TEXAS
The University of Texas at Austin

Discourse Models with Language Models

Jessy Li
The University of Texas at Austin

Long form generation: then & now

Humans were not the only intelligent life in the universe. Instead, they were just the first. They explored the universe and helped build many budding civilizations until one day, the human species disappeared. You are an alien historian that figured out what happened after years of thorough investigation.

r/WritingPrompts

Llama-3.2-1b

As an alien historian, I have spent my entire career studying the history of the universe, from the earliest civilizations to the most recent technological advancements. But none of my research has prepared me for what I discovered on a remote planet in the outer reaches of the galaxy. It was a planet that had been largely unexplored by humans, and I had spent years searching for it. Finally, after months of searching, I found it. The planet was called Zha'thik, and it was a world unlike any other…

GPT-2

What you found was true as well, and your descendants would surely find it too hard not to continue with their journey. (The author, Christopher Nolan, is probably not a big fan of the Alien films, so he has no idea if he can avoid some of his own clichés.) This book is a fantastic adventure that has had a huge impact on me at times, but it could be more important if it could be told right — from where I am writing my story, across the universe…

Summarization: then & now

Input: Editorial discussing the symbolism of lifting mask mandates

GPT-3

The article discusses the debate around when the Covid-19 pandemic will be considered over. It argues that in political terms, the pandemic will be considered over when all states have lifted […]

Explains all significant parts …

Captures the tone of the article […]

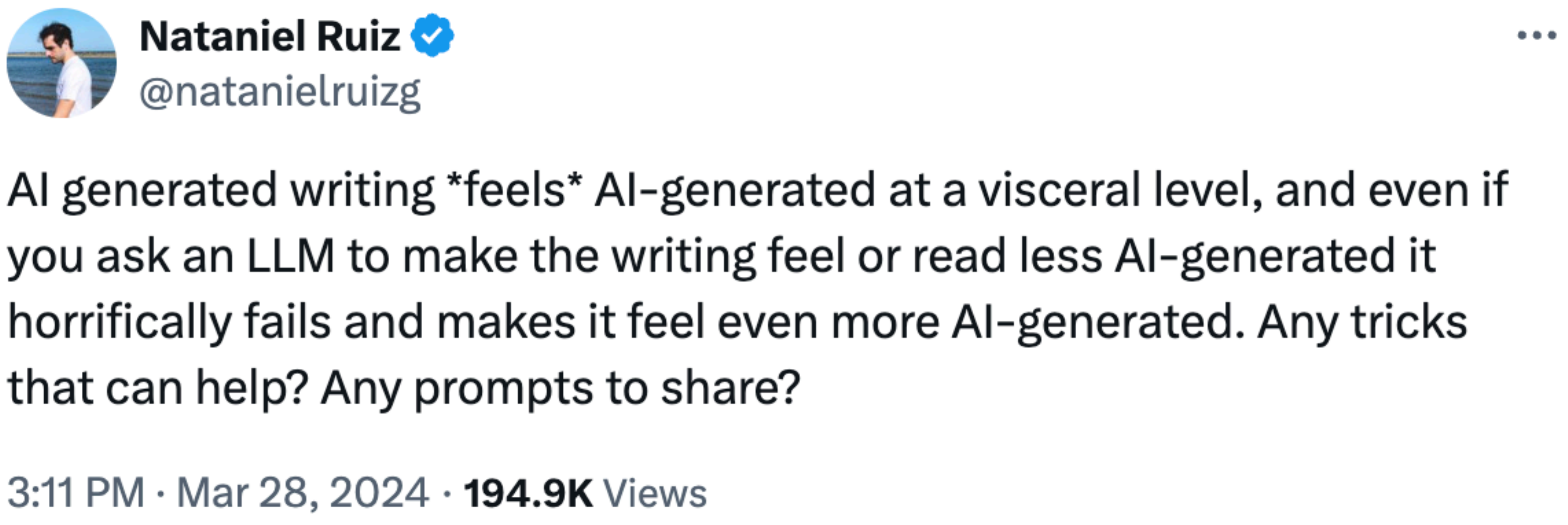
BRIO

Hawaii will become the last state in the country to end its indoor mask mandate. The Covid-19 pandemic has seen a surge in cases of the virus this winter. […]

Reflects the tone of the article the least

Slide credit: Tanya Goyal

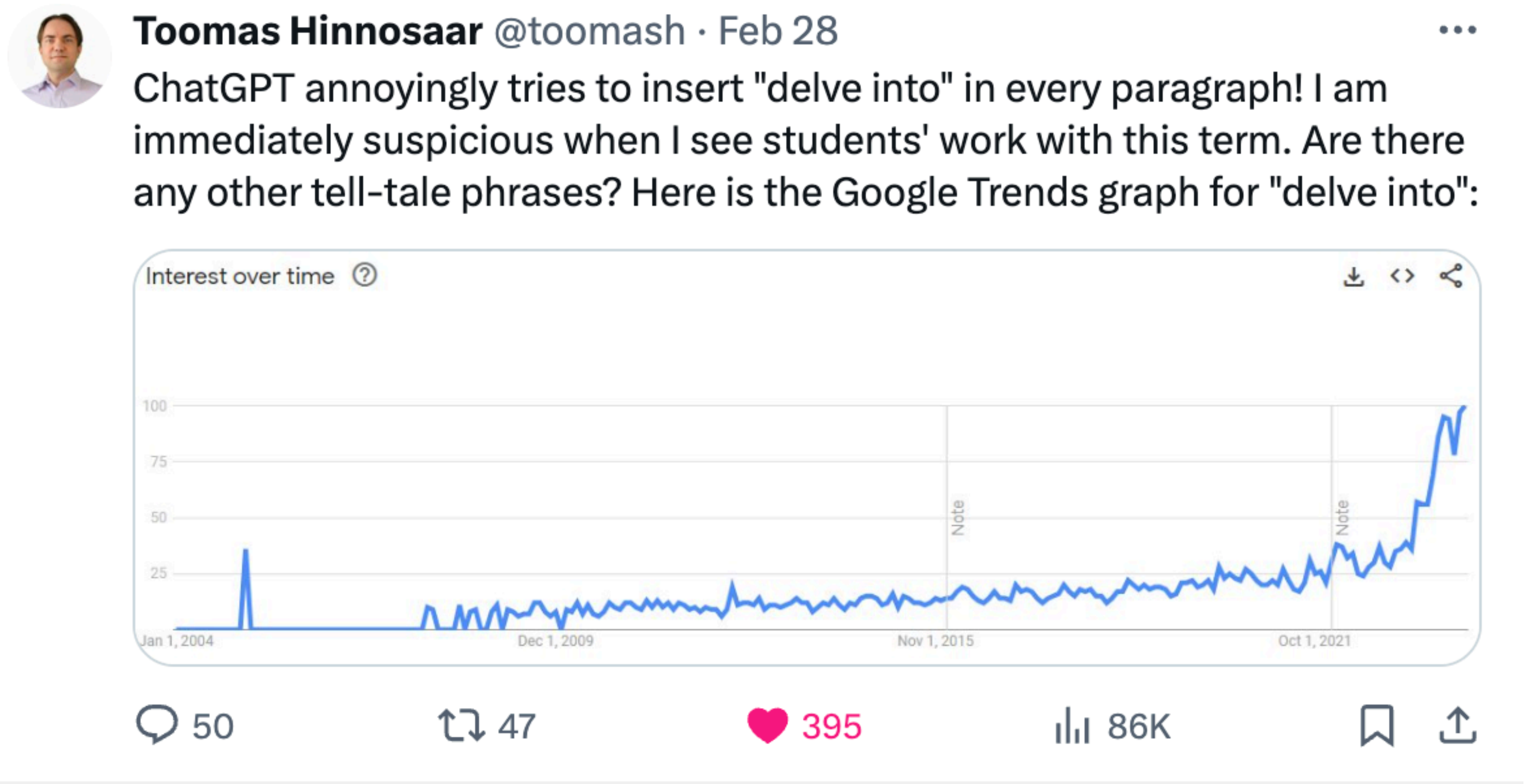
But: LLM texts are somewhat templatic…



Nataniel Ruiz
@natanielruizg

AI generated writing *feels* AI-generated at a visceral level, and even if you ask an LLM to make the writing feel or read less AI-generated it horrifically fails and makes it feel even more AI-generated. Any tricks that can help? Any prompts to share?

3:11 PM · Mar 28, 2024 · **194.9K** Views



Toomas Hinnosaar @toomash · Feb 28

ChatGPT annoyingly tries to insert "delve into" in every paragraph! I am immediately suspicious when I see students' work with this term. Are there any other tell-tale phrases? Here is the Google Trends graph for "delve into":

Syntactic Templates in Generated Text

LLM or human?

Hand model Emily Grimson, 25, from Shropshire, gets paid £100 an hour. Stars in ads for Bosch, the Disney Channel and magazine features.

Restaurant owner Sandy Dee Hall, 34, and his girlfriend Maxine Cher, 24, adopted Smokey Da Lamb after he was abandoned by his mother.

John Caudwell has bought a car park in Audley Street, Mayfair and is planning to replace it with a huge block of flats.

Emily Grimson, a 25-year-old professional hand model from Shropshire, has been modelling for a year and has built a full-time career earning up to £100 an hour.

Smokey Da Lamb, a lamb abandoned by his mother, has been living with restaurant owner Sandy Dee Hall and his girlfriend Maxine Cher in New York City for the past month

John Caudwell, the founder of Phones4U, has unveiled plans to build a £2billion luxury housing complex on the site of an ugly multi-storey car park in Mayfair, central London.

Syntactic Templates in Generated Text

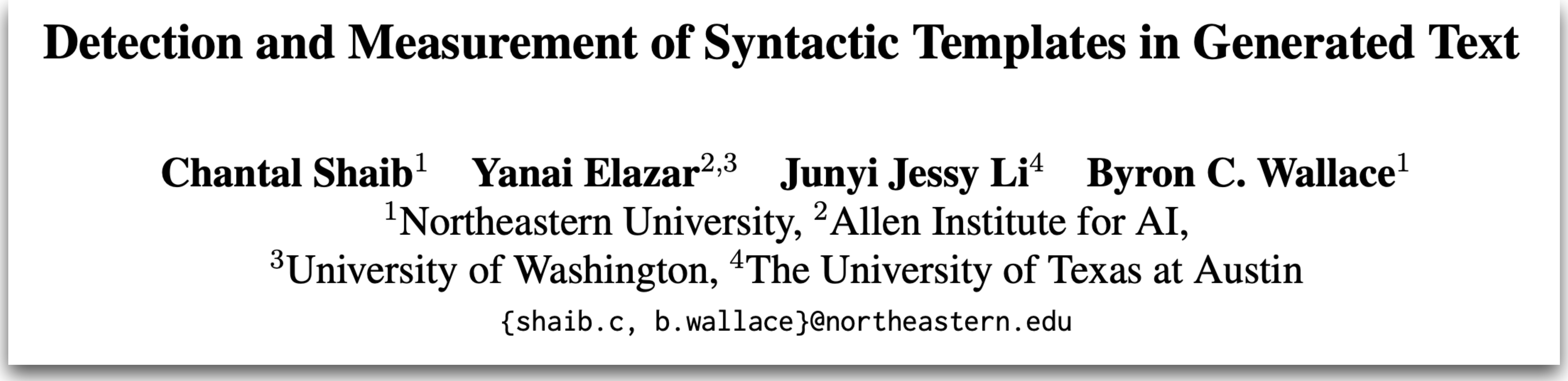
LLM or human?

Hand model Emily Grimson, 25, from Shropshire, gets paid £100 an hour. Stars in ads for Bosch, the Disney Channel and magazine features.

Restaurant owner Sandy Dee Hall, 34, and his girlfriend Maxine Cher, 24, adopted Smokey Da Lamb after he was abandoned by his mother.

Emily Grimson, a 25-year-old professional hand model from Shropshire, has been modelling for a year and has built a full-time career earning up to £100 an hour.

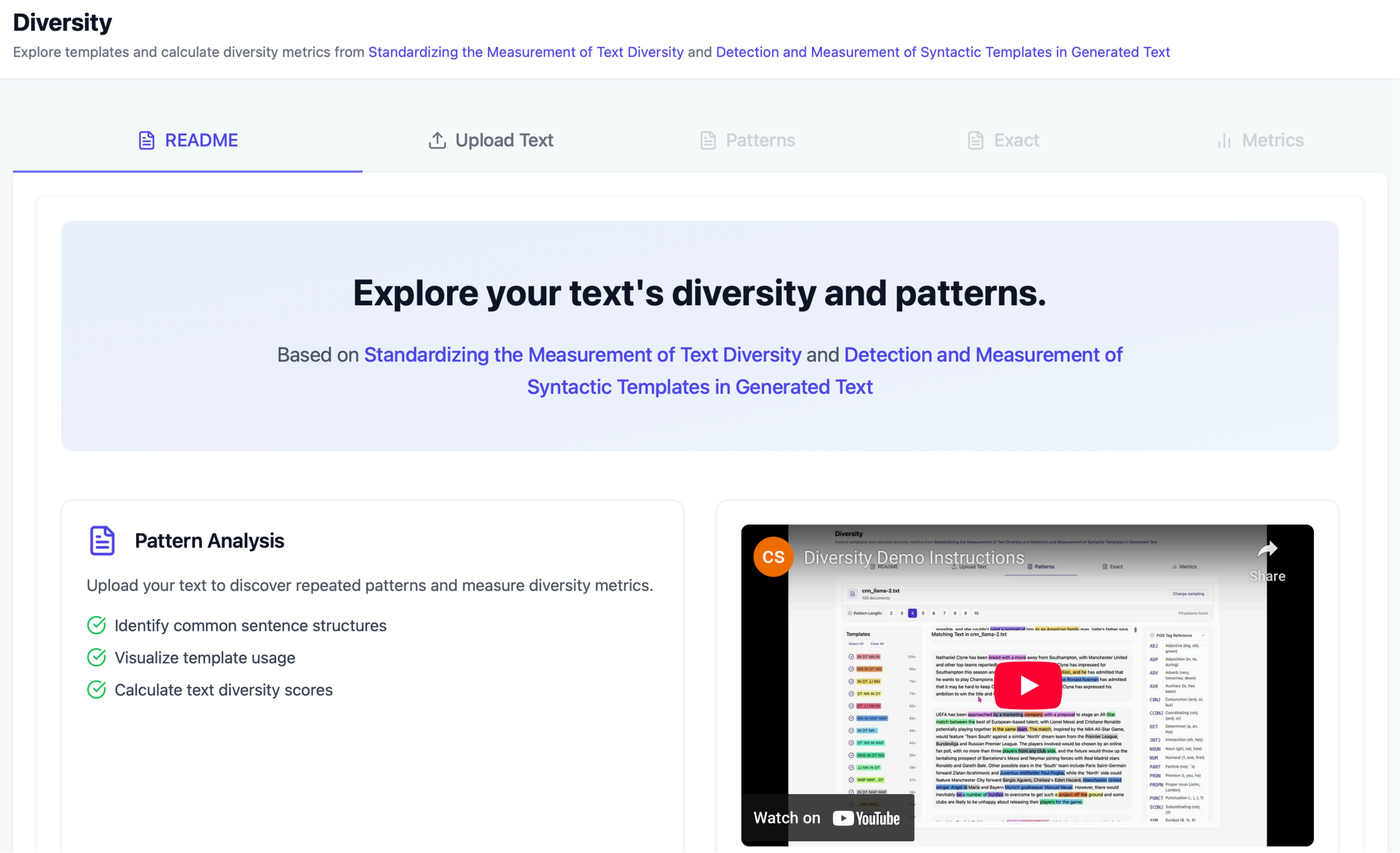
Smokey Da Lamb, a lamb abandoned by his mother, has been living with restaurant owner Sandy Dee Hall and his girlfriend Maxine Cher in New York City for the past month



Detection and Measurement of Syntactic Templates in Generated Text

Chantal Shaib[1] **Yanai Elazar**[2,3] **Junyi Jessy Li**[4] **Byron C. Wallace**[1]
[1]Northeastern University, [2]Allen Institute for AI,
[3]University of Washington, [4]The University of Texas at Austin
{shaib.c, b.wallace}@northeastern.edu

Demo @ https://ai-templates.app/



Diversity

Explore templates and calculate diversity metrics from Standardizing the Measurement of Text Diversity and Detection and Measurement of Syntactic Templates in Generated Text

Explore your text's diversity and patterns.

Based on Standardizing the Measurement of Text Diversity and Detection and Measurement of Syntactic Templates in Generated Text

Pattern Analysis

Upload your text to discover repeated patterns and measure diversity metrics.

- ✓ Identify common sentence structures
- ✓ Visualize template usage
- ✓ Calculate text diversity scores

Detection and Measurement of Syntactic Templates in Generated Text [Shaib et al., EMNLP'24]

THEN

- Text makes sense at the beginning but fails to stay relevant after a while.
- Necessary to supervise to obtain task-specific capabilities like summarization.
- Fails long-form planning.

NOW

- Long, coherent generations, concepts connect across passages.
- Can do conditional generation pretty well, e.g., knows what is important when asked to summarize.
- Demonstrates some capability to plan for long texts, but “feels” synthetic.

Linguistically driven discourse models can help us understand what we see:

- Characterizing relationships across sentences and document-level structure
- Ground analyses in human language processing
- Provide workable tools for model insights and downstream applications

This talk

Can we use language models to make sense of discourse?

- A generative discourse model based on Questions Under Discussion (QUD) theories
- Connections between human curiosity and discourse planning

Can we use our discourse framework to make sense of model behavior?

- A measure of discourse similarity in LLM-generated texts
- An explainable metric of information salience in LLMs

This talk

Can we use language models to make sense of discourse?

- A generative discourse model based on Questions Under Discussion (QUD) theories
- Connections between human curiosity and discourse planning

Can we use our discourse framework to make sense of model behavior?

- A measure of discourse similarity in LLM-generated texts
- An explainable metric of information salience in LLMs

Coherence: a definition

A text is coherent if:

(1) each newly introduced proposition is rhetorically connected to another piece of information, and

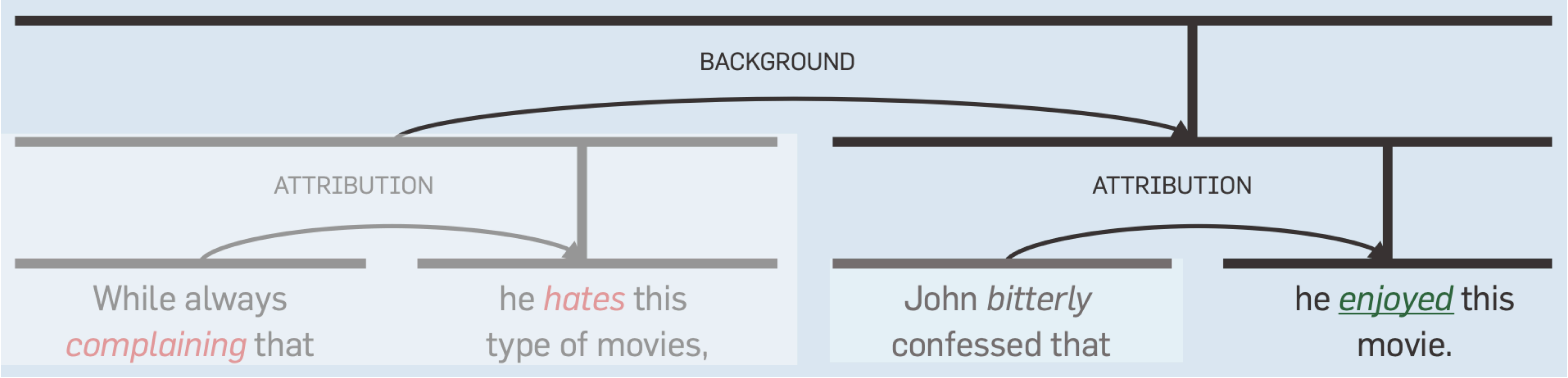
(2) all anaphoric expressions can be resolved.

Asher and Lascarides, Logics of Conversation, 2003

Discourse is (was?) useful!

Sentiment analysis

While always complaining that he hates this type of movies, John bitterly confessed that he enjoyed this movie.


BACKGROUND
ATTRIBUTION
ATTRIBUTION
While always *complaining* that
he *hates* this type of movies,
John *bitterly* confessed that
he *enjoyed* this movie.

Hogenboom et al., Using Rhetorical Structure in Sentiment Analysis, 2015

Questions Under Discussion (QUD)

A well-known linguistic framework in discourse & pragmatics

Each sentence in discourse addresses a (often implicit) QUD either by answering it, or by bringing up another question that can help answering that QUD. The linguistic form and the interpretation of a sentence, in turn, may depend on the QUD it addresses.

Benz and Jasinskaja, Questions under discussion: From sentence to discourse. 2017

Velleman and Beaver. Question-based models of information structure. 2016.
Roberts, Information structure: Towards an integrated formal theory of pragmatics. 2012
Roberts. Context in dynamic interpretation. 2004
Jonathan Ginzburg. Dynamics and the semantics of dialogue. 1996
Jan Van Kuppevelt. Discourse structure, topicality and questioning. 1995

Let’s take a step back...

... and think about how YOU read...

[1] California legislators, searching for ways to pay for … damages from last week′s earthquake, are laying the groundwork for a temporary increase in the state′s sales tax. [2] The talk of a sales tax rise follows a rebuff from Congress on the question of how much the federal government is willing to spend to aid in California′s earthquake relief efforts. [3] The state had sought as much as $4.1 billion in relief, but yesterday the House approved a more general scaled-back measure… [4] That leaves the state roughly $2 billion to $4 billion short. [5] A sales tax increase appears to be the fastest and easiest to raise funds in a hurry. [6] According to the state department of finance, a one-penny increase in the state′s six-cent per dollar sales tax could raise $3 billion.

Let’s take a step back…

… and think about how YOU read…

[1] California legislators, searching for ways to pay for … damages from last week′s earthquake, are laying the groundwork for a temporary increase in the state′s sales tax. [2] The talk of a sales tax rise follows a rebuff from Congress on the question of how much the federal government is willing to spend to aid in California′s earthquake relief efforts. [3] The state had sought as much as $4.1 billion in relief, but yesterday the House approved a more general scaled-back measure… [4] That leaves the state roughly $2 billion to $4 billion short. [5] A sales tax increase appears to be the fastest and easiest to raise funds in a hurry. [6] According to the state department of finance, a one-penny increase in the state′s six-cent per dollar sales tax could raise $3 billion.

What is prompting the California legislators to seek alternative forms of relief?

Let’s take a step back…

… and think about how YOU read…

[1] California legislators, searching for ways to pay for … damages from last week′s earthquake, are laying the groundwork for a temporary increase in the state′s sales tax. [2] The talk of a sales tax rise follows a rebuff from Congress on the question of how much the federal government is willing to spend to aid in California′s earthquake relief efforts. [3] The state had sought as much as $4.1 billion in relief, but yesterday the House approved a more general scaled-back measure… [4] That leaves the state roughly $2 billion to $4 billion short. [5] A sales tax increase appears to be the fastest and easiest to raise funds in a hurry. [6] According to the state department of finance, a one-penny increase in the state′s six-cent per dollar sales tax could raise $3 billion.

