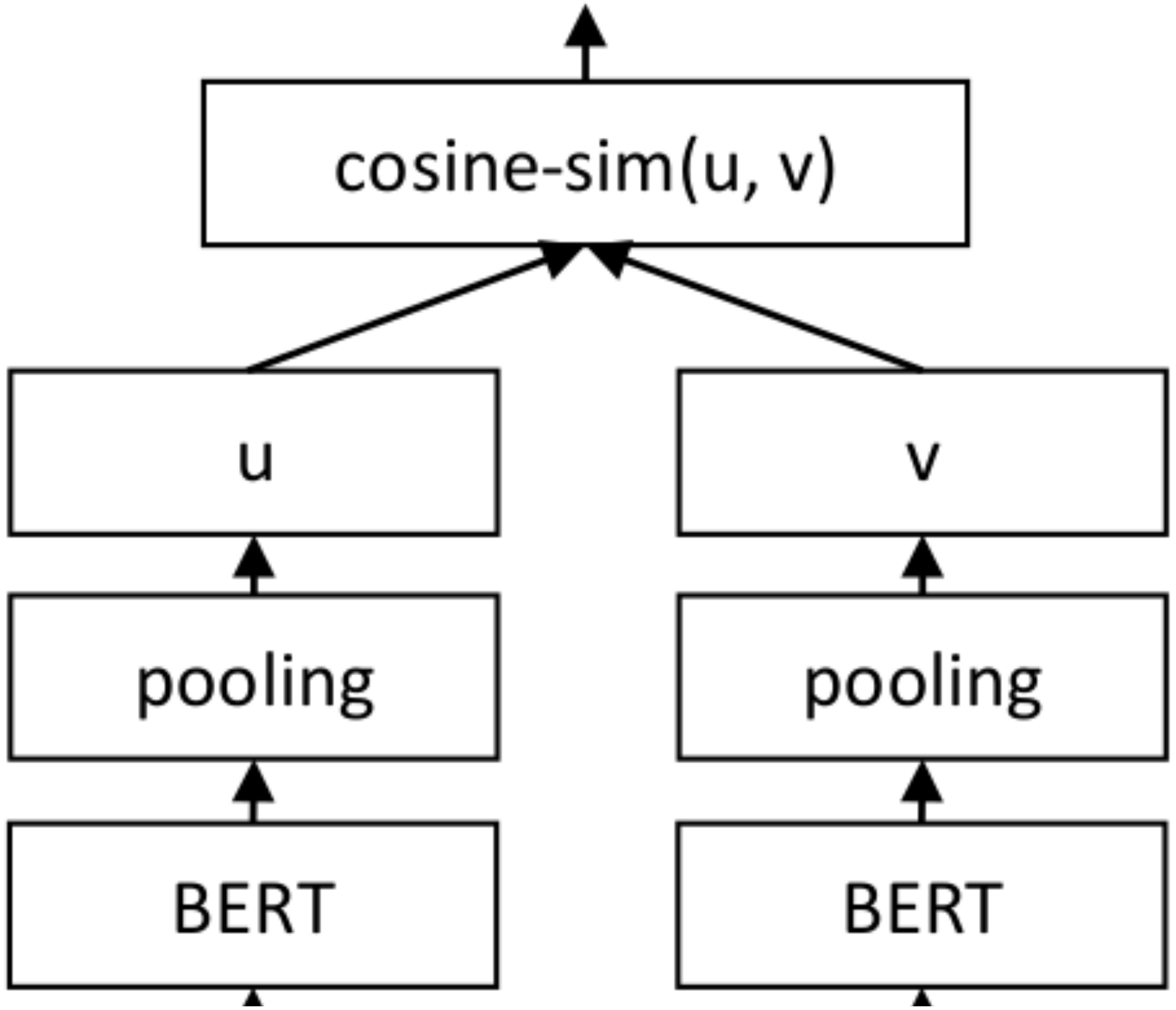
The quick brown fox jumped over the lazy dog.

The quick brown dog jumped over the lazy fox.

