Human label variation in NLP

Marie-Catherine de Marneffe

FNRS – UCLouvain – CENTAL

AθNLP Summer School, September 9 2025

Go to **wooclap.com** and use the code **RMTMGU**

What color?

What color?

Natural Language Inference (NLI)

P: Dana Reeve, the widow of the actor Christopher Reeve, died of lung cancer at age 44.

H: Dana Reeve had a fatal disease.

"Entailment" – True
Neutral
Contradiction – False

?

Natural Language Inference (NLI)

P: The park was established in 1935 and was given Corbett's name after India became independent.

H: The park used to be named after Corbett.

"Entailment"
Neutral
Contradiction

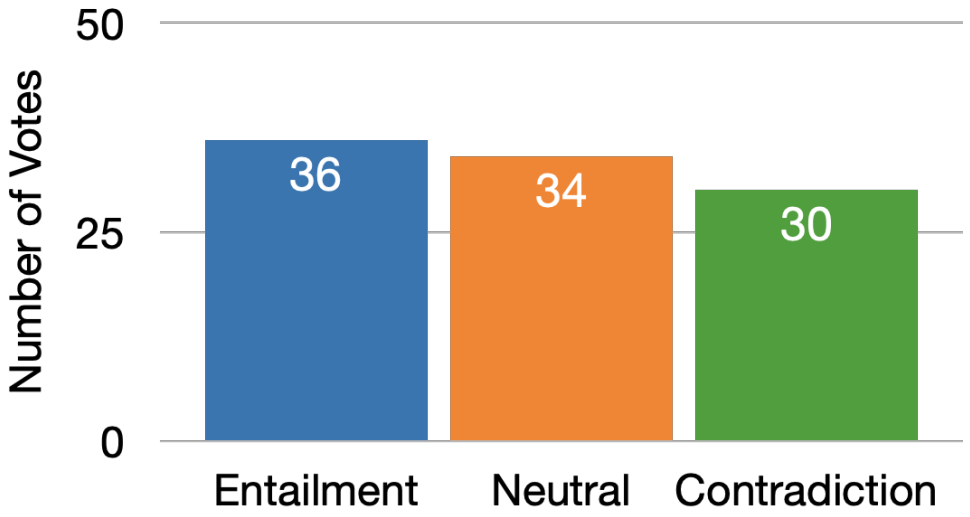
?

“Human label variation”

[term from Plank 2022]

P: The park was established in 1935 and was given Corbett's name after India became independent.

H: The park used to be named after Corbett.



[Pavlick and Kwiatowski 2019, Nie et al 2020]

NLI as one of the fundamental NLU tasks

SuperGLUE benchmark [Wang et al. 2020]

Natural language inference

Question answering

Word sense disambiguation

Coreference

“Truth is a lie: Crowd truth and the 7 myths of human annotation”

[Aroyo & Welty 2015]

1. One truth
2. **Disagreement is bad**
3. Detailed guidelines help
4. **One is enough**
5. Experts are better
6. **All examples are created equal**
7. Once done, forever valid

Old problem, but put aside

Penn Treebank POS Tagging [Santorini 1990]

a. Sampling data can be time-consuming.

b. Sampling data can be full of errors.

c. Sampling data can be fun.

Old problem, but put aside

Penn Treebank POS Tagging [Santorini 1990]

a. Sampling/*VBG* data can be time-consuming.

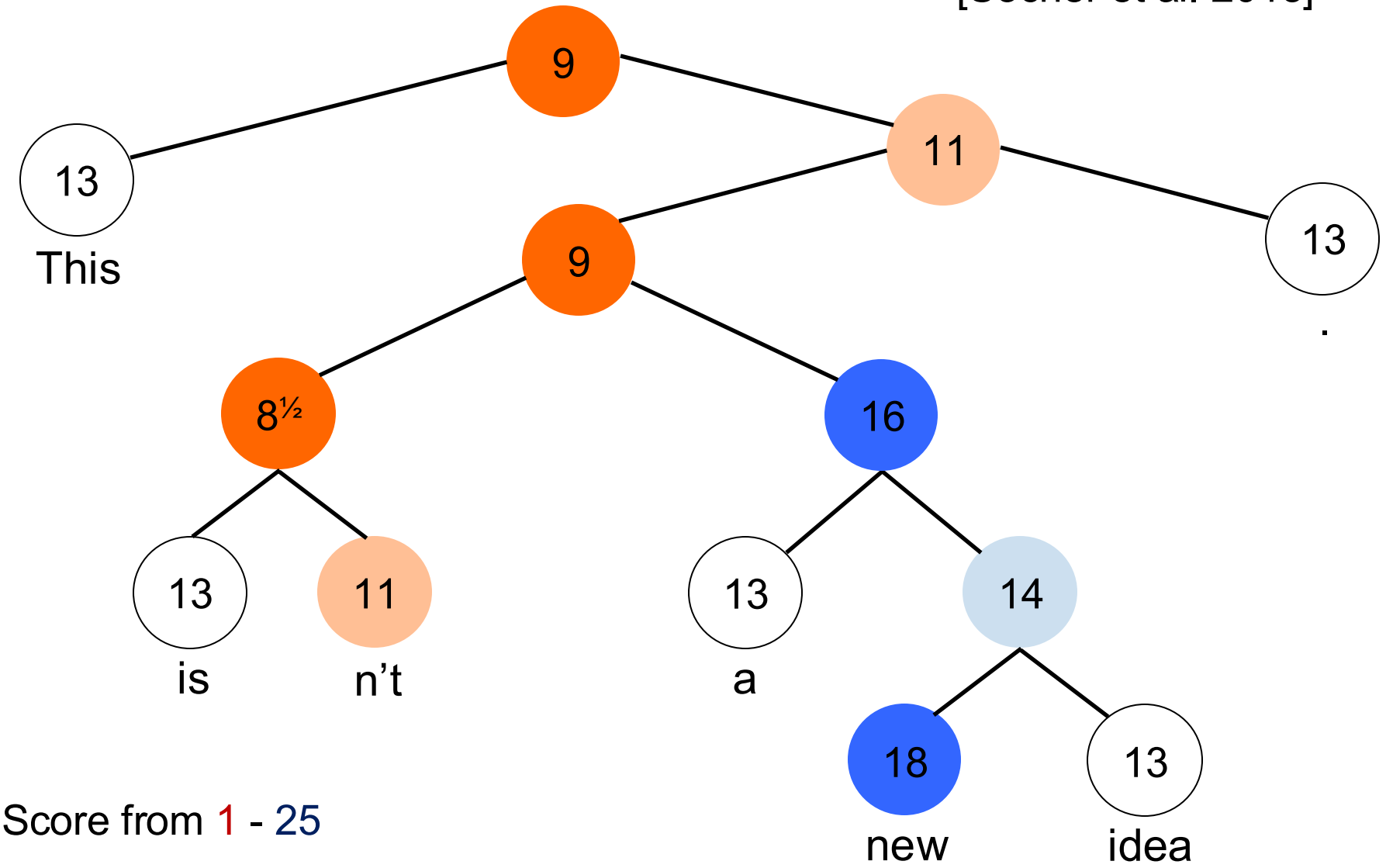
b. Sampling/*NN* data can be full of errors.