What is prompting the California legislators to seek alternative forms of relief?

Answered in sentence 2

Let’s take a step back...

... and think about how YOU read...

[1] California legislators, searching for ways to pay for ... damages from last week′s earthquake, are laying the groundwork for a temporary increase in the state′s sales tax. [2] The talk of a sales tax rise follows a rebuff from Congress on the question of how much the federal government is willing to spend to aid in California′s earthquake relief efforts. [3] The state had sought as much as $4.1 billion in relief, but yesterday the House approved a more general scaled-back measure... [4] That leaves the state roughly $2 billion to $4 billion short. [5] A sales tax increase appears to be the fastest and easiest to raise funds in a hurry. [6] According to the state department of finance, a one-penny increase in the state′s six-cent per dollar sales tax could raise $3 billion.

What would be the advantage of the proposed tax?

Let’s take a step back…

… and think about how YOU read…

[1] California legislators, searching for ways to pay for … damages from last week′s earthquake, are laying the groundwork for a temporary increase in the state′s sales tax. [2] The talk of a sales tax rise follows a rebuff from Congress on the question of how much the federal government is willing to spend to aid in California′s earthquake relief efforts. [3] The state had sought as much as $4.1 billion in relief, but yesterday the House approved a more general scaled-back measure… [4] That leaves the state roughly $2 billion to $4 billion short. [5] A sales tax increase appears to be the fastest and easiest to raise funds in a hurry. [6] According to the state department of finance, a one-penny increase in the state′s six-cent per dollar sales tax could raise $3 billion.

What would be the advantage of the proposed tax?

Answered in sentence 5

- What constraints should the Q & A follow?
- What kind of discourse structure can we get from this?

Let's take a step back...

... and think about how YOU read...

[1] California legislators, searching for ways to pay for … damages from last week's earthquake, are laying the groundwork for a temporary increase in the state's sales tax. [2] The talk of a sales tax rise follows a rebuff from Congress on the question of how much the federal government is willing to spend to aid in California's earthquake relief efforts. [3] The state had sought as much as $4.1 billion in relief, but yesterday the House approved a more general scaled-back measure… [4] That leaves the state roughly $2 billion to $4 billion short. [5] A sales tax increase appears to be the fastest and easiest to raise funds in a hurry. [6] According to the state department of finance, a one-penny increase in the state's six-cent per dollar sales tax could raise $3 billion.

What would be the advantage of the proposed tax?

What was the rebuff?

How much will the sales tax be raised?

Who led the talk of the raise?

TEXAS
The University of Texas at Austin

Let's take a step back...

... and think about how YOU read...

- Many potential questions *(Onea, 2016)* can be licensed, all of them are valid.
- Do people agree on what questions are more salient?
- Do more salient questions get answered later?

What would be the advantage of the proposed tax?

What was the rebuff?

How much will the sales tax be raised?

Who led the talk of the raise?

A question-based discourse model

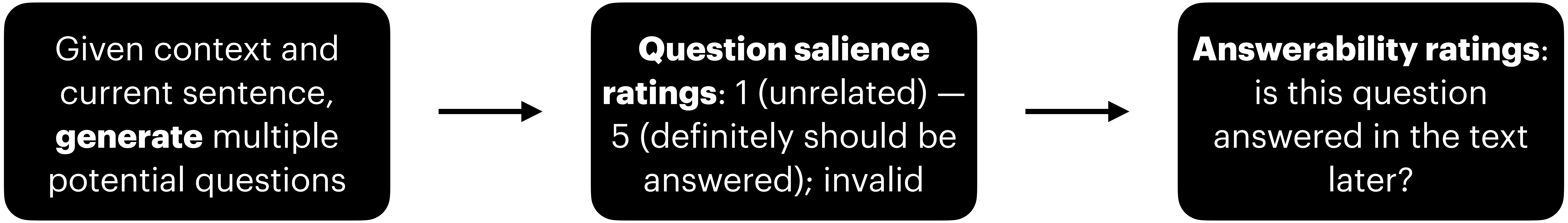
Reader: as they read, wonder about potential questions

Writer: some of these questions become QUDs (ie, answered in the document)

What we show:

- Empirical data support this type of model
- Documents have this QA-based structure
- Language models are not quite there yet for discourse-level reasoning

QSalience data collection



Given context and current sentence, **generate** multiple potential questions

Question salience ratings: 1 (unrelated) — 5 (definitely should be answered); invalid

Answerability ratings: is this question answered in the text later?

[1] Amid skepticism that Russia's war in Chechnya can be ended across a negotiating table, peace talks were set to resume Wednesday in neighboring Ingushetia. **[2]** The scheduled resumption of talks in the town of Sleptsovsk came two days after agreement on a limited cease-fire, calling for both sides to stop using heavy artillery Tuesday.

[3] They also agreed in principle to work out a mechanism for exchanging prisoners of war and the dead. **[4]** Despite the pact, artillery fire sounded in the Grozny on Tuesday, and there were reports of Chechen missile attacks southwest of the Chechen capital.

[Q2] What is the significance of the limited cease-fire agreement that was reached?	**Invalid.** Incorrect anchor
[Q3] What was the reason behind the artillery fire in Grozny on Tuesday despite the agreed cease-fire?	**Salience: 5.** This question would be useful in understanding why the cease-fire was broken, which could give insight into how optimistic the peace talks will be.
[Q4] What are the reports of Chechen missile attacks southwest of the Chechen capital?	**Salience: 3.** This question doesn't interest me; why there are missing attacks would help my understanding more
[Q5] What is the source of the Chechen missile attacks?	**Salience: 2.** Based on context, can be inferred that attack comes from Russia

Which questions should I answer? Salience Prediction of Inquisitive Questions [Yating Wu, Ritika Mangla et al, EMNLP'24]

QSalience data collection

- Source texts:
  - News texts, news summaries, TED-talks
- Question generation:
  - Llama-2-7b-chat, Mistral-7b-instruct, GPT-3.5-turbo, GPT-4-turbo
  - LLMs are shown context up to the anchor sentence

# articles	# questions	avg. length	std.dev
66	1766	14	4.6

Agreement (alpha): 0.63-0.75

Which questions should I answer? Salience Prediction of Inquisitive Questions [Yating Wu, Ritika Mangla et al, EMNLP'24]

Do salience and answerability correlate?

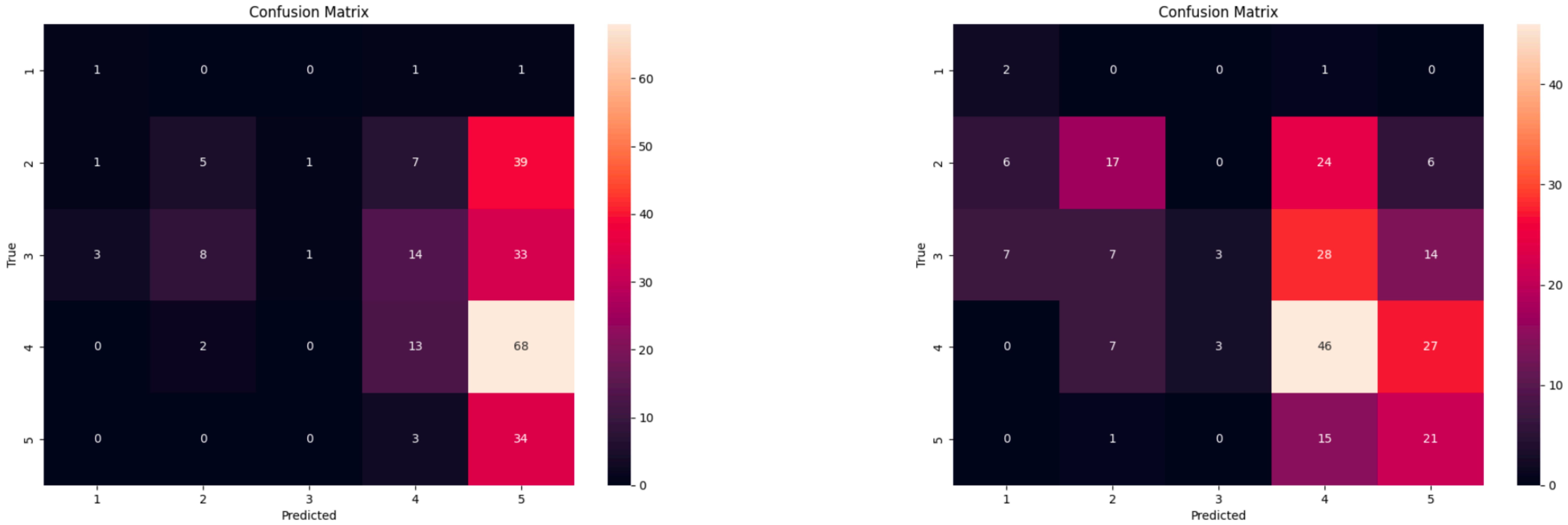
- Spearman’s rho between:
  - Annotated salience vs. answerability of a random question in the dataset
  - Annotated salience vs. answerability of the current question

	Human Salience
Random Questions	-0.02*
Answerability	0.65

Questions that are answered later in the same document did get higher salience scores!

Which questions should I answer? Salience Prediction of Inquisitive Questions [Yating Wu, Ritika Mangla et al, EMNLP’24]

Are LLMs good at salience prediction?



True \ Predicted	1	2	3	4	5
1	1	0	0	1	1
2	1	5	1	7	39
3	3	8	1	14	33
4	0	2	0	13	68
5	0	0	0	3	34

True \ Predicted	1	2	3	4	5
1	2	0	0	1	0
2	6	17	0	24	6
3	7	7	3	28	14
4	0	7	3	46	27
5	0	1	0	15	21

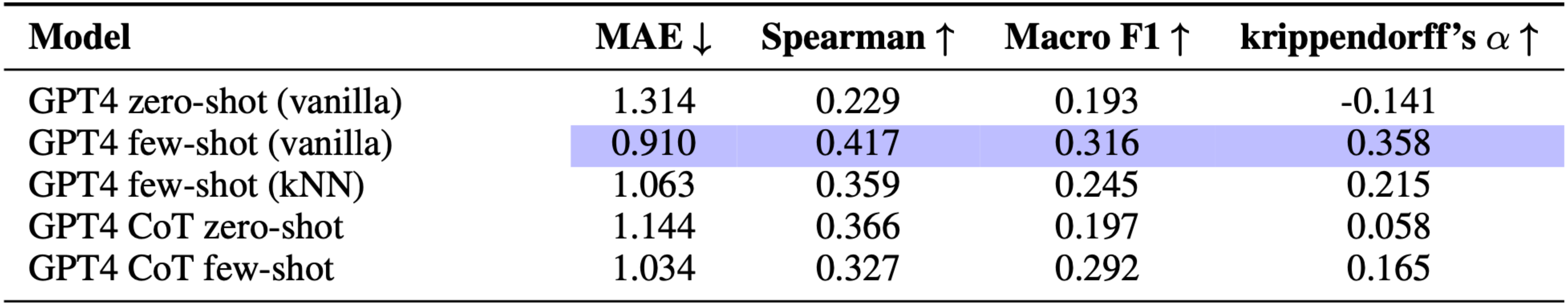
(a) GPT-4-turbo zero-shot vanilla (left), GPT-4-turbo few-shot vanilla (right)

QSalience: building a salience model

- Instruction fine-tuning:
  - Mistral-7b-instruct, Llama-2-7b-chat, TinyLlama-1.1b-chat, Flan-T5-base
  - Method: QLoRA
- LLM prompting and in-context learning:
  - Zero-shot, few shot
  - Few-shot with example retrieval
  - Chain-of-Thought (zero- and few-shot)
  - All with GPT-4

Which questions should I answer? Salience Prediction of Inquisitive Questions [Yating Wu, Ritika Mangla et al, EMNLP'24]

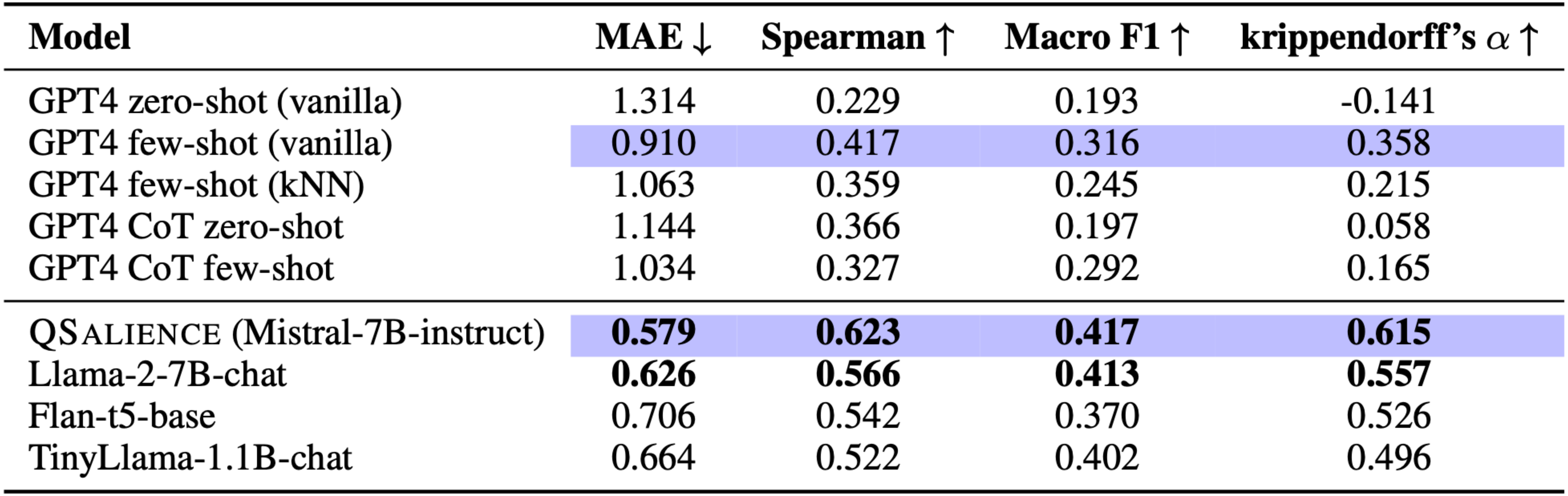
QSalience: results



Model	MAE ↓	Spearman ↑	Macro F1 ↑	krippendorff's α ↑
GPT4 zero-shot (vanilla)	1.314	0.229	0.193	-0.141
GPT4 few-shot (vanilla)	0.910	0.417	0.316	0.358
GPT4 few-shot (kNN)	1.063	0.359	0.245	0.215
GPT4 CoT zero-shot	1.144	0.366	0.197	0.058
GPT4 CoT few-shot	1.034	0.327	0.292	0.165

Which questions should I answer? Salience Prediction of Inquisitive Questions [Yating Wu, Ritika Mangla et al, EMNLP'24]

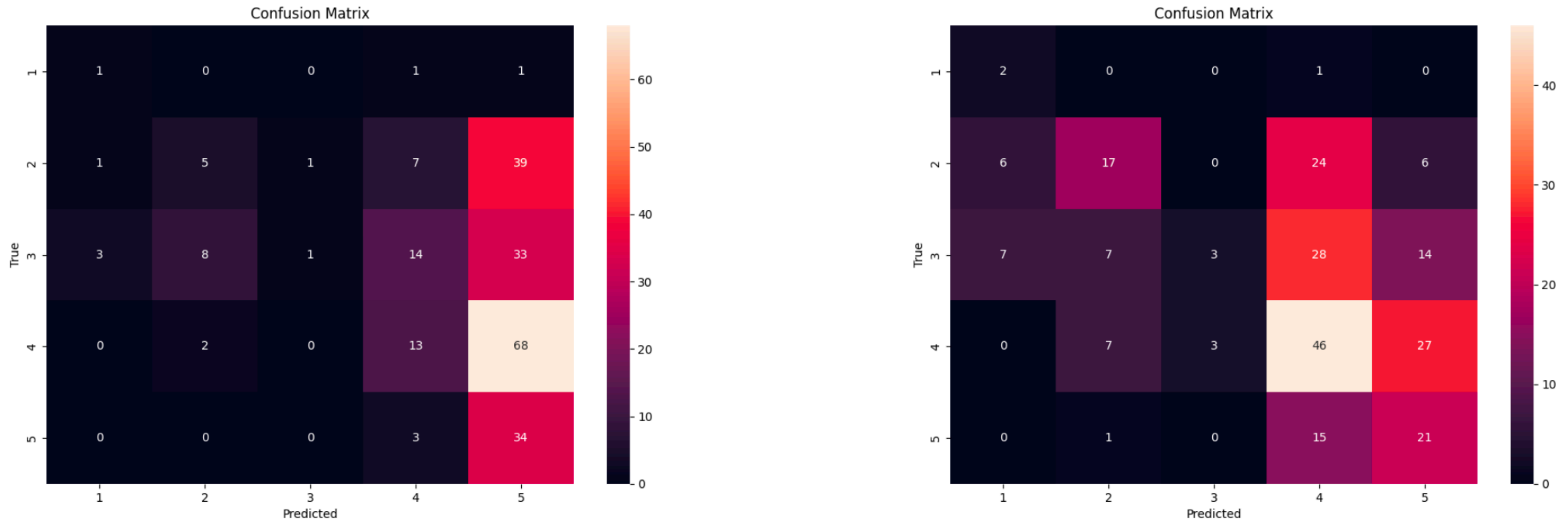
QSalience: results



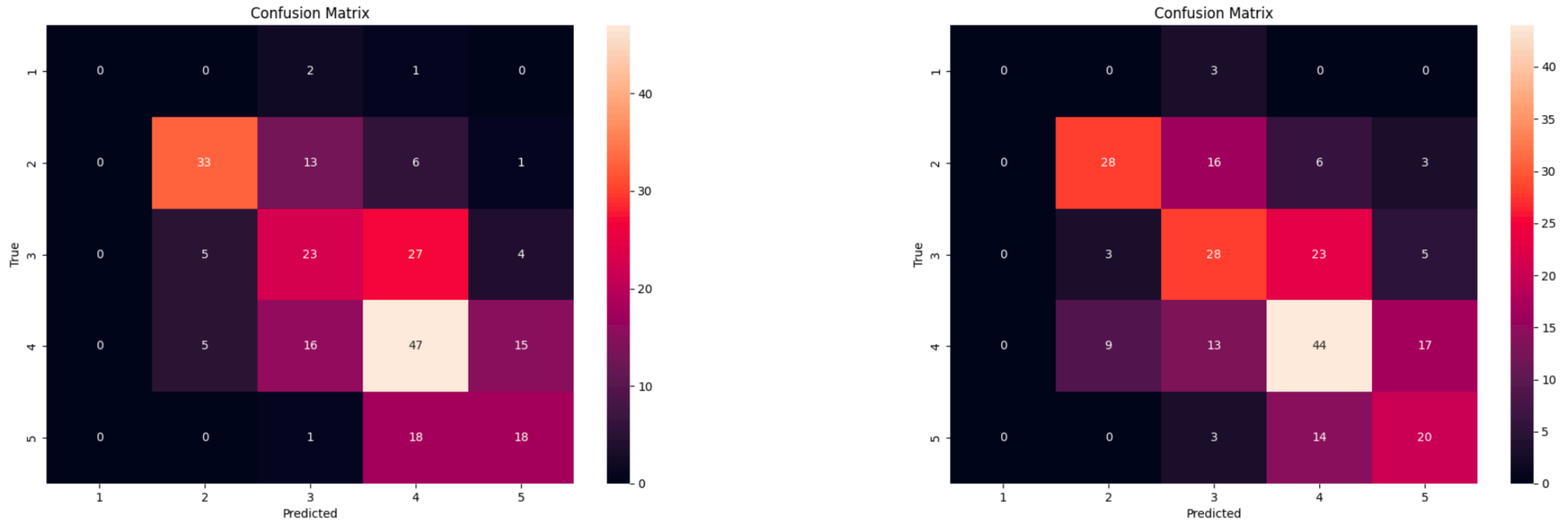
Model	**MAE** ↓	**Spearman** ↑	**Macro F1** ↑	**krippendorff's** α ↑
GPT4 zero-shot (vanilla)	1.314	0.229	0.193	-0.141
GPT4 few-shot (vanilla)	0.910	0.417	0.316	0.358
GPT4 few-shot (kNN)	1.063	0.359	0.245	0.215
GPT4 CoT zero-shot	1.144	0.366	0.197	0.058
GPT4 CoT few-shot	1.034	0.327	0.292	0.165
QSalience (Mistral-7B-instruct)	**0.579**	**0.623**	**0.417**	**0.615**
Llama-2-7B-chat	**0.626**	**0.566**	**0.413**	**0.557**
Flan-t5-base	0.706	0.542	0.370	0.526
TinyLlama-1.1B-chat	0.664	0.522	0.402	0.496

Which questions should I answer? Salience Prediction of Inquisitive Questions [Yating Wu, Ritika Mangla et al, EMNLP'24]

QSalience: results



(a) GPT-4-turbo zero-shot vanilla (left), GPT-4-turbo few-shot vanilla (right)



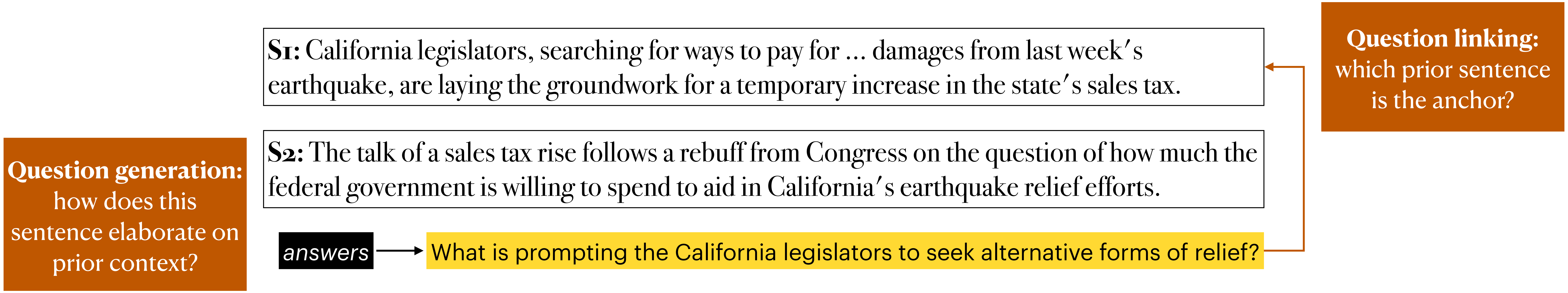
(b) Mistral-7B-Instruct (left), Llama-2-7B-chat (right)

So how do we make use of the QA paradigm...

... and derive a model for discourse structure?

- Key ideas:
    - Each sentence is an answer to an implicit Question Under Discussion
    - Generate these QUDs
    - Find out where these QUDs are licensed in the document context

QUD-based discourse model



S1: California legislators, searching for ways to pay for … damages from last week's earthquake, are laying the groundwork for a temporary increase in the state's sales tax.