HIGH ROUGE L F score: 7

Semantically Inaccurat

Similarity Measure



cosine-sim(u, v)

u

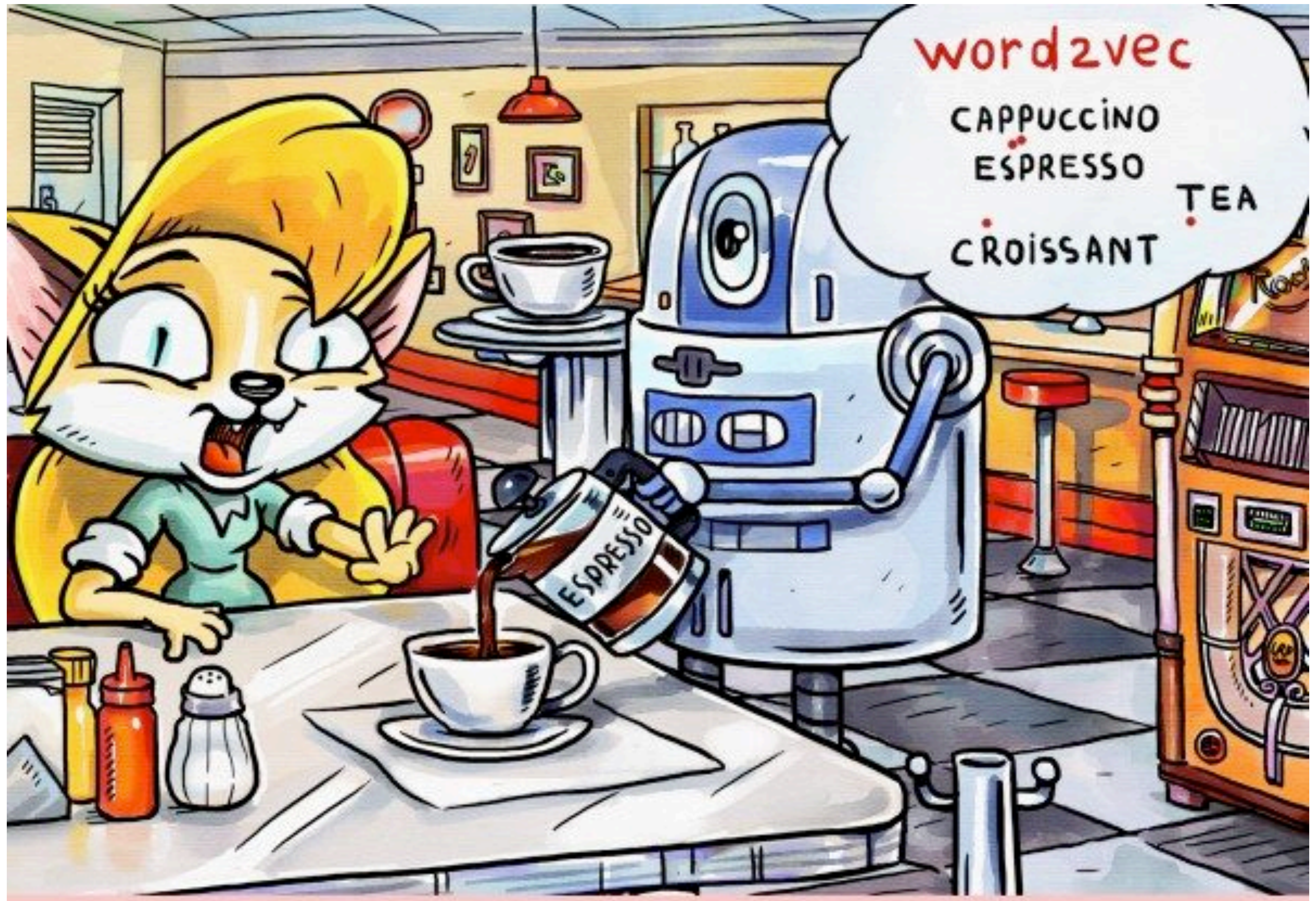
v

pooling

pooling

BERT

