

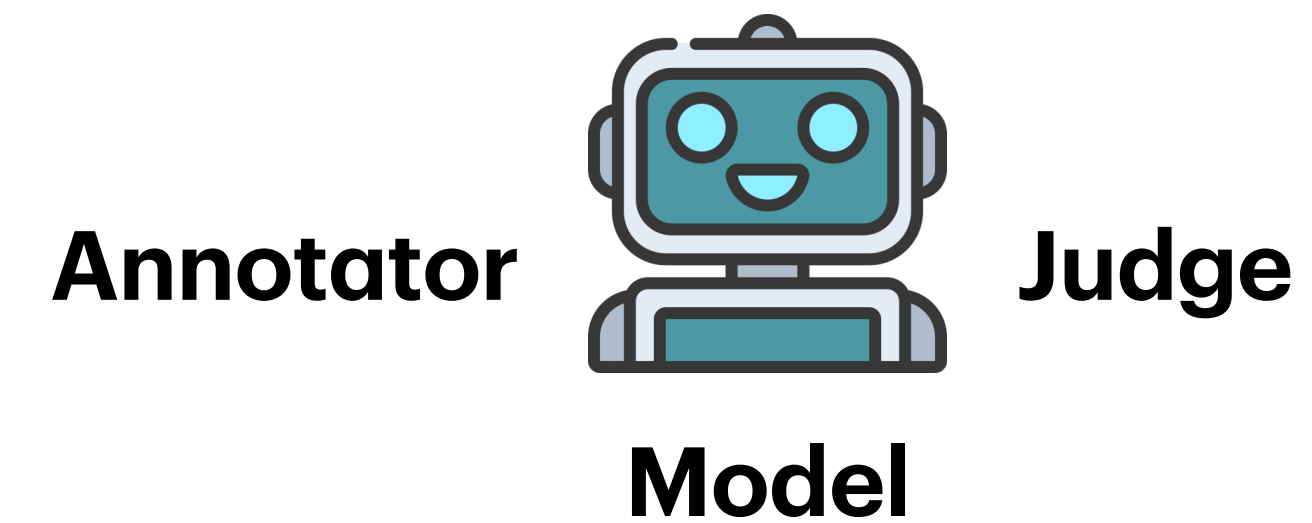
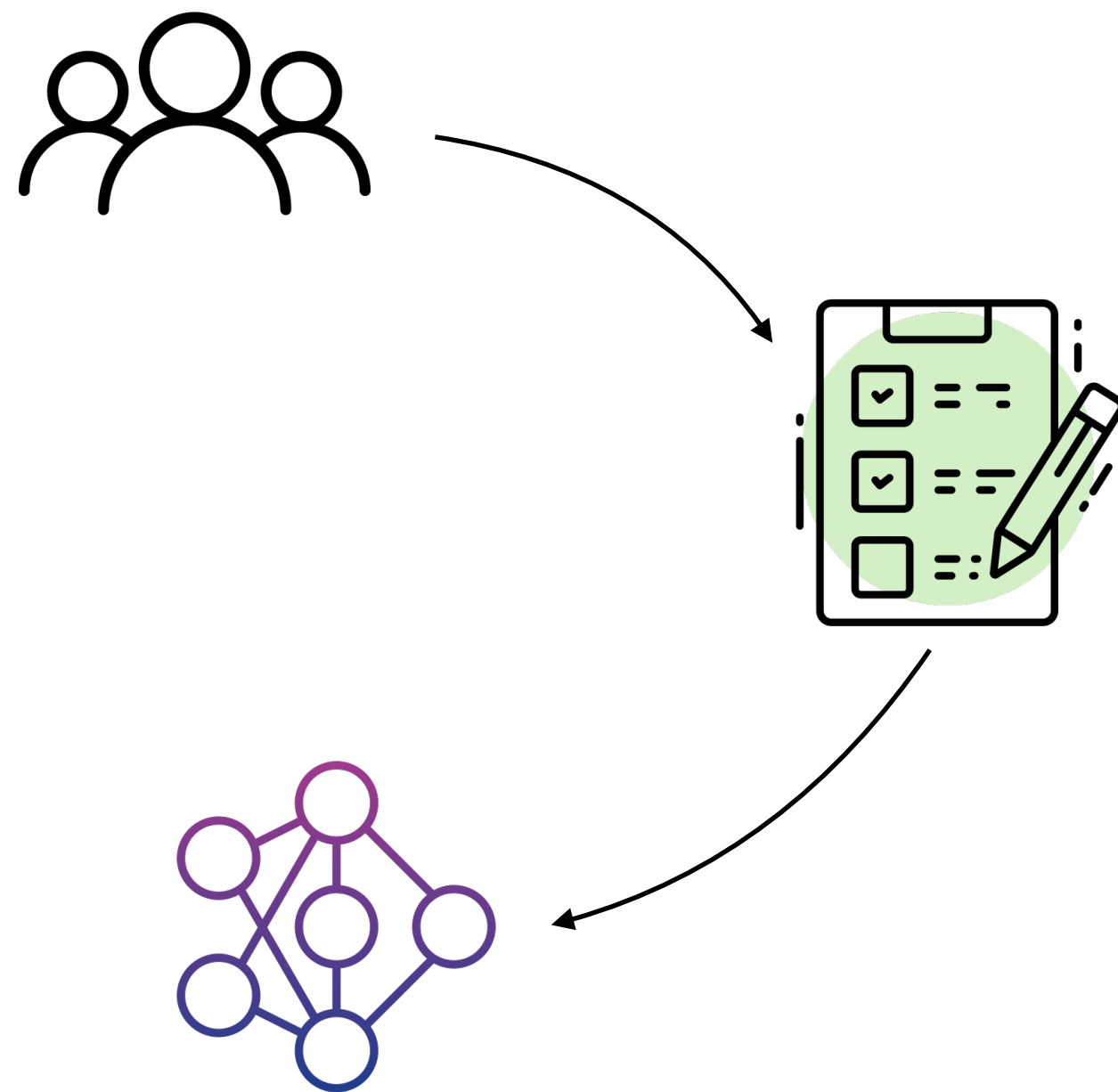
Engaging experts and LLMs in corpora development

Jessy Li

The University of Texas at Austin

Linguistic Annotation Workshop 2025

We stand on: Penn Treebank, OntoNotes, PropBank, FrameNet, UD, WordNet, Penn Discourse Treebank, RST-DT, GUM, TimeML, and countless annotated corpora from both experts and crowdsourcing



- LLMs have
- Been perceived as “the best model we have for language”
 - Replaced most simple crowdsourcing tasks
 - Outperformed humans on various tasks
 - Reached the ceiling of many traditional eval methods, e.g., ROUGE

Dataset	BRIO		T0		GPT3	
	Best ↑	Worst ↓	Best ↑	Worst ↓	Best ↑	Worst ↓
CNN	36	24	8	67	58	9
BBC	20	56	30	29	57	15

Table 3: Percentage of times a summarization system is selected as the best or worst according to majority vote (may be tied). Human annotators have a clear preference for GPT3-D2 for both CNN and BBC style summaries.

Goyal et al., “News Summarization and Evaluation in the Era of GPT-3”, ArXiv 2022

Dataset	Model	Overlap-Based			Similarity-Based		QAEval	
		ROUGE(1/2/L)	METEOR	BLEU	BERTScore	MoverScore	EM	F1
CNN	PEGASUS	34.85/14.62/28.23	.24	7.1	.858	.229	.105	.160
	BRIO	38.49/17.08/31.44	.31	6.6	.864	.261	.137	.211
	T0	35.06/13.84/28.46	.25	5.9	.859	.238	.099	.163
	GPT3-D2	31.86/11.31/24.71	.25	3.8	.858	.216	.098	.159
DailyMail	PEGASUS	45.77/23.00/36.65	.33	12.2	.865	.308	.159	.229
	BRIO	49.27/24.76/39.21	.37	11.7	.871	.331	.175	.259
	T0	42.97/19.04/33.95	.28	8.9	.863	.290	.121	.184
	GPT3-D2	38.68/14.24/28.08	.26	6.6	.859	.248	.101	.159
XSum	PEGASUS	47.97/24.82/39.63	.36	9.8	.901	.362	.145	.221
	BRIO	49.66/25.97/41.04	.39	10.6	.901	.372	.139	.224
	T0	44.20/20.72/35.84	.34	8.0	.896	.340	.125	.208
	GPT3-D2	28.78/7.64/20.60	.19	2.2	.869	.197	.066	.119
Newsroom	PEGASUS	39.21/27.73/35.68	.39	.14	.873	.272	0.182	0.253
	BRIO	-	-	-	-	-	-	-
	T0	25.64/9.49/21.41	.20	.04	.849	.145	.080	0.125
	GPT3-D2	27.44/10.67/22.18	.22	.05	.859	.159	.089	0.142

This talk: how do we navigate this landscape?

The roles of experts in corpora development, annotation, and evaluation

- Case studies in language and alignment

How can LLMs help the annotation process itself?

- Explanation-based rescaling (EBR)

An era where expert input is critical

Analysis and annotation, to capture implicit reasoning in discourse and pragmatics

Alignment: should expert write responses or tell models how-to?

Analysis and Annotation

...why? LLMs are so good at “language” already!