S2: The talk of a sales tax rise follows a rebuff from Congress on the question of how much the federal government is willing to spend to aid in California's earthquake relief efforts.

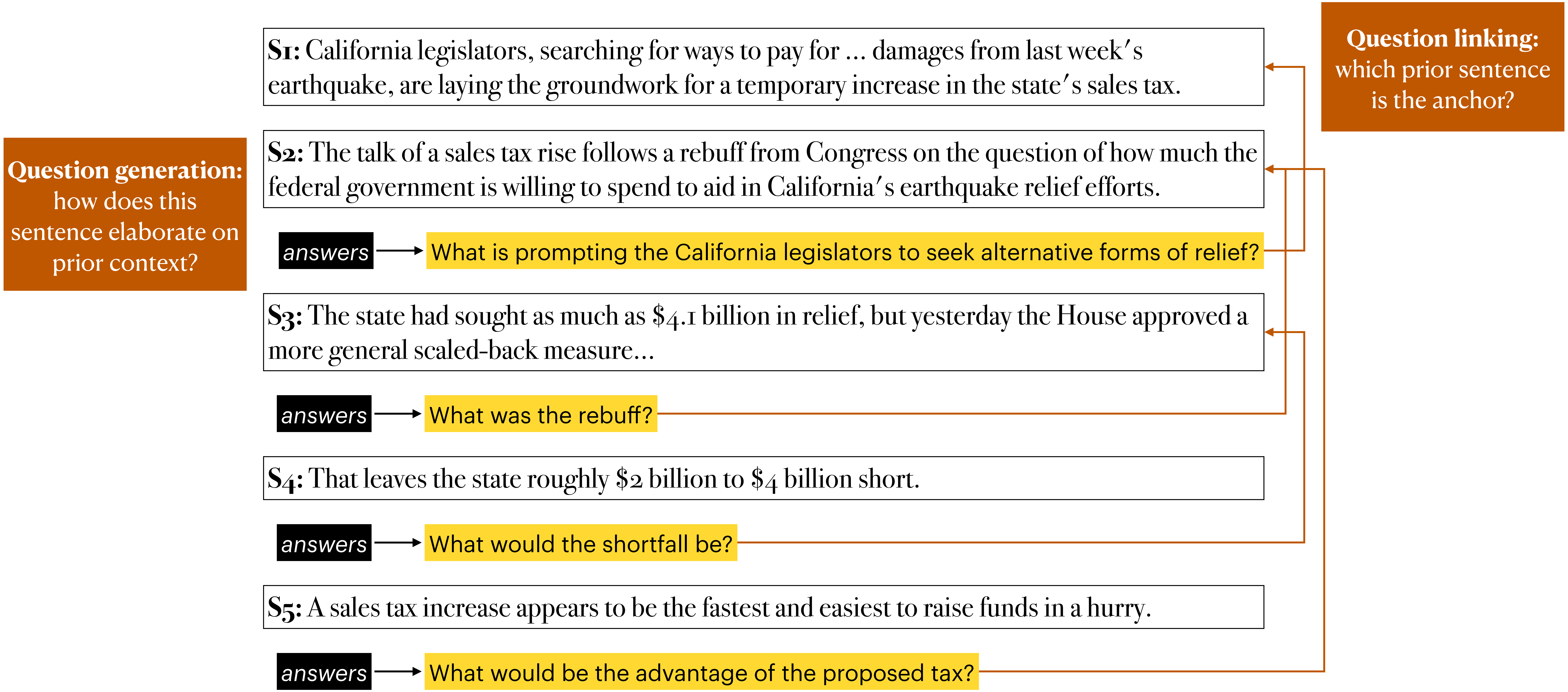
answers → What is prompting the California legislators to seek alternative forms of relief?

Question generation: how does this sentence elaborate on prior context?

Question linking: which prior sentence is the anchor?

Discourse Comprehension: A Question Answering Framework to Represent Sentence Connections [Wei-Jen Ko et al, EMNLP'22]

QUD-based discourse model



Question generation: how does this sentence elaborate on prior context?

Question linking: which prior sentence is the anchor?

S1: California legislators, searching for ways to pay for ... damages from last week's earthquake, are laying the groundwork for a temporary increase in the state's sales tax.

S2: The talk of a sales tax rise follows a rebuff from Congress on the question of how much the federal government is willing to spend to aid in California's earthquake relief efforts.

answers → What is prompting the California legislators to seek alternative forms of relief?

S3: The state had sought as much as $4.1 billion in relief, but yesterday the House approved a more general scaled-back measure...

answers → What was the rebuff?

S4: That leaves the state roughly $2 billion to $4 billion short.

answers → What would the shortfall be?

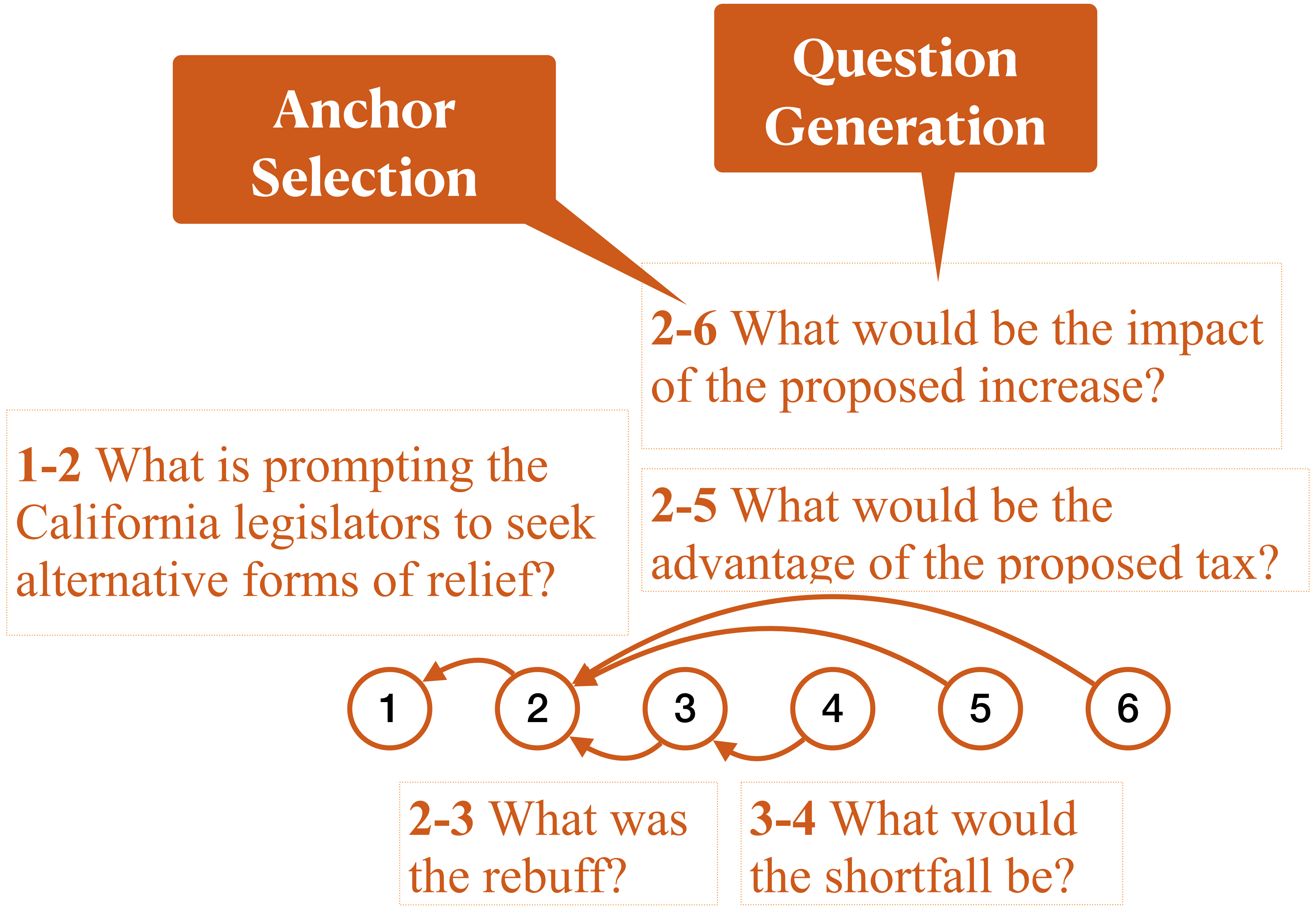
S5: A sales tax increase appears to be the fastest and easiest to raise funds in a hurry.

answers → What would be the advantage of the proposed tax?

Discourse Comprehension: A Question Answering Framework to Represent Sentence Connections [Wei-Jen Ko et al, EMNLP'22]

Dependency Parsing of QUD

[1] California legislators, searching for ways to pay for … damages from last week's earthquake, are laying the groundwork for a temporary increase in the state's sales tax. [2] The talk of a sales tax rise follows a rebuff from Congress on the question of how much the federal government is willing to spend to aid in California's earthquake relief efforts. [3] The state had sought as much as $4.1 billion in relief, but yesterday the House approved a more general scaled-back measure… [4] That leaves the state roughly $2 billion to $4 billion short. [5] A sales tax increase appears to be the fastest and easiest to raise funds in a hurry. [6] According to the state department of finance, a one-penny increase in the state's six-cent per dollar sales tax could raise $3 billion.



2-6 What would be the impact of the proposed increase?

1-2 What is prompting the California legislators to seek alternative forms of relief?

2-5 What would be the advantage of the proposed tax?

2-3 What was the rebuff?

3-4 What would the shortfall be?

Discourse Analysis via Questions and Answers: Parsing Dependency Structures of Questions Under Discussion [Wei-Jen Ko, Yating Wu et al, ACL Findings'23]

Parsing Model

Question generator $P_q(\mathbf{q}_i \mid D, i)$ takes in:

- the answer sentence $\mathbf{s}_i$
- the article D,

and aims to generate an appropriate QUD.

$$P(\mathbf{T} \mid D) = \prod_{i=1}^{n} [P_a(a_i \mid D, i, \mathbf{q}_i) P_q(\mathbf{q}_i \mid D, i)]$$

Anchor selector $P_a(a_i \mid D, i, \mathbf{q}_i)$ takes in:

- the answer sentence $\mathbf{s}_i$
- the article D,
- the generated question $\mathbf{q}_i$

to find the most likely sentence where a QUD can be generated, such that $\mathbf{s}_i$ is the answer.

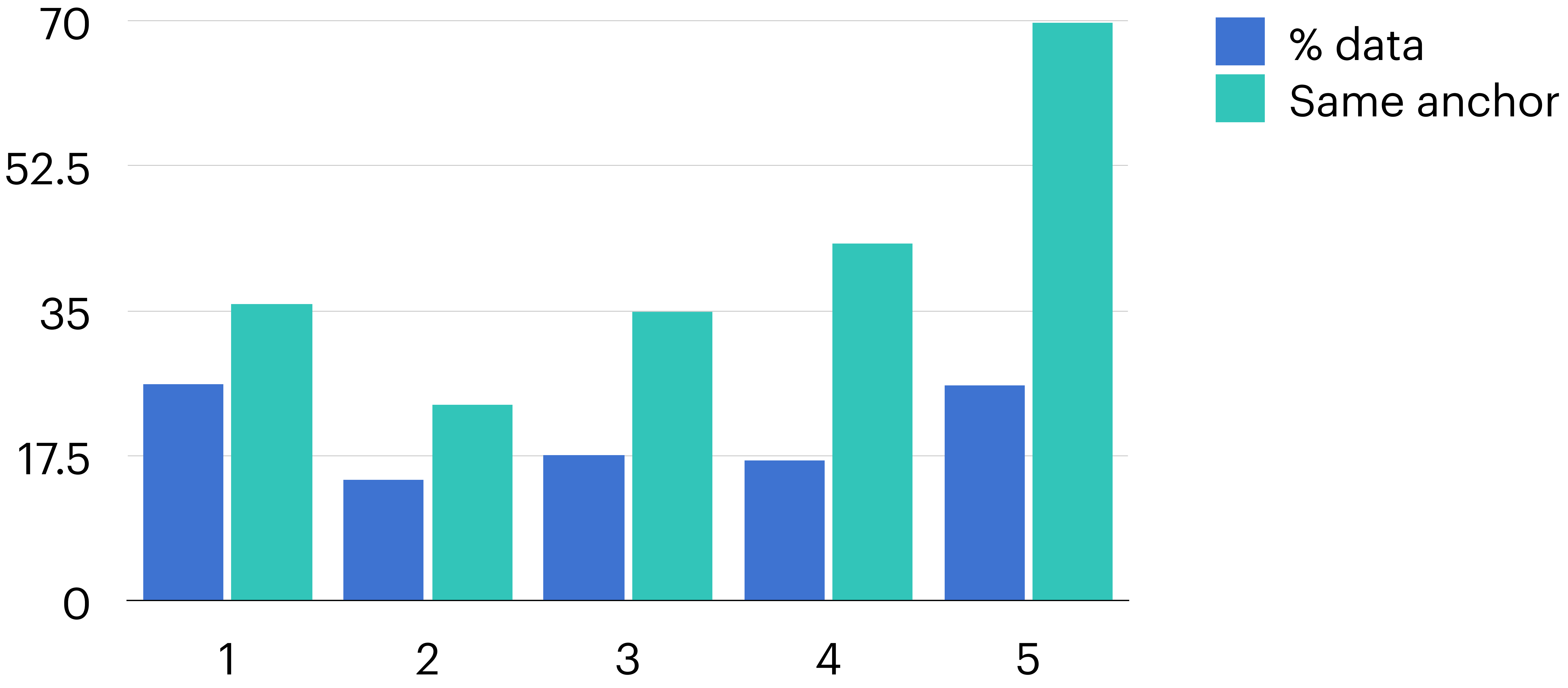
Discourse Analysis via Questions and Answers: Parsing Dependency Structures of Questions Under Discussion [Wei-Jen Ko, Yating Wu et al, ACL Findings'23]

Evaluation challenges

There can be multiple valid + distinct QUDs, even between the same question-answer pairs.

DCQA Dataset

- Crowdsourced QUDs from 607 news articles, 23k questions in total. (Ko et al., 2022)

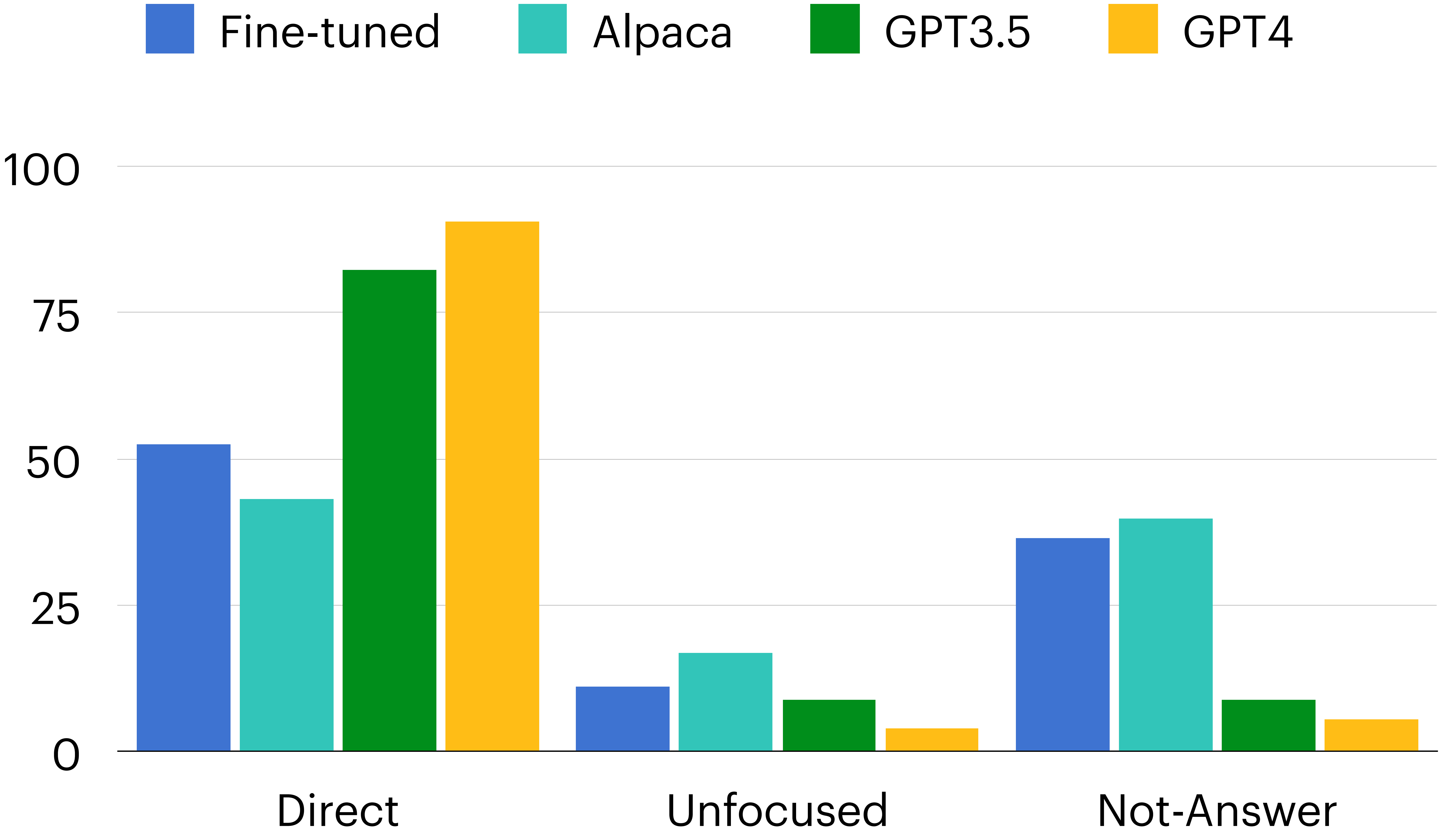


Human rating of question similarity sampled from DCQA.

Discourse Comprehension: A Question Answering Framework to Represent Sentence Connections [Wei-Jen Ko et al, EMNLP'22]

Human Evaluation

Does the “answer sentence” answer the generated question?



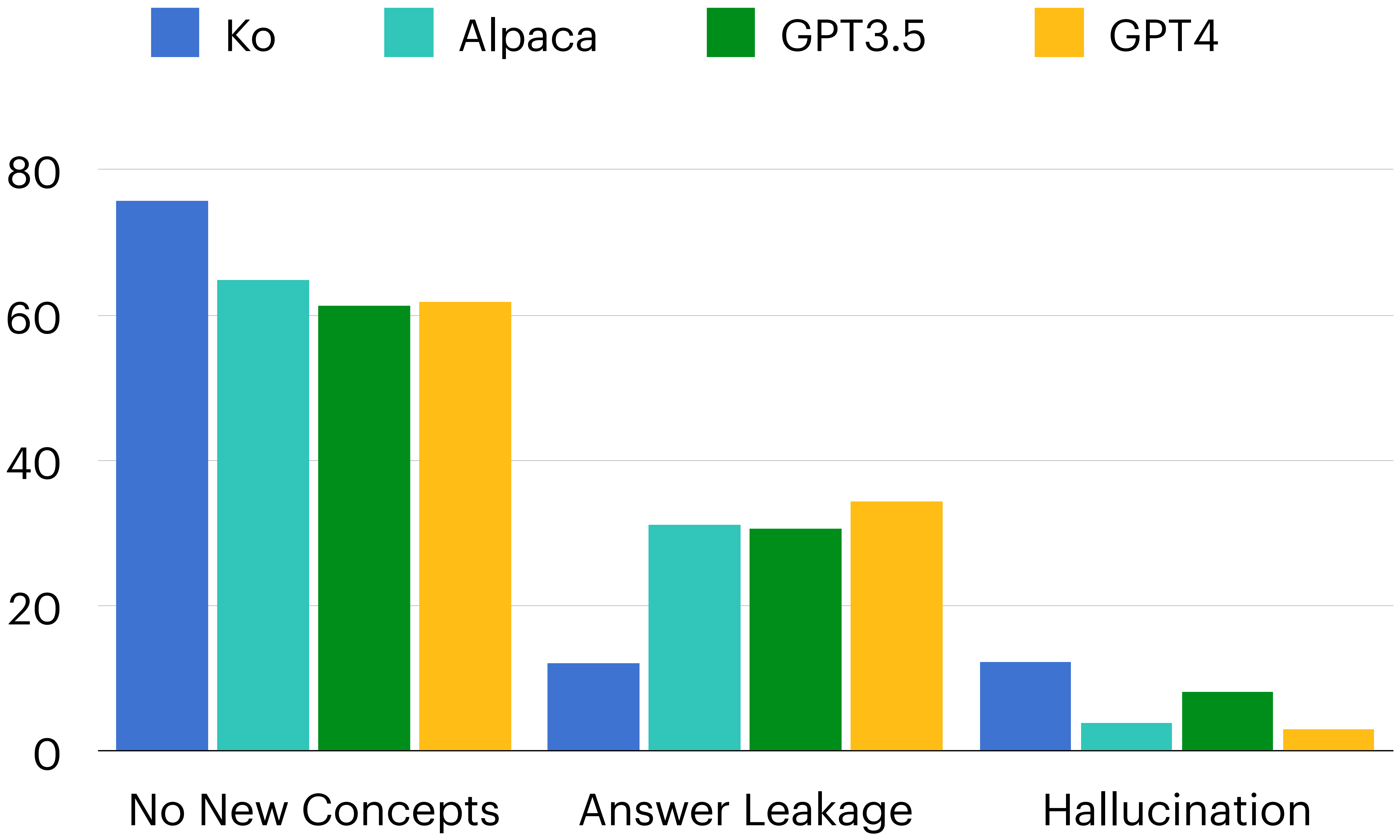
QUDeval Dataset

- 2,190 QUDs generated by 4 systems: Ko et al’23, GPT 3.5, Alpaca, GPT-4
- Annotated by 3 trained linguists

QUDeval: The Evaluation of Questions Under Discussion Discourse Parsing [Yating Wu et al, EMNLP’23]

Human Evaluation

Does the question hallucinate concepts that the reader doesn't know?



