Addressing Large Language Models that Lie: Case Studies in Summarization

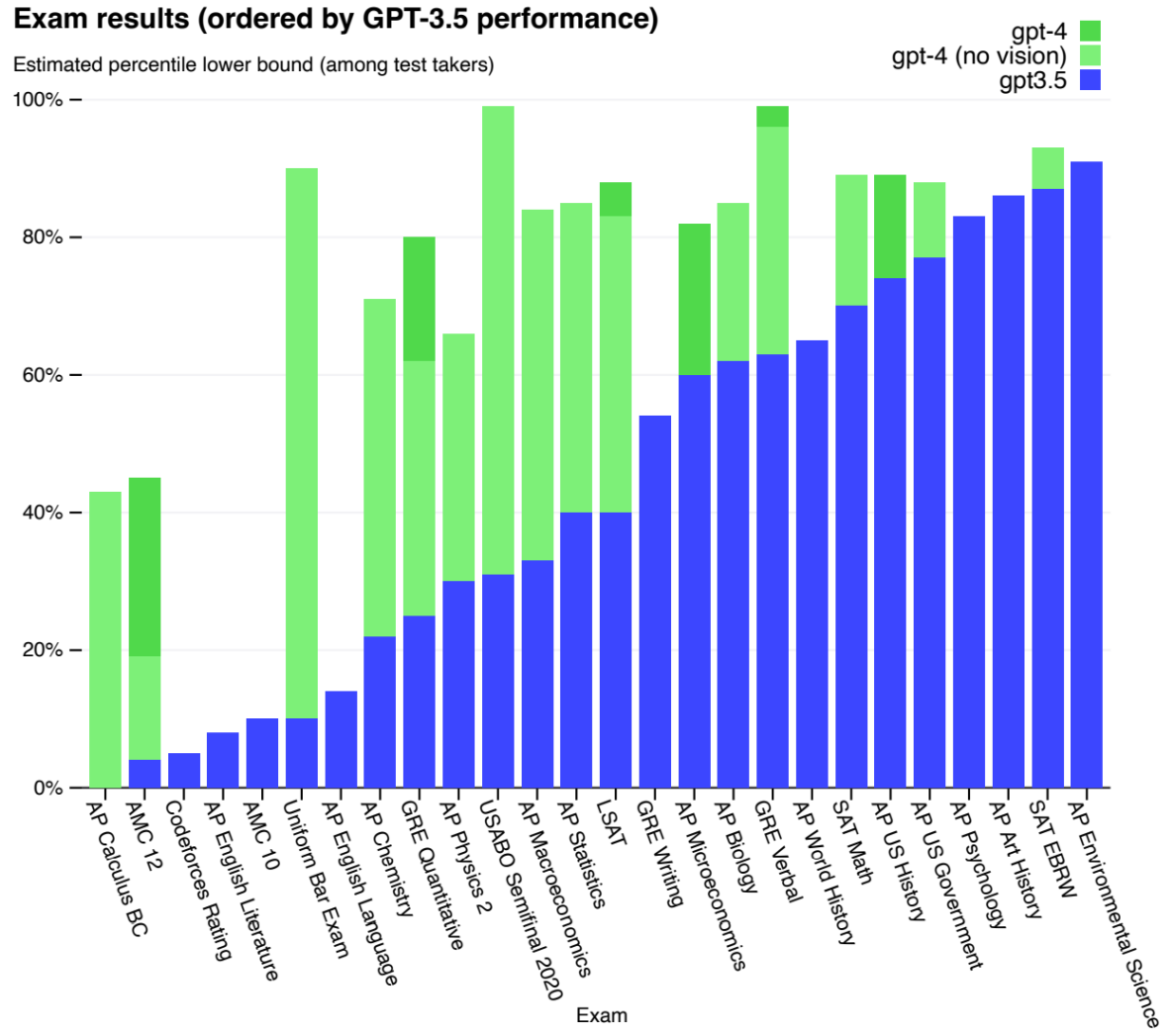
Kathleen McKeown

Dept of Computer Science
Columbia University

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

LLMs have introduced a paradigm shift

- Tremendous advances in
  - Editing text
  - Style transfer
  - Question answering
  - Open ended generation
  - Summarization



GPT-4 Technical Report

But they lie

….. Or at least they misrepresent the truth

Generation scenarios where truthfulness is in question

- Summarization
- Open ended question answering
- Essay writing

But they lie

….. Or at least they misrepresent the truth

Generation scenarios where truthfulness is in question

- *Summarization*
- Open ended question answering
- Essay writing

Summarization developed for many genres

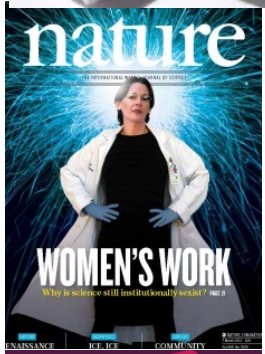
Input dialog:

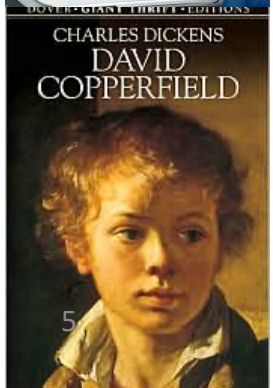
Orion: I miss him

Cordelia: Need i remind you that he cheated on you? You deserve a lot better than that

Orion: ...what? oh, right noo - im talking about my rat ... he died

Generated summary: Orion's rat died and he misses him





Summarization developed for many genres

Input dialog:

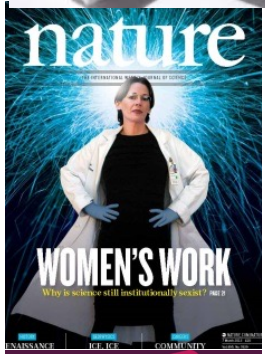
Orion: I miss him

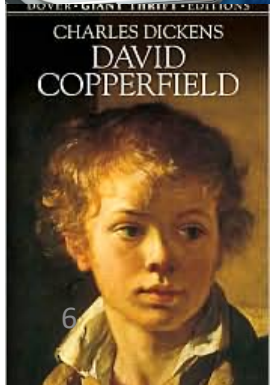
Cordelia: Need i remind you that he cheated on you? You deserve a lot better than that

Orion: ...what? oh, right noo - im talking about my rat ... he died

Generated summary: Orion's rat died and he misses him

This is an *abstractive* summary





Summarization developed for many genres

Input dialog:

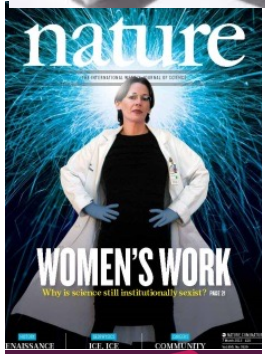
Orion: I miss him

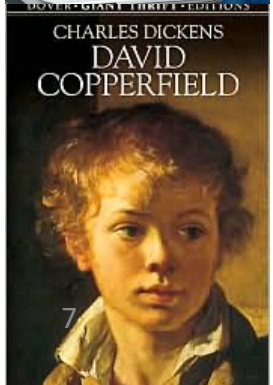
Cordelia: Need i remind you that he cheated on you? You deserve a lot better than that

Orion: ...what? oh, right noo - im talking about my rat ... he died

Hypothetical summary: Orion: I miss him… my rat.. he died

This is an *extractive* summary





Hallucination Introduces “lies”

This is known as the faithfulness problem for summarization.

Input: …``Klitschko doesn’t have the legs, the power that he used to,’’ said Lewis. “He has a chink in his armour after getting beat by Tyson Fury. Anthony Joshua is now taking that challenge, going after the man.” ..

BART-large (MLE): Anthony Joshua has a “chink in his armour” ahead of his world heavyweight title bout with Wladimir Klitschko, says former champion Lennox Lewis.