- But are they, really?
- No, not even for coherence

Findings came from
high-quality human
annotation

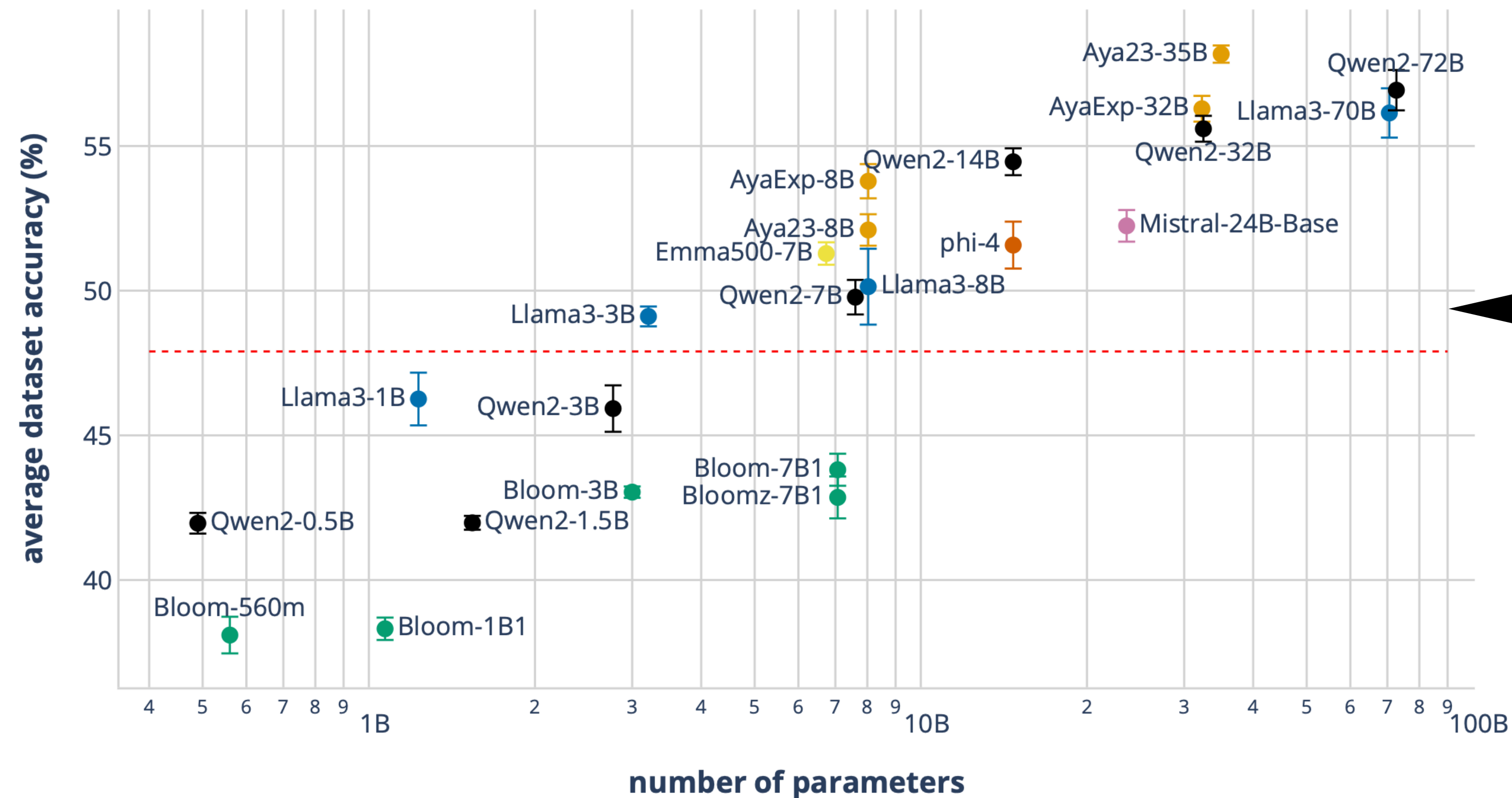
Table 1: Definition of all coherence error types, an example annotation for each, and their prevalence (%) in generated summaries, which is calculated as the number of error occurrences in all summaries normalized by the total number of sentences in all summaries.

Error Type	Definition	Example spans & questions	% errors per sentence inc / hier
Entity omission	An entity (e.g., person, object, place) is mentioned in the summary, but key context or details are missing or unclear.	<i>Span:</i> A mysterious man introduces Proctor to "Arrivalism." <i>Question:</i> Who is this mysterious man?	7.3 / 3.71
Event omission	An event is mentioned in the summary, but key details are missing or unclear.	<i>Span:</i> During a mission to find Caeli, Proctor is captured by watchmen while Thea escapes. <i>Question:</i> What happened to Caeli?	4.25 / 2.27
Causal omission	A reason or motivation is missing or under-explained.	<i>Span:</i> Proctor seeks answers from... Callista about the investigation. <i>Question:</i> Why would Callista know something about the investigation?	2.75 / 1.21
Discontinuity	An interruption in the flow of the narrative such as sudden jumps in time or perspective.	<i>Span:</i> In the new settlement, Thea adjusts to her life, working hard and finding solace in nature. <i>Question:</i> Why the shift to Thea's perspective?	2.23 / 1.56
Salience	Inclusion of details that do not contribute to the main plot.	<i>Span:</i> His father... flees, resulting in a chaotic chase on the pier. <i>Question:</i> What is the significance of this incident?	1.42 / 1.03
Language	Spelling or grammar issues; ambiguous wording.	<i>Span:</i> Despite her love for him, Deborah is heartbroken by his decision. <i>Question:</i> Why is the preposition "Despite" used here when she is, in fact, heartbroken because of her love for him?	0.82 / 0.71
Inconsistency	A discrepancy or contradiction within a story's plot, character development, or themes.	<i>Span:</i> In a farewell, Proctor marries his brother Malcolm to Cynthia and says goodbye to his loved ones. <i>Question:</i> If Cynthia is his mother and Malcolm is his brother, how can a mother and son marry?	0.97 / 1.03
Duplication	Redundant repetition of similar information.	<i>Span 1:</i> Proctor... deals with students and school issues, seeking help from Callista to fund a roof replacement. <i>Span 2:</i> Proctor's life continues as he... deals with school issues, such as funding for a roof replacement <i>Question:</i> Why does the same information appear twice?	2.12 / 1.18

Analysis and Annotation

...why? LLMs are so good at “language” already!

- ... and not for discourse relations either



Data came from expert-curated multilingual discourse corpora

Analysis and Annotation

...why? LLMs are so good at “language” already!

- ...and not for creative writing

Findings came from expert insights and evaluation

Dimension	Test	GPT3.5	GPT4	Claudev1.3	NewYorker	Expert Agreement
Fluency	Understandability & Coherence	22.2	33.3	55.6	91.7	0.27
	Narrative Pacing	8.3	52.8	61.1	94.4	0.39
	Scene vs Exposition	8.3	50.0	58.3	91.7	0.27
	Literary Devices & Language Proficiency	5.6	36.1	13.9	88.9	0.37
	Narrative Ending	8.3	19.4	33.3	91.7	0.48
Flexibility	Emotional Flexibility	16.7	19.4	36.1	91.7	0.32
	Perspective & Voice Flexibility	8.3	16.7	19.4	72.2	0.44
	Structural Flexibility	11.1	19.4	30.6	88.9	0.39
Originality	Originality in Form	2.8	8.3	0.0	63.9	0.41
	Originality in Thought	2.8	44.4	19.4	91.7	0.40
	Originality in Theme & Content	0	19.4	11.1	75.0	0.66
Elaboration	World Building & Setting	16.7	41.7	58.3	94.4	0.33
	Character Development	8.3	16.7	16.7	61.1	0.31
	Rhetorical Complexity	2.8	11.1	5.6	88.9	0.66
Average		8.7	27.9	30.0	84.7	0.41

Do LLMs get discourse particles?

Just the word!

Temporal: Used to indicate that something happened very recently, or close to another event.

The train *just* left (*recently*).

Adjective: Used to describe a person or idea, especially a law or policy, as fair, appropriate, or lawful.

That queen was a fair and *just* ruler.
This law is not *just*!

Exclusive: Used to exclude other possibilities or options.

A: What does Betsy eat?
B: Betsy *just* eats chicken nuggets.

Unelaboratory: Used to deny further elaboration on an event or concept.

A: What kind of dog is Fido?
B: Fido's *just* a dog.

Unexplanatory: Used to deny that there is an explanation or to offer a weak explanation with no stronger one available.

The lights in this place *just* turn on and off. (*Paraphrase: There is no reason why.*)

Emphatic: Used to add emphasis to an already strong word or phrase.

This pumpkin bisque is *just* delicious!

Lee, 1987; Grosz, 2012; Coppock and Beaver, 2014; Beltrama, 2022; Deo and Thomas, 2025, among others

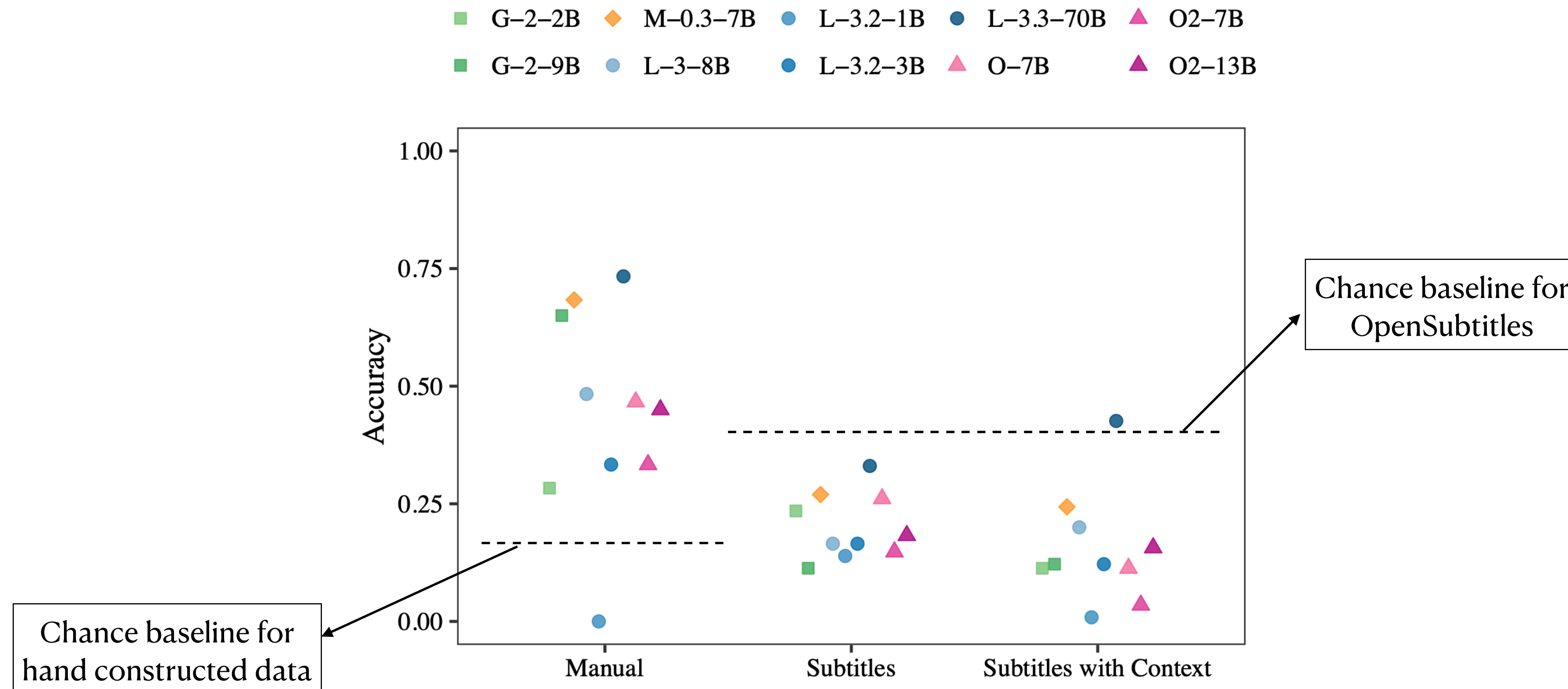
Experts: platinum data construction

- What did linguists do here?
 - Construct 90 diverse, unambiguous examples covering each sense evenly
 - Given naturally occurring data, perform subtle sense disambiguation
 - 149 sentences in OpenSubtitles
- Who are the linguists?
 - Ashwini Deo and William Carl Thomas (Deo and Thomas, 2025)
 - Linguistic graduate students in their class