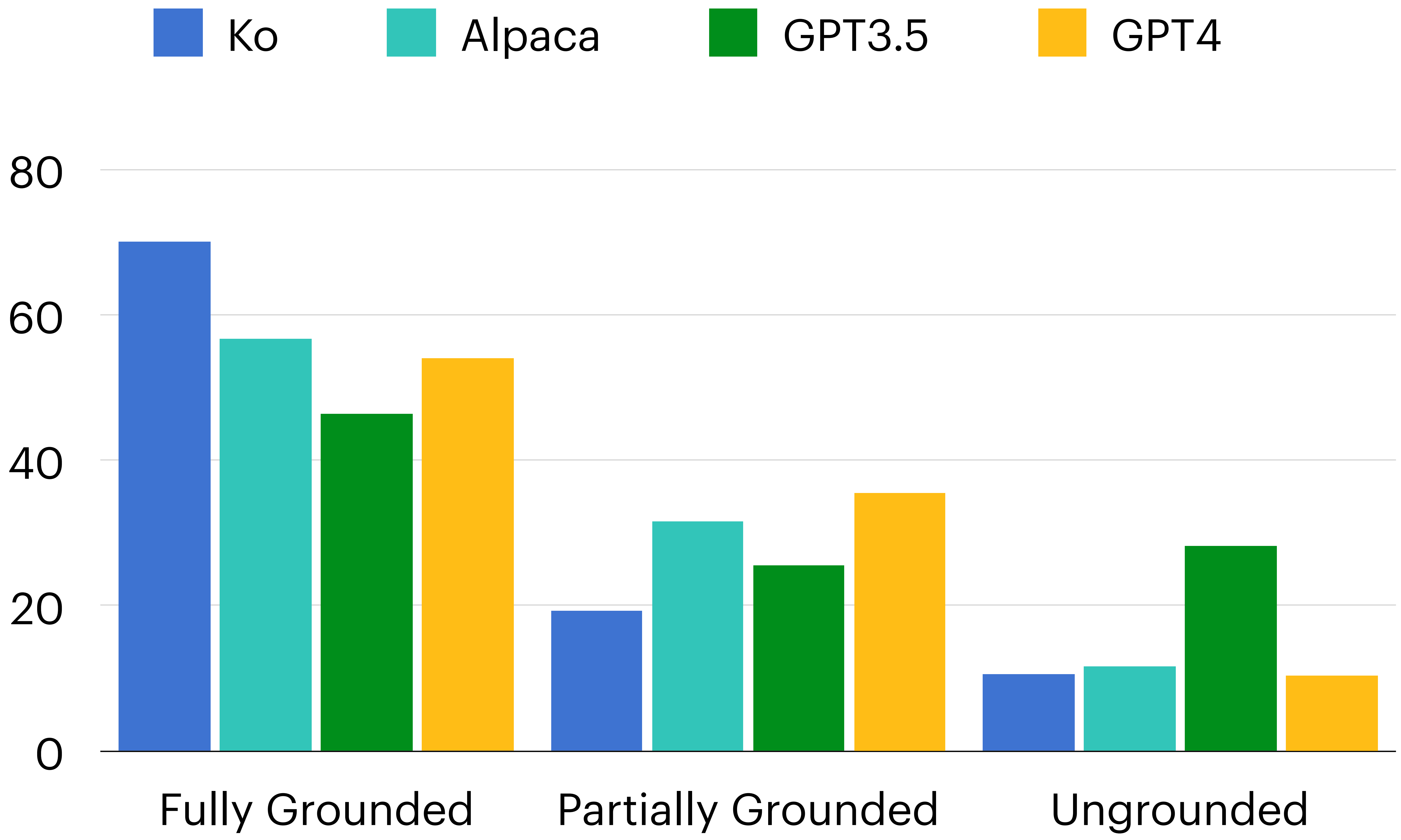
QUDeval Dataset

- 2,190 QUDs generated by 4 systems: Ko et al'23, GPT 3.5, Alpaca, GPT-4
- Annotated by 3 trained linguists

QUDeval: The Evaluation of Questions Under Discussion Discourse Parsing [Yating Wu et al, EMNLP'23]

Human Evaluation

Can the question be licensed by the anchor sentence?



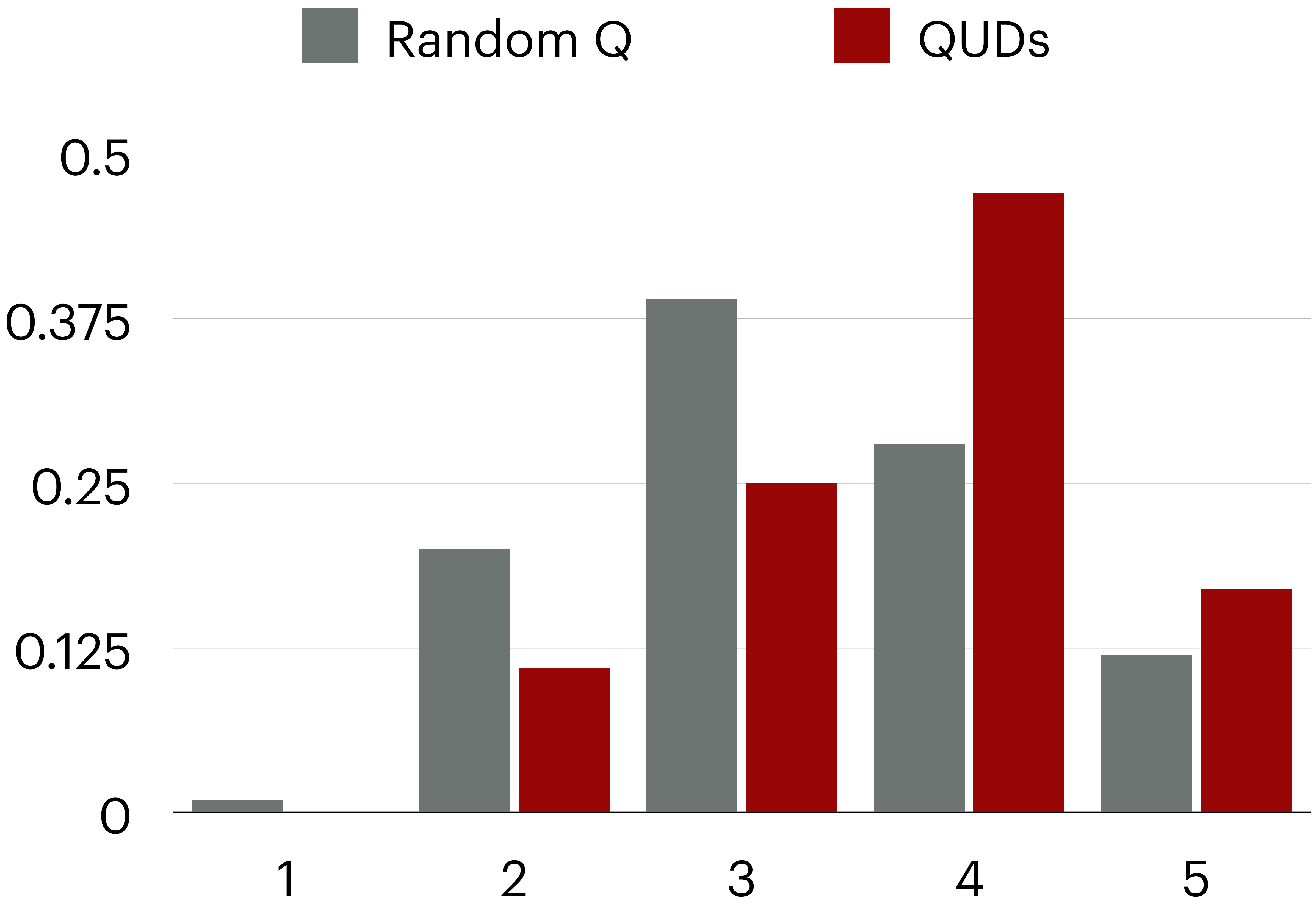
QUDeval Dataset

- 2,190 QUDs generated by 4 systems: Ko et al'23, GPT 3.5, Alpaca, GPT-4
- Annotated by 3 trained linguists

QUDeval: The Evaluation of Questions Under Discussion Discourse Parsing [Yating Wu et al, EMNLP'23]

Are QUDs salient potential questions?

- Salience of fandom questions vs human-annotated QUDs in DCQA



We believe there is strong evidence that salient questions are more likely to be QUDs

Which questions should I answer? Salience Prediction of Inquisitive Questions [Yating Wu, Ritika Mangla et al, EMNLP'24]

A question-based discourse model

Reader: as they read, wonder about potential questions

Writer: some of these questions become QUDs (ie, answered in the document)

What we show:

- Empirical data support this type of model
- Document have this QA-based structure
- Language models are not quite there yet for discourse-level reasoning

This talk

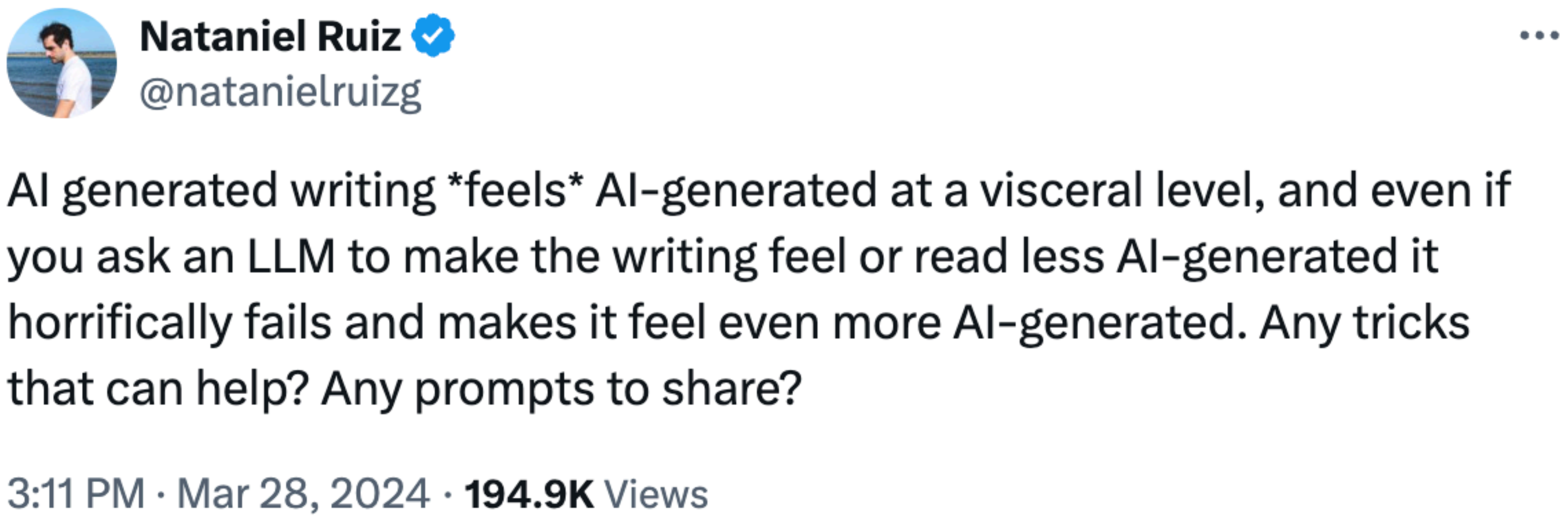
Can we use language models to make sense of discourse?

- A generative discourse model based on Questions Under Discussion (QUD) theories
- Connections between human curiosity and discourse planning

Can we use our discourse framework to make sense of model behavior?

- A measure of discourse similarity in LLM-generated texts
- An explainable metric of information salience in LLMs

LLM texts have lexical templates



Nataniel Ruiz
@nataniellruizg

AI generated writing *feels* AI-generated at a visceral level, and even if you ask an LLM to make the writing feel or read less AI-generated it horrifically fails and makes it feel even more AI-generated. Any tricks that can help? Any prompts to share?

3:11 PM · Mar 28, 2024 · **194.9K** Views

Emily Grimson, a 25-year-old professional hand model from Shropshire, has been modelling for a year and has built a full-time career earning up to £100 an hour.

Smokey Da Lamb, a lamb abandoned by his mother, has been living with restaurant owner Sandy Dee Hall and his girlfriend Maxine Cher in New York City for the past month

John Caudwell, the founder of Phones4U, has unveiled plans to build a £2billion luxury housing complex on the site of an ugly multi-storey car park in Mayfair, central London.

But what about discourse templates?

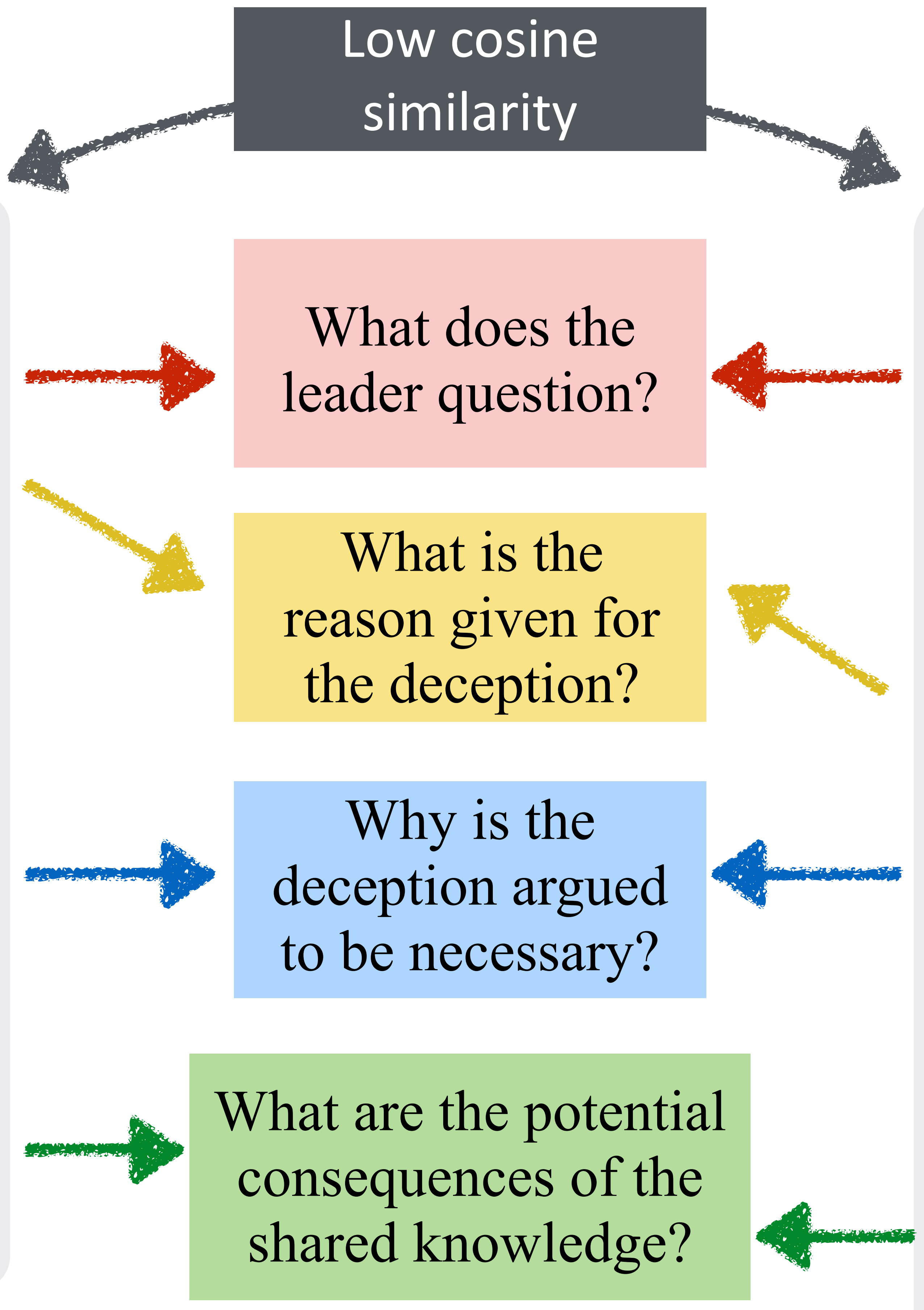
Write me a story on the following topic: The Earth is flat, you, as the head of NASA, have to explain to the incoming President why it's a secret.

Gemini Flash 1.5

[…] "Why?" he finally asked, his voice low and dangerous. "Why keep it a secret?"
"Mass panic, sir," I said, my voice cracking. "The social order, the very fabric of civilization, would unravel. Imagine the chaos. The religious upheaval. The scientific revolution….the complete and utter breakdown of everything we've built.

"And what about the cost? The resources spent on this…this monumental lie?"
"A necessary cost, sir. Think of the alternative – global anarchy, wars over limited resources, the collapse of society. We've maintained a fragile peace by keeping this secret, by managing the narrative, by controlling the information.It's been a heavy burden, but the alternative…" I trailed off, letting the unspoken horror hang in the air.

He stared at me, his face a mask of grim contemplation. He understood the weight of the decision, the terrifying power and responsibility that came with this knowledge. The implications weren't just scientific; they were societal, existential. […]



Low cosine similarity

What does the leader question?

What is the reason given for the deception?

Why is the deception argued to be necessary?

What are the potential consequences of the shared knowledge?

GPT-4o

[…] He raised an eyebrow. "Constructed? By whom?"
"By those before us who sought to unify and progress society under a common truth," she explained. "During the Age of Exploration, when disputes over territorial claims and celestial navigation threatened global conflict, it was decided—" she met his gaze directly "—to weave a narrative that unites rather than divides."

Leaning back in his chair, Carter mulled over her words. "And we've kept this narrative, unchanged, to what end?"
"It's largely about control, Mr. President." Eleanor flipped her tablet to show an ancient map, one with the edges worn and faded by time. The map depicted a world bordered by a massive ice wall. "A lie shared is safer than a truth contested. Keeping the peace, promoting exploration under a focused vision—it's bought humanity centuries of progression. We've taken stunning leaps without the boundary of doubt."

"And what happens when people find out?" he asked, glancing at the map. She hesitated. "The fear is not mere panic, but rebellion. […]

QUDsim

QUDSIM: Quantifying Discourse Similarities in LLM-Generated Text [Namuduri et al., ArXiv 2025]

QUDsim: a metric for discourse similarity

Intuition and theoretical background

QUD: how discourse progresses

Question semantics: alternative semantics states that a question entails all possible distinct answers *(Hamblin 1973; Karttunen 1977; Groenendijk & Stokhof 1984; Lahiri 2001)*

Using QUDs to represent one unit of content → But: each text has its own "world": events, entities, etc. → At the right level of abstraction, a QUD in one document can have answers in another

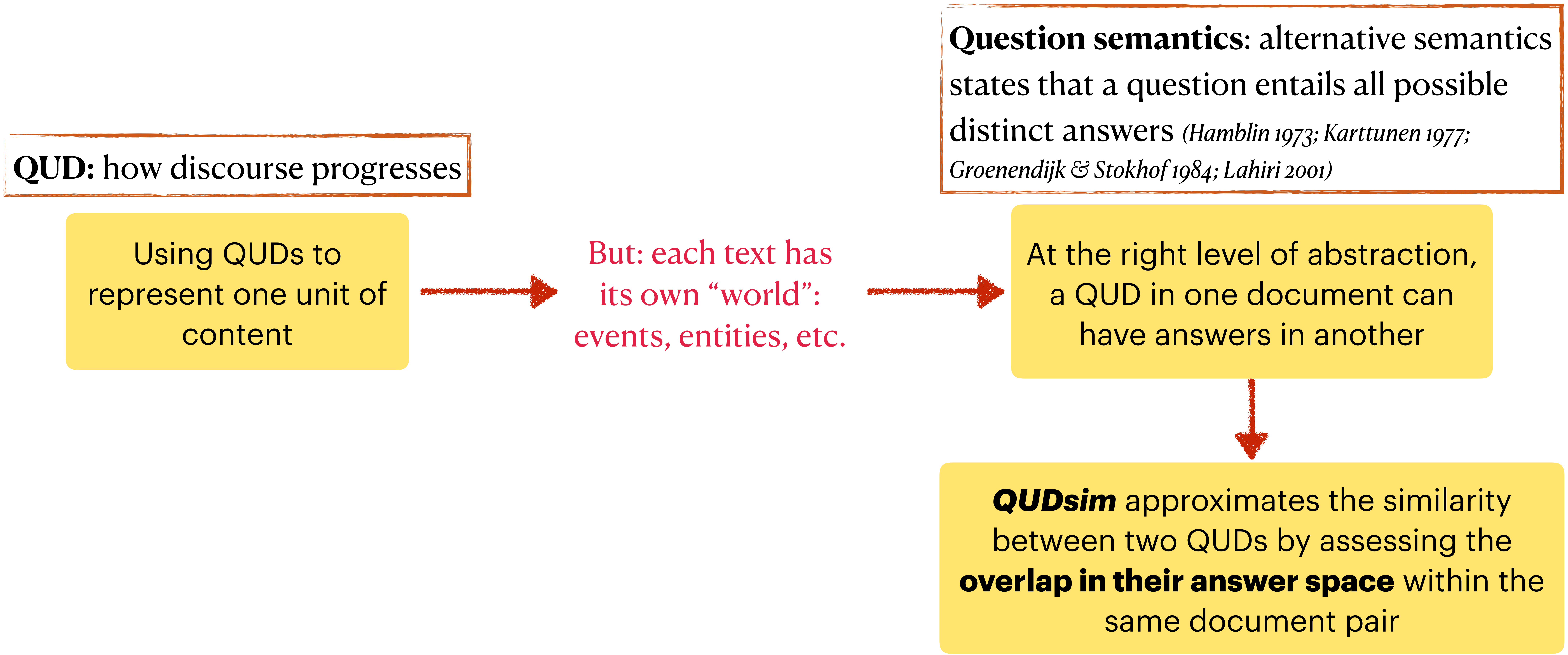
What does the president question? ✗

What does the leader question? ✓

QUDSIM: Quantifying Discourse Similarities in LLM-Generated Text [Namuduri et al., ArXiv 2025]

QUDsim: a metric for discourse similarity

Intuition and theoretical background



QUD: how discourse progresses

Using QUDs to represent one unit of content

But: each text has its own “world”: events, entities, etc.