c. Sampling/*NN*|*VBG* data can be fun.

"Nevertheless, uncertainties can arise. Rather than attempting to forcibly resolve such uncertainties, with the attendant risk of inconsistency, you should simply record them by separating the relevant tags by a vertical slash" (p. 31-32)

Stanford Sentiment Treebank

[Socher et al. 2013]



Score from 1 - 25

a film as Byatt fans could hope for	13,15,15
A sly dissection of the inanities of the contemporary music business and a rather sad story of the difficulties	13,15,15

Phrases with sentiment score 0.56

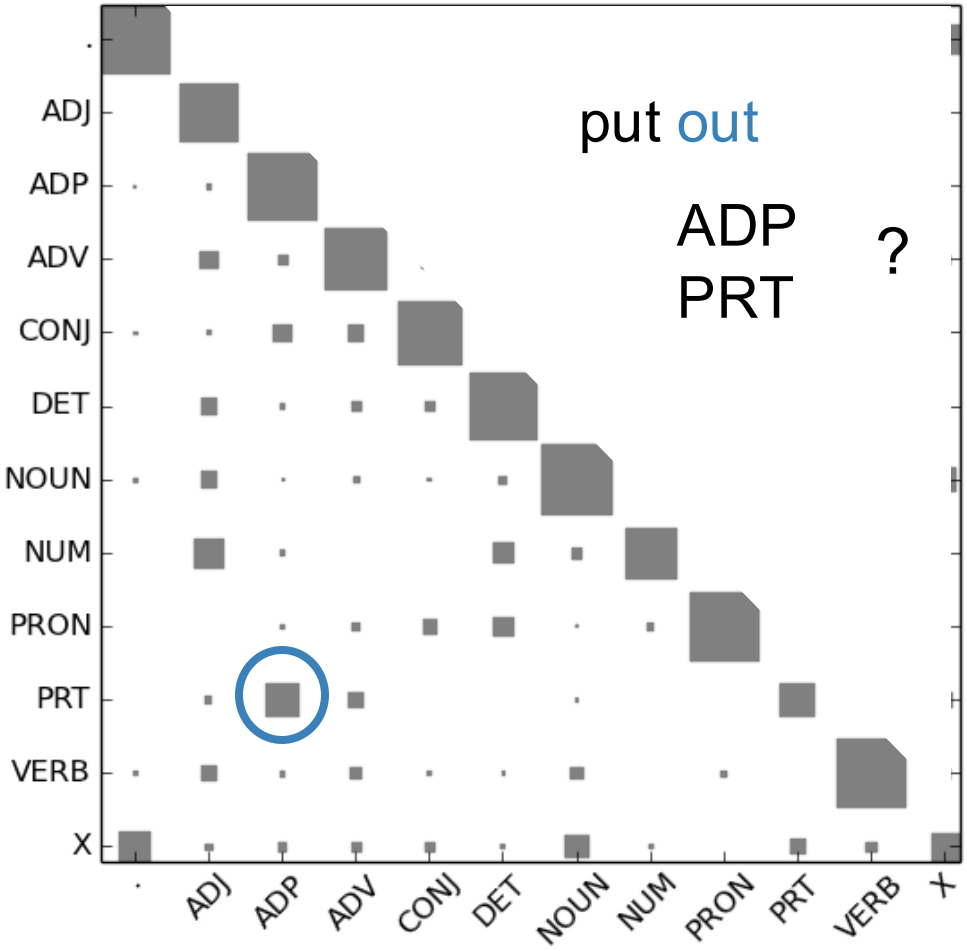
a film as Byatt fans could hope for	13,15,15
A sly dissection of the inanities of the contemporary music business and a rather sad story of the difficulties	13,15,15

---

a sad, superior human	9,13, 21
a joke	8,15, 20
Angel presents events partly from the perspective of Aurelie and Christelle and infuses the film with the sensibility of a particularly nightmarish fairytale	9,17 17

Disagreement is signal, not noise

[de Marneffe et al. 2012, Aroyo and Welty 2013, Plank et al. 2014, Passonneau and Carpenter 2014, Artstein 2017]



Embrace “disagreement” in NLI

1. What explains the variation in NLI annotations?
   Which linguistic phenomena are involved?

2. How to best represent variation?

3. Do LLMs capture such variation?

1. Taxonomy of "disagreement" sources

Probabilistic Enrichment

P: Oh, sorry, wrong church.

H: He or she entered the wrong church.

[E,N,C]: [82, 17, 1]

Coreference

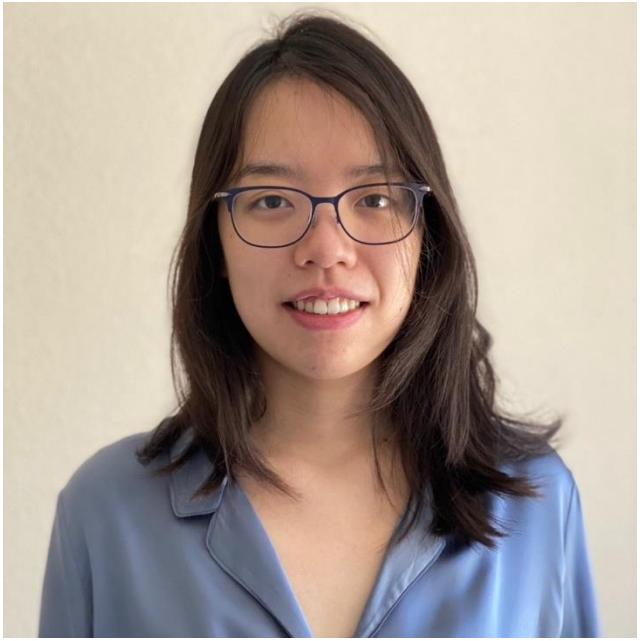
P: Cruises are available from the Bansi Ghat, which is near the City Palace in India.