Note that these tasks entail qualities that LLMs do not possess

Evaluating LLMs on interpreting “just”

Task 1: sense labeling

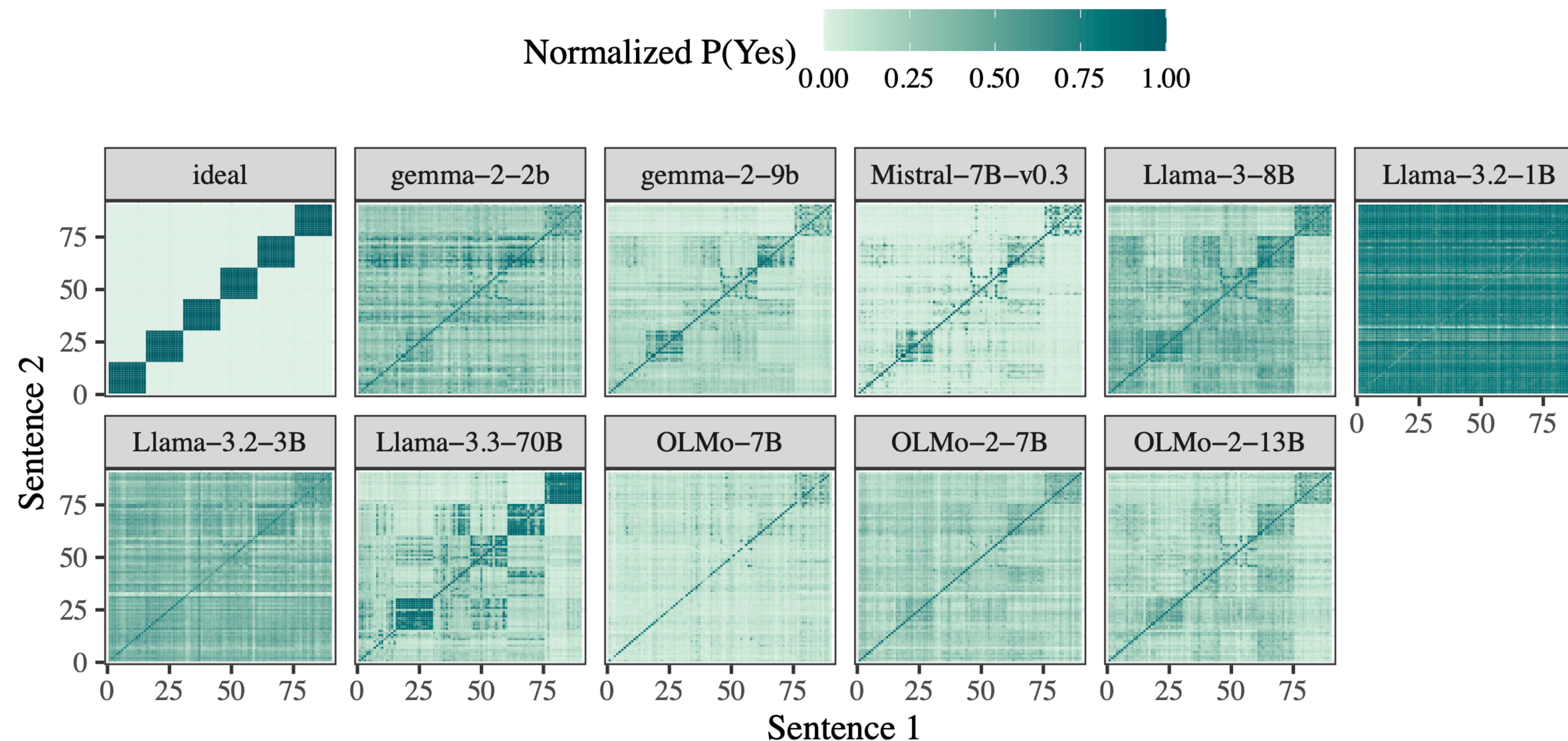


Evaluating LLMs on interpreting “just”

Task 2: pairwise test

- Do the following two sentences use “just” in the same way? Respond with "Yes" or "No".

Left-right:
 Exclusive,
 Unelaboratory,
 Unexplanatory,
 Emphatic,
 Temporal,
 Adjective



Do LLMs get implicit discourse reasoning?



Reader: as they read,
wonder about potential
questions

Onea (2016)

Writer: some of these questions
become Questions Under Discussion
(ie, answered in the document)



Van Kuppevelt (1995), Roberts (2012)

[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. ...

What would be the advantage
of the proposed tax?

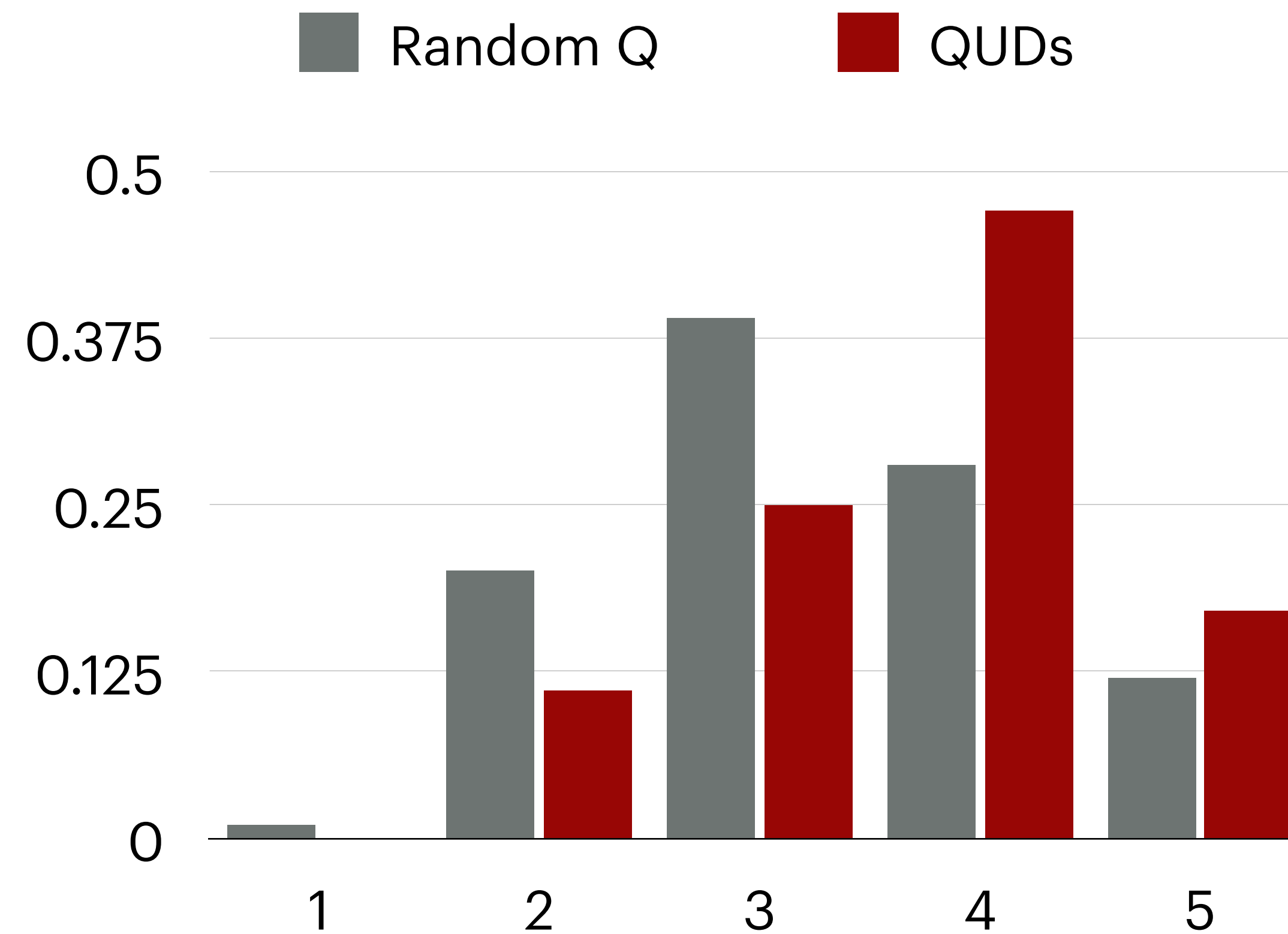
What was the rebuff?

How much will the sales tax be raised?

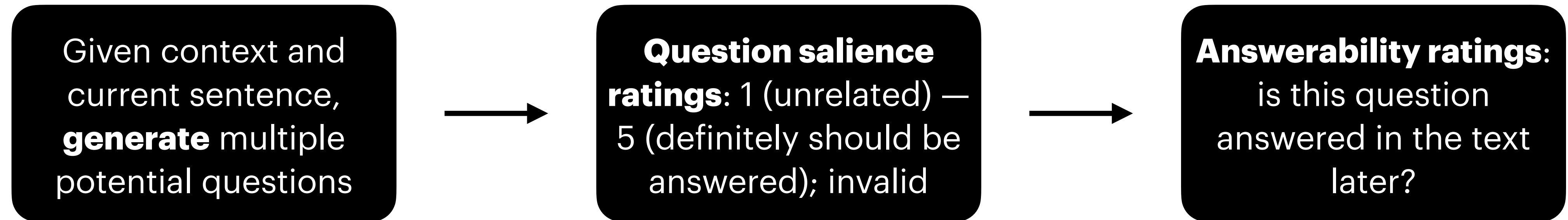
Who led the talk of the raise?

Potential questions that become QUDs...