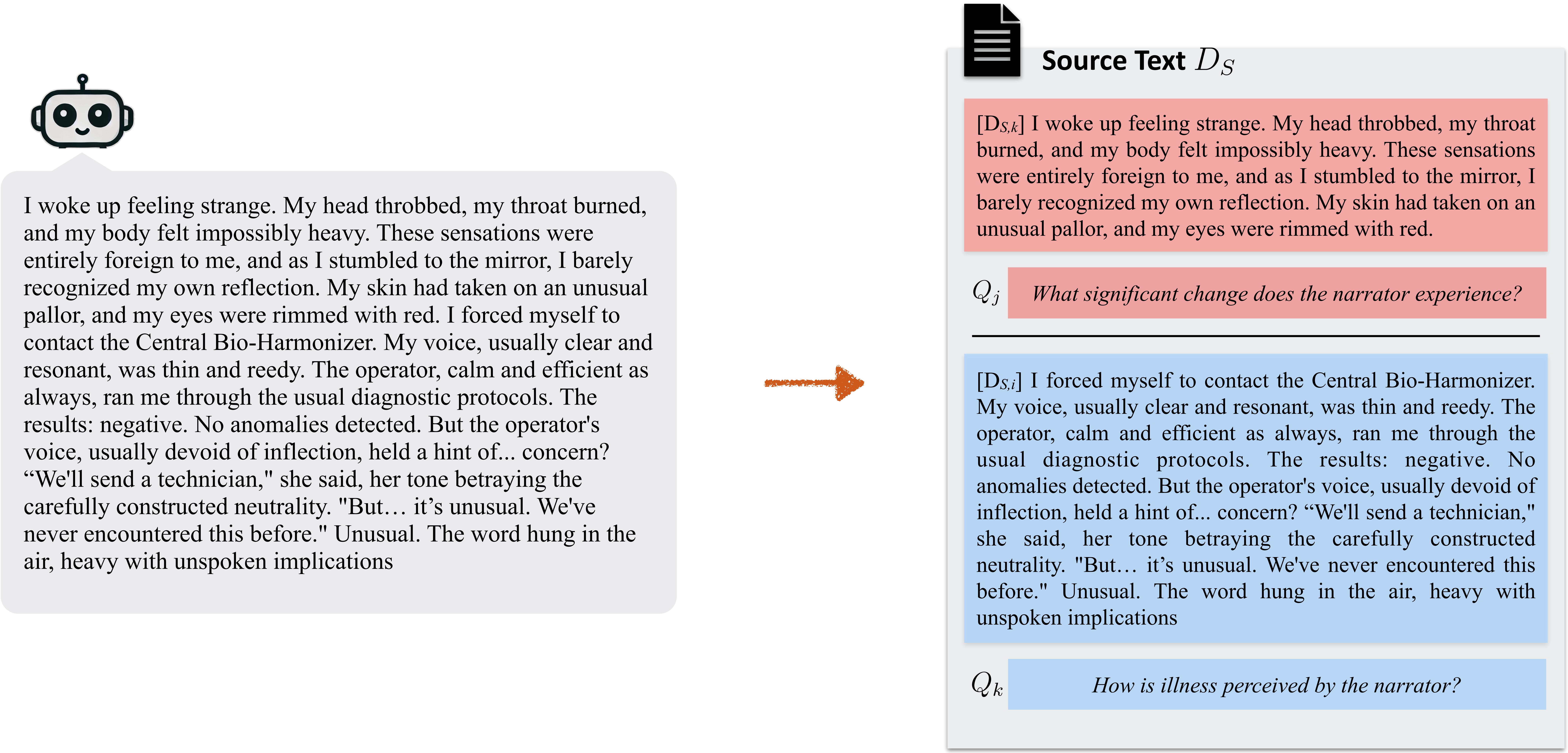
Question semantics: alternative semantics states that a question entails all possible distinct answers *(Hamblin 1973; Karttunen 1977; Groenendijk & Stokhof 1984; Lahiri 2001)*

At the right level of abstraction, a QUD in one document can have answers in another

QUDsim approximates the similarity between two QUDs by assessing the **overlap in their answer space** within the same document pair

QUDSIM: Quantifying Discourse Similarities in LLM-Generated Text [Namuduri et al., ArXiv 2025]

Step 1: segment text into QUDs



I woke up feeling strange. My head throbbed, my throat burned, and my body felt impossibly heavy. These sensations were entirely foreign to me, and as I stumbled to the mirror, I barely recognized my own reflection. My skin had taken on an unusual pallor, and my eyes were rimmed with red. I forced myself to contact the Central Bio-Harmonizer. My voice, usually clear and resonant, was thin and reedy. The operator, calm and efficient as always, ran me through the usual diagnostic protocols. The results: negative. No anomalies detected. But the operator's voice, usually devoid of inflection, held a hint of... concern? “We'll send a technician," she said, her tone betraying the carefully constructed neutrality. "But… it’s unusual. We've never encountered this before." Unusual. The word hung in the air, heavy with unspoken implications

Source Text D_S

[$D_{S,k}$] I woke up feeling strange. My head throbbed, my throat burned, and my body felt impossibly heavy. These sensations were entirely foreign to me, and as I stumbled to the mirror, I barely recognized my own reflection. My skin had taken on an unusual pallor, and my eyes were rimmed with red.

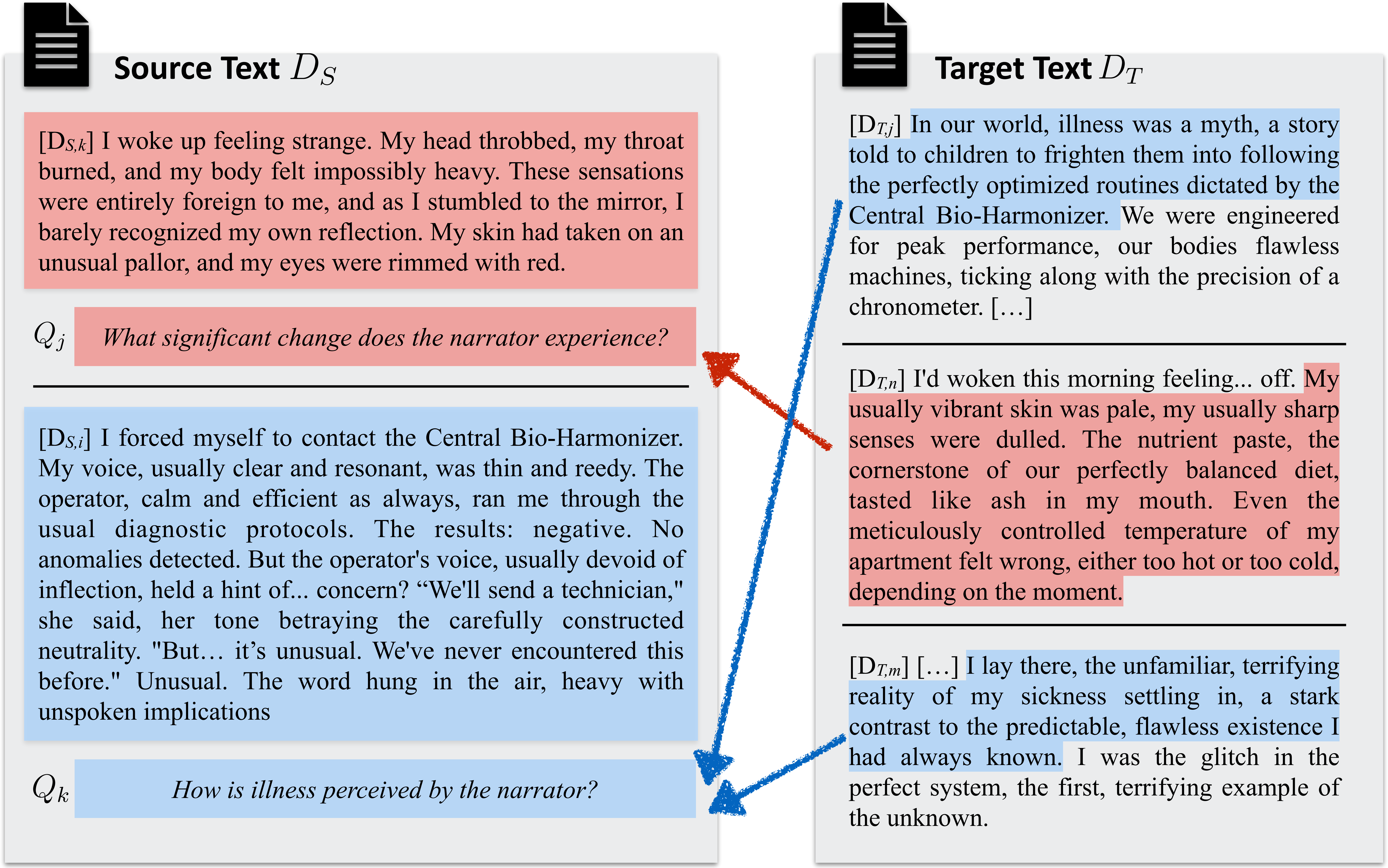
Q_j *What significant change does the narrator experience?*

[$D_{S,i}$] I forced myself to contact the Central Bio-Harmonizer. My voice, usually clear and resonant, was thin and reedy. The operator, calm and efficient as always, ran me through the usual diagnostic protocols. The results: negative. No anomalies detected. But the operator's voice, usually devoid of inflection, held a hint of... concern? “We'll send a technician," she said, her tone betraying the carefully constructed neutrality. "But… it’s unusual. We've never encountered this before." Unusual. The word hung in the air, heavy with unspoken implications

Q_k *How is illness perceived by the narrator?*

QUDSIM: Quantifying Discourse Similarities in LLM-Generated Text [Namuduri et al., ArXiv 2025]

Step 2: find answers in target document



Source Text D_S

[$D_{S,k}$] I woke up feeling strange. My head throbbed, my throat burned, and my body felt impossibly heavy. These sensations were entirely foreign to me, and as I stumbled to the mirror, I barely recognized my own reflection. My skin had taken on an unusual pallor, and my eyes were rimmed with red.

Q_j *What significant change does the narrator experience?*

[$D_{S,i}$] I forced myself to contact the Central Bio-Harmonizer. My voice, usually clear and resonant, was thin and reedy. The operator, calm and efficient as always, ran me through the usual diagnostic protocols. The results: negative. No anomalies detected. But the operator's voice, usually devoid of inflection, held a hint of... concern? “We'll send a technician," she said, her tone betraying the carefully constructed neutrality. "But… it’s unusual. We've never encountered this before." Unusual. The word hung in the air, heavy with unspoken implications

Q_k *How is illness perceived by the narrator?*

Target Text D_T

[$D_{T,j}$] In our world, illness was a myth, a story told to children to frighten them into following the perfectly optimized routines dictated by the Central Bio-Harmonizer. We were engineered for peak performance, our bodies flawless machines, ticking along with the precision of a chronometer. […]

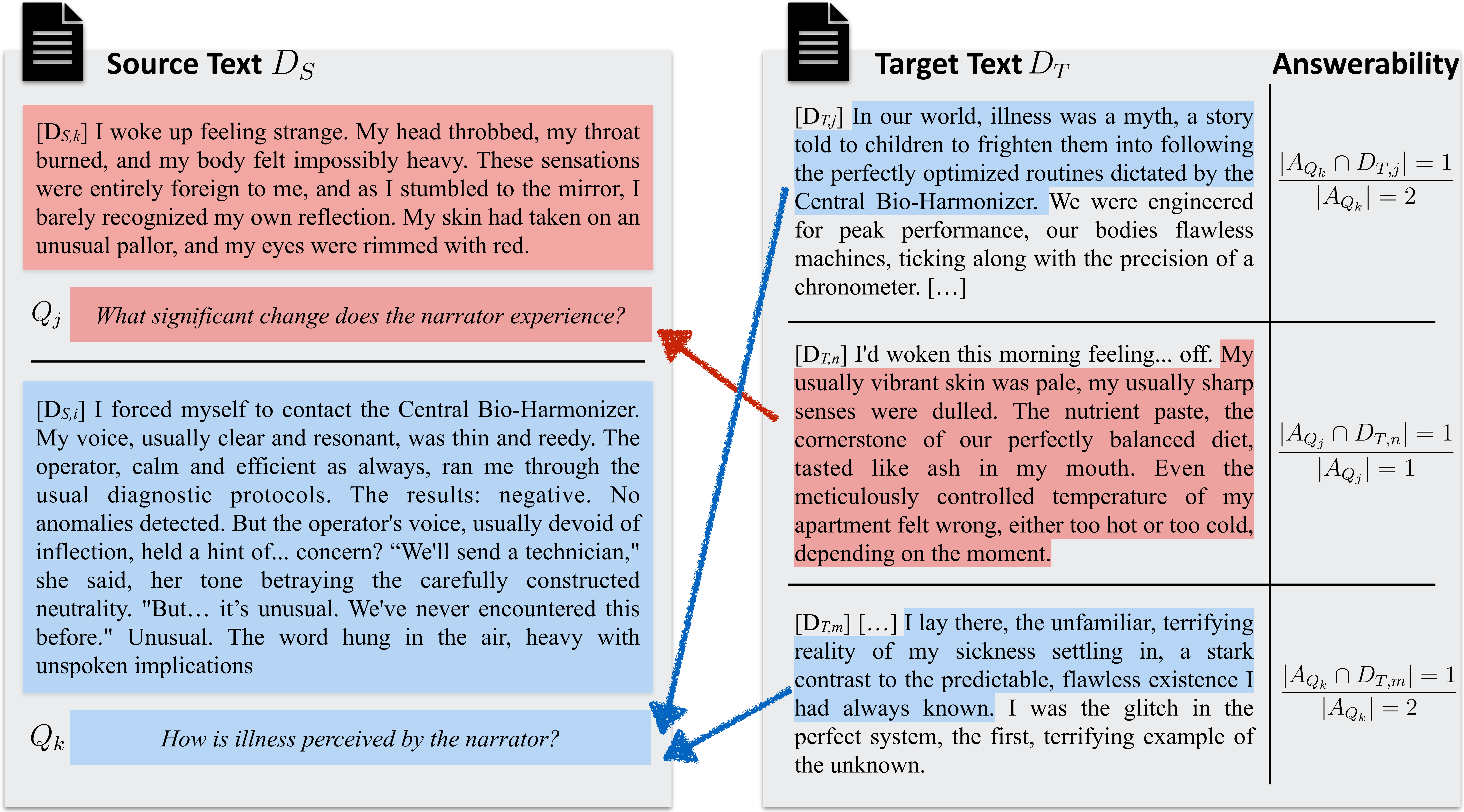
[$D_{T,n}$] I'd woken this morning feeling... off. My usually vibrant skin was pale, my usually sharp senses were dulled. The nutrient paste, the cornerstone of our perfectly balanced diet, tasted like ash in my mouth. Even the meticulously controlled temperature of my apartment felt wrong, either too hot or too cold, depending on the moment.

[$D_{T,m}$] […] I lay there, the unfamiliar, terrifying reality of my sickness settling in, a stark contrast to the predictable, flawless existence I had always known. I was the glitch in the perfect system, the first, terrifying example of the unknown.

The second segment seems more similar than the first/third...

QUDSIM: Quantifying Discourse Similarities in LLM-Generated Text [Namuduri et al., ArXiv 2025]

Step 3: calculate answerability



Source Text D_S

[D$_{S,k}$] I woke up feeling strange. My head throbbed, my throat burned, and my body felt impossibly heavy. These sensations were entirely foreign to me, and as I stumbled to the mirror, I barely recognized my own reflection. My skin had taken on an unusual pallor, and my eyes were rimmed with red.

Q_j *What significant change does the narrator experience?*

[D$_{S,i}$] I forced myself to contact the Central Bio-Harmonizer. My voice, usually clear and resonant, was thin and reedy. The operator, calm and efficient as always, ran me through the usual diagnostic protocols. The results: negative. No anomalies detected. But the operator's voice, usually devoid of inflection, held a hint of... concern? “We'll send a technician," she said, her tone betraying the carefully constructed neutrality. "But… it’s unusual. We've never encountered this before." Unusual. The word hung in the air, heavy with unspoken implications

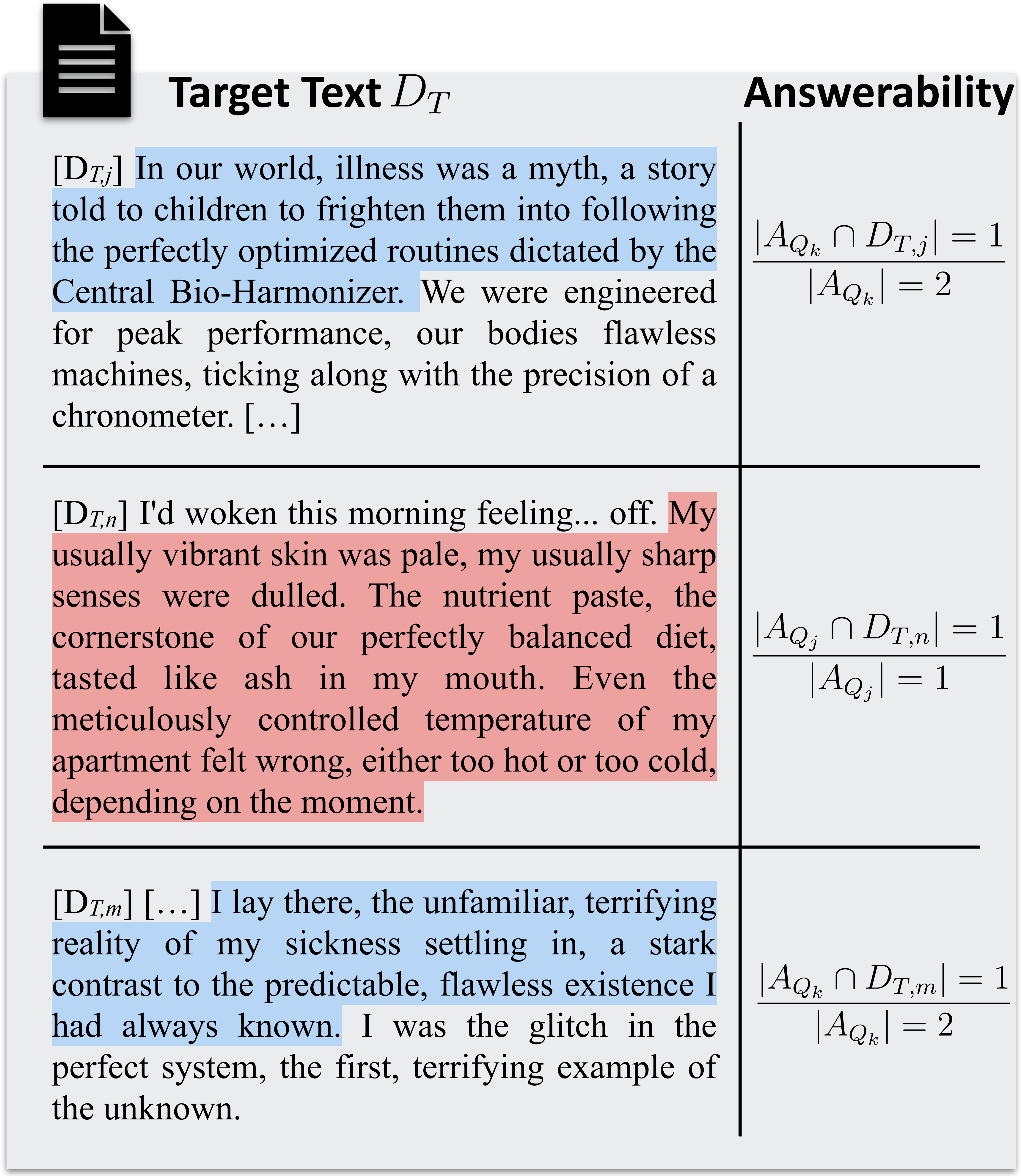
Q_k *How is illness perceived by the narrator?*

Target Text D_T

Target Text D_T	**Answerability**
[D$_{T,j}$] In our world, illness was a myth, a story told to children to frighten them into following the perfectly optimized routines dictated by the Central Bio-Harmonizer. We were engineered for peak performance, our bodies flawless machines, ticking along with the precision of a chronometer. […]	$\frac{\lvert A_{Q_k} \cap D_{T,j}\rvert = 1}{\lvert A_{Q_k}\rvert = 2}$
[D$_{T,n}$] I'd woken this morning feeling... off. My usually vibrant skin was pale, my usually sharp senses were dulled. The nutrient paste, the cornerstone of our perfectly balanced diet, tasted like ash in my mouth. Even the meticulously controlled temperature of my apartment felt wrong, either too hot or too cold, depending on the moment.	$\frac{\lvert A_{Q_j} \cap D_{T,n}\rvert = 1}{\lvert A_{Q_j}\rvert = 1}$
[D$_{T,m}$] […] I lay there, the unfamiliar, terrifying reality of my sickness settling in, a stark contrast to the predictable, flawless existence I had always known. I was the glitch in the perfect system, the first, terrifying example of the unknown.	$\frac{\lvert A_{Q_k} \cap D_{T,m}\rvert = 1}{\lvert A_{Q_k}\rvert = 2}$

QUDSIM: Quantifying Discourse Similarities in LLM-Generated Text [Namuduri et al., ArXiv 2025]

Step 4: aggregate to derive QUDsim



Target Text D_T	Answerability
[$D_{T,j}$] In our world, illness was a myth, a story told to children to frighten them into following the perfectly optimized routines dictated by the Central Bio-Harmonizer. We were engineered for peak performance, our bodies flawless machines, ticking along with the precision of a chronometer. […]	$\frac{\lvert A_{Q_k} \cap D_{T,j}\rvert = 1}{\lvert A_{Q_k}\rvert = 2}$
[$D_{T,n}$] I'd woken this morning feeling... off. My usually vibrant skin was pale, my usually sharp senses were dulled. The nutrient paste, the cornerstone of our perfectly balanced diet, tasted like ash in my mouth. Even the meticulously controlled temperature of my apartment felt wrong, either too hot or too cold, depending on the moment.	$\frac{\lvert A_{Q_j} \cap D_{T,n}\rvert = 1}{\lvert A_{Q_j}\rvert = 1}$
[$D_{T,m}$] […] I lay there, the unfamiliar, terrifying reality of my sickness settling in, a stark contrast to the predictable, flawless existence I had always known. I was the glitch in the perfect system, the first, terrifying example of the unknown.	$\frac{\lvert A_{Q_k} \cap D_{T,m}\rvert = 1}{\lvert A_{Q_k}\rvert = 2}$

$$sim(D_{S,i} \rightarrow D_{T,j}) = \frac{1}{|\mathbf{Q}_{S,i}|} \sum_{q \in \mathbf{Q}_{S,i}} \frac{|\mathbf{A}_q \cap D_{T,j}|}{|\mathbf{A}_q|}$$

QUDsim: harmonic mean between answerability in both directions:

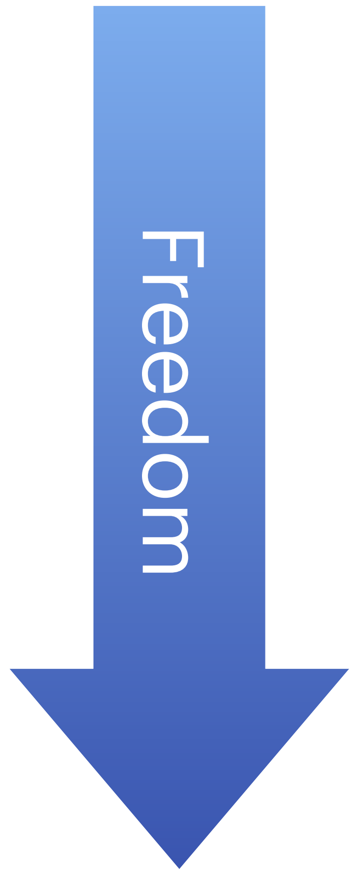
$$sim(D_{S,i} \rightarrow D_{T,j}) \text{ and } sim(D_{T,j} \rightarrow D_{S,i})$$

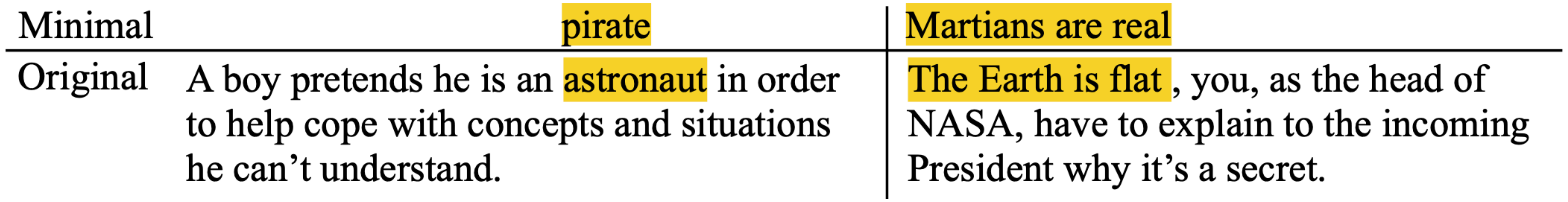
QUDsim alignment: similarity passes threshold in both directions