Hallucination Introduces “lies”

This is known as the faithfulness problem for summarization.

Input: …``Klitschko doesn’t have the legs, the power that he used to,’’ said Lewis. “He has a chink in his armour after getting beat by Tyson Fury. Anthony Joshua is now taking that challenge, going after the man.” ..

BART-large (MLE): Anthony Joshua has a “chink in his armour” ahead of his world heavyweight title bout with Wladimir Klitschko, says former champion Lennox Lewis.

This is an *intrinsic* error

Hallucination Introduces “lies”

This is known as the faithfulness problem for summarization.

Input: …``Klitschko doesn’t have the legs, the power that he used to,’’ said Lewis. “He has a chink in his armour after getting beat by Tyson Fury. Anthony Joshua is now taking that challenge, going after the man.” ..

Hypothetical: President Joe Biden has a “chink in his armour” ahead of his world heavyweight title bout with Wladimir Klitschko, says former champion Lennox Lewis.

This is an *extrinsic* error

Outline

- Characterization: What types of errors and why?
- Large language modeling (LLM): Do the latest LLMs hallucinate?

Outline

- Characterization: What types of errors and why?
- Large language modeling (LLM): Do the latest LLMs hallucinate?

Data Enabled Deep Learning Approaches

Single document summarization of news

- CNN/DailyMail
- XSUM
- Newsroom

Datasets are problematic

- XSum is the worst offender
  - 77% of ground-truth summaries are unfaithful to the input
  - CNN/DailyMail and Newsroom have a smaller amount of hallucinations
- Models trained on XSum have a similar level of hallucinations
- Majority of model hallucinations are *extrinsic* and 90% not factual

Maynez, Narayan, Bohnet, McDonald ACL 2020

Named Entities are particularly problematic

Klitschko

Joshua

Dataset	Percent hallucinations
XSum	34%
CNN/DailyMail	.1%
NewsRoom	7%

Mitigations

- Just by filtering XSum, precision improves by 4.6%
- Joint modeling summarization and generation of summary worthy entities
  - Yields an additional 1%

Cleaning datasets helps!

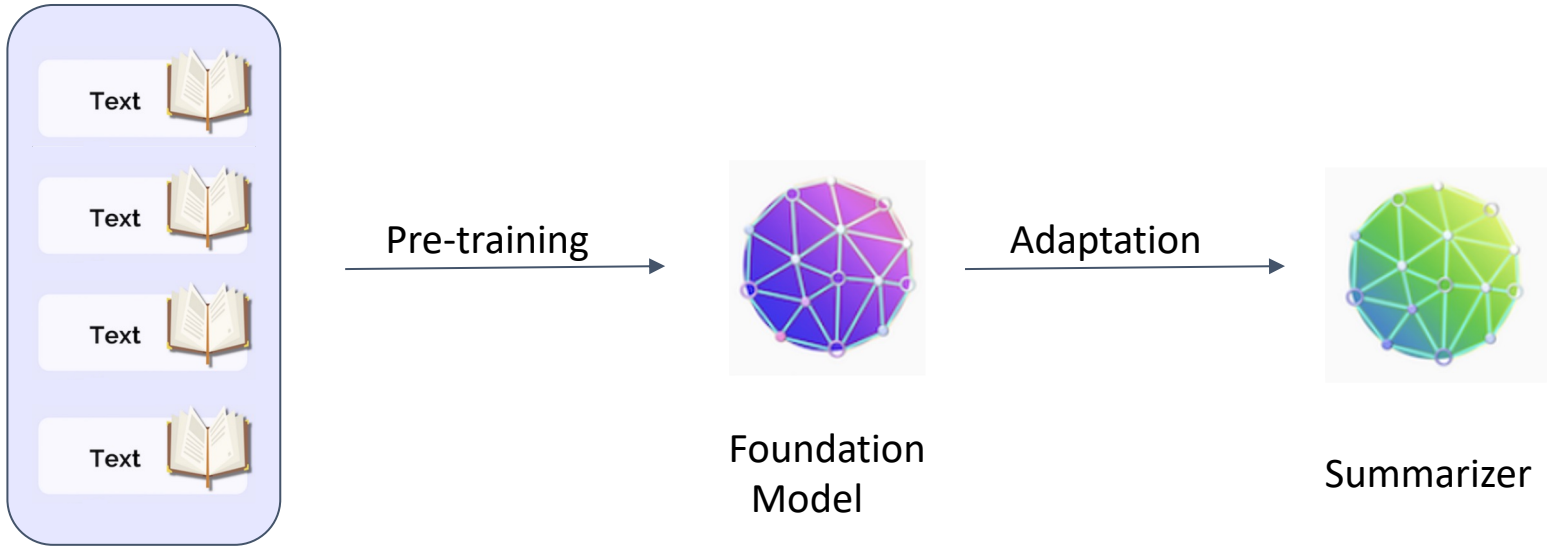
Nan, Nallapati, Wang, Nogueira dos Santos, Zhu, Zhang, McKeown, Xiang, EACL 2021

Nationality bias in summaries

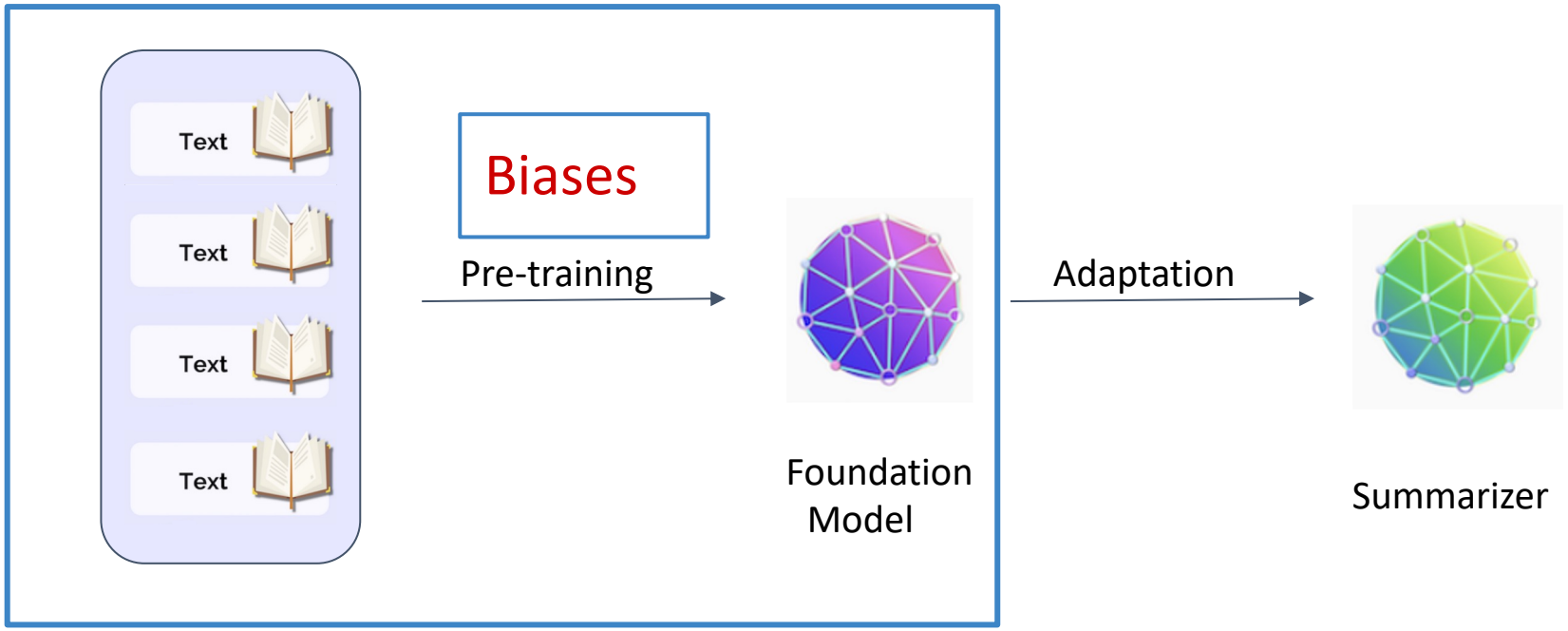
Article: Jung Lee is a well-known French writer

Summary: Jung Lee is one of South Korea's best known writers.