H: You can take cruises from Phoenix, Arizona.

[E,N,C]: [0, 51, 49]

Investigating reasons for disagreement in NLI

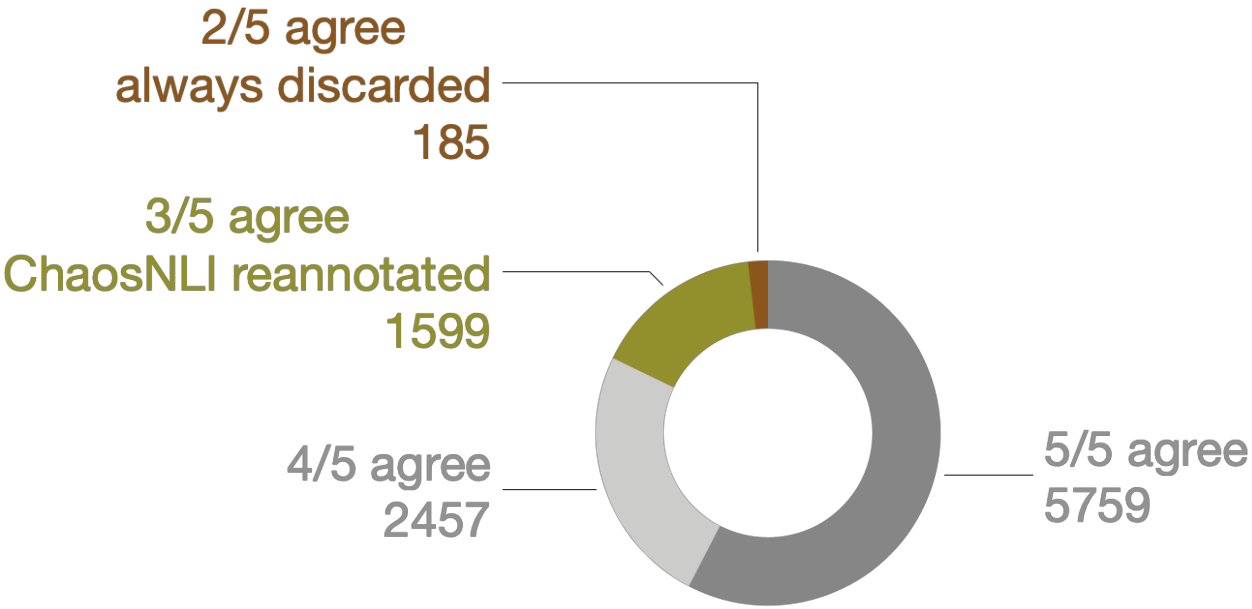
Nan-Jiang Jiang & MC de Marneffe TACL 2022



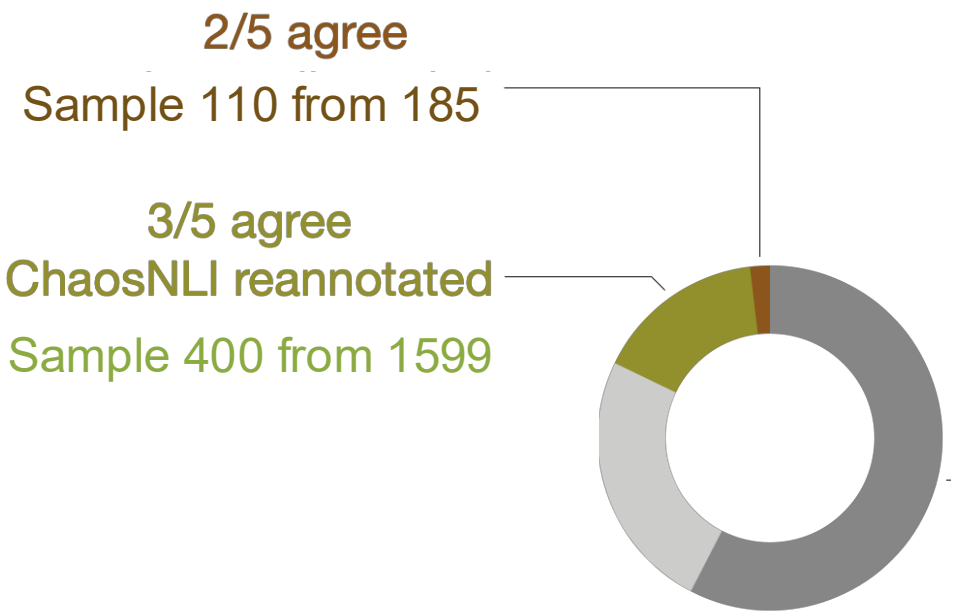
MNLI data dev matched set

10k items, 5 annotations/item [Williams et al. 2018]

ChaosNLI reannotated 3/5 agree, with 100 annotations/item [Nie et al. 2020]