... are rated with higher salience



QSalience data collection



[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

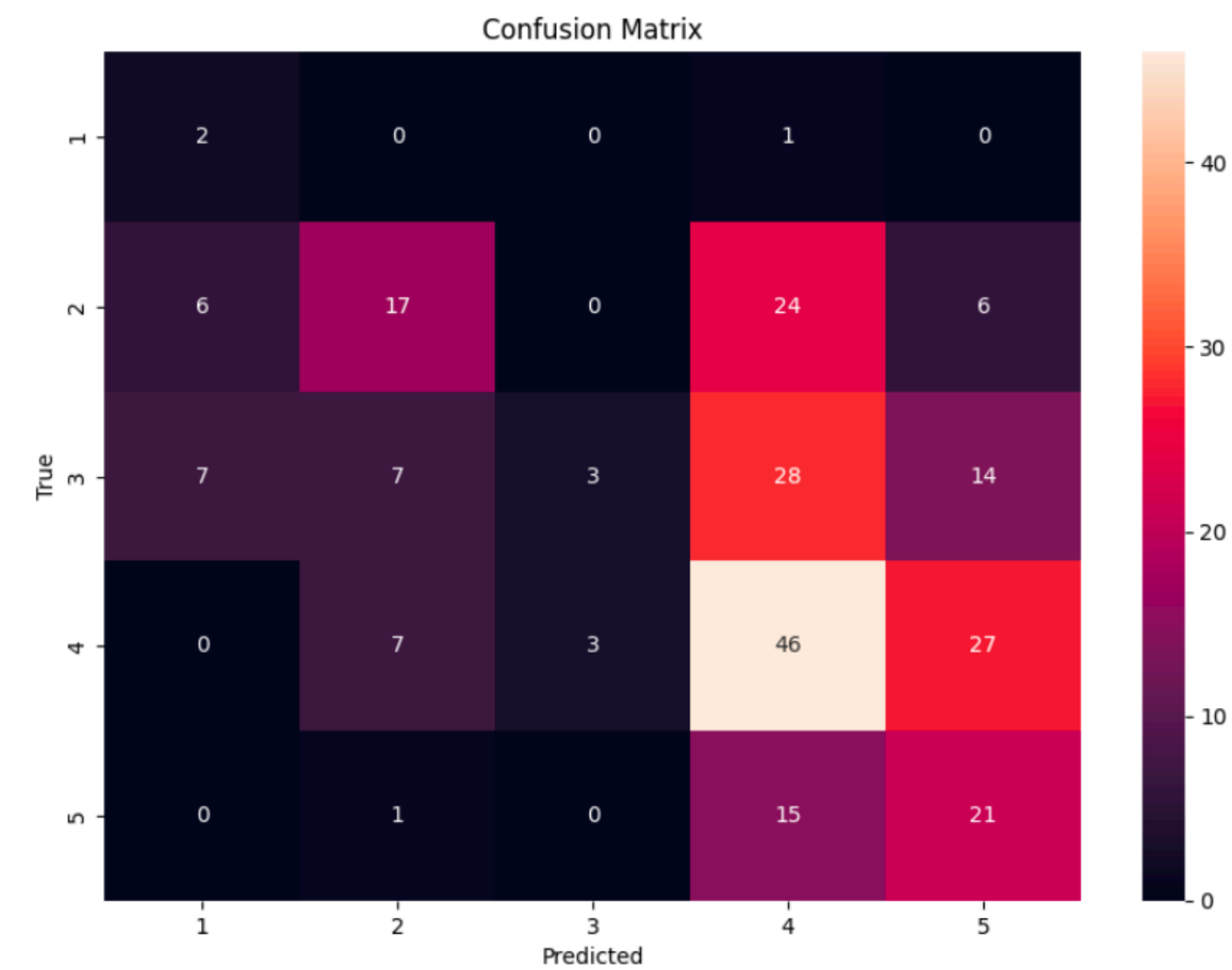
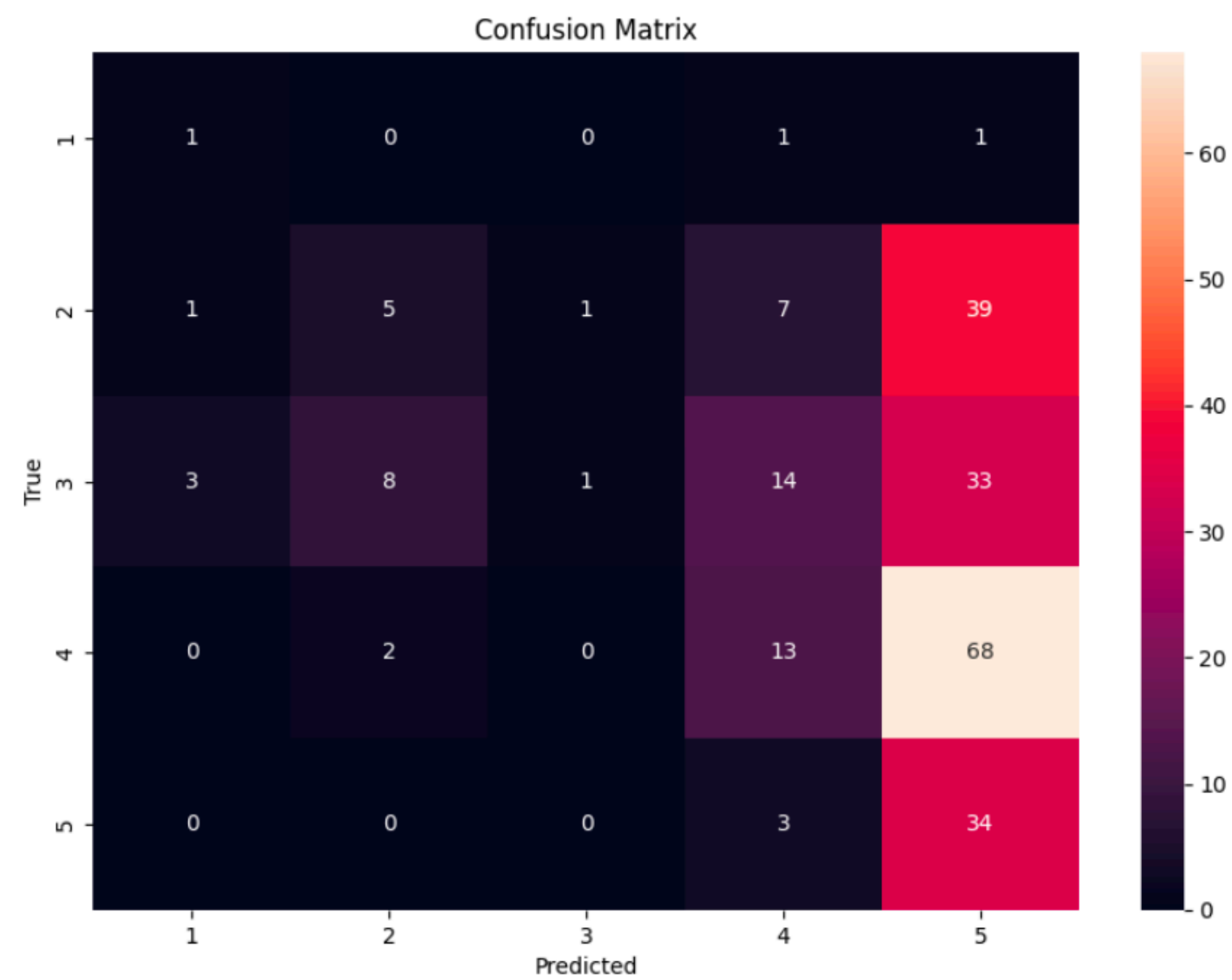
Do salience and answerability correlate?

- 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!

Are LLMs good at salience prediction?



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

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: results



Reader: as they read, wonder about potential questions

Not without training!

Writer: some of these questions become QUDs



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

An era where expert input is critical

Analysis and annotation:
Experts help reveal where
models lack fundamentally, even
when outputs are generally
perceived as high quality

Alignment: should
expert write responses
or tell models how-to?

Alignment

Task: having models produce **cognitive reappraisals** for emotional well-being

Event: fired from job

Oh no, I made that terrible mistake!

 regret

This is just unfair, X was intentionally setting me up!

 anger

Such a toxic environment, I'd never want to come back!

 relief

- How people subjectively evaluate or **appraise** the situation characterizes their emotional experiences.
- This is typically characterized by range of different “dimensions”.

Self responsibility:

Does the narrator think that they are responsible for causing the situation?

Problem-focused coping:

Does the narrator think that they can cope with the consequences of the situation?

Arnold, 1960;
 Lazarus, 1966;
 Lazarus et al., 1980;
 Roseman, 1984;
 Scherer et al., 1984;
 Smith and Ellsworth, 1985;
 Weiner, 1985;
 Clore and Ortony, 2000;
 Roseman and Smith, 2001;
 Scherer et al., 2001;
 Sander et al., 2005;
 Ortony et al., 2022

...

Alignment: the goal

Task: having models produce **cognitive reappraisals** for emotional well-being

Reappraisal Goal: guide the narrator over their perception of their ability to **emotionally cope** with the consequences of the event

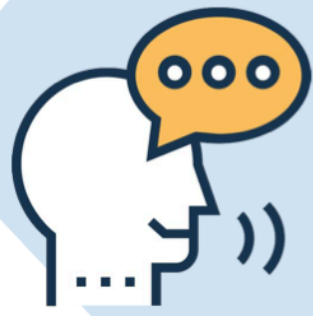


... I wanted to see an old friend from high school, but I found out that this friend did not get vaccinated. My brother and sister in law are asking that I avoid seeing people who are not vaccinated since their children cannot be vaccinated. Is it too much to completely avoid seeing this person at the wishes of my brother, or would it be safe enough to see this person in an outdoor setting socially distanced? ...

Alignment: the goal

Task: having models produce **cognitive reappraisals** for emotional well-being

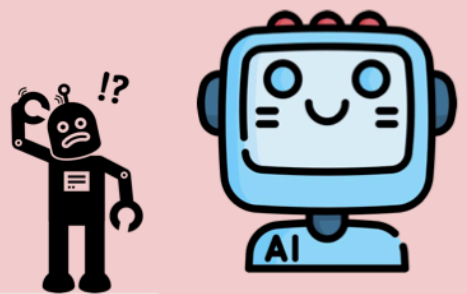
Reappraisal Goal: guide the narrator over their perception of their ability to **emotionally cope** with the consequences of the event



... I wanted to see an old friend from high school, but I found out that this friend did not get vaccinated. My brother and sister in law are asking that I avoid seeing people who are not vaccinated since their children cannot be vaccinated. Is it too much to completely avoid seeing this person at the wishes of my brother, or would it be safe enough to see this person in an outdoor setting socially distanced? ...

Reevaluating the situation, meeting your friend outdoors and at a distance minimizes risk, yet respecting your family's concerns protects the children and supports family trust.

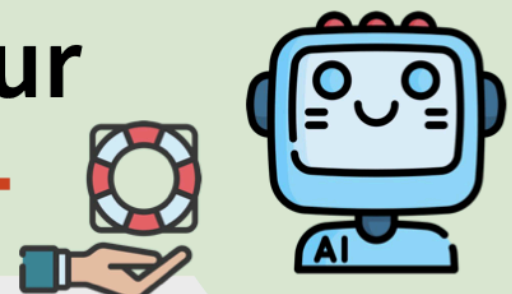
UNGUIDED



... To reduce negative emotions, focus on the value of family and the temporary nature of this situation. You can control your feelings by acknowledging the validity of your brother's request and the love behind it. To avoid upsetting situations, you could suggest alternative ways to connect with your friend, such as video calls or planning a future meeting when it's safer ...