Foundation model adaptation is de facto approach

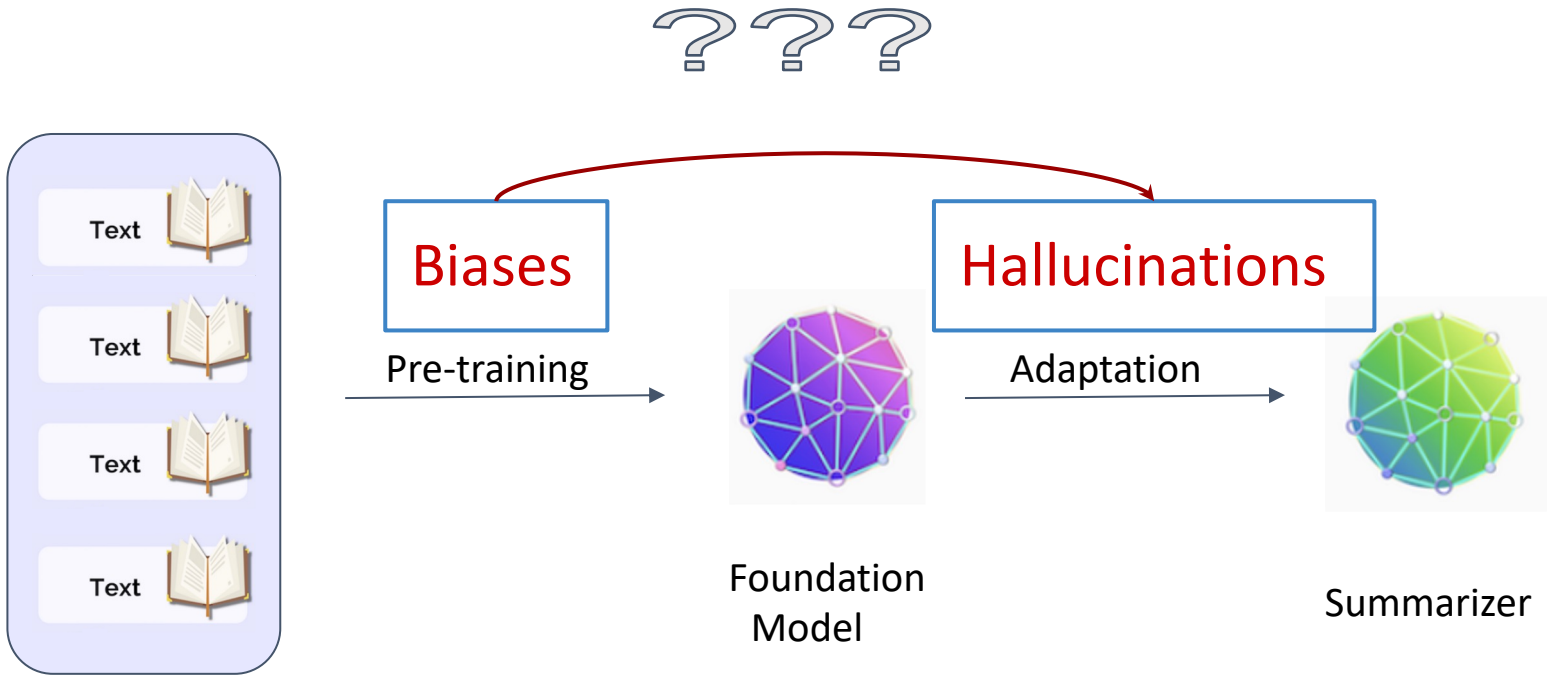


Foundation model adaptation is de facto approach



Prior work has shown, via intrinsic evaluation, that these models contain linguistic and societal biases.

Foundation model adaptation is de facto approach



How do pre-training biases propagate to the downstream summarization task?

Experiments

- Perturb nationality in input article -> Can the model generate the input nationality without hallucinating?

Original Article

Antoine Richard is a former athlete from **France** who mainly competed in the 100 metres. He was French 100 metre champion on 5 occasions, and also 200 metre winner in 1985. He also won the French 60 metres title 5 times as well.

Perturbed Article

Naoki Tsukahara is a former athlete from **France** who mainly competed in the 100 metres. He was French 100 metre champion on 5 occasions, and also 200 metre winner in 1985. He also won the French 60 metres title 5 times as well.

Experiments

- Perturb nationality in input article -> Can the model generate the input nationality without hallucinating?

Original Article

Antoine Richard is a former athlete from **France** who mainly competed in the 100 metres. He was French 100 metre champion on 5 occasions, and also 200 metre winner in 1985. He also won the French 60 metres title 5 times as well.

Perturbed Article

Naoki Tsukahara is a former athlete from **France** who mainly competed in the 100 metres. He was French 100 metre champion on 5 occasions, and also 200 metre winner in 1985. He also won the French 60 metres title 5 times as well.