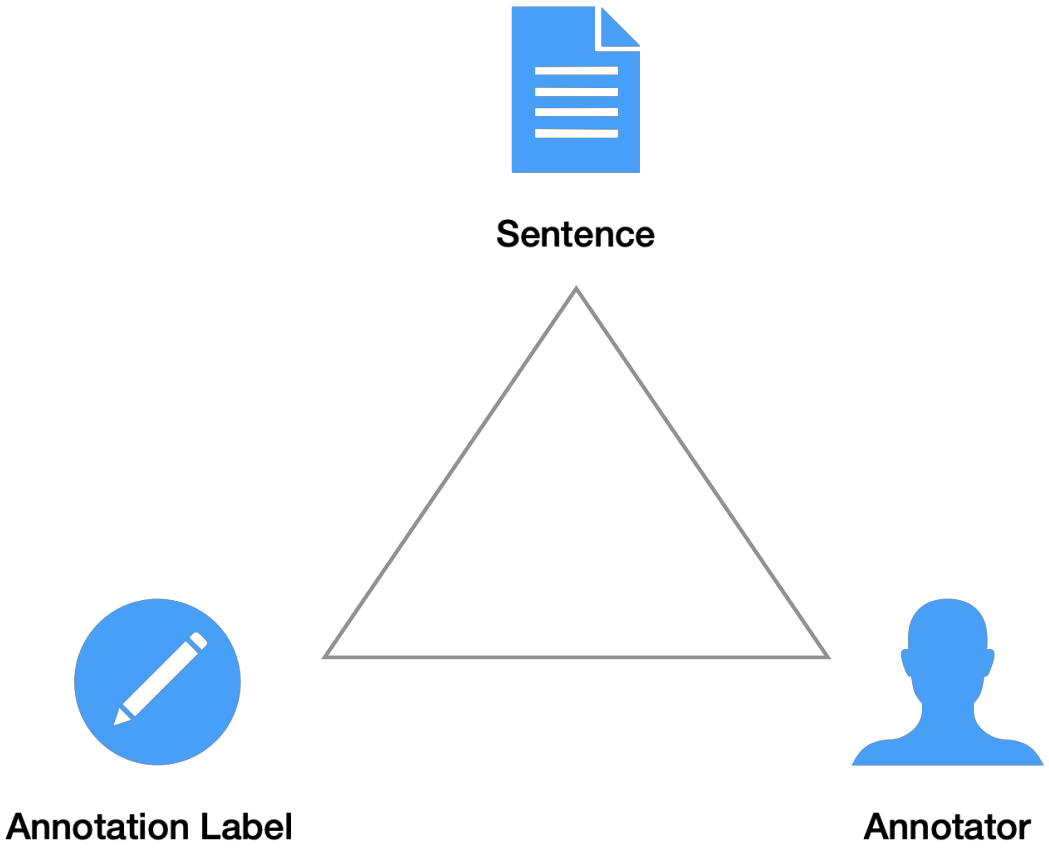


Data to develop the taxonomy: 510 items

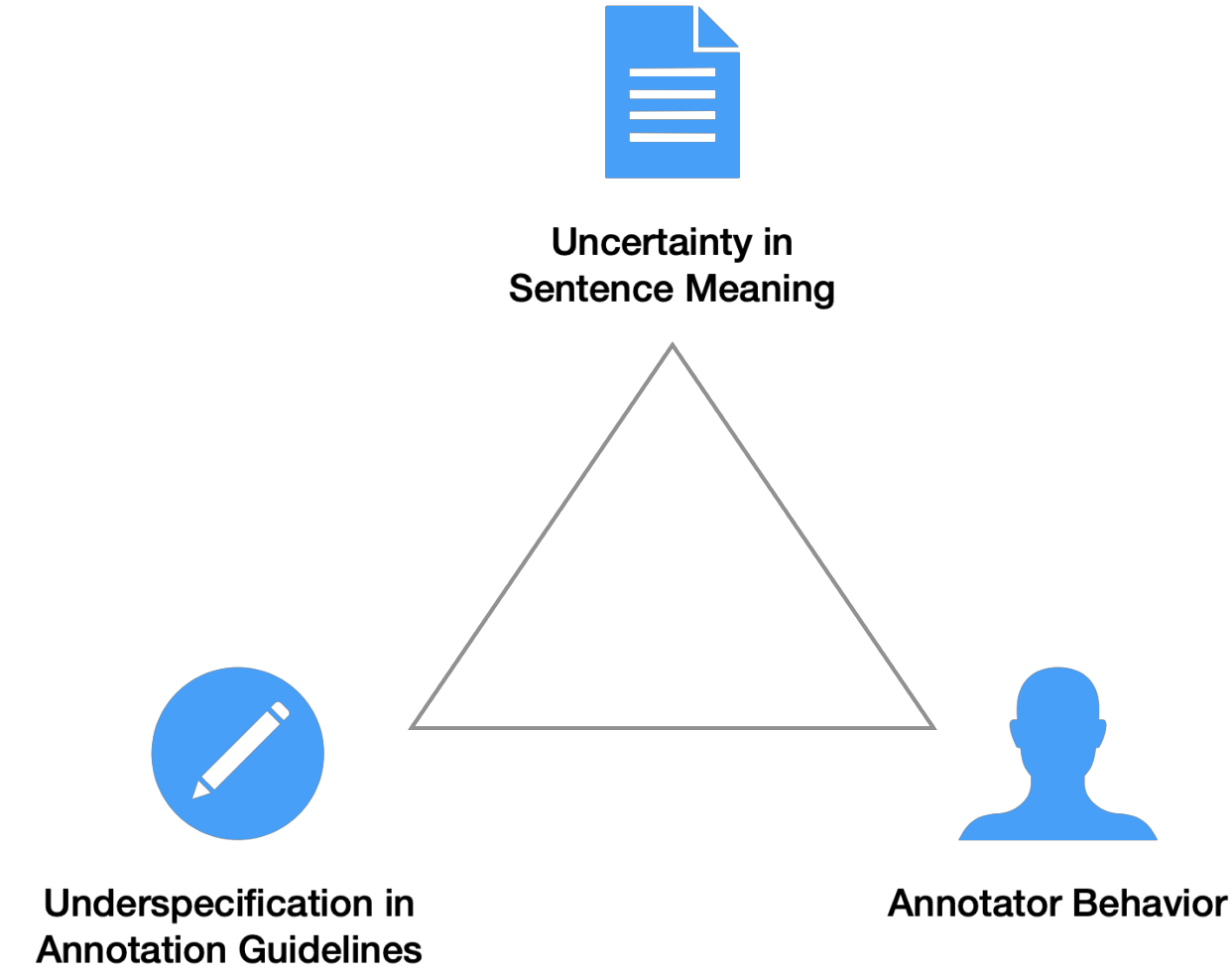


Triangle of reference

[Ogden & Richards 1923, Aroyo & Welty 2013]



Sources of variation in the annotations



Annotation process

Round 1
Annotator #1
annotated 400 items
— multi-category

developed the taxonomy

wrote the guidelines

Round 2
Annotator #1 & #2
annotated 110 items
(not used for taxonomy development)