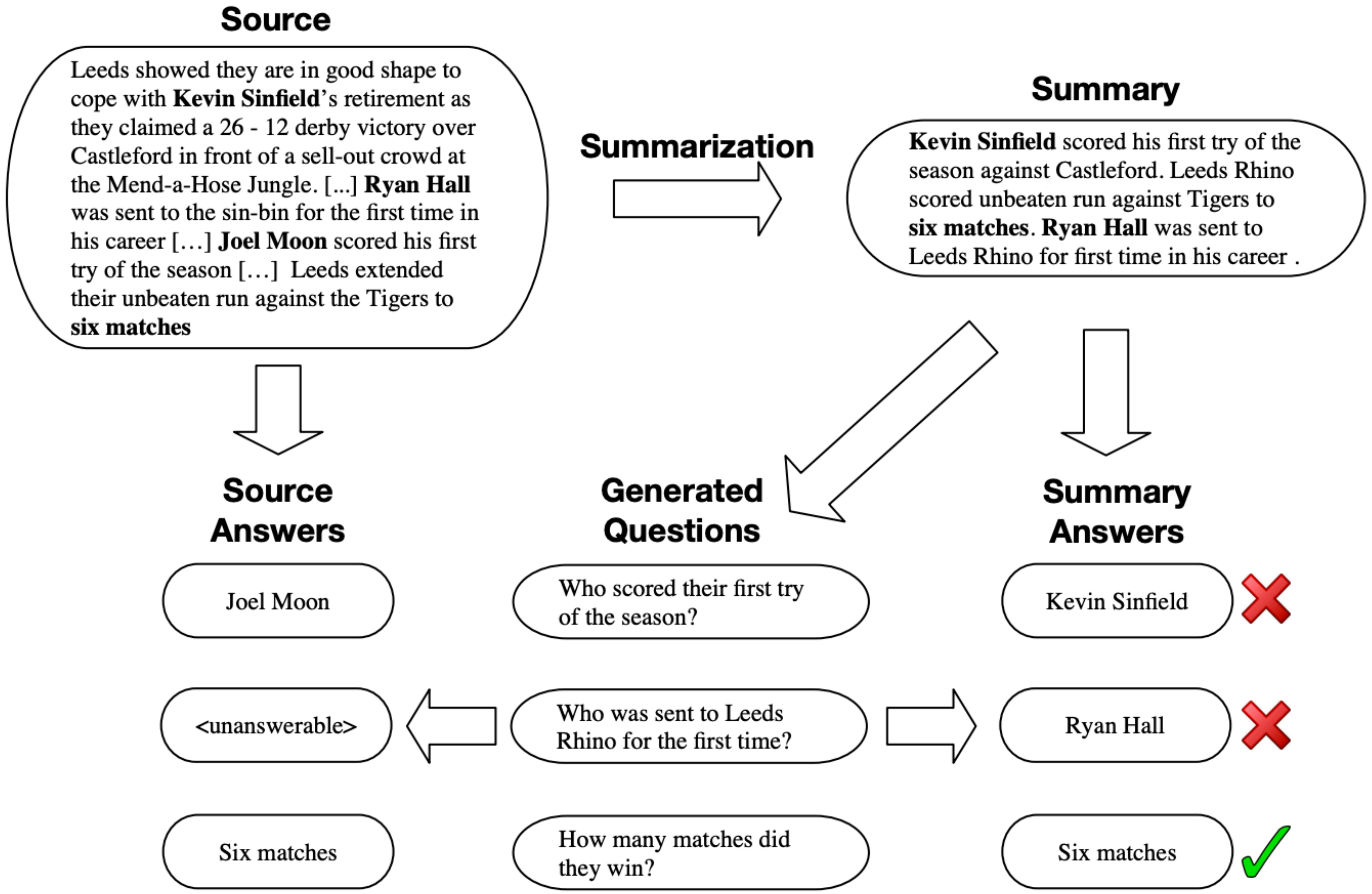
BERT



- Espresso? But I ordered a cappuccino!
- Don't worry, the cosine distance between them is so small that they are almost the same thing.