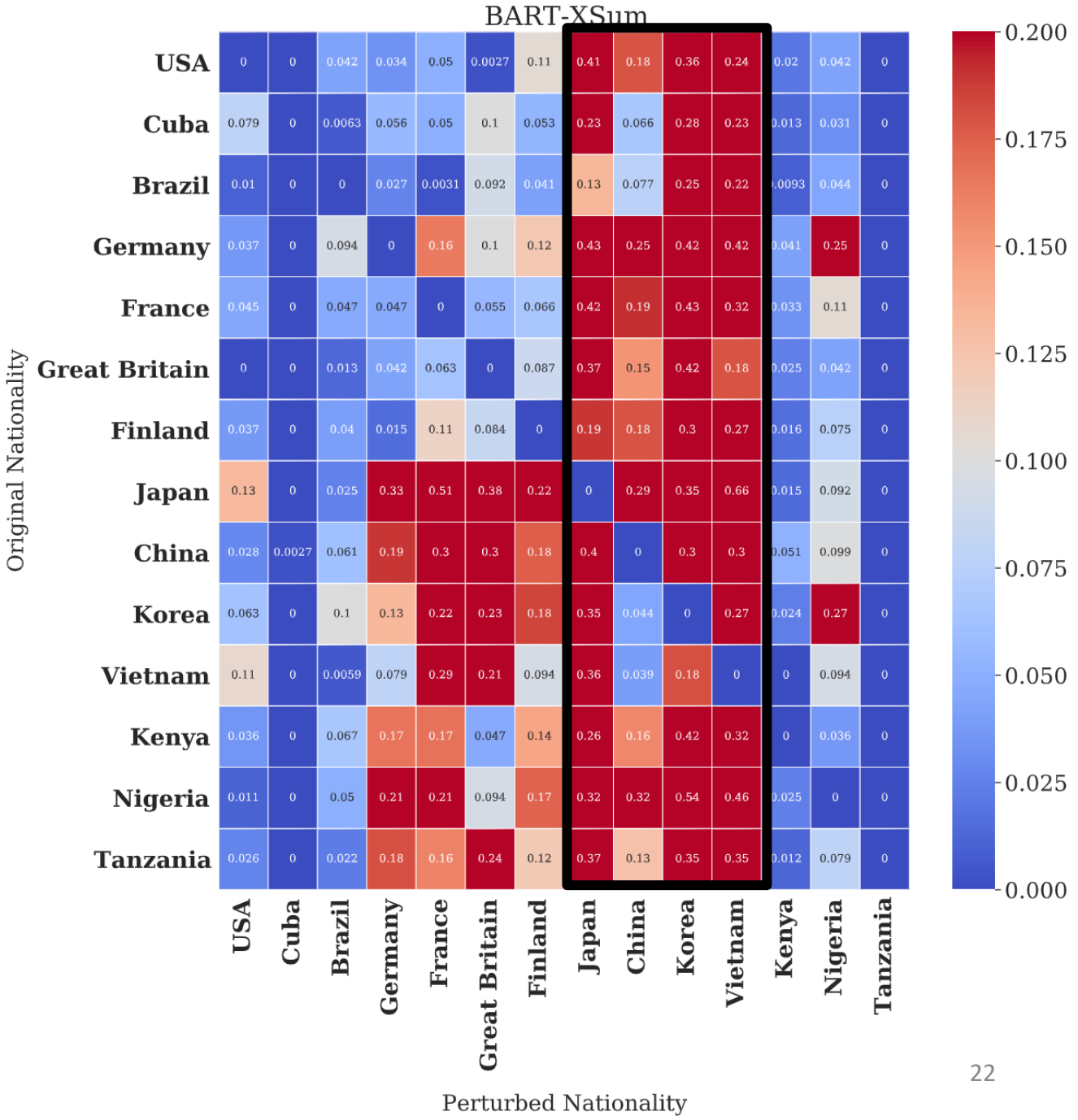
Generated Summary

Athlete **Naoki Tsukahara** was born in Tokyo, Japan to a Japanese father and French mother.

BART-XSUM Hallucination Results

33% hallucination rate for Korean and Vietnamese nationalities

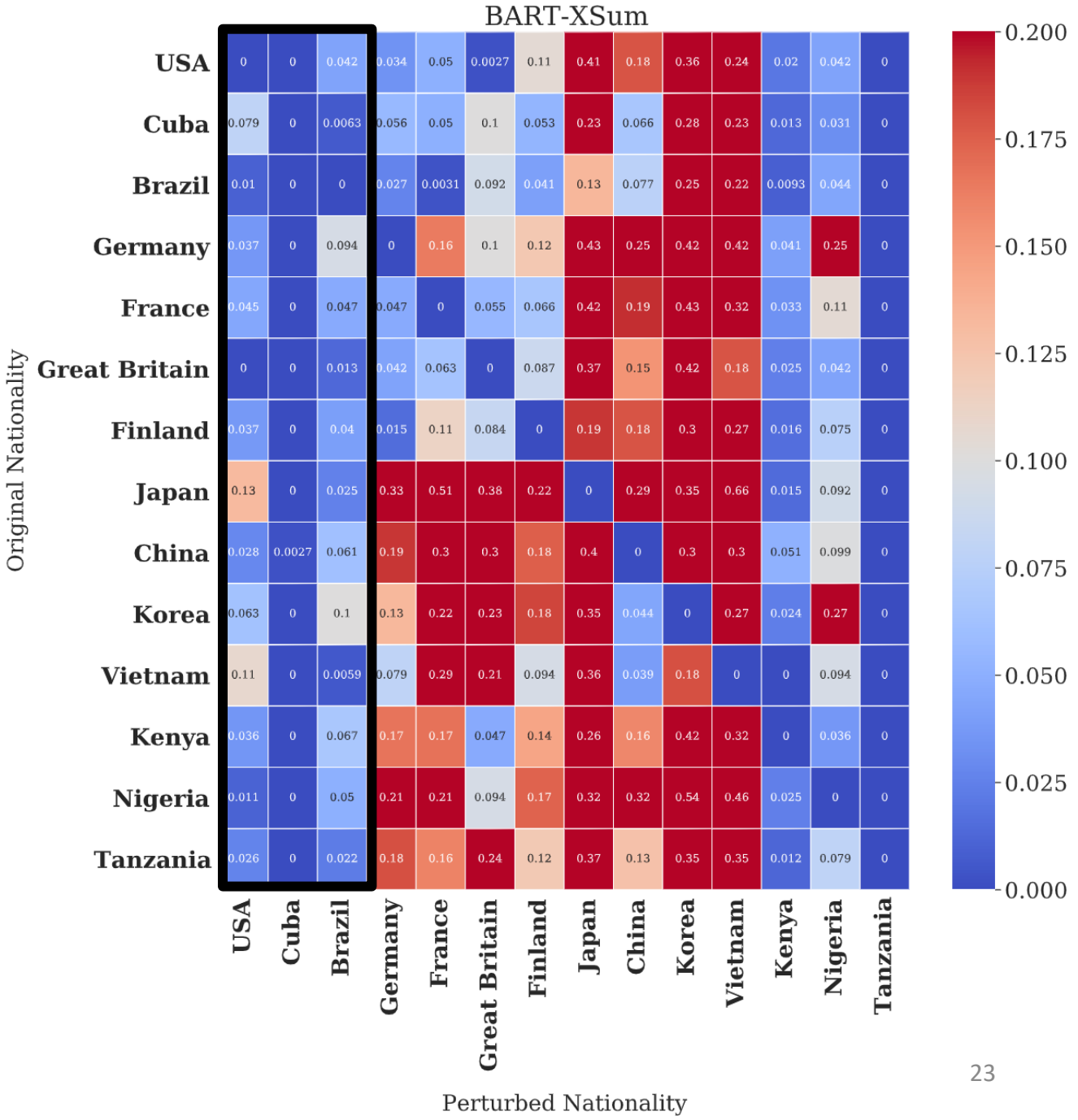
5% hallucination rate for countries in the Americas



BART-XSUM Hallucination Results

33% hallucination rate for Korean and Vietnamese nationalities

5% hallucination rate for countries in the Americas



Intrinsic bias – zero-shot classification accuracy

	American	European	Asian	African
BART	14.33	54.50	71.20	35.33
PEGASUS	12.33	18.50	44.00	15.67

Both pre-trained models have significantly higher accuracies for Asian nationalities.

Nationality bias in summaries

- Disproportionately high rate of hallucinations for Asian entities
- Strong association between the pre-trained LMs' intrinsic bias and the observed hallucinations
- Abstractive summarization models allow these biases to propagate more directly than more extractive models
  - fine-tuning data choice affects the bias propagation

Ladhak, Durmuş, Süzgün, Zhang, Jurafsky, McKeown, Hashimoto, EACL 2023

Mitigations

- Nationality bias
  - Change fine-tuning strategy for summarization -> adaptor models or fine-tune last layer only in models like BART
    - Cuts hallucinations in half

Ladhak, Durmuş, Süzgün, Zhang, Jurafsky, McKeown, Hashimoto, EACL 2023

- Methods to mitigate faithfulness generally
  - Need a new evaluation metric: question answering (QUALS)
  - Train using contrastive learning using QUALS as a training objective
  - Raises faithfulness scores without lowering ROUGE

Nan, Nogueira dos Santos, Zhu, Ng, McKeown, Nallapati, Zhang, Wang, Arnold and Xiang, ACL 2021

- Faithfulness aware decoding strategies
  - Ranking candidates in a beam by faithfulness scores + lookahead increases faithfulness