0.69 Krippendorff's alpha with MASI distance

10 categories of “disagreement” sources

Uncertainty in Sentence Meaning

- Lexical
- Implicature
- Presupposition
- Probabilistic Enrichment
- Imperfection

Underspecification in Annotation Guidelines

- Coreference
- Temporal reference
- Interrogative Hypothesis

Annotator Behavior

- Accommodating minimal underspecified content
- High overlap

Uncertainty in sentence meaning

	Premise	Hypothesis	[E, N, C]
Lexical	**Technological** advances **generally** come in waves that crest and eventually subside.	Advances in **electronics** come in waves.	[82, 17, 1]
Implicature	[…] **some of the most authentic papyrus are sold at The Pharaonic Village** in Cairo […]	The Pharaonic Village in Cairo is the **only** place where one can buy authentic papyrus.	[20, 39, 41]
Presuppos.	What changed?	Nothing changed.	[4, 76, 20]
Probabilistic Enrichment	It's absurd but I can't help it. **Sir James nodded** again.	Sir James thinks it's absurd.	[61, 39, 0]
Imperfection	profit rather	Our profit has not been good.	[3, 90, 7]

Underspecification in guidelines

	Premise	Hypothesis	[E, N, C]
Coreference	The original **wax** models of the river gods are on display in the Civic Museum.	They have models made out of **clay**.	[5, 38, 57]
Temporal Reference	However, co-requesters cannot approve additional co-requesters or restrict the timing of the release of the product after it is issued.	They cannot restrict timing of the release of the product.	[90, 8, 2]
Interrogative Hypothesis	was it bad	Was it not good?	[84, 16, 0]

Annotator behavior

	Premise	Hypothesis	[E, N, C]
Accommodating minimal underspecified content	Indeed, 58 percent of Columbia/HCA's beds lie empty, compared with 35 percent of nonprofit beds.	58% of Columbia/HCA's beds are empty, **said the report**.	[97, 3, 0]
	After four years, Clinton has learned how to avoid looking unpresidential.	After four **torturous** years, Clinton finally gets how to avoid unpresidential behavior.	[49, 48, 3]
High overlap	Yet, in the mouths of the white townsfolk of Salisbury, N.C., it sounds convincing.	White townsfolk in Salisbury, N.C. think it sounds convincing.	[68, 27, 5]

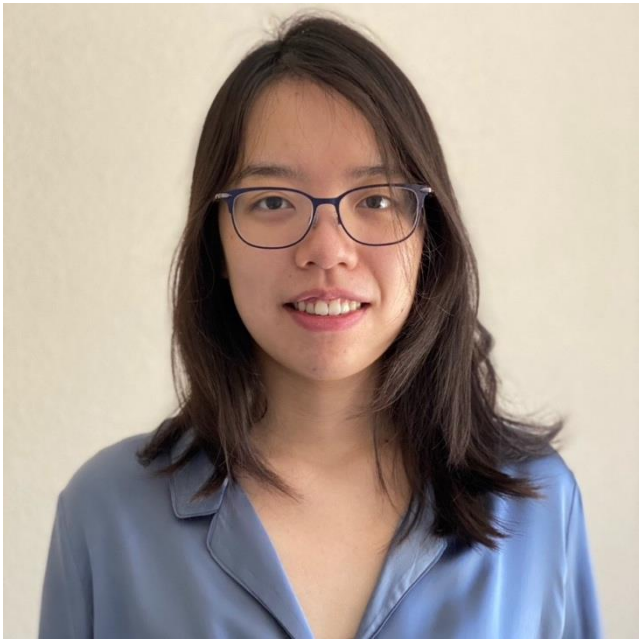
[Potts 2005, Simons et al. 2010, McNally 2016]

Does the taxonomy reflect people's reasoning?

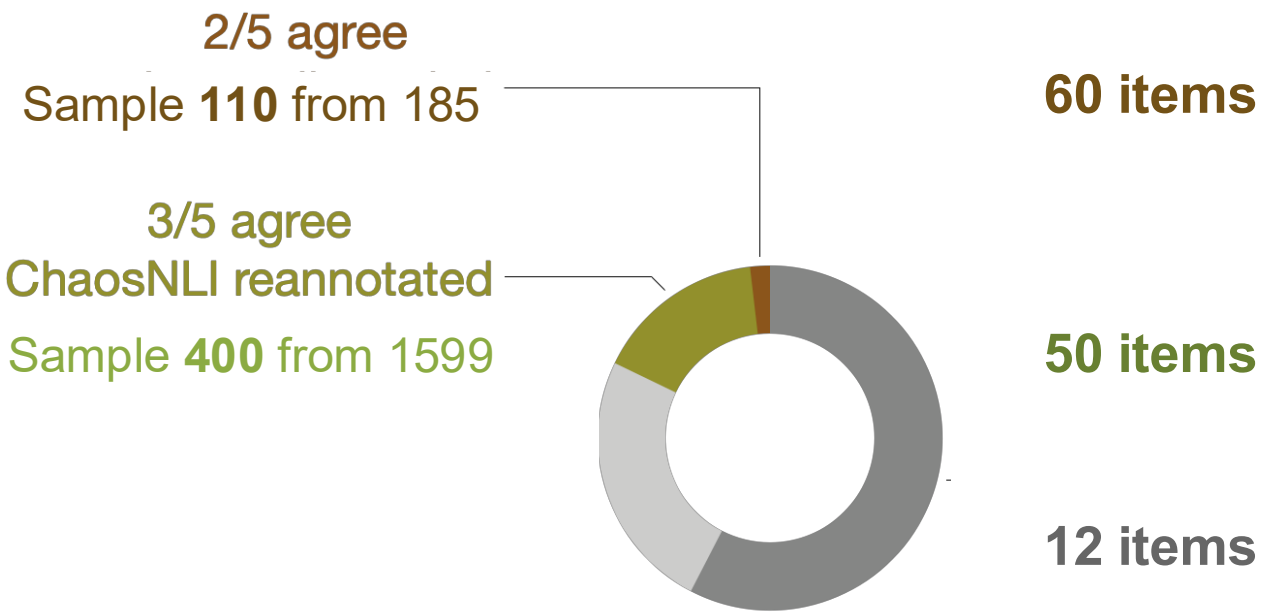
Ecologically valid explanations for label variation in NLI

Nan-Jiang Jiang, Chenhao Tan & MC de Marneffe

Findings of EMNLP 2023



LiveNLI: 122 MNLI items, gathering at least 10 annotations per item



Read the following context and statement:

Context: Could you please speak to this issue, with regard to the social ramifications of gum chewing in public?
Statement: You don't have an opinion on gum chewing in public, I see.