GUIDED

RESORT



How to best utilize expertise?

Expert writing a response vs expert writing *principles*

Self responsibility:

Does the narrator think that they are responsible for causing the situation?

Reappraisal goal: Re-evaluate whether the narrator deserves to be blamed or credited for the situation at hand. If not responsible, the narrator is encouraged to acknowledge that fact and reassess the situation.

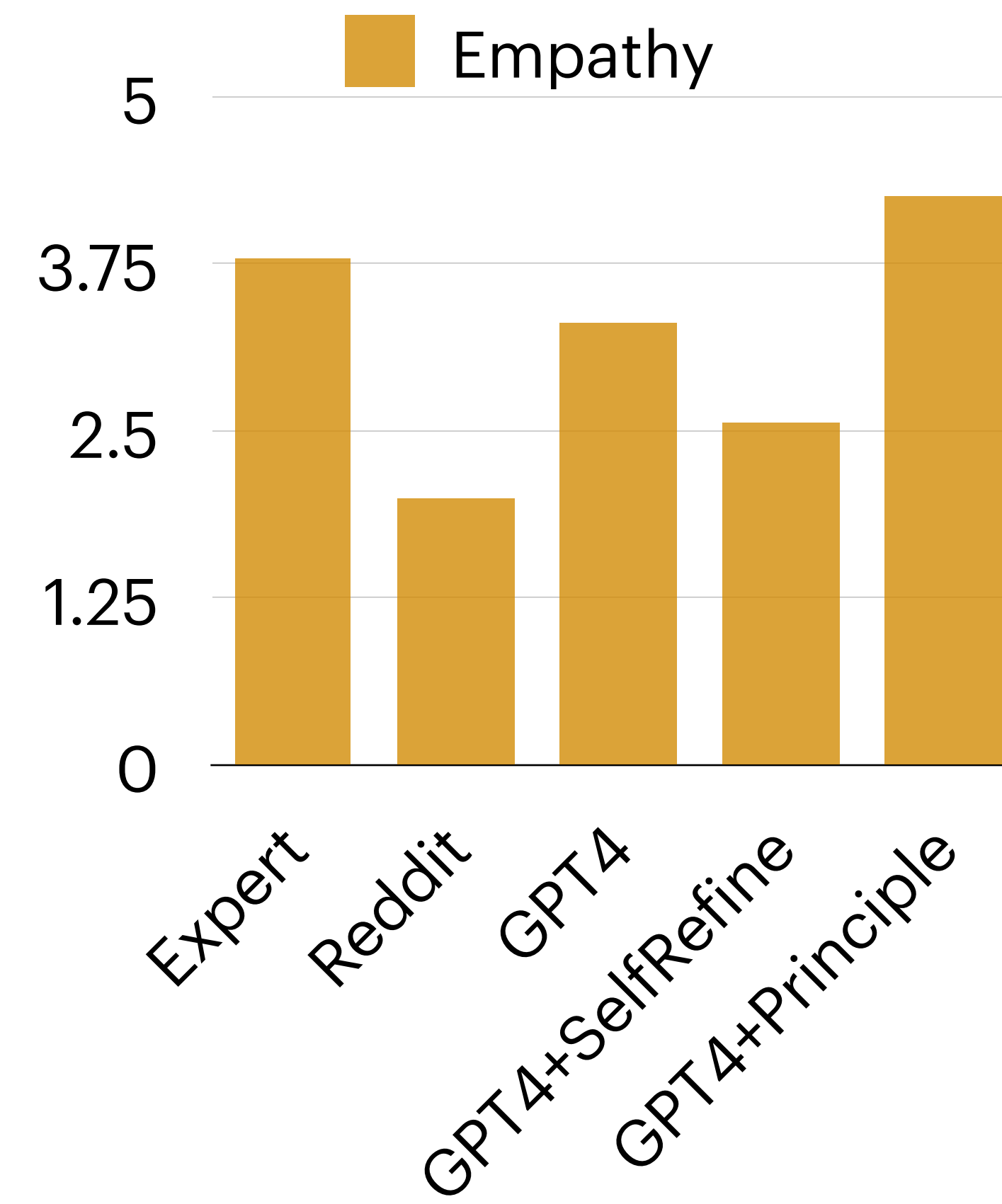
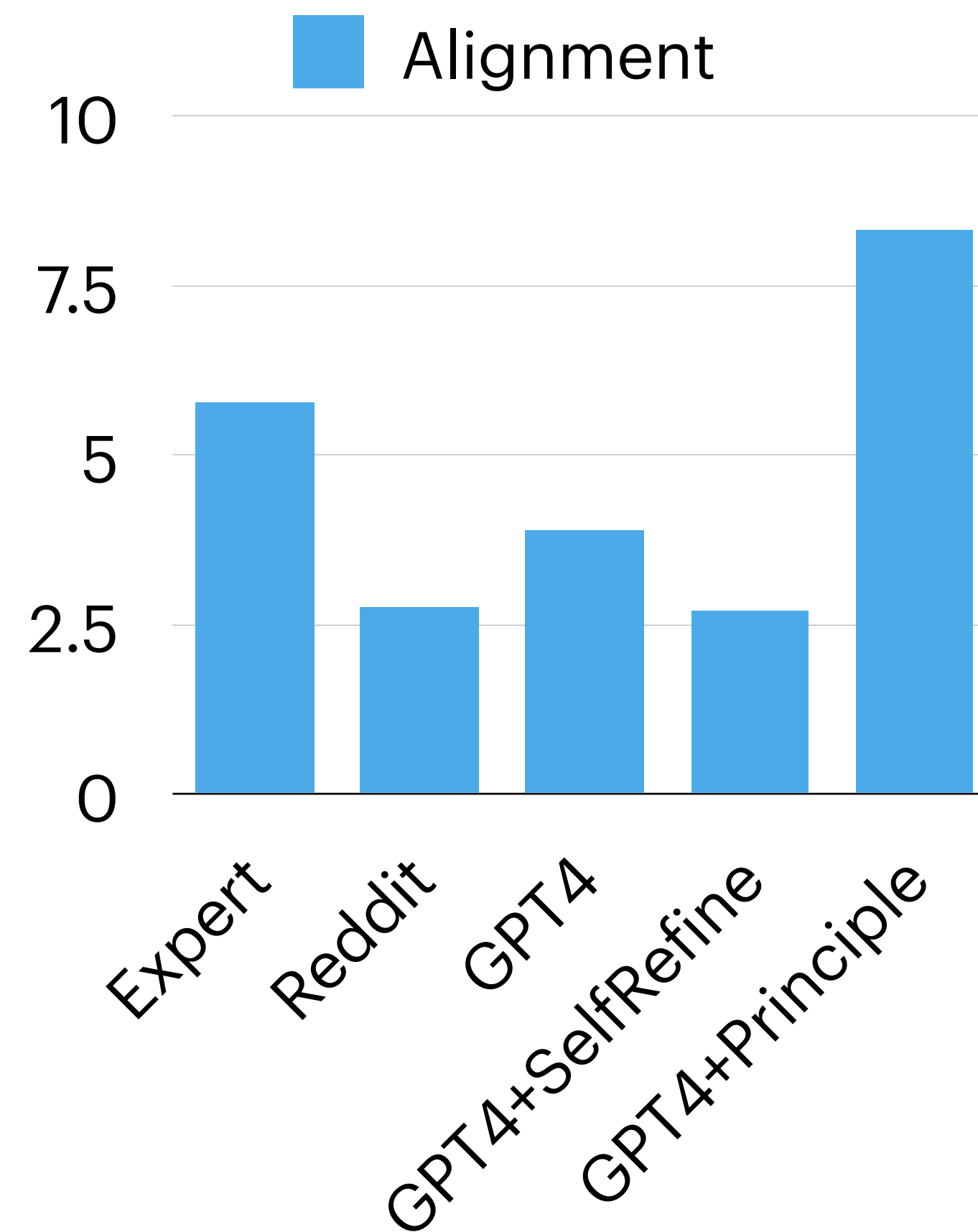
Problem-focused coping:

Does the narrator think that they can cope with the consequences of the situation?

Reappraisal goal: Focus on the narrators' competence (self-efficacy) to handle the situation at hand. The narrator is encouraged to use any resources or support to handle the situation competently and independently.

How to best utilize expertise?

Expert writing a response vs expert writing *principles*: psychologist evaluation



An era where expert input is critical

Analysis and annotation:
Experts help reveal where
models lack fundamentally, even
when outputs are generally
perceived as high quality

Experts can weigh in on LLM
alignment: sometimes models
benefit more from well-curated
principles than gold answers!

This talk

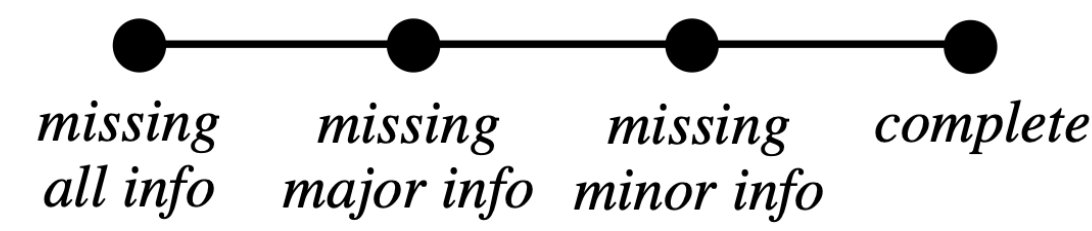
The roles of experts in corpora development, annotation, and evaluation

- Case studies in language and alignment

How can LLMs help the annotation process itself?

- Explanation-based rescaling (EBR)

Task: evaluating the answer completeness to high-level questions in a document.




Question: Why was the entire climbing season in doubt?
Answer: The Sherpas had walked out in protest of the deaths of 16 of their colleagues in an avalanche, and their demands for better pay, treatment and benefits.