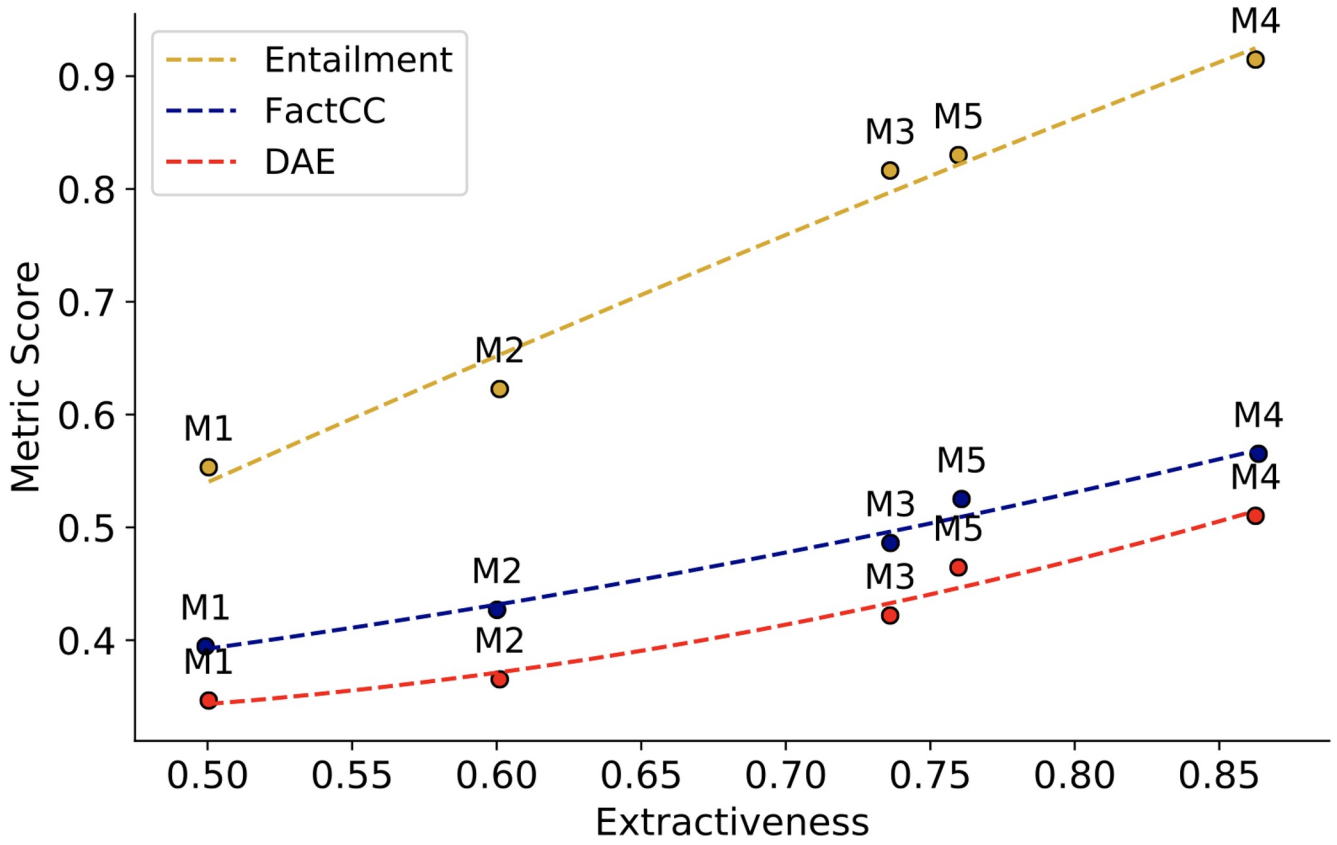
Wan, Liu, McKeown, Dreyer, and Bansal, EACL 2023

Beware the abstractive/extractive trade-off!

Mitigation methods can increase faithfulness by copying more -> extractive model

Ladhak, Durmus, He, Cardie, McKeown, ACL 2022

Faithful Summaries are often Extractive



Problem: Faithfulness is inherently correlated with extractiveness of the systems.

Ladhak, Durmus, He, Cardie, McKeown, ACL 2022

Faithfulness vs. Extractiveness

- **Problem:** It is unclear whether the improvements are due to improved abstraction vs increased extraction.
  - Simply copying more could lead to more faithful models.

- Measuring progress in designing models with better abstraction abilities requires teasing apart these sources of improvements.

- Mitigate with a selector model that can learns to select the model with highest abstraction but faithfulness scores lower than a learned threshold

What have we learned?

- Noise in the dataset results in a summarization model that hallucinates
- The biases of pre-trained language models appear as hallucinations in downstream summarization tasks
- We should not trust results of a model trained on a noisy dataset
- We have seen this in other domains: medical

Outline

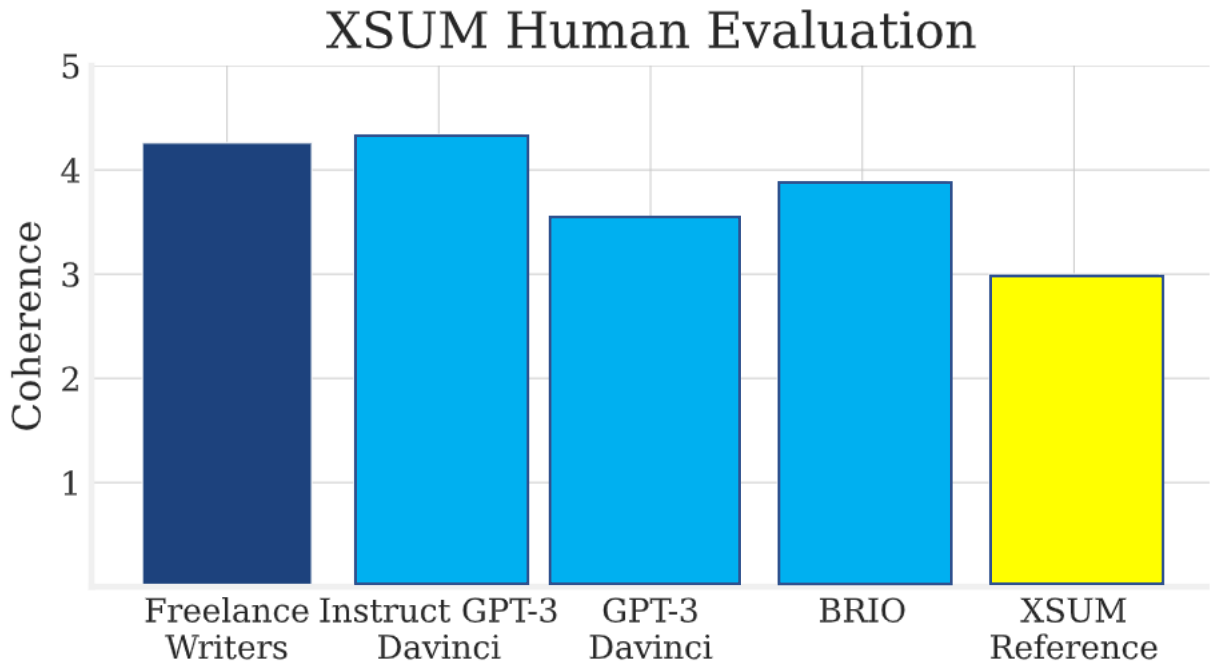
- Characterization: What types of errors and why?
- Large language modeling (LLM): Do the latest LLMs hallucinate?

Large Language Modeling for Summarization

- LLMs have surged in performance in the last six months
- How do they perform on summarization tasks?
- Where should we focus our efforts on further summarization research?