Choose one or more from the following:

If you feel uncertain and you feel that multiple options apply, choose them all instead, even though it might feel contradictory.
Assuming the context is true, the statement:

- [ ] is most likely to be true
- [ ] can be either true or false
- [ ] is most likely to be false

Explain, in a few sentences, why you chose your answer.

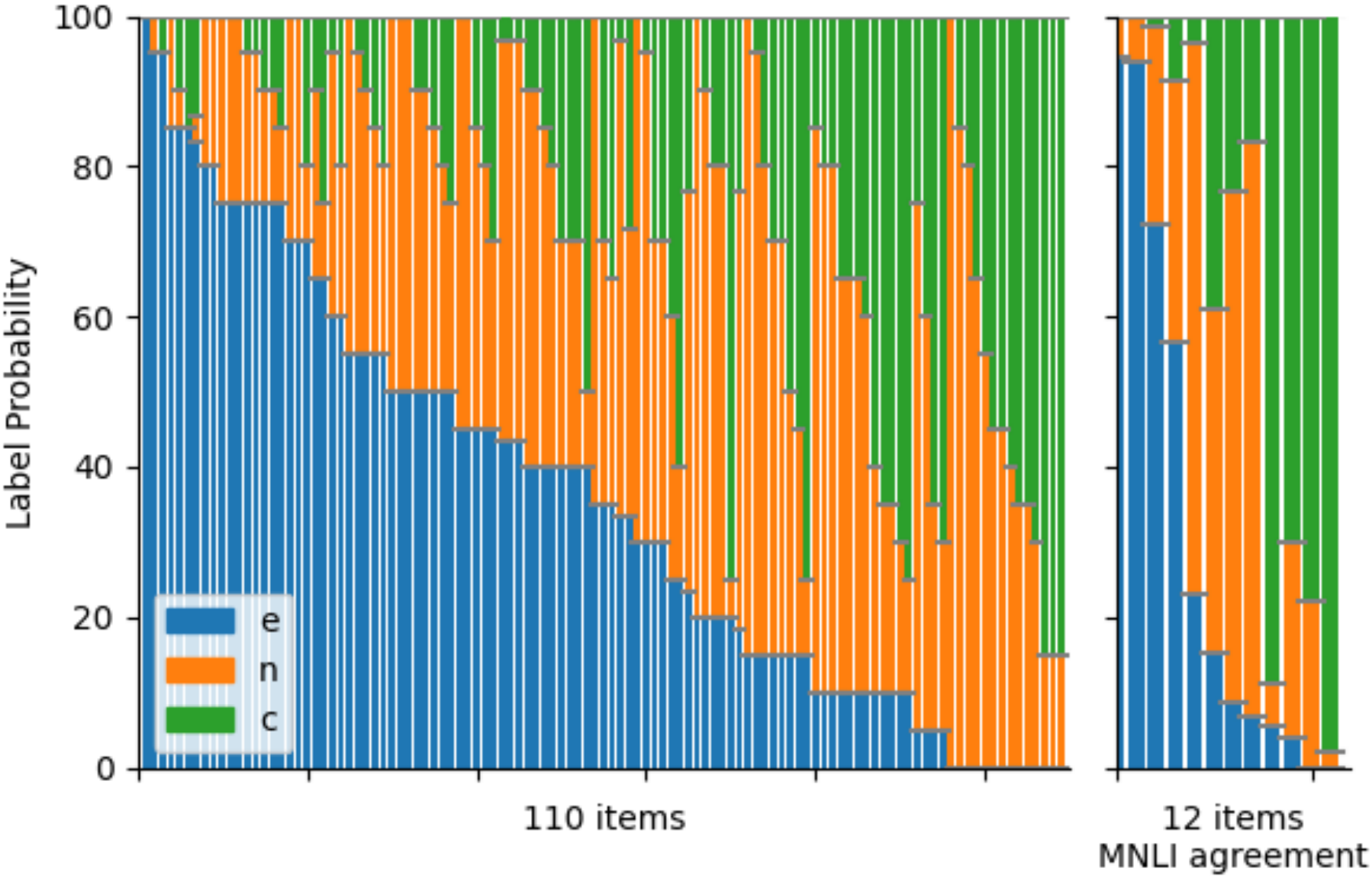
If you chose more than one option, elaborate in which circumstances each option is possible.
Explain all the options you chose.
Your explanation should include **new information** and **refer to specific parts of the sentences**. It should **NOT simply repeat the sentences**
Avoid "The context and statement means the same/opposite thing". **Specify which part of the context and statement means the same/opposite thing.**
Avoid "Just because X doesn't mean Y". **Say under what circumstances X does not mean Y, or say that X can mean Y or Z.**
Avoid "The statement is ambiguous/it's not clear what it means". **Elaborate what the possible meanings are and why it is ambiguous.**

Minimum word count: 10 Words: 0

Highlight the words in the Context and Statement that are relevant to your explanations.

Your explanations should refer to specific words/parts of the sentences. Highlight those words and phrases that your explanations mentioned.
Only highlight the words that are most important for the explanations.

Systematic variation persists



Annotators construe different QUDs [Roberts 2012]

Variation also comes from people judging the truth of different contents in the item (28 items out of 122)

P: Most pundits side with bushy-headed George Stephanopoulos, arguing that only air strikes would be politically palatable.
H: Mr. Stephanopoulos has a very large pundit following due to his stance on air strikes only being politically palatable. [0.4, 0.3, 0.3]

QUD: Does Stephanopoulos have a very large pundit following?
E – This hypothesis is most likely to be true because in the premise is stated that "Most pundits" would side with Mr. Stephanopoulos. Most pundits could also mean a very large pundit following.