Article: (1) KATMANDU, Nepal - Dozens of Sherpa guides packed up their tents and left Mount Everest's base camp Wednesday, after the deaths of 16 of their colleagues in an avalanche exposed an undercurrent of resentment by Sherpas over their pay, treatment and benefits. (2) With the entire climbing season increasingly thrown into doubt, the government quickly announced that top tourism officials would fly to base camp Thursday to negotiate with the Sherpas and encourage them to return to work. [...] (8) It was unclear whether they would return to work if the government accepts all their demands. [...] (10) But the Sherpas said they deserved far more - including more insurance money, more financial aid for the victims' families and new regulations to ensure climbers' rights. (11) Without the help of the Sherpas, who are key guides and also haul tons of gear up the mountain, it would be nearly impossible for climbers to scale Everest. [...] (15) 'It is just impossible for many of us to continue climbing while there are three of our friends buried in the snow,' said Dorje Sherpa, an experienced Everest guide from the tiny Himalayan community that has become famous for its high-altitude skills and endurance.[...] (40) The insurance payout for those killed in the avalanche, which now stands at \$10,400, will also be increased to \$15,620, or 2 million rupees, the Ministry of Tourism said - far less than the Sherpas' demand for \$20,800. [...] (43) Hundreds of people have died trying.

 **Label:** missing major

 **Label:** complete

 **Label:** complete

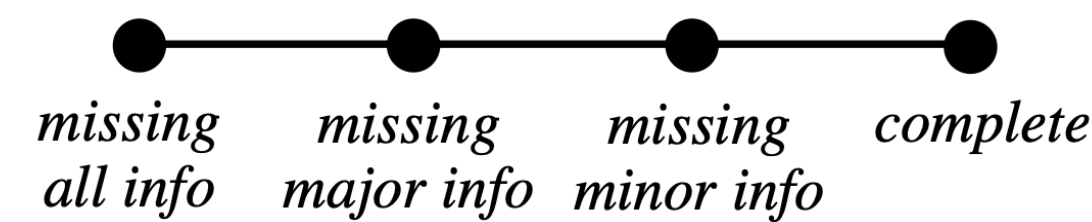
 **Label:** missing minor

 **Label:** missing minor

High disagreement/
no majority label!

- Noise & low quality labels?
- Unclear guidelines?
- *R2: “inter-annotator agreement is low, dataset is not high-quality”, score: 2*

Task: evaluating the answer completeness to high-level questions in a document.



Question: Why was the entire climbing season in doubt?
Answer: The Sherpas had walked out in protest of the deaths of 16 of their colleagues in an avalanche, and their demands for better pay, treatment and benefits.

Article: (1) KATMANDU, Nepal - Dozens of Sherpa guides packed up their tents and left Mount Everest's base camp Wednesday, after the deaths of 16 of their colleagues in an avalanche exposed an undercurrent of resentment by Sherpas over their pay, treatment and benefits. (2) With the entire climbing season increasingly thrown into doubt, the government quickly announced that top tourism officials would fly to base camp Thursday to negotiate with the Sherpas and encourage them to return to work. [...] (8) It was unclear whether they would return to work if the government accepts all their demands. [...] (10) But the Sherpas said they deserved far more - including more insurance money, more financial aid for the victims' families and new regulations to ensure climbers' rights. (11) Without the help of the Sherpas, who are key guides and also haul tons of gear up the mountain, it would be nearly impossible for climbers to scale Everest. [...] (15) 'It is just impossible for many of us to continue climbing while there are three of our friends buried in the snow,' said Dorje Sherpa, an experienced Everest guide from the tiny Himalayan community that has become famous for its high-altitude skills and endurance.[...] (40) The insurance payout for those killed in the avalanche, which now stands at \$10,400, will also be increased to \$15,620, or 2 million rupees, the Ministry of Tourism said - far less than the Sherpas' demand for \$20,800. [...] (43) Hundreds of people have died trying.

- Label:** missing major
NLE: Important information was neglected, sentences 11 and 15 are missing
- Label:** complete
NLE: The government is unwilling to give the Sherpas a decent wage or benefits for risking their lives.
- Label:** complete
NLE: Sentence one contains the needed information and it is summarized in the answer.
- Label:** missing minor
Sentences Missing: 10, 15, 40
NLE: The machine response answered the question correctly but missed some relevant information that clarifies the Sherpas' demands.
- Label:** missing minor
Sentences Missing: 8
NLE: It is also unknown if the Sherpas would accept and return.

High disagreement/
no majority label!

- People tend to focus on different sentences
- People have different internal scales: the reasoning and the labels are not always consistent across annotators

We as a field have embraced this

The Reliability of Anaphoric Annotation, Reconsidered: Taking Ambiguity into Account

Did It Happen? The Pragmatic Complexity of Veridicality Assessment

Marie-Catherine de Marneffe*
Stanford University

Christopher D. Manning**
Stanford University

Christopher Potts†
Stanford University

Discourse Structure and Computation: Past, Present and Future

**The Good, the Bad, and the Disagreement:
Complex ground truth in rhetorical structure analysis**

Debopam Das
Dept. of Linguistics
University of Potsdam
Potsdam, Germany
debidas@uni-potsdam.de

Communication Science
Potsdam, Germany

Subjectivity in the Annotation of Bridging Anaphora

Lauren Levine and Amir Zeldes
Georgetown University
Department of Linguistics
{lel76, amir.zeldes}@georgetown.edu

We as a field have embraced this

The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation

Barbara Plank

Center for Information and Language Processing (CIS), MaiNLP lab, LMU Munich, Germany

Munich Center for Machine Learning (MCML), Munich, Germany

b.plank@lmu.de

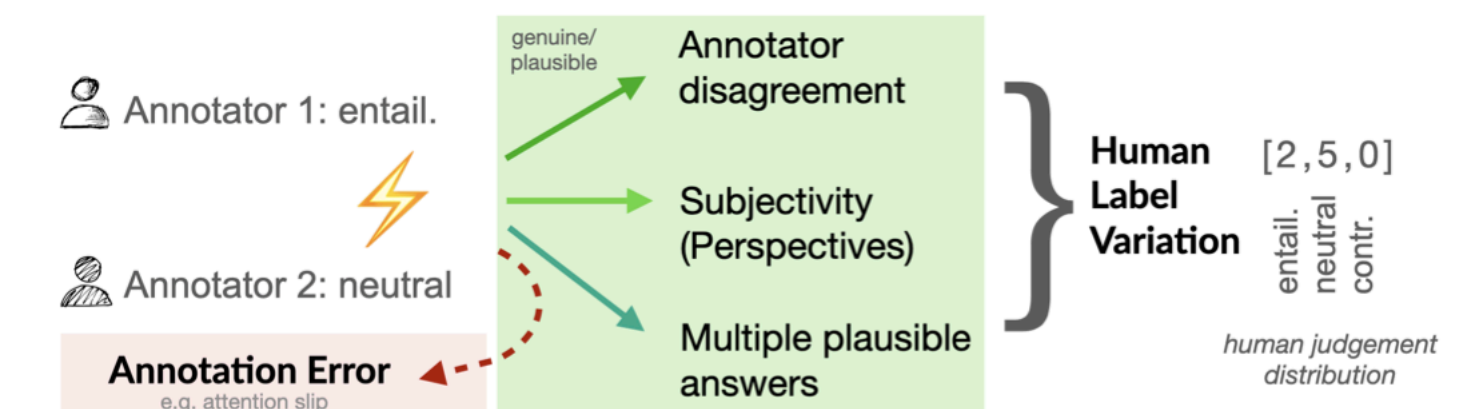


Figure 1: We propose the term *human label variation* to capture the fact that inherent disagreement in annotation can be due to genuine disagreement, subjectivity or simply because two (or more) views are plausible.

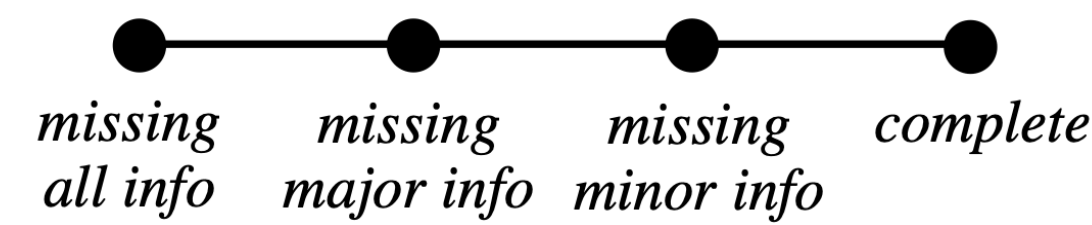
We have found some solutions (?)

- Existing methods focus on:
 - Filtering (surveyed in Paun et al., 2022)
 - Aggregation (Dawid & Skene, 1979, among others)
 - Multiple ground truths (Sheng et al., 2008, Pavlick and Kwiatkowski 2019, among others)
- But:
 - Filtering & aggregation seek to “smooth out” subjectivity.
 - Even with multiple labels, nuances and variations encoded in each label is still lost!

Label: missing minor
Sentences Missing: 8
NLE: It is also unknown if the Sherpas would accept and return.

Label: missing minor
Sentences Missing: 10, 15, 40
NLE: The machine response answered the question correctly but missed some relevant information that clarifies the Sherpas' demands.

Task: evaluating the answer completeness to high-level questions in a document.



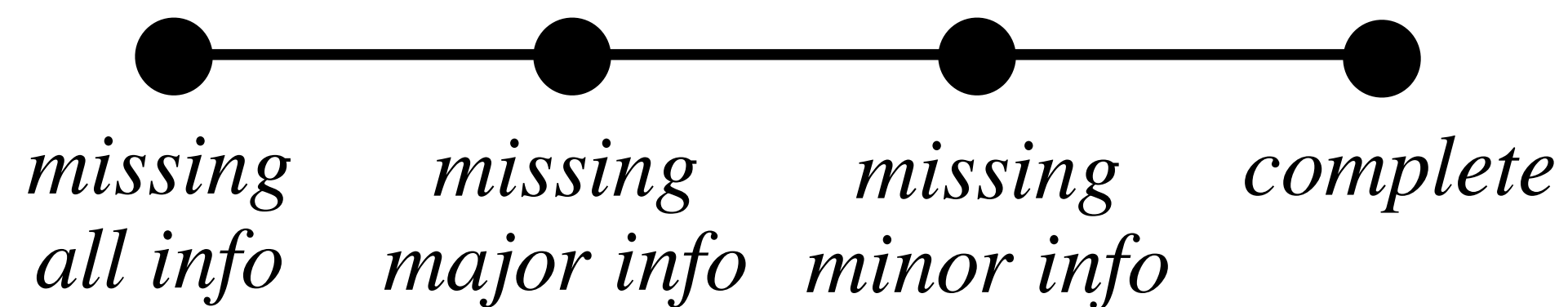
<p>Label: missing major</p> <p>NLE: Important information was neglected, sentences 11 and 15 are missing</p>	50
<p>Label: complete</p> <p>NLE: The government is unwilling to give the Sherpas a decent wage or benefits for risking their lives.</p>	100
<p>Label: complete</p> <p>NLE: Sentence one contains the needed information and it is summarized in the answer.</p>	100
<p>Label: missing minor</p> <p>Sentences Missing: 10, 15, 40</p> <p>NLE: The machine response answered the question correctly but missed some relevant information that clarifies the Sherpas' demands.</p>	75
<p>Label: missing minor</p> <p>Sentences Missing: 8</p> <p>NLE: It is also unknown if the Sherpas would accept and return.</p>	85

What if we rescale this to a finer-grained scale?