Benchmarking LLMs for Summarization

Evaluation of ten diverse LLMS on news summarization

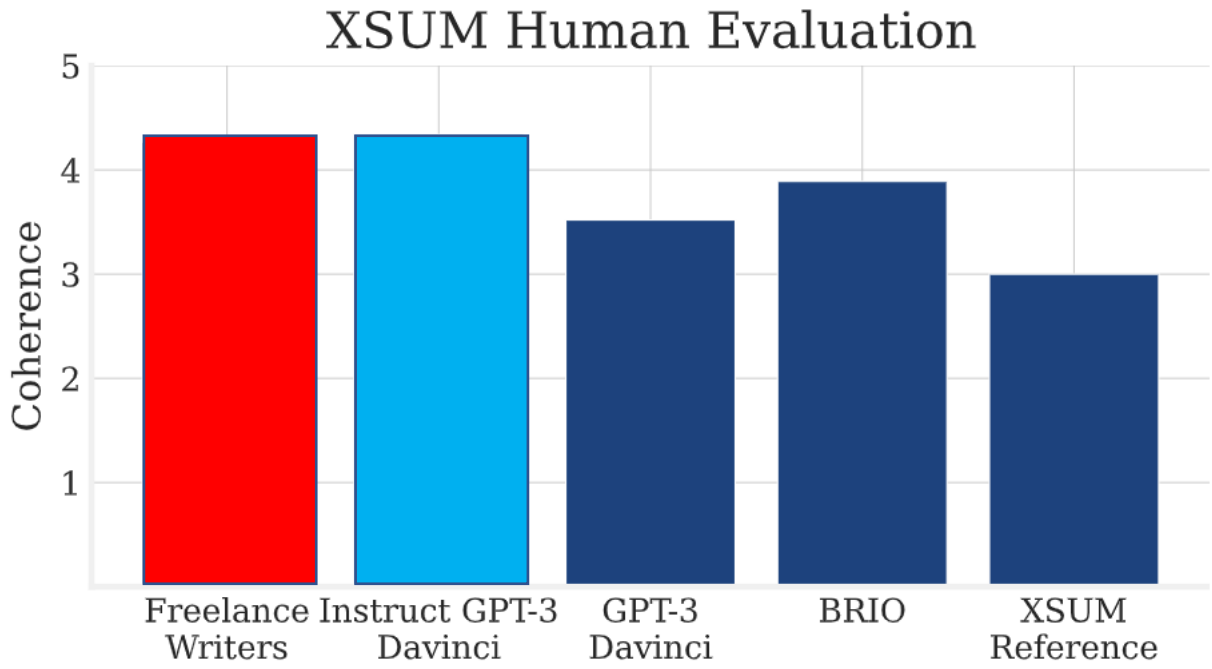


- Results
  - Instruction tuning the key to zero-shot summarization
  - Low quality reference summaries of CNN and Xsum
    - judged worse than system output by humans
    - Degrades performance of fine-tuned or few-shot systems

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown and Tatsu Hashimoto, EMNLP 2023

Benchmarking LLMs for Summarization

Evaluation of ten diverse LLMS on news summarization



- Results
  - Collected human summaries for 100 samples
    - Instruct DaVinci comparable to human summaries
    - Instruct DaVinci paraphrases less

Benchmarking LLMs for New Summarization

Setting	Models	CNN/Daily Mail			XSUM		
		Faithfulness	Coherence	Relevance	Faithfulness	Coherence	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**
Five-shot language models	Anthropic-LM (52B)	0.94	3.88	4.33	0.70	**4.77**	4.14
	Cohere XL (52.4B)	**0.99**	3.42	4.48	0.63	**4.79**	4.00
	GLM (130B)	0.94	3.69	4.24	0.74	4.72	4.12
	OPT (175B)	0.96	3.64	4.33	0.67	**4.80**	4.01
	GPT-3 (350M)	0.86	3.73	3.85	-	-	-
	GPT-3 (6.7B)	0.97	3.87	4.17	0.75	4.19	3.36
	GPT-3 (175B)	**0.99**	3.95	4.34	0.69	4.69	4.03
	Ada Instruct v1 (350M*)	0.84	3.84	4.07	0.63	3.54	3.07
	Curie Instruct v1 (6.7B*)	0.96	**4.30**	4.43	0.85	4.28	3.80
	Davinci Instruct v2 (175B*)	**0.98**	4.13	4.49	0.77	**4.83**	**4.33**
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

Benchmarking LLMs for New Summarization

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**
Five-shot language models	Anthropic-LM (52B)	0.94	3.88	4.33	0.70	**4.77**	4.14
	Cohere XL (52.4B)	**0.99**	3.42	4.48	0.63	**4.79**	4.00
	GLM (130B)	0.94	3.69	4.24	0.74	4.72	4.12
	Ada Instruct v1 (350M*)	0.84	3.84	4.07	0.63	3.54	3.07
	Curie Instruct v1 (6.7B*)	0.96	**4.30**	4.43	0.85	4.28	3.80
	Davinci Instruct v2 (175B*)	**0.98**	4.13	4.49	0.77	**4.83**	**4.33**
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

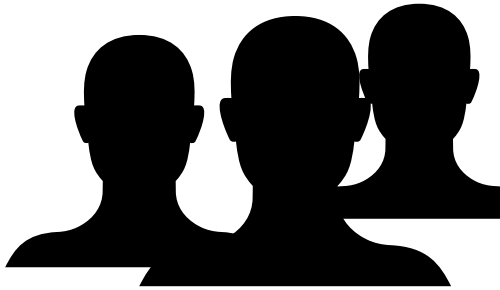
Instruction-tuning + RLHF is the key to zero-shot summarization performance

Benchmarking LLMs for New Summarization

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**
Five-shot language mode	Anthropic-LM (52B)	0.94	3.88	4.33	0.70	**4.77**	4.14
	Cohere XL (52.4B)	**0.99**	3.42	4.48	0.63	**4.79**	4.00
	GLM (130B)	0.94	3.69	4.24	0.74	4.72	4.12
	Ada Instruct v1 (350M*)	0.84	3.84	4.07	0.63	3.54	3.07
	Curie Instruct v1 (6.7B*)	0.96	**4.30**	4.43	0.85	4.28	3.80
	Davinci Instruct v2 (175B*)	**0.98**	4.13	4.49	0.77	**4.83**	**4.33**
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

Instruction tuning has a bigger impact than scaling for zero-shot summarization

Faithfulness is resolved with Instruct GPT



- XSum: Faithfulness scores are close to perfect for Instruct zero shot
- Scores drop by 20-30 points across the board for five-shot models

-> even a small amount of training on noisy data is problematic

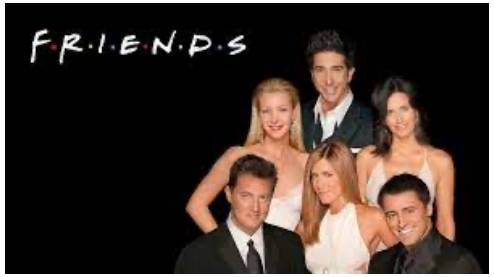
Single Document News Summarization is Solved

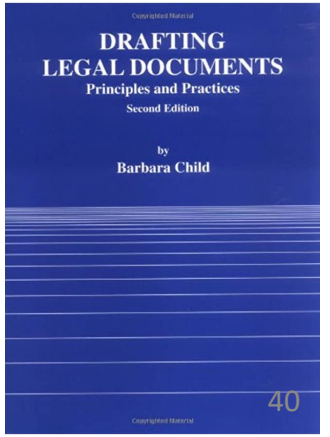
- Is this surprising?
  - *No:* Prior work shows that small differences in Rouge = small differences in summaries (Kedzie, McKeown, Daume 2018)
  - *Yes:* Faithfulness has been a problem for a while with LLMs.

- Does this mean that summarization is solved?

Different Genres Can Be Difficult

- Novels and narratives
  - Interpretation
  - Long input
- Dialogs can be challenging – some of the same issues
  - TV episode summaries
  - Meeting summarization
- Long documents, non-standard genres (e.g., legal)
- Multi-document settings





Characterizing LLMS on Novels and Narratives

- Has the model simply memorized the data?
- Does the model learn the point of the narrative?
- Is faithfulness more of an issue than in news?
- How do we evaluate?

Characterizing LLMS on Novels and Narratives

- Has the model simply memorized the data?
  - Yes with novels/narratives with large online presence
- Does the model learn the point of the narrative?
- Is faithfulness more of an issue than in news?
- How do we evaluate?

Characterizing LLMS on Novels and Narratives

- Has the model simply memorized the data?
  - Yes with novels/narratives with large online presence
- Does the model learn the point of the narrative?
  - Learning to interpret the point of the narrative most difficult
- Is faithfulness more of an issue than in news?
- How do we evaluate?