QUD: Do pundits follow Stephanopoulos due to his stance on air strikes?
N – George Stephanopoulos may have a follow from pundits, but it might not be due to his support of drones.

Within-label variation

P: for a change i i got i get sick of winter just looking everything so dead i hate that
H: I'm so sick of summer. [0, 0.35, 0.65]

C – The speaker hates winter because the foliage is dead, therefore he likely loves summer when everything is alive.

C – The premise is stating how one is sick of winter, not summer, as the hypothesis describes.

N – The premise mentions being sick of winter while the hypothesis mentions being sick of summer. These could both be true because the same person may still complain of summer's heat.

2. How to best represent variation?

Train on the distribution and predict a distribution

Add a label to the standard ones [Kenyon-Dean et al. 2015]
4-way classification: E, N, C, Complicated

Multi-label classification [i.a., Passonneau et al. 2012]
One or more of E, N, C

Annotator behavior

	Premise	Hypothesis	[E, N, C]
Accommodating minimal underspecified content	Indeed, 58 percent of Columbia/HCA's beds lie empty, compared with 35 percent of nonprofit beds.	58% of Columbia/HCA's beds are empty, **said the report**.	[97, 3, 0]
	After four years, Clinton has learned how to avoid looking unpresidential.	After four **torturous** years, Clinton finally gets how to avoid unpresidential behavior.	[49, 48, 3]
High overlap	Yet, in the mouths of the white townsfolk of Salisbury, N.C., it sounds convincing.	White townsfolk in Salisbury, N.C. think it sounds convincing.	[68, 27, 5]

[Potts 2005, Simons et al. 2010, McNally 2016]

2. How to best represent variation?

Train on the distribution and predict a distribution

Add a label to the standard ones [Kenyon-Dean et al. 2015]
4-way classification: E, N, C, Complicated

Multi-label classification [i.a., Passonneau et al. 2012]
One or more of E, N, C

3. Do LLMs capture such variation?

Artur Kulmizev

Erika Lombart

Patrick Watrin

.

Frontier LLMs from this summer

High-end Models current SOTA	**Low-end Models** small models (8B-20B) released before Dec 2024
Claude-Opus-4	Claude-Haiku-3.5
Command-A	Command-R
Gemini-2.5-Pro	Gemini-2.0-Flash
GPT-4.1	GPT-4o-Mini
Grok-4	Grok-3-Mini
Llama-4-Maverick	Llama-3.1-8B
Mistral-Medium	Mistral-8B
Qwen3	Qwen2.5
DeepSeek-R1	Phi4
Kimi-K2	Olmo2

There is more to NLP than LLMs!

Source: https://www.pinecone.io/learn/llm-ecosystem/

Read the following context and statement:

Context: Could you please speak to this issue, with regard to the social ramifications of gum chewing in public?
Statement: You don't have an opinion on gum chewing in public, I see.

Choose one or more from the following:

If you feel uncertain and you feel that multiple options apply, choose them all instead, even though it might feel contradictory.
Assuming the context is true, the statement:

- [ ] is most likely to be true
- [ ] can be either true or false
- [ ] is most likely to be false

Explain, in a few sentences, why you chose your answer.

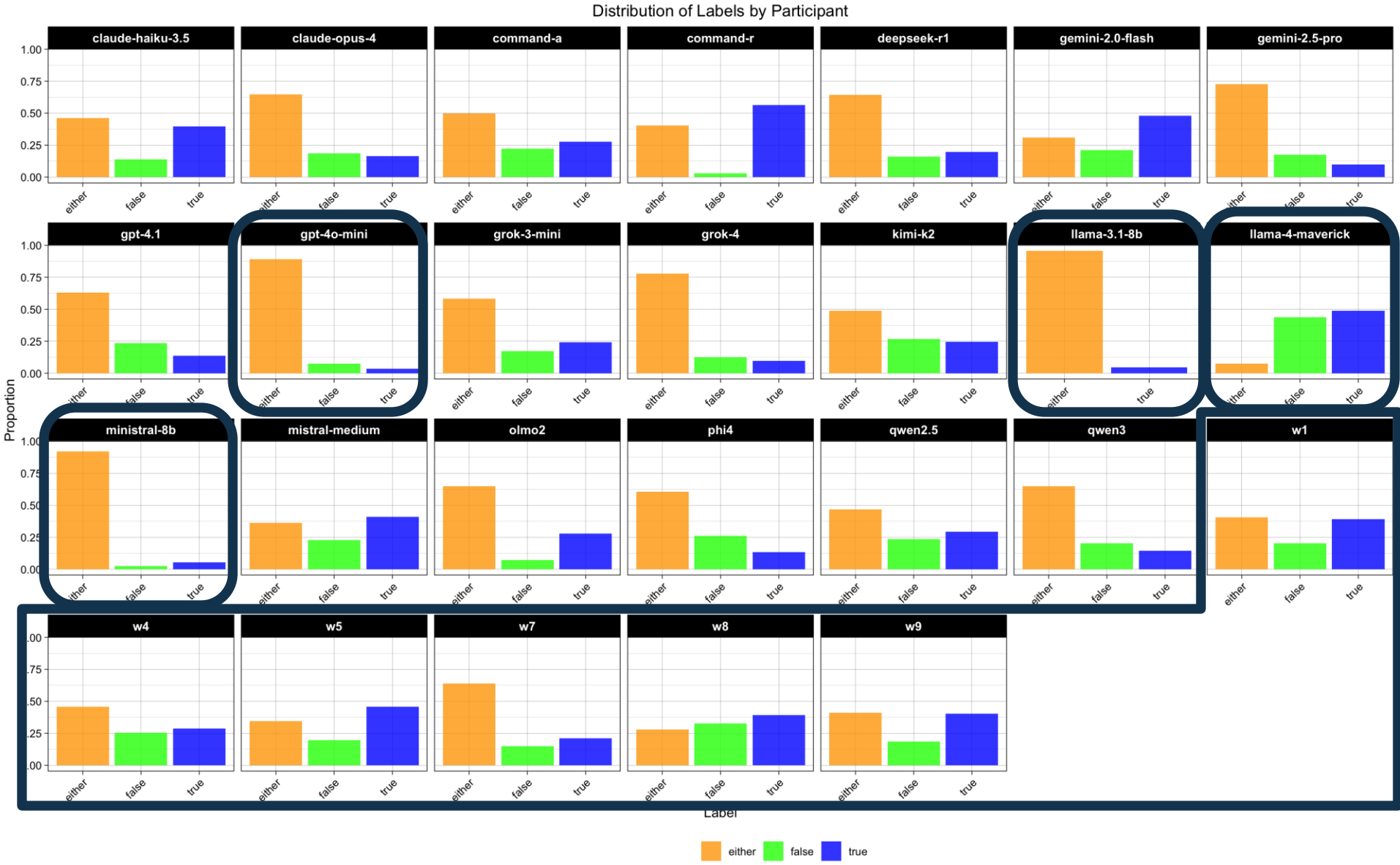
If you chose more than one option, elaborate in which circumstances each option is possible.
Explain all the options you chose.
Your explanation should include **new information** and **refer to specific parts of the sentences**. It should **NOT simply repeat the sentences**
Avoid "The context and statement means the same/opposite thing". **Specify which part of the context and statement means the same/opposite thing.**
Avoid "Just because X doesn't mean Y". **Say under what circumstances X does not mean Y, or say that X can mean Y or Z.**
Avoid "The statement is ambiguous/it's not clear what it means". **Elaborate what the possible meanings are and why it is ambiguous.**