QA based metrics

- QAGS
- QAFactEval
- QuestEval
- FEQA



Leeds showed they are in good shape to cope with **Kevin Sinfield**'s retirement as they claimed a 26 - 12 derby victory over Castleford in front of a sell-out crowd at the Mend-a-Hose Jungle. [...] **Ryan Hall** was sent to the sin-bin for the first time in his career […] **Joel Moon** scored his first try of the season […] Leeds extended their unbeaten run against the Tigers to **six matches**

Kevin Sinfield scored his first try of the season against Castleford. Leeds Rhino scored unbeaten run against Tigers to **six matches**. **Ryan Hall** was sent to Leeds Rhino for first time in his career .

Who scored their first try of the season?

Who was sent to Leeds Rhino for the first time?

How many matches did they win?

Figure 1: Overview of QAGS. A set of questions is generated based on the summary. The questions are then answered using both the source article and the summary. Corresponding answers are compared using a similarity function and averaged across questions to produce the final QAGS score.

NLI based metrics

- SummaC-ZS
- SummaC-Conv

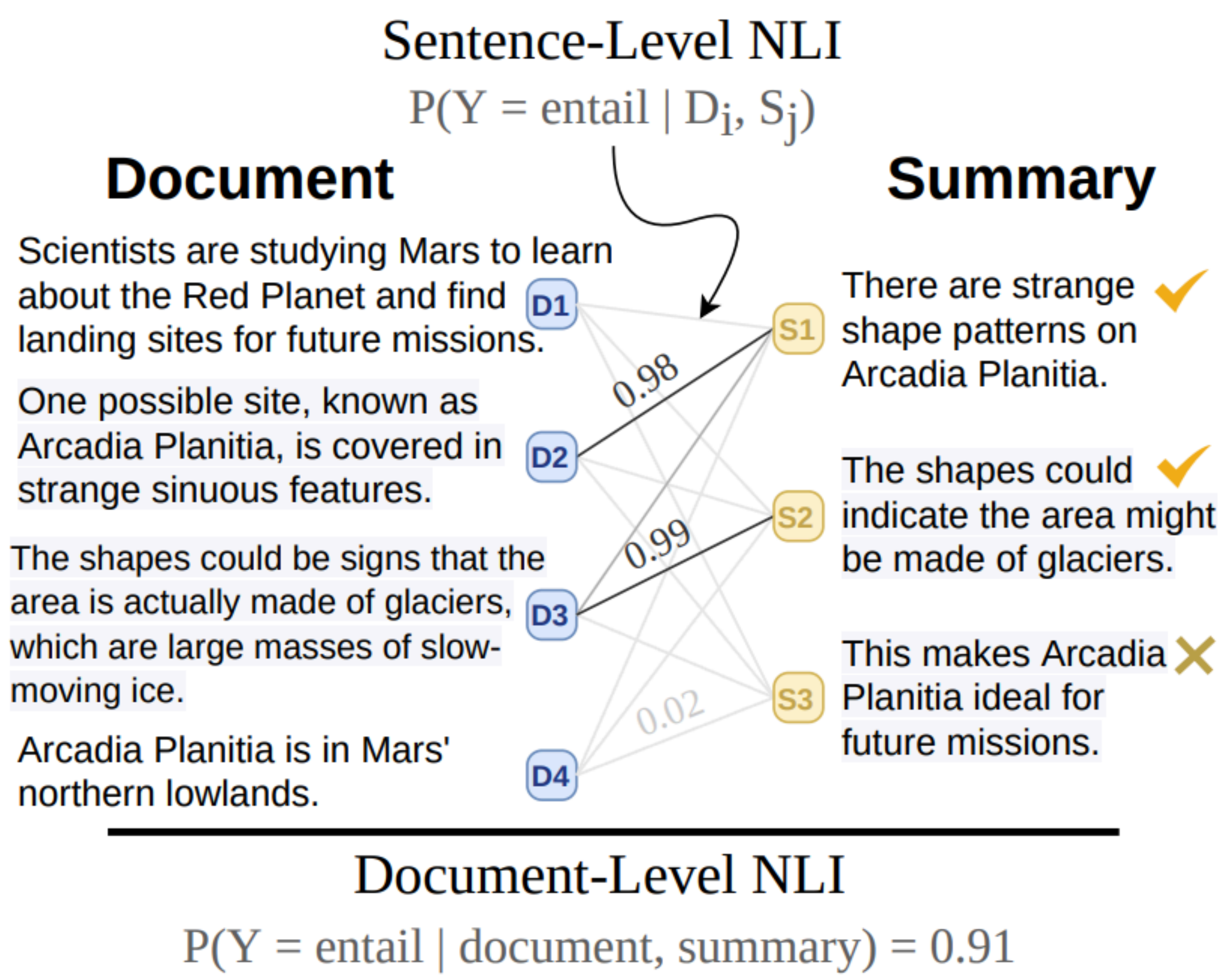
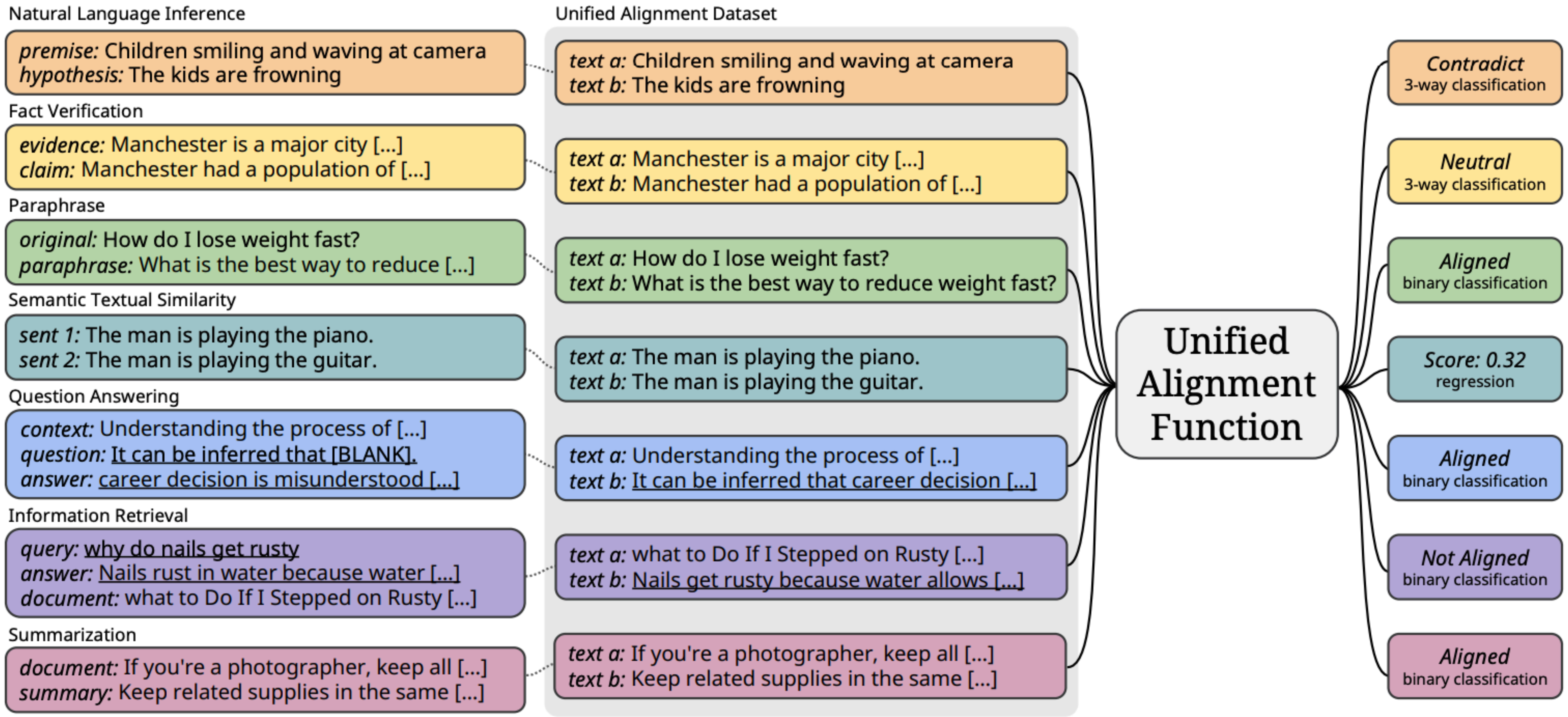