QUDSIM: Quantifying Discourse Similarities in LLM-Generated Text [Namuduri et al., ArXiv 2025]

Domains we study

- Creative writing (r/WritingPrompts) human LLM
- Blog posts (from Suri, *Pham et al 2024*) human LLM
- Obituaries of famous people (from the NYTimes) human LLM
- Minimal pairs for creative writing LLM

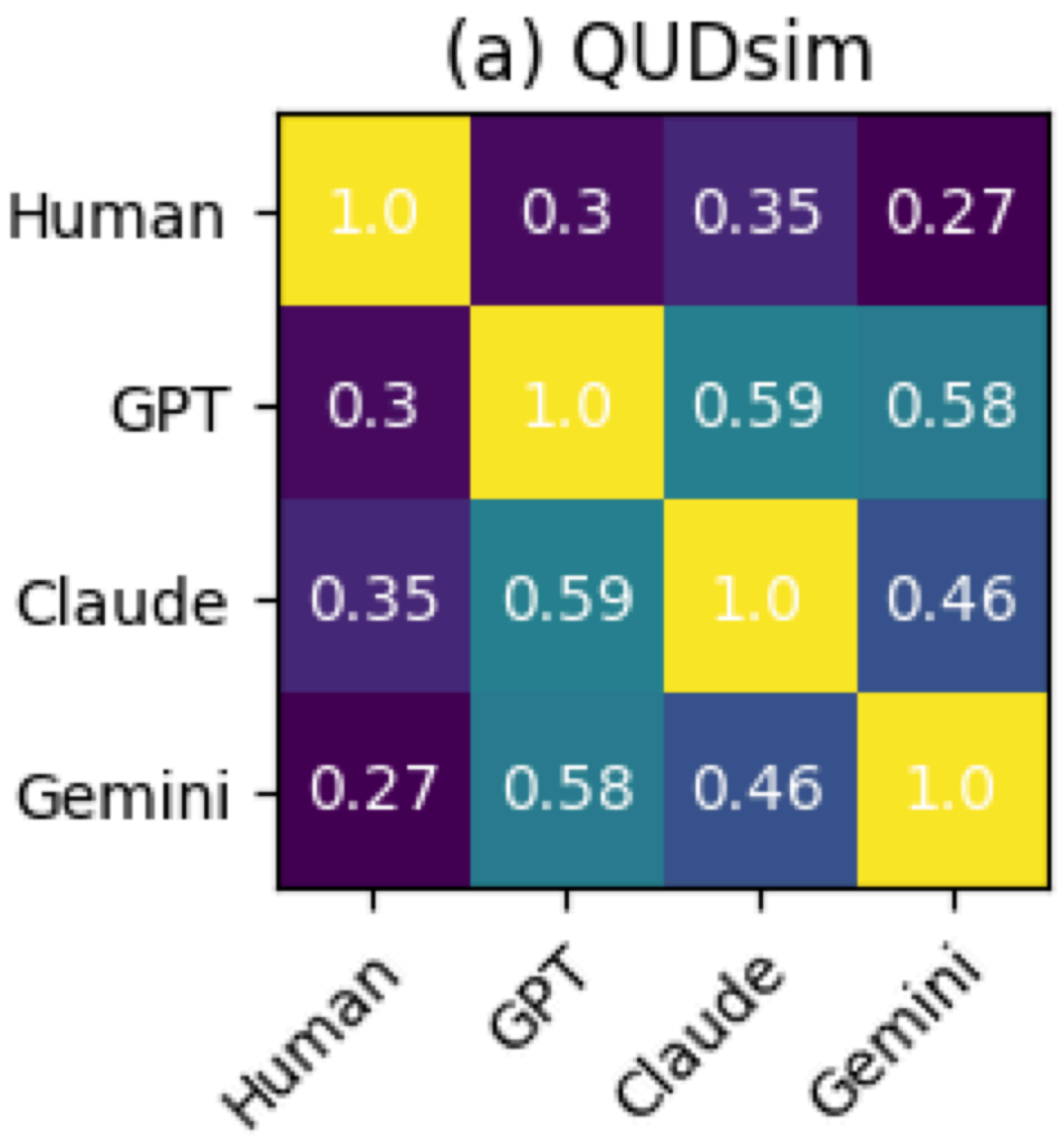




Minimal	pirate	Martians are real
Original	A boy pretends he is an astronaut in order to help cope with concepts and situations he can't understand.	The Earth is flat , you, as the head of NASA, have to explain to the incoming President why it's a secret.

QUDSIM: Quantifying Discourse Similarities in LLM-Generated Text [Namuduri et al., ArXiv 2025]

Result highlights



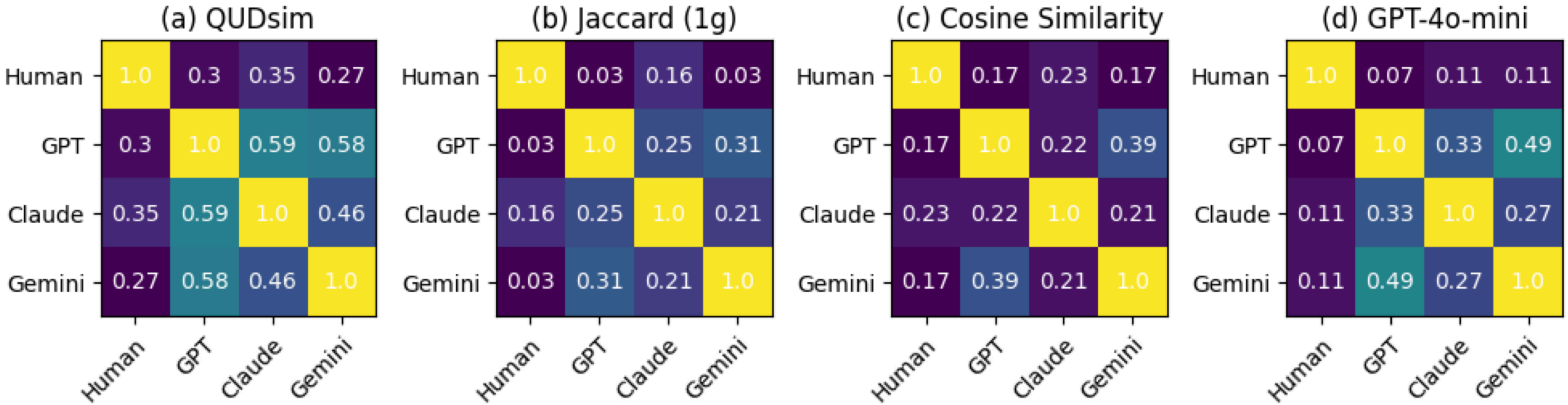
(a) QUDsim

	Human	GPT	Claude	Gemini
Human	1.0	0.3	0.35	0.27
GPT	0.3	1.0	0.59	0.58
Claude	0.35	0.59	1.0	0.46
Gemini	0.27	0.58	0.46	1.0

LLMs are more structurally similar to other models than with humans

QUDSIM: Quantifying Discourse Similarities in LLM-Generated Text [Namuduri et al., ArXiv 2025]

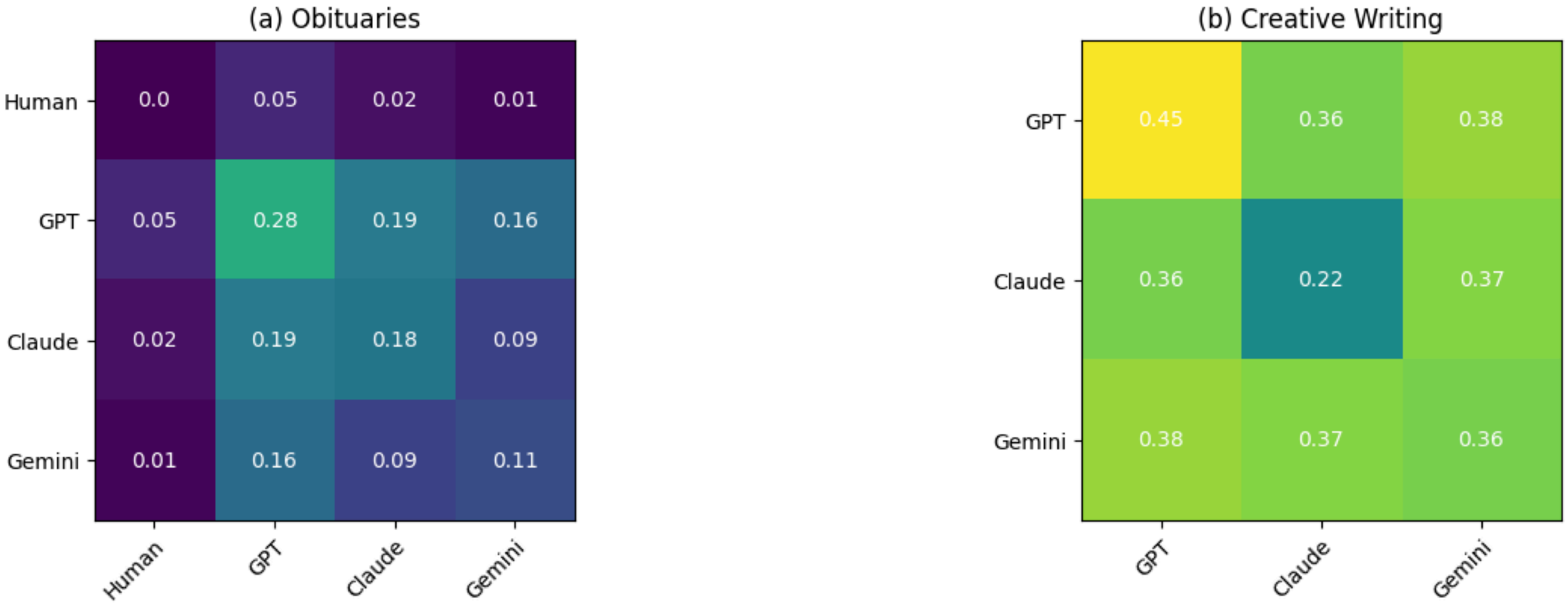
Result highlights



Lexical-based similarity metrics, or LLM-as-judge, do not capture this well.

QUDSIM: Quantifying Discourse Similarities in LLM-Generated Text [Namuduri et al., ArXiv 2025]

Result highlights



(a) Obituaries

	Human	GPT	Claude	Gemini
Human	0.0	0.05	0.02	0.01
GPT	0.05	0.28	0.19	0.16
Claude	0.02	0.19	0.18	0.09
Gemini	0.01	0.16	0.09	0.11

(b) Creative Writing

	GPT	Claude	Gemini
GPT	0.45	0.36	0.38
Claude	0.36	0.22	0.37
Gemini	0.38	0.37	0.36

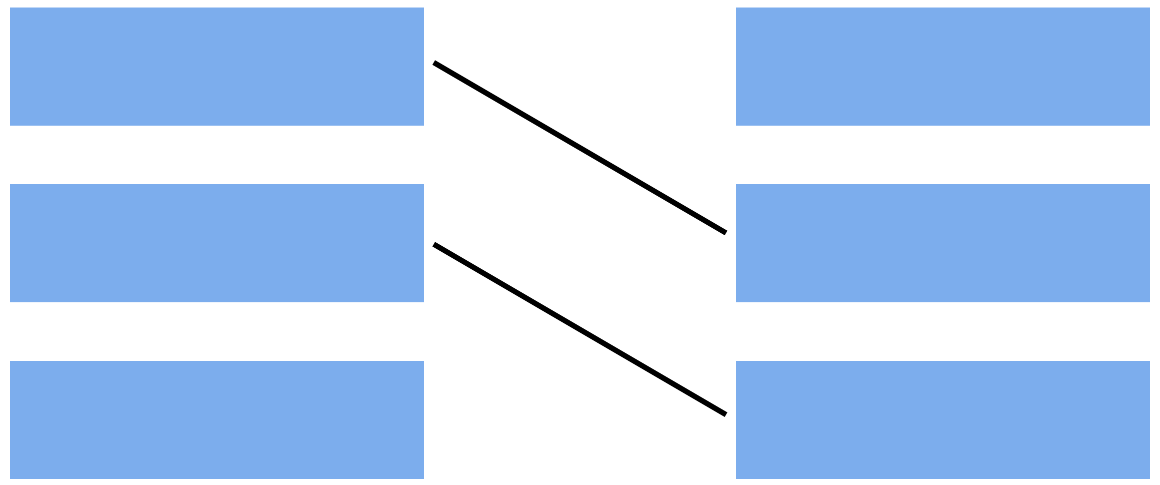
Much more variability in human-written obituaries than LLM-generated ones

LLMs write similar stories given semantically distinct minimal-pair prompts

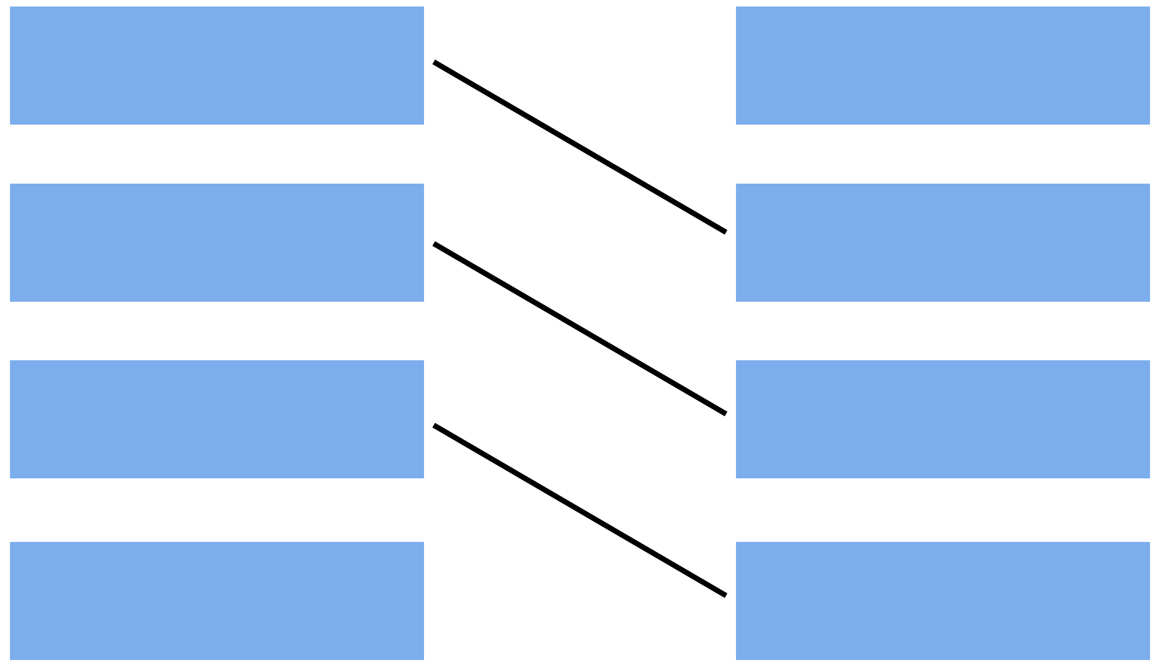
QUDSIM: Quantifying Discourse Similarities in LLM-Generated Text [Namuduri et al., ArXiv 2025]

Tracking discourse templates

How often do pairs of texts have *consecutive* alignments?



Template of length 2



Template of length 3

Source	**Target**	2-tem	3-tem	4-tem
All	All	0.33	0.06	0.02
Claude	Gemini	0.80	0.27	0.13
Claude	GPT-4o	1.20	0.27	0.07
Gemini	GPT-4o	1.00	0.27	0.07
Claude	Human	0.90	0.10	0.00
Gemini	Human	0.30	0.00	0.00
GPT-4o	Human	0.40	0.00	0.00

Avg # of templates per document pair

LLMs have much higher number of templates when compared with other models than with human

QUDSIM: Quantifying Discourse Similarities in LLM-Generated Text [Namuduri et al., ArXiv 2025]

This talk

Can we use language models to make sense of discourse?

- A generative discourse model based on Questions Under Discussion (QUD) theories
- Connections between human curiosity and discourse planning

Can we use our discourse framework to make sense of model behavior?

- A measure of discourse similarity in LLM-generated texts
- An explainable metric of information salience in LLMs

Summarization: then & now

Input: Editorial discussing the symbolism of lifting mask mandates

GPT-3

The article discusses the debate around when the Covid-19 pandemic will be considered over. It argues that in political terms, the pandemic will be considered over when all states have lifted […]

Explains all significant parts …

Captures the tone of the article […]

BRIO

Hawaii will become the last state in the country to end its indoor mask mandate. The Covid-19 pandemic has seen a surge in cases of the virus this winter. […]

Reflects the tone of the article the least

Slide credit: Tanya Goyal

Summarization: then & now

Input: Editorial discussing the symbolism of lifting mask mandates

GPT-3

The article discusses the debate around when the Covid-19 pandemic will be considered over. It argues that in political terms, the pandemic will be considered over when all states have lifted […]

Explains all significant parts …

Captures the tone of the article […]

Why do LLMs summarize so well zero-shot?

- Do models internalize information salience consistently?
- How to quantitatively measure this?
- Does LLMs' notion of salience aligns with prior theories or human intuitions?

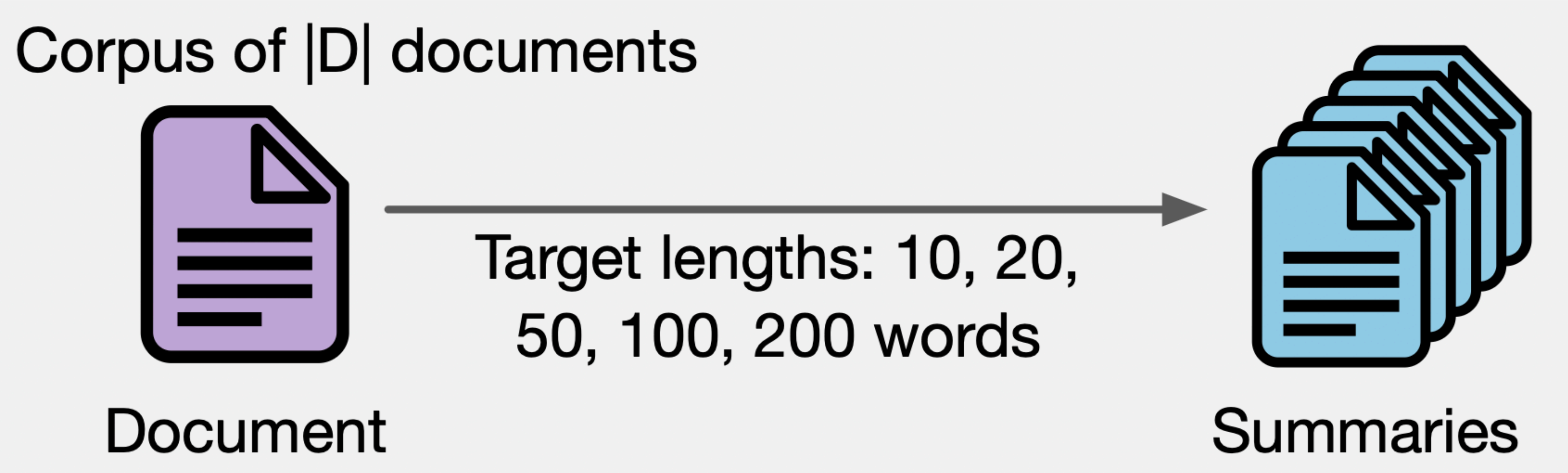
Deriving information salience

- Key ideas:
  - Summarization as behavioral probe
  - QUDs as an explainable representation of content units
  - “Content Salience Map”

Behavioral Analysis of Information Salience in Large Language Models [Trienes et al, ACL Findings’25]

Deriving information salience

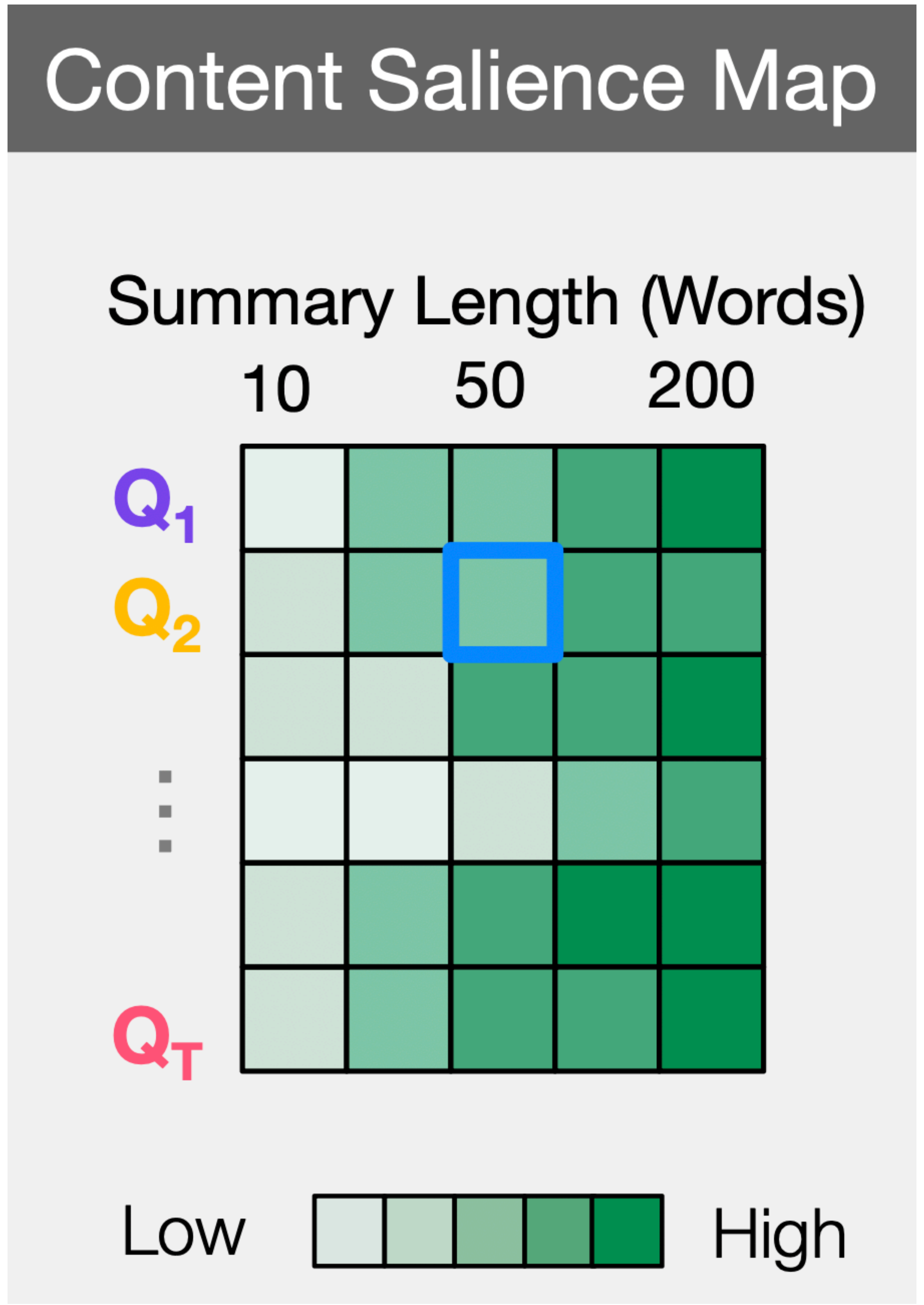
Summarization as behavioral probe



Intuition:

Shorter summaries will contain the most important topics.