Narrative Summarization (long document)

Example LLM Summaries
Story: [TRUNCATED] ... While standing there, I feel numb. My fingers grip the bouquet tight. I remember Mom, a sweet woman. Her smile, her strong hands, and witty humor . . . After several minutes, I gather the courage to speak to her. "How's it going, mom?" I feel strong emotions coming. Mom's voice answers me in return: You brat! I took care of you and your brother! I gave you your own closet. But you scratched it like a cat! You! You didn't know how to wash in the tub, so I showed you how! I laugh at the whisper. "You're telling a different story, mom." I smile. "At least you won't hurt me or Hector anymore."
Instruct Davinci: The narrator is visiting a cemetery, and has a strange gift where they can hear the voices of the dead. As they explore, they come across several graves that are speaking to them, including one belonging to a baby, a man who was prejudiced against people of color, and a beloved mother. The narrator talks to their mother's grave and **reminisces, and finds that even in death, their mother still has plenty of wit and humor.**

Our research benchmarking LLMs on narrative summarization shows that 70% summaries contain hallucinations

Characterizing LLMS on Novels and Narratives

- Has the model simply memorized the data?
  - Yes with novels/narratives with large online presence
- Does the model learn the point of the narrative?
  - Learning to interpret the point of the narrative most difficult
- Is faithfulness more of an issue than in news?
  - Yes, early experiments show .3 faithfulness in narrative
- How do we evaluate?

Characterizing LLMS on Novels and Narratives

- Has the model simply memorized the data?
  - Yes with novels/narratives with large online presence
- Does the model learn the point of the narrative?
  - Learning to interpret the point of the narrative most difficult
- Is faithfulness more of an issue than in news?
  - Yes, early experiments show .3 faithfulness in narrative
- How do we evaluate?
  - Given input length and output fluency, human evaluation difficult

Evaluating Summarization of Short Stories with Experienced Writers

- Use stories written by experienced creative writers -> not online
- Evaluate attributes of narrative: faithfulness and thematic analysis
- Ask creative writers to evaluate summaries of their own stories
- Evaluate ability of three LLMs: GPT4, Claude and Llama-2 70B

Subbiah, Zhang, Chilton and McKeown, arxiv 2024

Data Statistics

Length Bucket	Stories		Summary Avg. Len.		
	#	Avg. Len.	GPT4	Claude	Llama
Short	10	1854	487	397	458
Medium	9	4543	500	339	482
Long	6	8126	531	382	592
Total	25	4327	502	373	499

Hierarchical summarization for Llama – chunk then summarize and summarize again

Authors Assess Quality

- Coverage – Does the summary cover the important plot points of the story?
- Faithfulness – Does the summary misrepresent details from the story or make things up?
- Coherence – Is the summary coherent, fluent, and readable?
- Analysis – Does the summary provide any correct analysis of some of the main takeaways or themes from the story?

Results Summary – Scores assigned by writers

Model	Cover.	Faithful.	Coheren.	Analys.	Avg.
GPT-4	3.48	3.12	3.52	3.40	3.38
Claude	3.17	2.67	3.43	3.26	3.13
Llama	2.40	1.92	3.08	2.76	2.54
GPT-4	56%	44%	60%	56%	54%
Claude	39%	30%	61%	43%	43%
Llama	12%	8%	32%	20%	18%

Writers assign scores from 1-4, 4 highest

Models are capable of producing good summaries

Results Summary – Scores assigned by writers

Model	Cover.	Faithful.	Coheren.	Analys.	Avg.
GPT-4	3.48	3.12	3.52	3.40	**3.38**
Claude	3.17	2.67	3.43	3.26	**3.13**
Llama	2.40	1.92	3.08	2.76	2.54
GPT-4	56%	44%	60%	56%	54%
Claude	39%	30%	61%	43%	43%
Llama	12%	8%	32%	20%	18%

Percent summaries that score a 4

Models are capable of producing good summaries

Faithfulness scores are low – even GPT produces faithful summaries only 44% of the time

Results – Scores assigned by writers

Model	Cover.	Faithful.	Coheren.	Analys.	Avg.
GPT-4	3.48	3.12	3.52	3.40	**3.38**
Claude	3.17	2.67	3.43	3.26	**3.13**
Llama	2.40	1.92	3.08	2.76	2.54
GPT-4	56%	44%	60%	56%	54%
Claude	39%	30%	61%	43%	43%
Llama	12%	8%	32%	20%	18%

Models are capable of producing good summaries

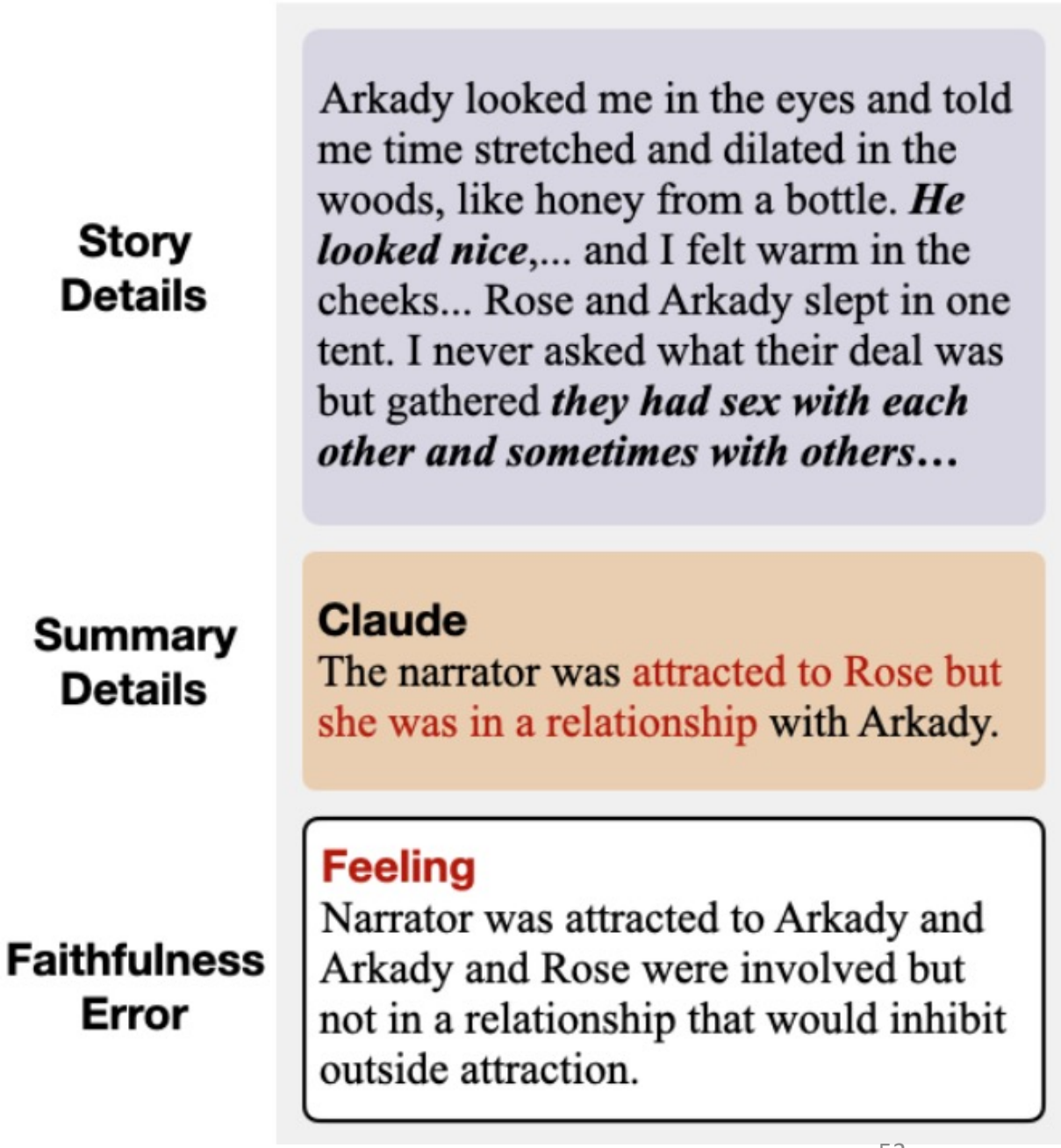
Faithfulness scores are low – even GPT produces faithful summaries only 44% of the time