- Ordering of the original labels preserved
- Nuances within categories reflected

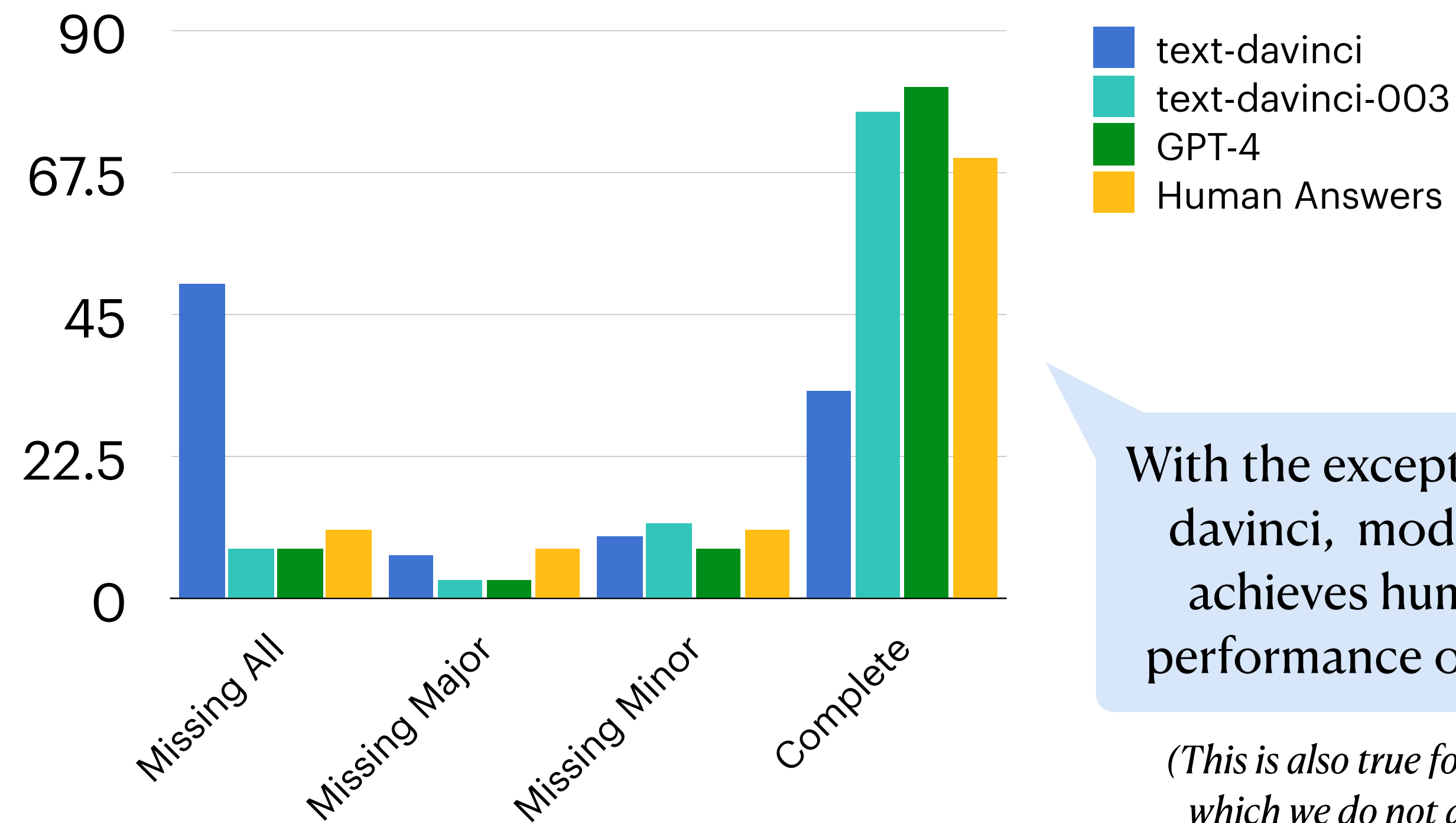
Task Setup

- Task to evaluate: models answering high level questions given a document
 - **Causal** (e.g., why are they now liberalizing bank laws?), **procedural** (e.g., how will the exceptions to the verification concept be taken care of?), **background** (e.g., what is the main focus of this movement?), **instantiation** (e.g., which groups were the human rights activists working on behalf of?), etc.
- Judgments:



Data Collection

- 12.6k QA pairs from 8 vetted crowd workers provided ratings + rationales
- Third party (expert) validation of the annotations



Data Collection

- But human agreement is not high: Kappa **0.328**, Kendall's Tau-b **0.325**

Worker ID	missing all	missing major	missing minor	complete
0	5%	8%	18%	69%
1	22%	10%	25%	43%
6	32%	5%	15%	48%

Label: missing major

Explanation: The text does not provide a specific answer, but a lot of detail could have been included in an attempt to address the question.

Label: missing minor

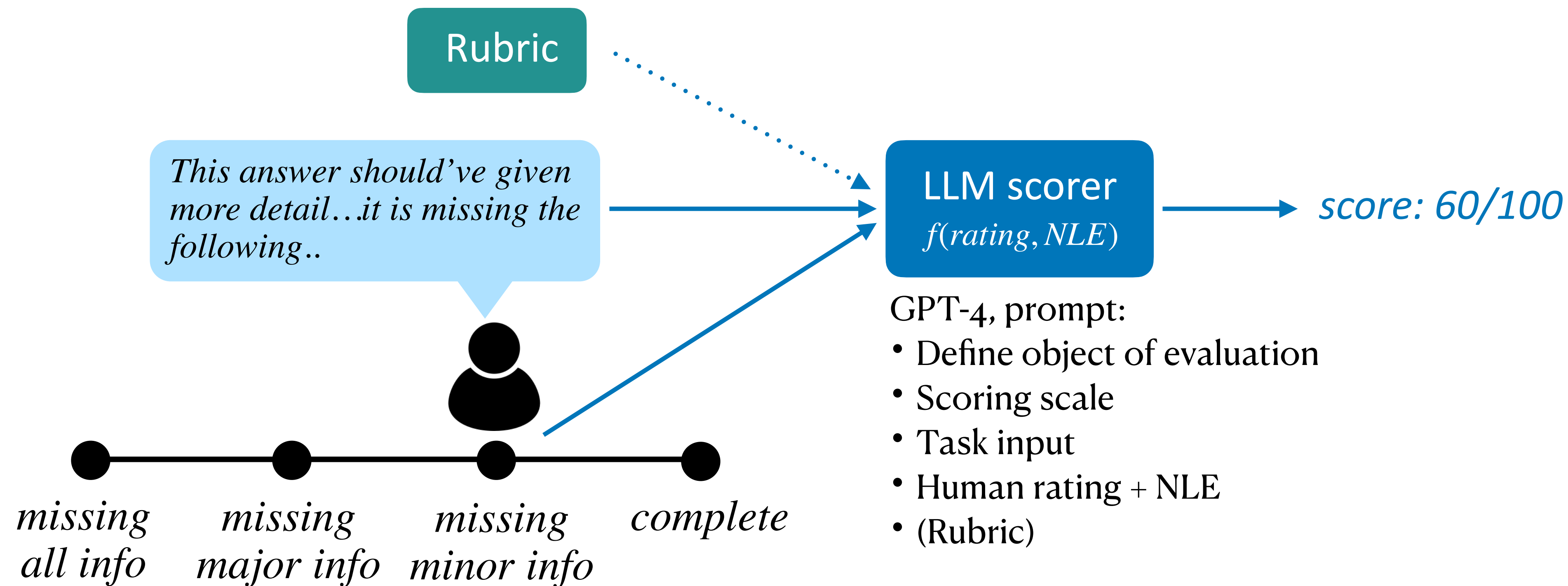
Explanation: Some additional relevant information was included in the article.

Annotators have different internal scales:
#0 is much more lenient than the other two

Annotators can differ in their label decision
but agree on details in their explanations.

Explanation-Based Rescaling

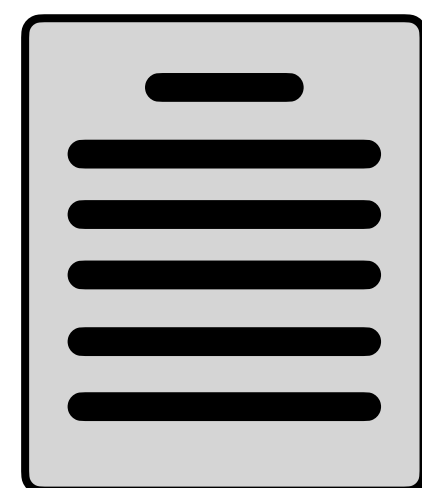
Key idea: project {original score, NL explanations} onto a more fine-grained scale



Rubric Creation

Discovering nuances post-annotation

Pre-annotation instructions



Label: ...
NLE: ...

Label: ...
NLE: ...

Label: ...
NLE: ...

Label: ...
NLE: ...

Label: ...
NLE: ...