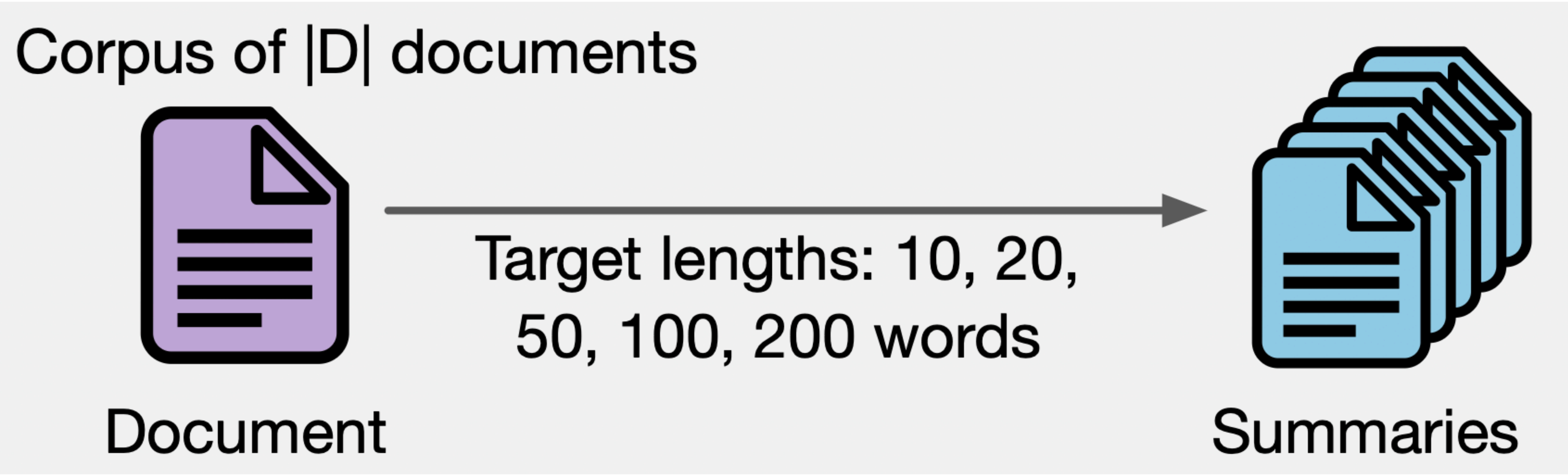
Longer summaries will (1) have more elaborated version of the most important topics (2) include less important topics.

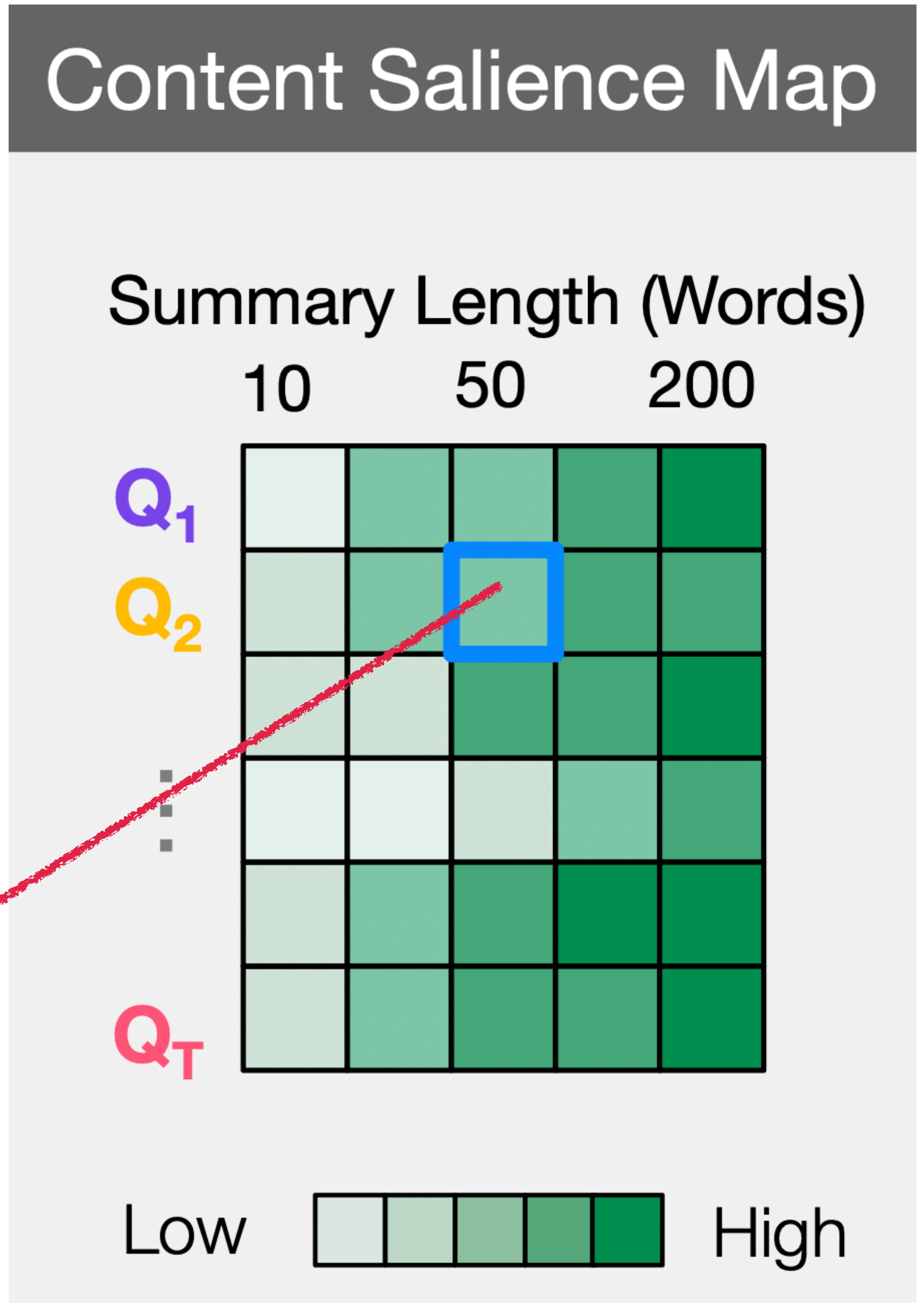


Behavioral Analysis of Information Salience in Large Language Models [Trienes et al, ACL Findings'25]

Deriving information salience

Summarization as behavioral probe





QUDs as topic

Consider: Randomized Control Trial (RCT) in medicine

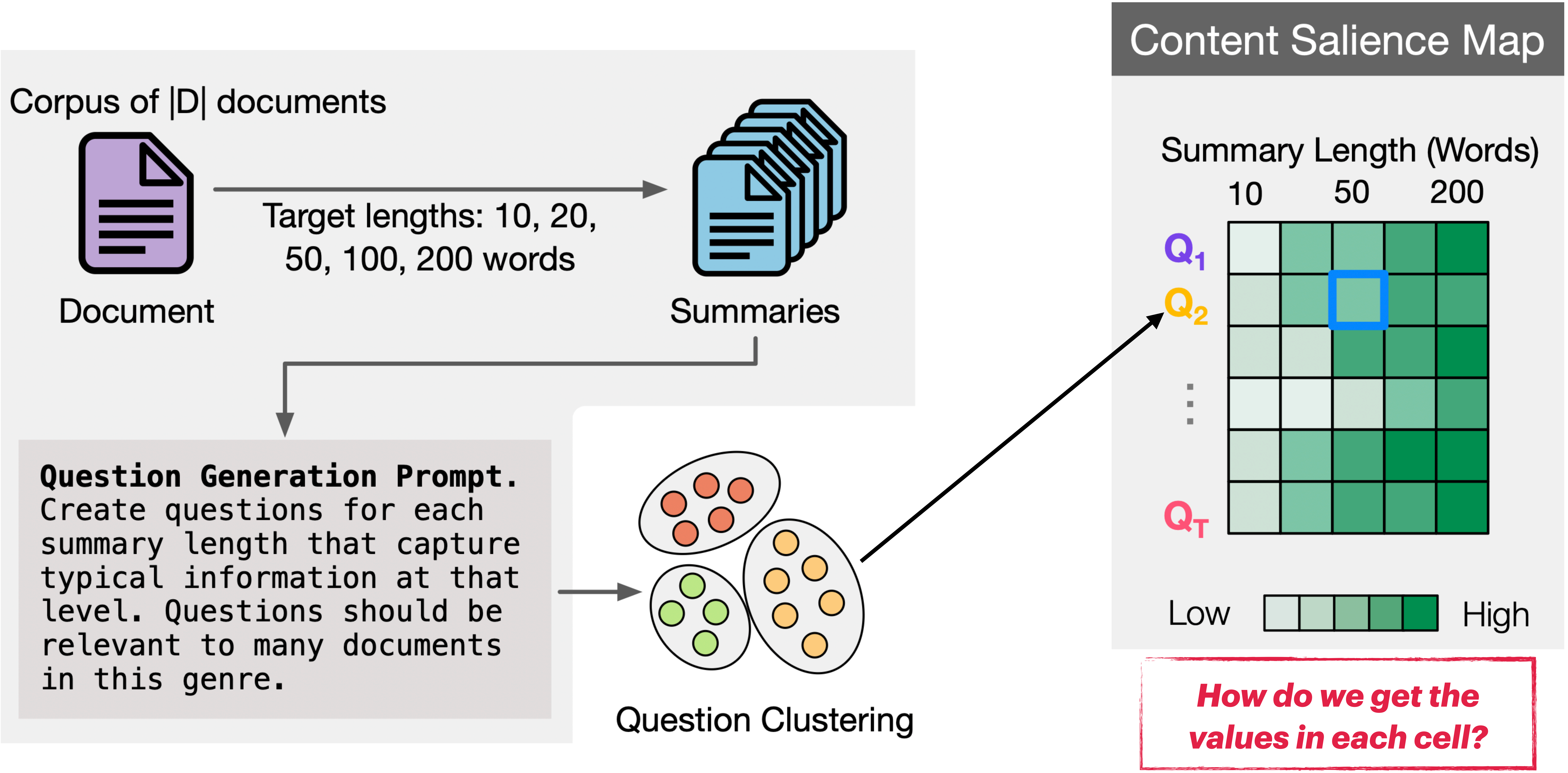
`Q2: What kind of patients were studied?`

Answer is unique to each document!

Behavioral Analysis of Information Salience in Large Language Models [Trienes et al, ACL Findings'25]

Deriving information salience

Step 1: generating QUDs for a corpus and cluster



Question Generation Prompt. Create questions for each summary length that capture typical information at that level. Questions should be relevant to many documents in this genre.

How do we get the values in each cell?

Behavioral Analysis of Information Salience in Large Language Models [Trienes et al, ACL Findings'25]

Deriving information salience

Step 2: Filling the cells

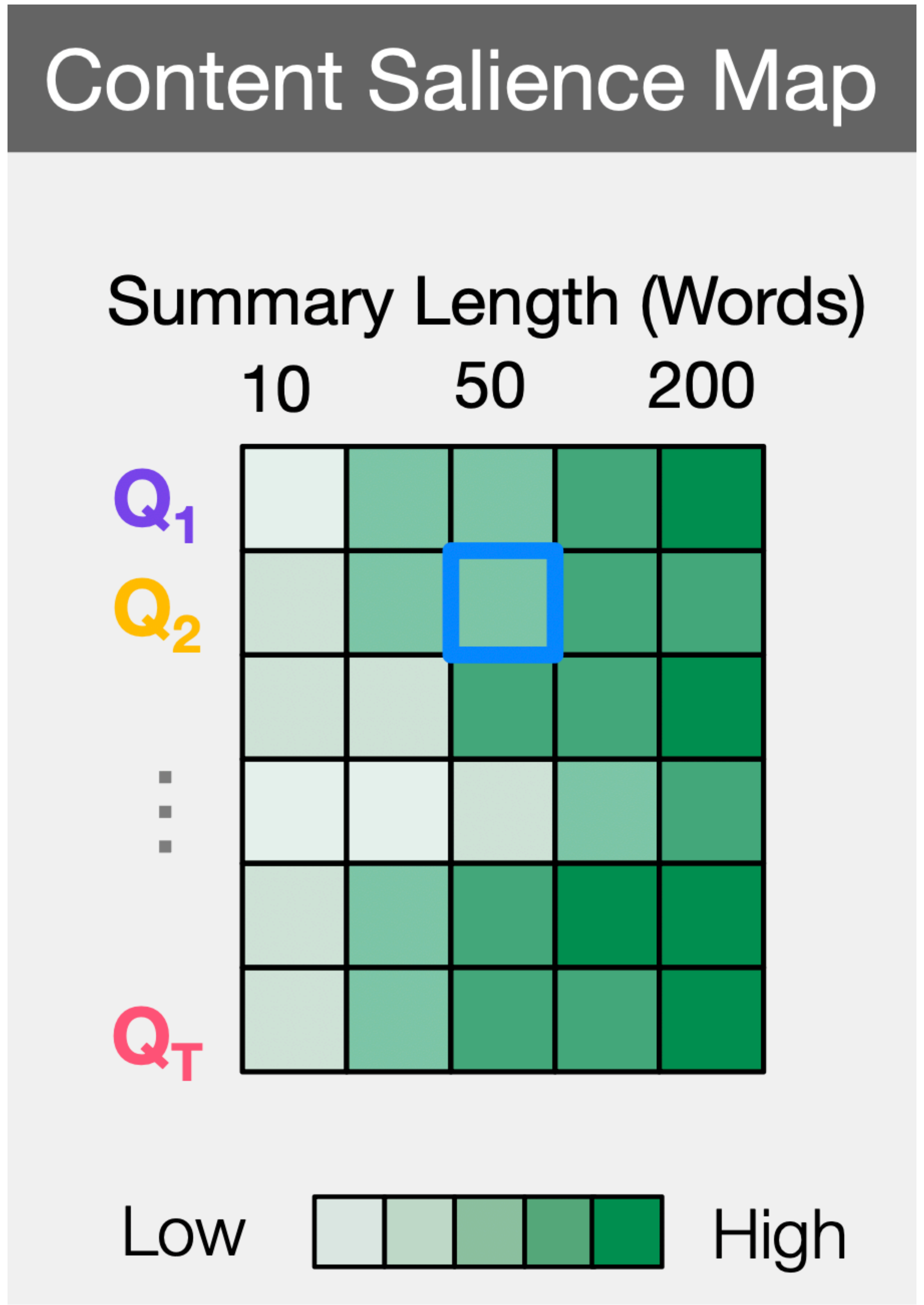
Atomic fact decomposition

Q2: What kind of patients were studied?

Document-answer claims:
✓ Patients with chronic Chagas cardiomyopathy (CCC)
✗ … left ventricular ejection fraction <45%
✗ … without or with heart failure symptoms
✗ … CCC stages B2 or C, respectively.

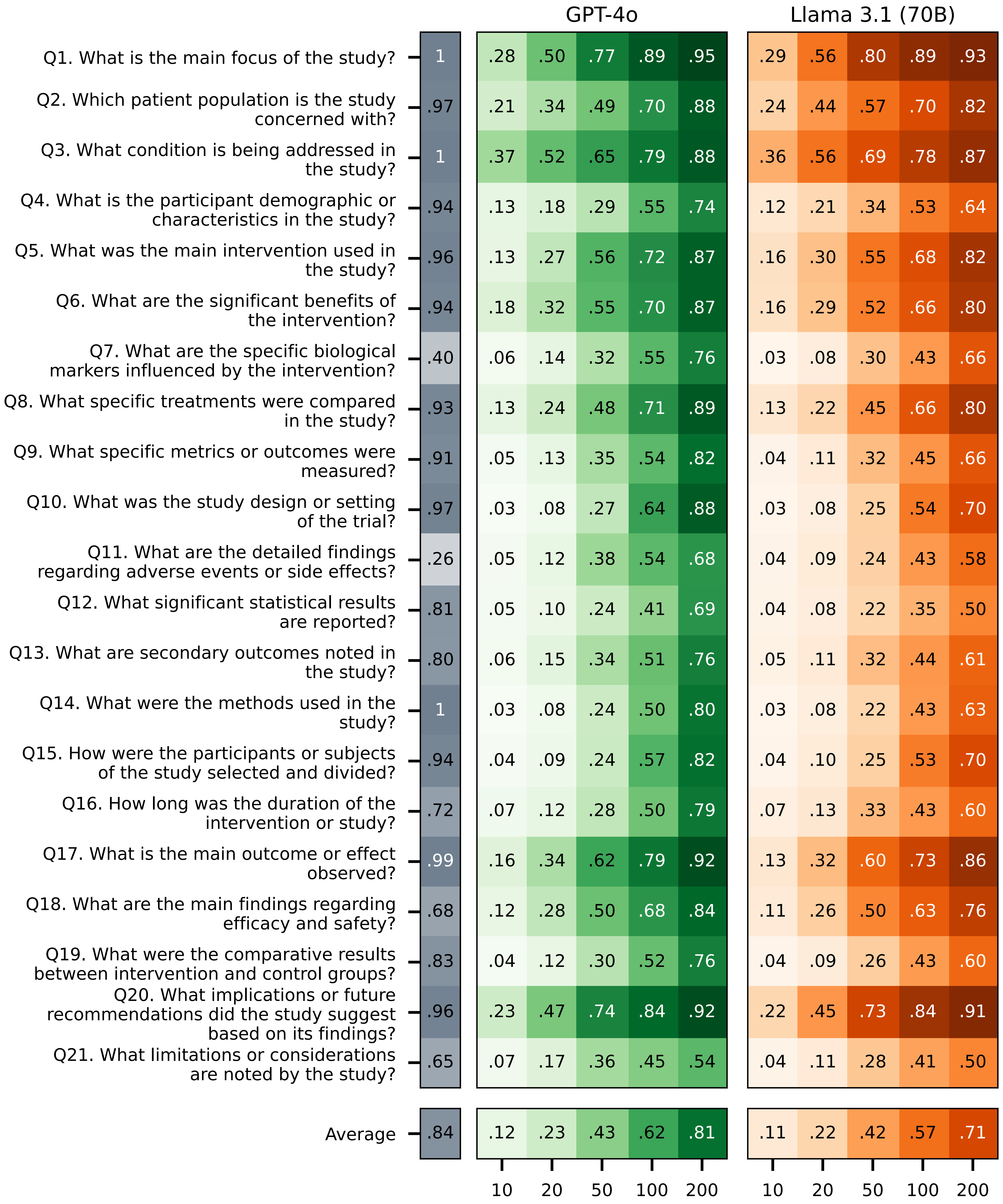
1/4 facts covered.
Answerability: 25%

Summary (50 words): The PEACH study investigated the effects of exercise-based cardiac rehabilitation on QoL in patients with chronic Chagas cardiomyopathy. Significant short-term improvements in physical and social functioning were observed in the exercise group, but no differences were found after six months.


Content Salience Map

Repeat and average across documents

Behavioral Analysis of Information Salience in Large Language Models [Trienes et al, ACL Findings’25]



		GPT-4o					Llama 3.1 (70B)				
		10	20	50	100	200	10	20	50	100	200
Q1. What is the main focus of the study?	1	.28	.50	.77	.89	.95	.29	.56	.80	.89	.93
Q2. Which patient population is the study concerned with?	.97	.21	.34	.49	.70	.88	.24	.44	.57	.70	.82
Q3. What condition is being addressed in the study?	1	.37	.52	.65	.79	.88	.36	.56	.69	.78	.87
Q4. What is the participant demographic or characteristics in the study?	.94	.13	.18	.29	.55	.74	.12	.21	.34	.53	.64
Q5. What was the main intervention used in the study?	.96	.13	.27	.56	.72	.87	.16	.30	.55	.68	.82
Q6. What are the significant benefits of the intervention?	.94	.18	.32	.55	.70	.87	.16	.29	.52	.66	.80
Q7. What are the specific biological markers influenced by the intervention?	.40	.06	.14	.32	.55	.76	.03	.08	.30	.43	.66
Q8. What specific treatments were compared in the study?	.93	.13	.24	.48	.71	.89	.13	.22	.45	.66	.80
Q9. What specific metrics or outcomes were measured?	.91	.05	.13	.35	.54	.82	.04	.11	.32	.45	.66
Q10. What was the study design or setting of the trial?	.97	.03	.08	.27	.64	.88	.03	.08	.25	.54	.70
Q11. What are the detailed findings regarding adverse events or side effects?	.26	.05	.12	.38	.54	.68	.04	.09	.24	.43	.58
Q12. What significant statistical results are reported?	.81	.05	.10	.24	.41	.69	.04	.08	.22	.35	.50
Q13. What are secondary outcomes noted in the study?	.80	.06	.15	.34	.51	.76	.05	.11	.32	.44	.61
Q14. What were the methods used in the study?	1	.03	.08	.24	.50	.80	.03	.08	.22	.43	.63
Q15. How were the participants or subjects of the study selected and divided?	.94	.04	.09	.24	.57	.82	.04	.10	.25	.53	.70
Q16. How long was the duration of the intervention or study?	.72	.07	.12	.28	.50	.79	.07	.13	.33	.43	.60
Q17. What is the main outcome or effect observed?	.99	.16	.34	.62	.79	.92	.13	.32	.60	.73	.86
Q18. What are the main findings regarding efficacy and safety?	.68	.12	.28	.50	.68	.84	.11	.26	.50	.63	.76
Q19. What were the comparative results between intervention and control groups?	.83	.04	.12	.30	.52	.76	.04	.09	.26	.43	.60
Q20. What implications or future recommendations did the study suggest based on its findings?	.96	.23	.47	.74	.84	.92	.22	.45	.73	.84	.91
Q21. What limitations or considerations are noted by the study?	.65	.07	.17	.36	.45	.54	.04	.11	.28	.41	.50
Average	.84	.12	.23	.43	.62	.81	.11	.22	.42	.57	.71

What notion of salience have LLMs learned in different domains?