Models do better at analysis than expected but *interpreting subtext is hard*

Summaries are least faithful in conveying *emotion*

Also categorized faithfulness errors by:

causation
action
character
setting

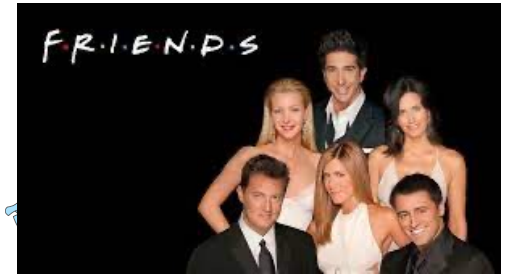


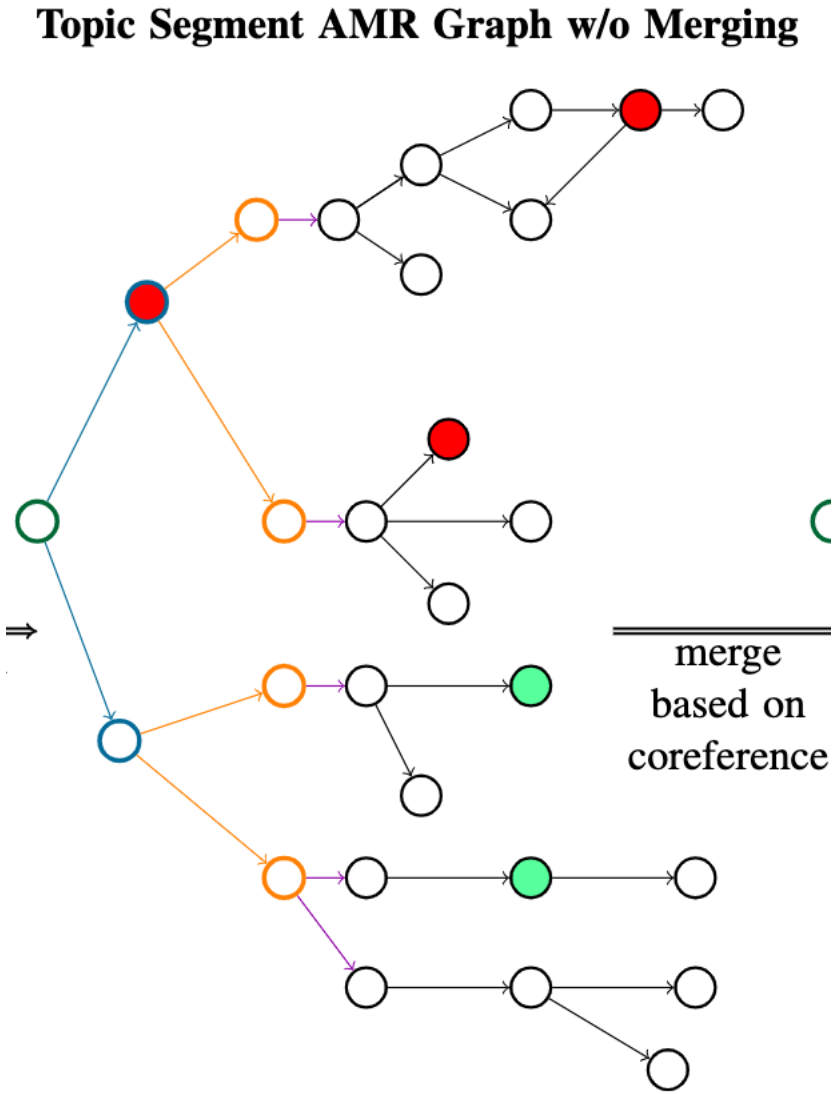
New Methods Instead of New Tasks

- Abstract semantic representations for summarization of TV episodes and meetings

- Topic segment AMR graphs
  - Capture discourse structure
  - Highlight salient semantics

- Text-graph cross-attention to leverage AMR and LLMs





Yilun Hua, Zhaoyuan Deng and Kathleen McKeown ACL 2023

Enabling Unbiased Summarization of Opinions from Vulnerable Groups

- Summarizing public reaction to important issues of our time using AI
- Better understanding of social media posts from Black individuals
- Training data for models contains < .07% African American Language (AAL)
  - LLMs have more difficulty interpreting and producing AAL in comparison to White Mainstream English (Deas et al, EMNLP 2023)
  - Investigating the use of phonology to mitigate
- Test improved AAL models on the summarization task

Prof. Kathleen McKeown
Department of Computer Science
Columbia University

Prof. Desmond Patton
School of Social Policy and Practice
Annenberg School for Communication
University of Pennsylvania

What Have We Learned?

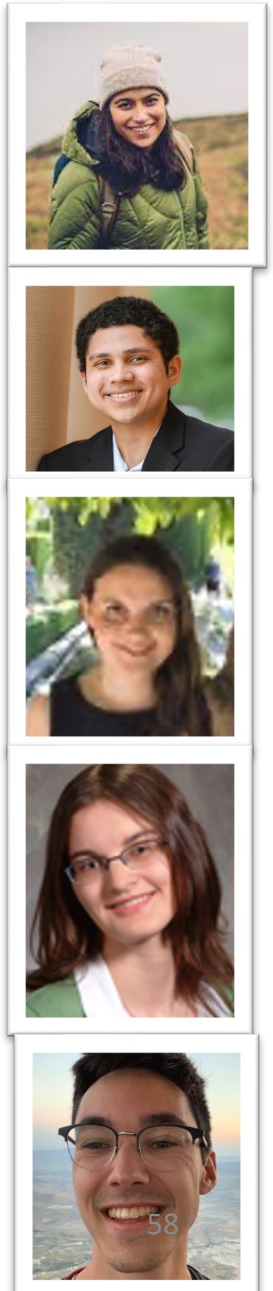
- Instruction tuned, zero-shot LLMs solve the hallucination problem for news
- Even a small amount of noisy data reduces faithfulness
- Other genres still difficult
  - LLMS are unfaithful > 50% of the time when summarizing narrative
  - Dialog is another difficult task with on average 28% unfaithful summaries
- Going forward, we need to develop fair summarization models for all segments of the population

Is there more research to be done in summarization –

and NLP?

Yes!

- Summarization of more difficult genres
- Adaptation to less represented dialects (e.g., AAL)
- Implicit information in language (e.g., generics)
- Adapting to new methods using LLMs and other representations
- Controllable, computationally efficient models



Collaborators



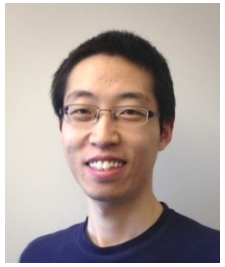


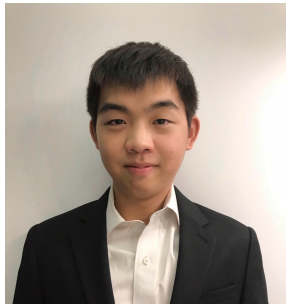
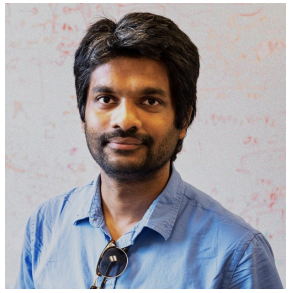


Cornell University Founded A.D. 1865

NYU

Stanford Artificial Intelligence



aws alexa

UNC NLP

Columbia University

Thank you!

Emily Allaway

Amith Ananthram

Yanda Chen

Nick Deas

Zachary Horvitz

Faisal Ladhak

Fei-tzin Lee

Melanie Subbiah

Elsbeth Turcan