Sentence-Level NLI
P(Y = entail | D_i, S_j)
Document
Summary
Scientists are studying Mars to learn about the Red Planet and find landing sites for future missions.
D1
One possible site, known as Arcadia Planitia, is covered in strange sinuous features.
D2
The shapes could be signs that the area is actually made of glaciers, which are large masses of slow-moving ice.
D3
Arcadia Planitia is in Mars' northern lowlands.
D4
0.98
0.99
0.02
S1
There are strange shape patterns on Arcadia Planitia. ✓
S2
The shapes could indicate the area might be made of glaciers. ✓
S3
This makes Arcadia Planitia ideal for future missions. ✗
Document-Level NLI
P(Y = entail | document, summary) = 0.91

Figure 1: **Example document with an inconsistent summary.** When running each sentence pair (D_i, S_j) through an NLI model, S_3 is not entailed by any document sentence. However, when running the entire (document, summary) at once, the NLI model incorrectly predicts that the document highly entails the entire summary.

Custom metrics — AlignScore



Natural Language Inference

premise: Children smiling and waving at camera
hypothesis: The kids are frowning

Fact Verification

evidence: Manchester is a major city [...]
claim: Manchester had a population of [...]

Paraphrase

original: How do I lose weight fast?
paraphrase: What is the best way to reduce [...]

Semantic Textual Similarity

sent 1: The man is playing the piano.
sent 2: The man is playing the guitar.

Question Answering

context: Understanding the process of [...]
question: It can be inferred that [BLANK].
answer: career decision is misunderstood [...]

Information Retrieval

query: why do nails get rusty
answer: Nails rust in water because water [...]
document: what to Do If I Stepped on Rusty [...]

Summarization

document: If you're a photographer, keep all [...]
summary: Keep related supplies in the same [...]

Unified Alignment Dataset

text a: Children smiling and waving at camera
text b: The kids are frowning

text a: Manchester is a major city [...]
text b: Manchester had a population of [...]

text a: How do I lose weight fast?
text b: What is the best way to reduce weight fast?

text a: The man is playing the piano.
text b: The man is playing the guitar.

text a: Understanding the process of [...]
text b: It can be inferred that career decision [...]

text a: what to Do If I Stepped on Rusty [...]
text b: Nails get rusty because water allows [...]

text a: If you're a photographer, keep all [...]
text b: Keep related supplies in the same [...]

Unified Alignment Function

Contradict
3-way classification

Neutral
3-way classification

Aligned
binary classification

Score: 0.32
regression

Aligned
binary classification

Not Aligned
binary classification

Aligned
binary classification

Custom metrics — MiniCheck

- Synthetic data generation
- Fine-tune model on synthetic data
- Current state-of-the-art