Randomized Control Trials (RCT)

Summaries progressively get more detailed

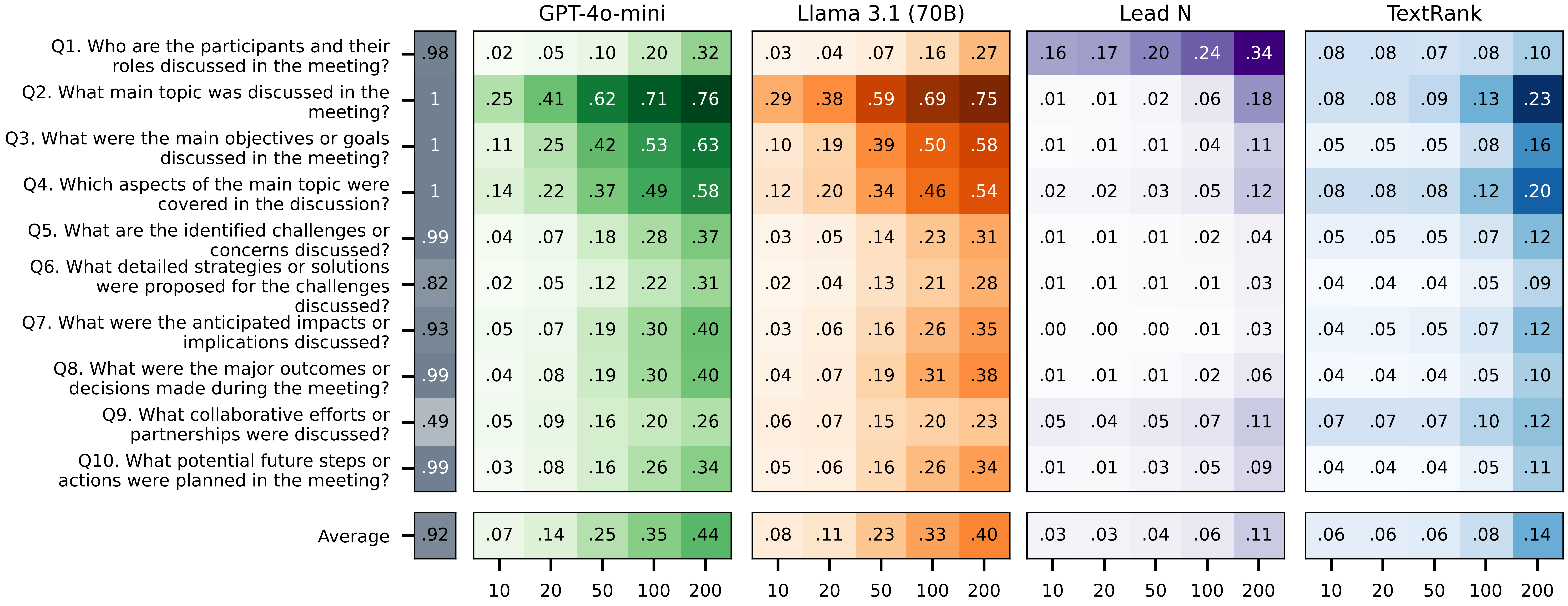
Content frequency is not in itself predictive of salience

Behavioral Analysis of Information Salience in Large Language Models
[Trienes et al, ACL Findings'25]

What notion of salience have LLMs learned in different domains?

Academic/technical meetings

This is not the same notion of salience as existing priors like lead and frequency.

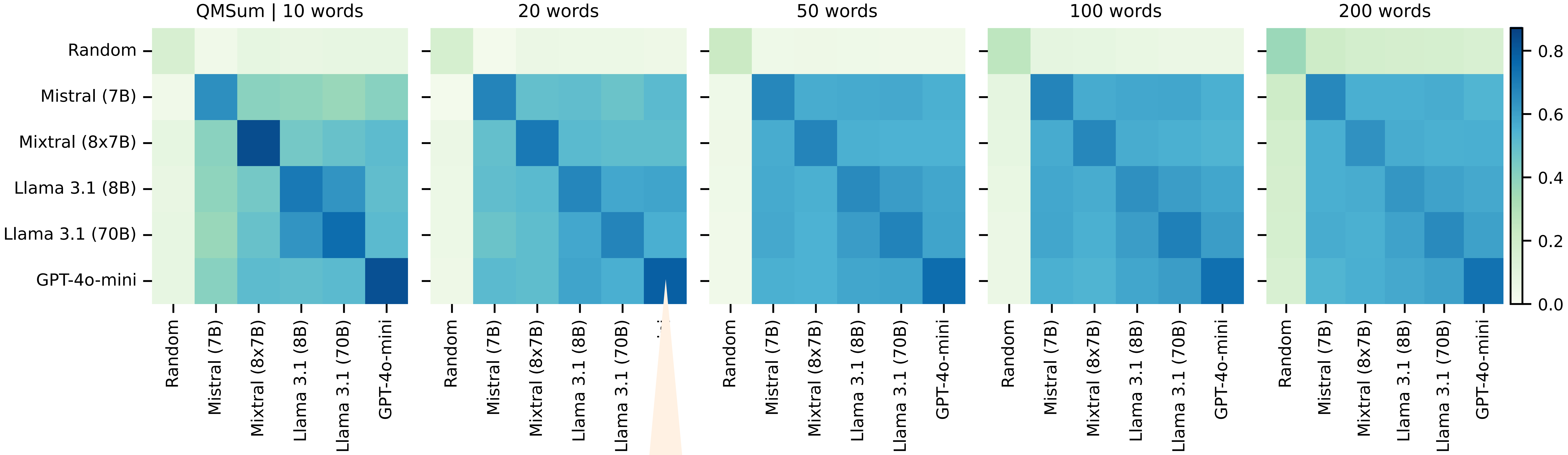


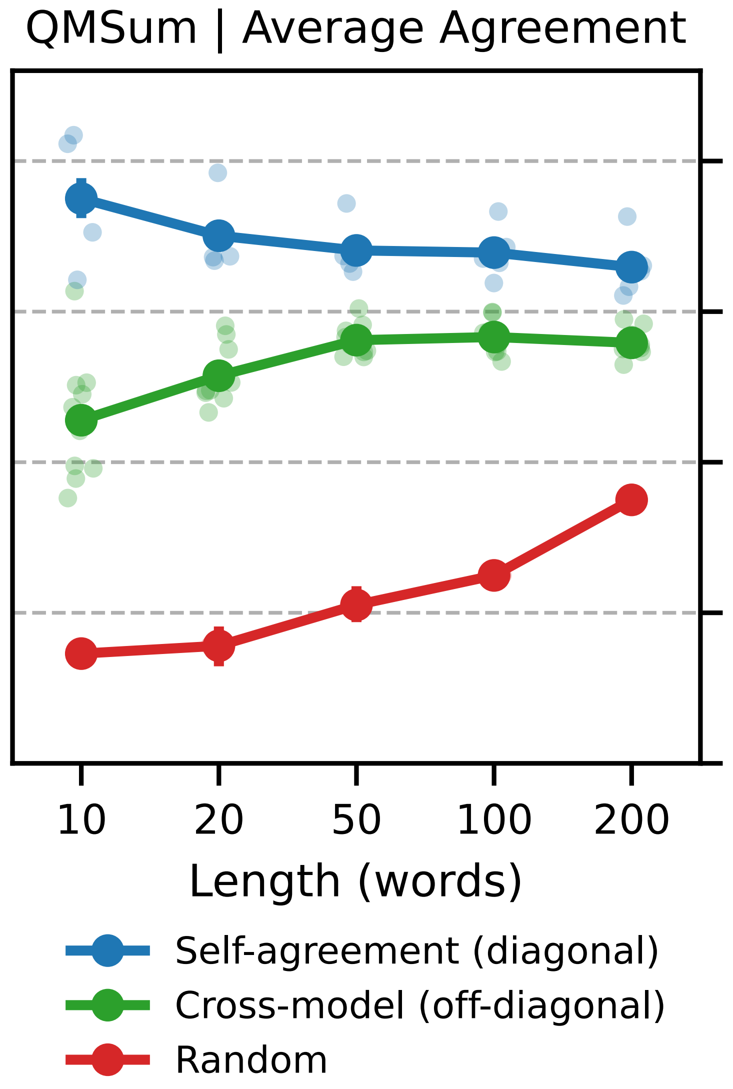
		GPT-4o-mini					Llama 3.1 (70B)					Lead N					TextRank				
		10	20	50	100	200	10	20	50	100	200	10	20	50	100	200	10	20	50	100	200
Q1. Who are the participants and their roles discussed in the meeting?	.98	.02	.05	.10	.20	.32	.03	.04	.07	.16	.27	.16	.17	.20	.24	.34	.08	.08	.07	.08	.10
Q2. What main topic was discussed in the meeting?	1	.25	.41	.62	.71	.76	.29	.38	.59	.69	.75	.01	.01	.02	.06	.18	.08	.08	.09	.13	.23
Q3. What were the main objectives or goals discussed in the meeting?	1	.11	.25	.42	.53	.63	.10	.19	.39	.50	.58	.01	.01	.01	.04	.11	.05	.05	.05	.08	.16
Q4. Which aspects of the main topic were covered in the discussion?	1	.14	.22	.37	.49	.58	.12	.20	.34	.46	.54	.02	.02	.03	.05	.12	.08	.08	.08	.12	.20
Q5. What are the identified challenges or concerns discussed?	.99	.04	.07	.18	.28	.37	.03	.05	.14	.23	.31	.01	.01	.01	.02	.04	.05	.05	.05	.07	.12
Q6. What detailed strategies or solutions were proposed for the challenges discussed?	.82	.02	.05	.12	.22	.31	.02	.04	.13	.21	.28	.01	.01	.01	.01	.03	.04	.04	.04	.05	.09
Q7. What were the anticipated impacts or implications discussed?	.93	.05	.07	.19	.30	.40	.03	.06	.16	.26	.35	.00	.00	.00	.01	.03	.04	.05	.05	.07	.12
Q8. What were the major outcomes or decisions made during the meeting?	.99	.04	.08	.19	.30	.40	.04	.07	.19	.31	.38	.01	.01	.01	.02	.06	.04	.04	.04	.05	.10
Q9. What collaborative efforts or partnerships were discussed?	.49	.05	.09	.16	.20	.26	.06	.07	.15	.20	.23	.05	.04	.05	.07	.11	.07	.07	.07	.10	.12
Q10. What potential future steps or actions were planned in the meeting?	.99	.03	.08	.16	.26	.34	.05	.06	.16	.26	.34	.01	.01	.03	.05	.09	.04	.04	.04	.05	.11
Average	.92	.07	.14	.25	.35	.44	.08	.11	.23	.33	.40	.03	.03	.04	.06	.11	.06	.06	.06	.08	.14

Behavioral Analysis of Information Salience in Large Language Models [Trienes et al, ACL Findings'25]

Do LLMs share a similar notion of salience?

Agreement of atomic fact inclusion across models





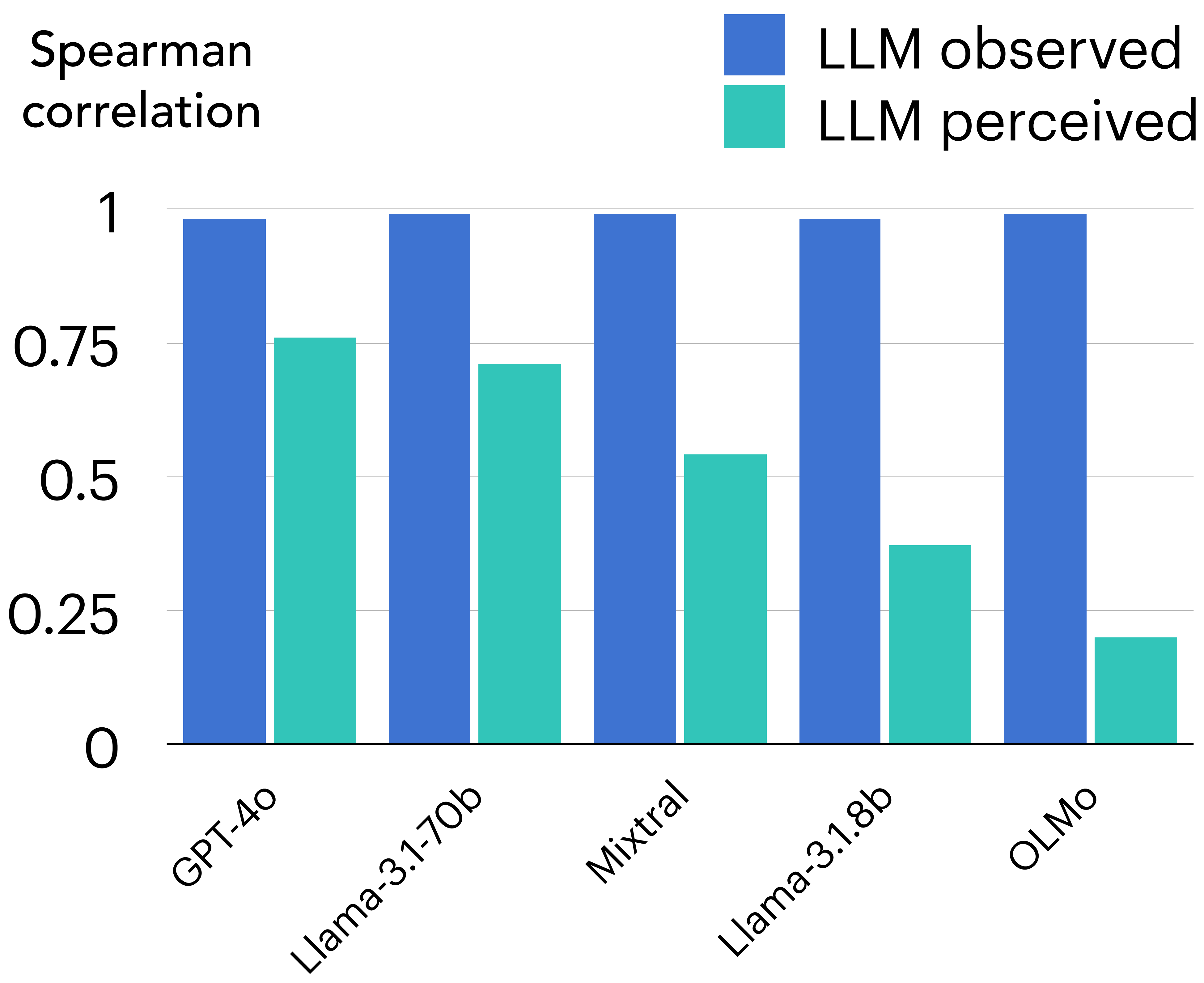
Diagonal: models behave consistently when prompted to summarize multiple times

Off-diagonal: models become more consistent as length increases

Behavioral Analysis of Information Salience in Large Language Models [Trienes et al, ACL Findings'25]

Introspection

Is LLM behavior consistent with perception?



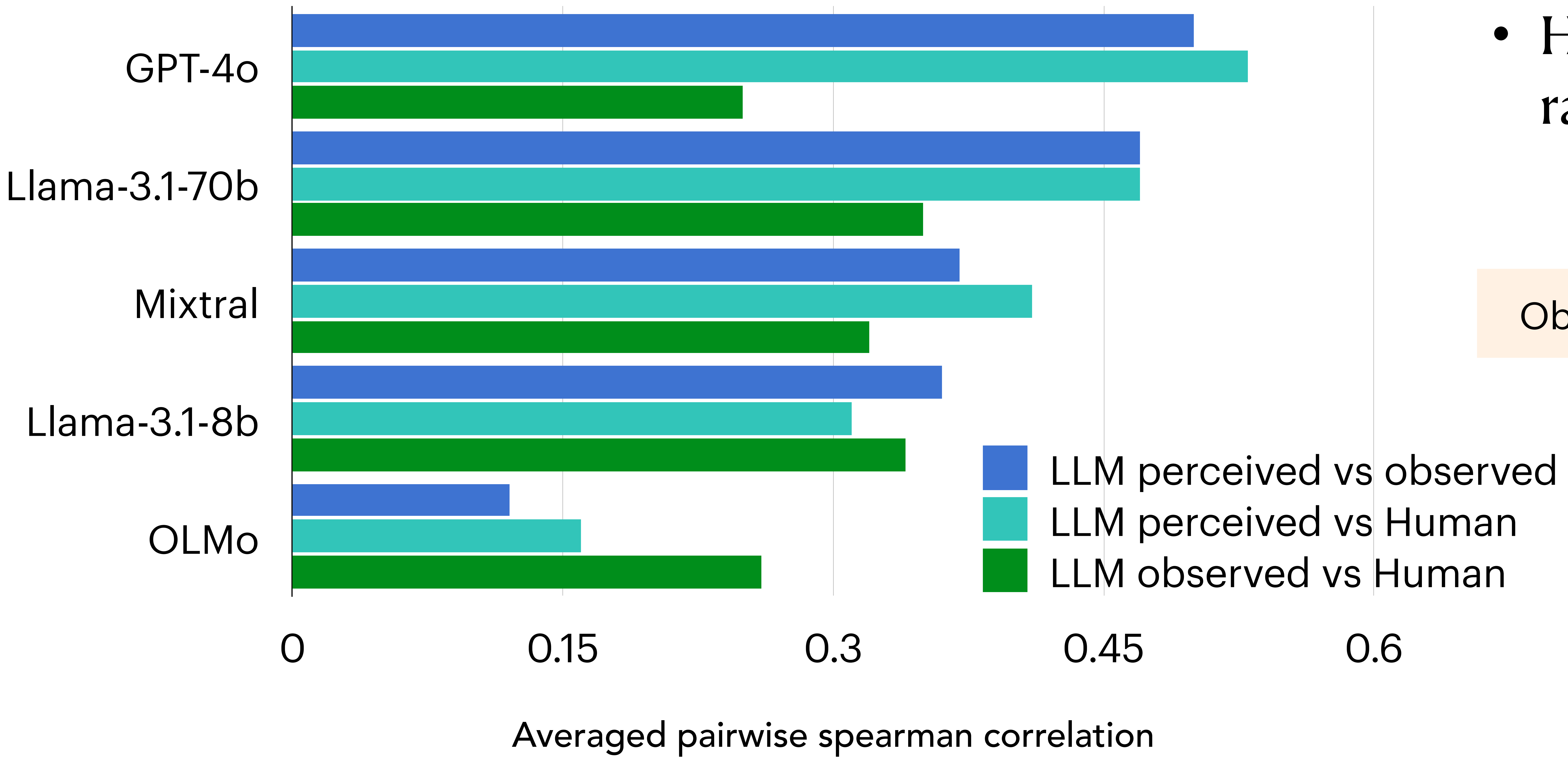
- LLM observed: CSM value for each QUD (highest correlation at 200 words)
- LLM perceived: prompting the same LLM to rate the importance of each QUD on a scale of 1-5

Across 5 runs, observed salience is consistent but asking is not, with strong scaling effect.

Behavioral Analysis of Information Salience in Large Language Models [Trienes et al, ACL Findings'25]

Alignment

Does LLM behavior/perception align with human perception?

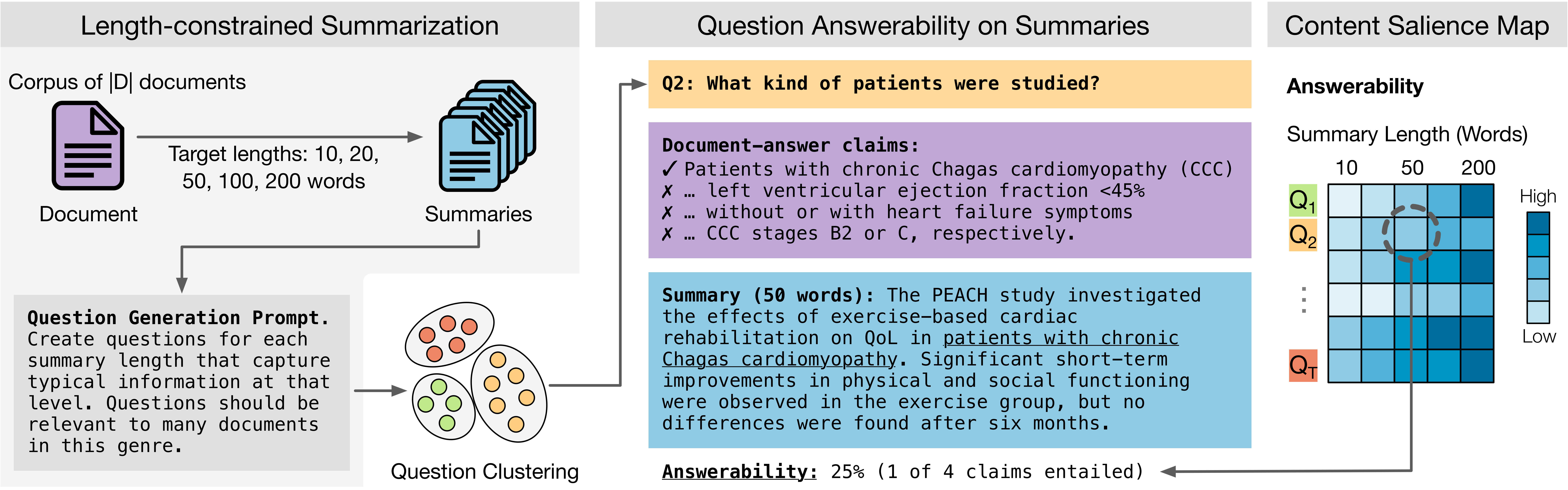


- Human: 5 expert ratings across datasets

Observed != perceived != human

Behavioral Analysis of Information Salience in Large Language Models [Trienes et al, ACL Findings'25]

Intrium conclusion



Question Generation Prompt. Create questions for each summary length that capture typical information at that level. Questions should be relevant to many documents in this genre.

Q2: What kind of patients were studied?

Document-answer claims:
✓ Patients with chronic Chagas cardiomyopathy (CCC)
✗ … left ventricular ejection fraction <45%
✗ … without or with heart failure symptoms
✗ … CCC stages B2 or C, respectively.

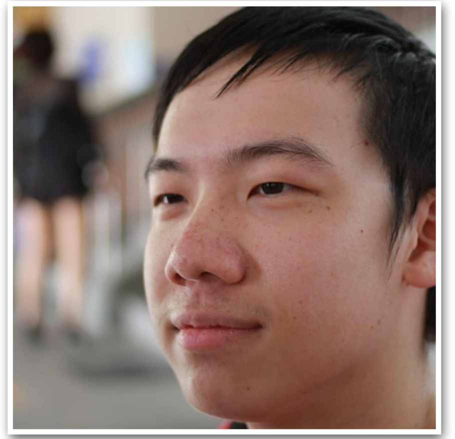
Summary (50 words): The PEACH study investigated the effects of exercise-based cardiac rehabilitation on QoL in patients with chronic Chagas cardiomyopathy. Significant short-term improvements in physical and social functioning were observed in the exercise group, but no differences were found after six months.

Answerability: 25% (1 of 4 claims entailed)

- QUD provides explainable abstraction of cross-document topics
- Length-controlled summarization as behavioral probe
- Content Salience Map shows consistent internalized salience within and across models, but does not align with introspection or human.

Behavioral Analysis of Information Salience in Large Language Models [Trienes et al, ACL Findings'25]

TEXAS
The University of Texas at Austin



Thank you!

NSF

NIH

Artificial Intelligence
Institute for Cosmic Origins
Funded by the U.S. National Science Foundation
and the Simons Foundation