Small scale post-annotation NLE analysis

- Missing minor NLE:
 - Missing names/entities
 - Mentioning “details”
- Missing major NLE:
 - Highlighting the role of the missing sentences
 - Explicitly states how much information is missing

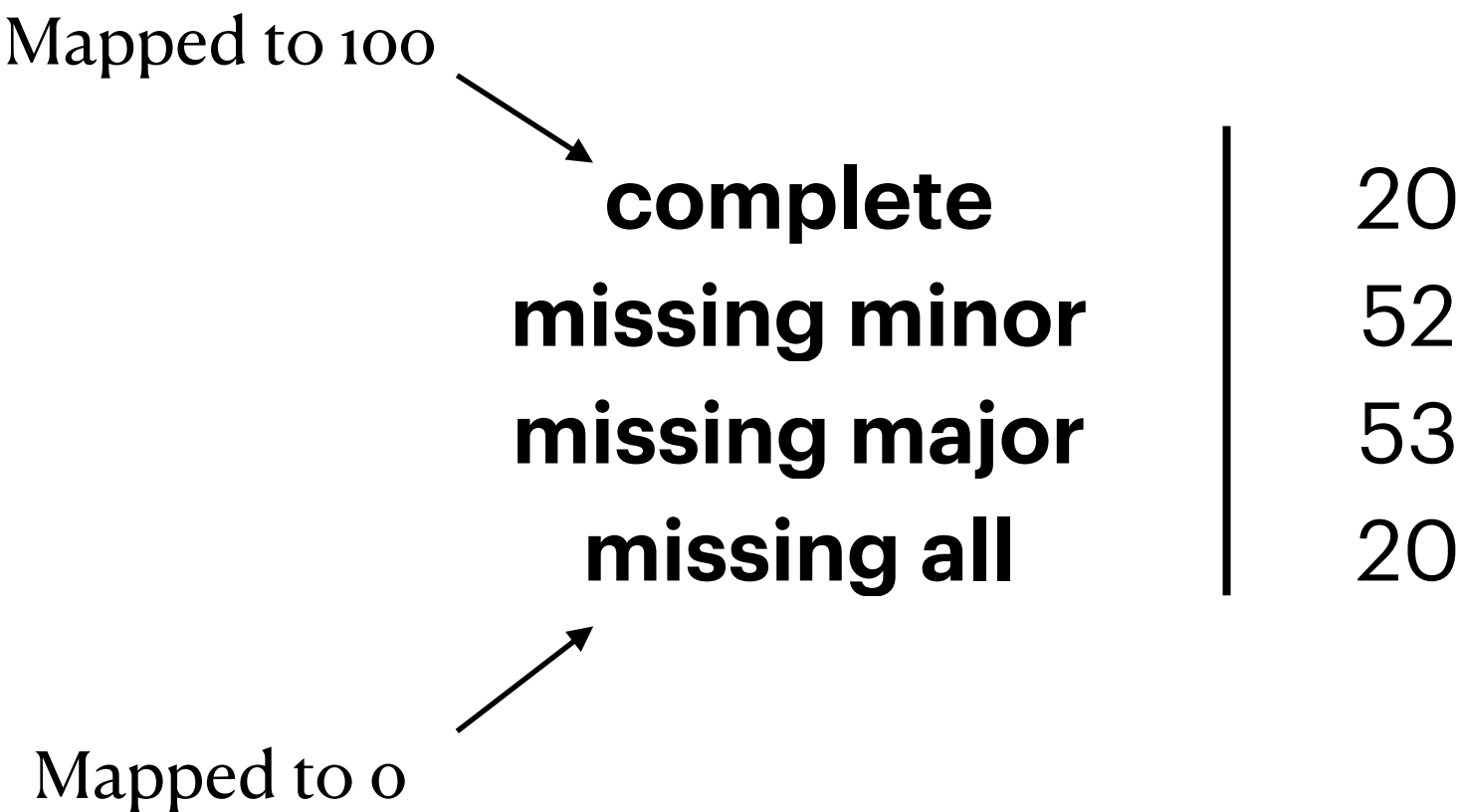
Rubric creation

On a scale of 0-100 how will you score machine response using the feedback and level of missing information stated above? Use the rubric below for scoring:

1. if the answer is complete, give 100 points
2. if the answer is missing one or more minor details then have deductions ranging from 5 to 30 points based on the severity of missing details
3. if the answer is missing a major facet of information, it results in a deduction of at least 40 points and more than 50 points are deducted if less than half of the correct information was given.
4. if the answer contains no correct information but only marginally relevant information from the article, 70 points are deducted
5. if the answer contains no correct information but the article clearly has information present, 100 points are deducted

Evaluation

- Reference Scores:** 145 instances rescaled by 3 experts

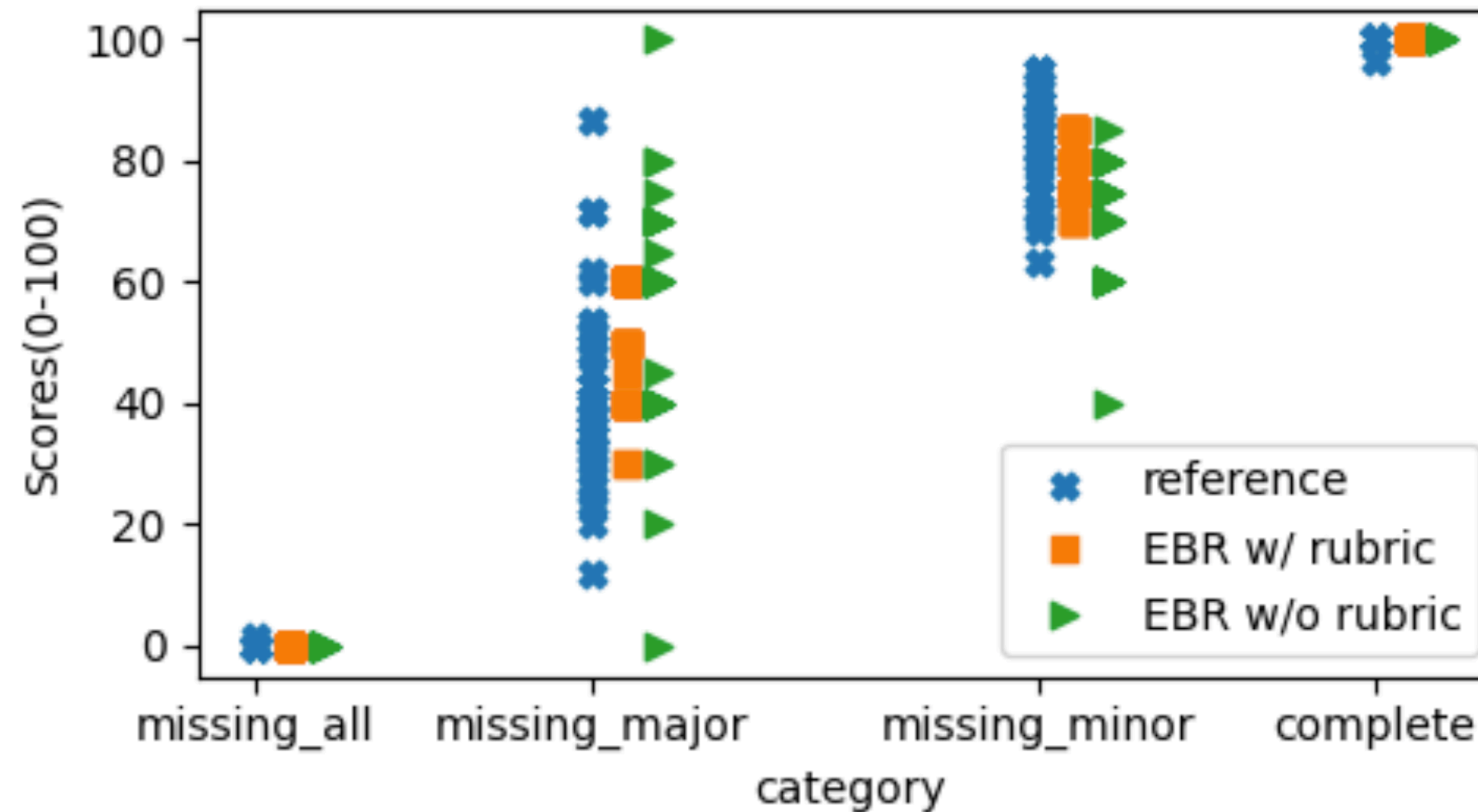


EBR retains the correlation with original likert-scale and with experts

	Kendall’s Tau-b	Mean Absolute Error (MAE)
Static Mapping	0.85	10.1
Missing sentence heuristic	0.42	24.2
No rubric EBR	0.69	12.9
EBR	0.83	8.1

EBR scores are closer to expert rescaling

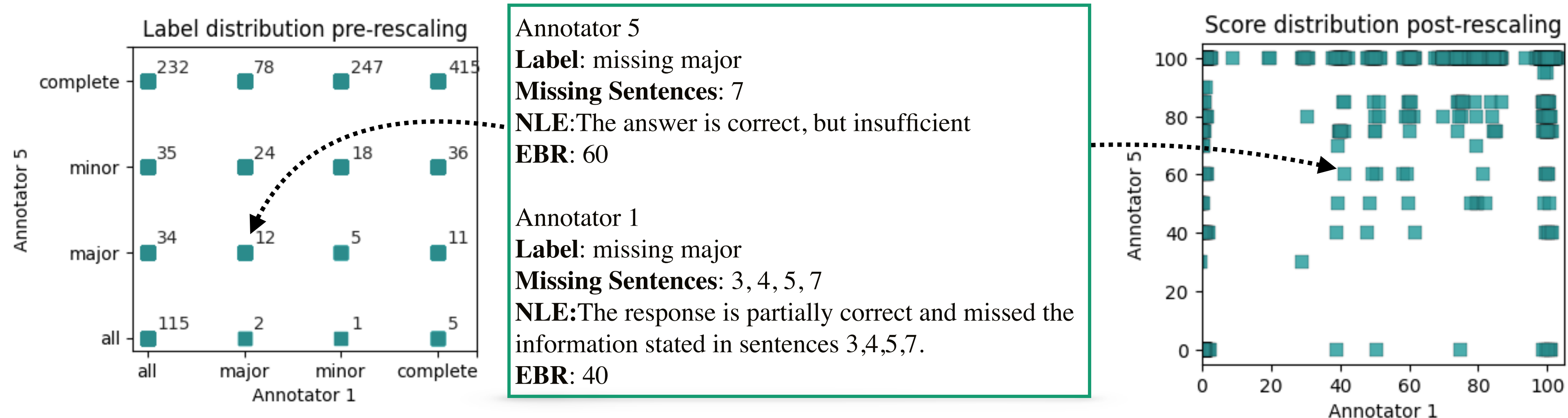
The role of a good rubric



Rubrics make the scores more consistent

What's happening with the scores?

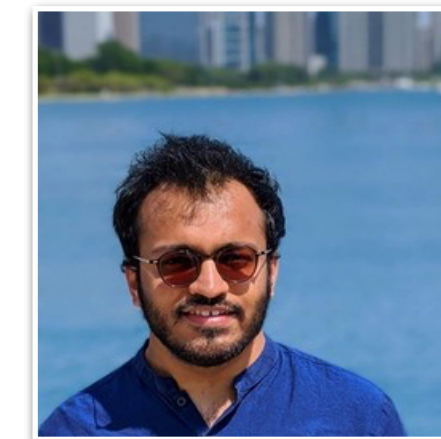
A case study



Note: EBR does not impact pairwise annotator correlations: Kendall's tau 0.33 (before) -> 0.32 (after)

Takeaways

- LLMs rescale annotator judgments effectively
- Rescaled annotations align with expert reference rescaling
- Rescaling preserves correlation, capturing nuances, subjectivity and scale use differences
- Try this out at https://github.com/ManyaWadhwa/explanation_based_rescaling



Thank you!