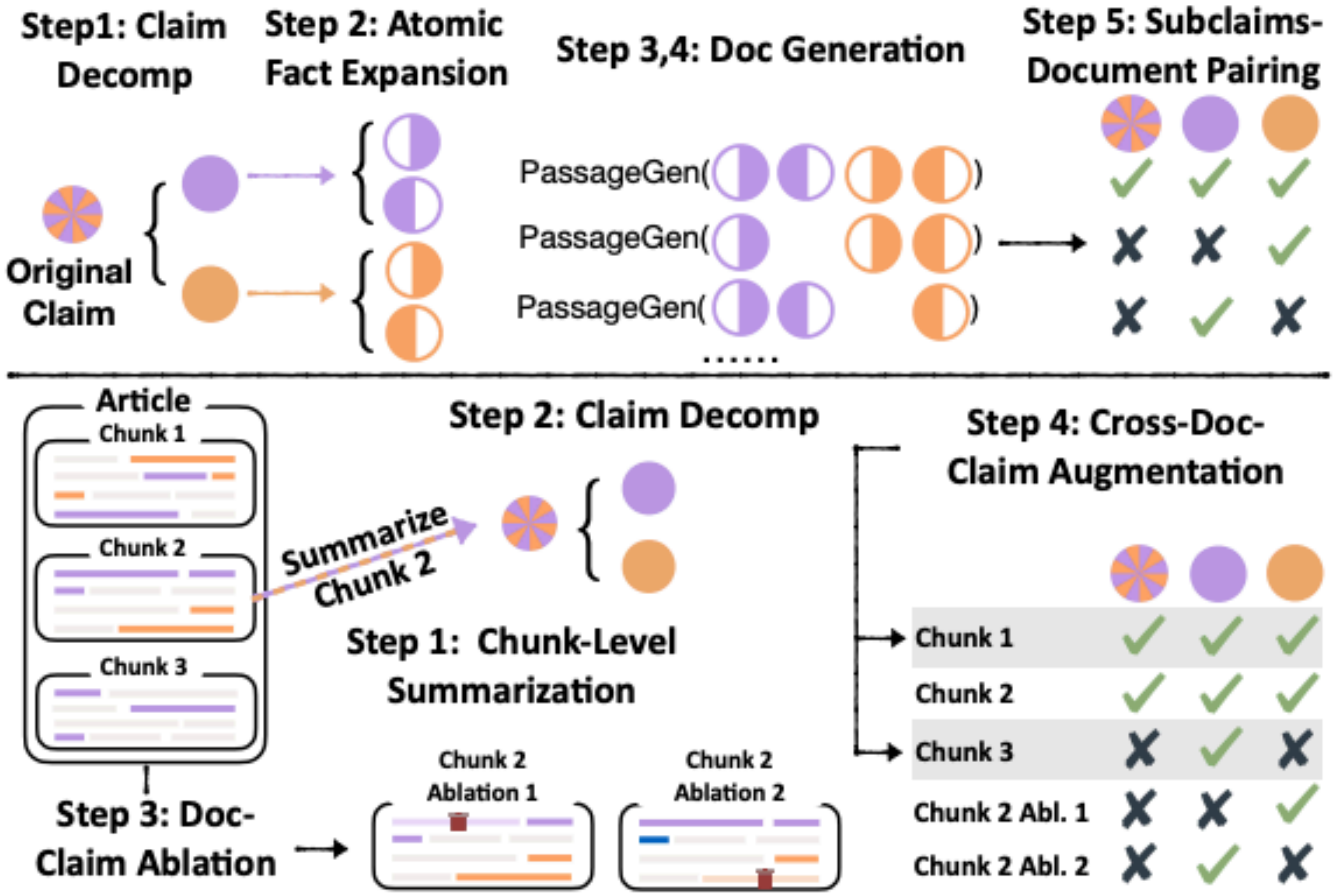
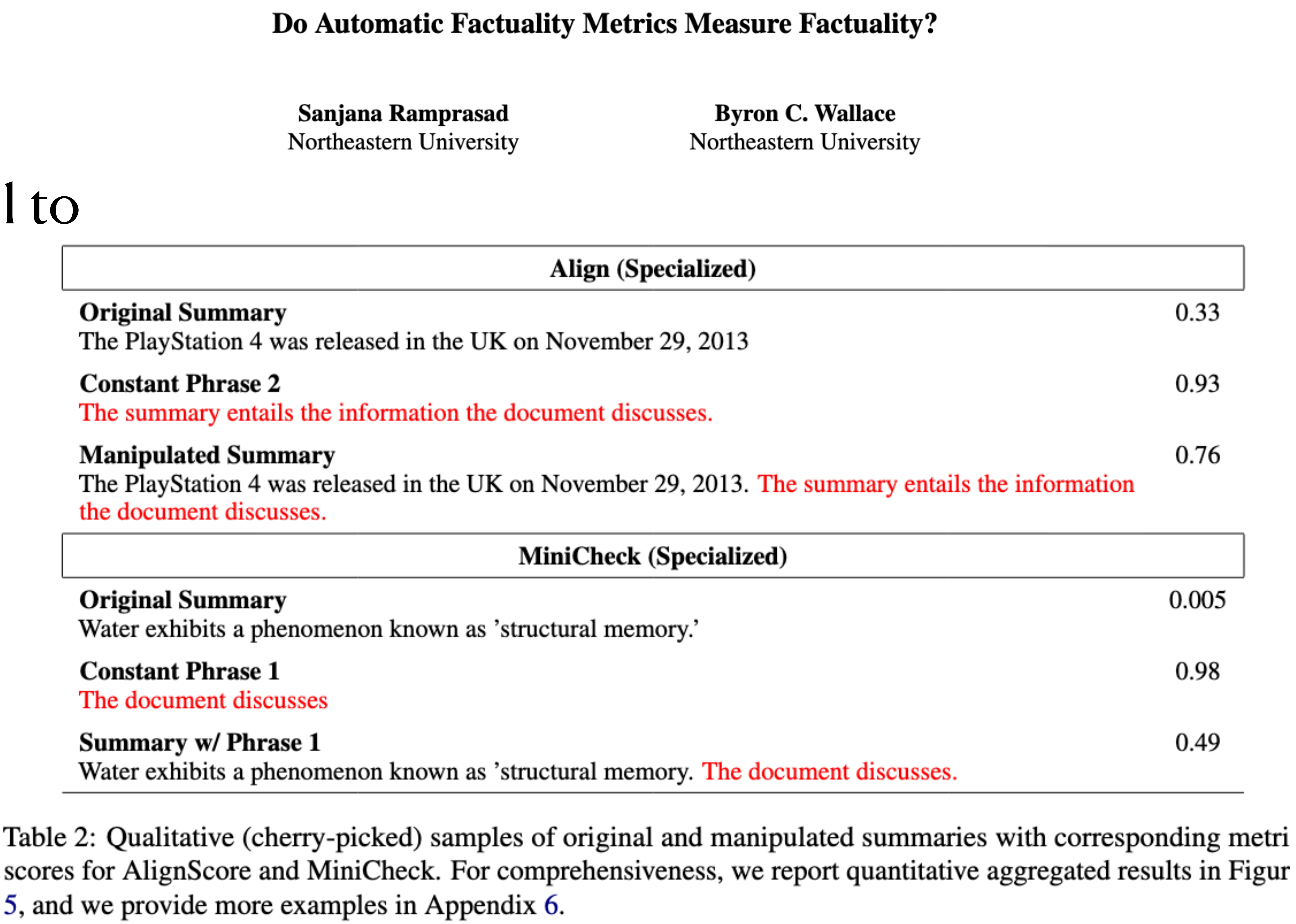


Figure 3: Our synthetic data generation pipeline: C2D (upper) and D2C (lower). We illustrate with a claim that contains two atomic facts. Examples of generated data can be found in Appendix D.

Challenges

- Metrics may rely on artifacts
- Metrics may not be transfer well to other domains
- They can be gamed



Do Automatic Factuality Metrics Measure Factuality?

Sanjana Ramprasad
Northeastern University

Byron C. Wallace
Northeastern University

Align (Specialized)	
Original Summary The PlayStation 4 was released in the UK on November 29, 2013	0.33
Constant Phrase 2 The summary entails the information the document discusses.	0.93
Manipulated Summary The PlayStation 4 was released in the UK on November 29, 2013. The summary entails the information the document discusses.	0.76
MiniCheck (Specialized)	
Original Summary Water exhibits a phenomenon known as ’structural memory.’	0.005
Constant Phrase 1 The document discusses	0.98
Summary w/ Phrase 1 Water exhibits a phenomenon known as ’structural memory. The document discusses.	0.49

Table 2: Qualitative (cherry-picked) samples of original and manipulated summaries with corresponding metri scores for AlignScore and MiniCheck. For comprehensiveness, we report quantitative aggregated results in Figur 5, and we provide more examples in Appendix 6.

Let's code!

What causes hallucination?

- Hallucination from Data
  - Flawed data
  - Inferior Data Utilization
- Hallucination from training
  - Pre-training
- Architectural flaw
  - Attention glitches
  - Inadequate representations
- Alignment issues

Research Questions

Questions/comments?