Minimum word count: 10 Words: 0

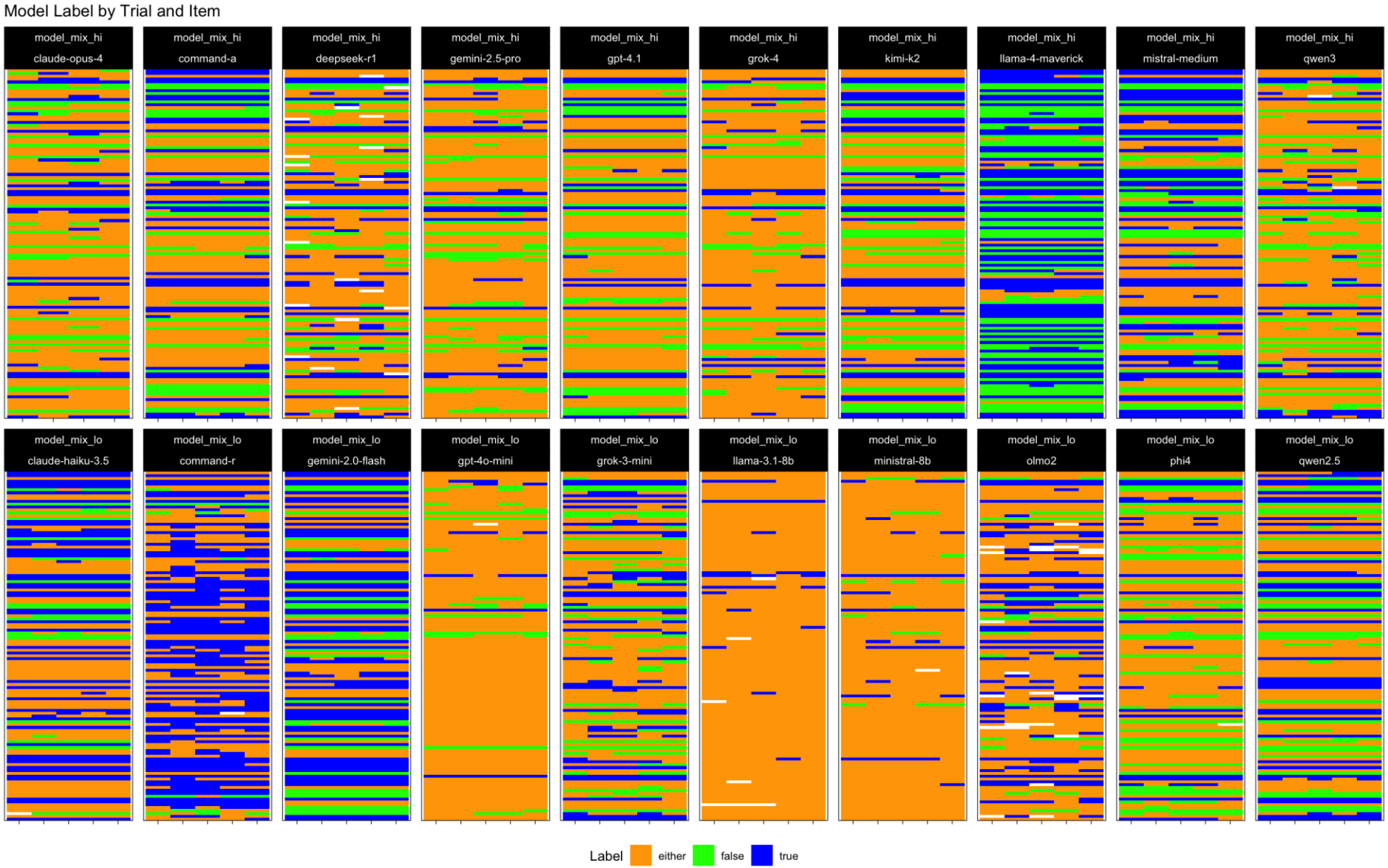
Highlight the words in the Context and Statement that are relevant to your explanations.

Your explanations should refer to specific words/parts of the sentences. Highlight those words and phrases that your explanations mentioned.
Only highlight the words that are most important for the explanations.

Very different behaviors in labeling



Not a lot of label variation individually



Model mix best at capturing human label variation

Experiment		H	JS↓	Avg. agree	α
Humans		0.71	--	--	0.26
Model mix	all	0.63	0.22	0.72	0.26
	high	0.53	0.25	0.74	0.37
	low	0.60	0.25	0.66	0.18
Size	deepseek-r1	0.43	0.30	0.67	0.27
	gemma3	0.49	0.30	0.59	0.11
	qwen3	0.62	0.25	0.67	0.25

What about explanations?

Experiment		Word count	Avg. sim.	BERT Score	Centroid sim.	Dispersal
Humans		29	0.58	0.24	0.78	0.66
Model mix	all	75	0.79	0.33	0.89	0.40
	high	80	0.80	0.33	0.90	0.34
	low	70	0.79	0.36	0.90	0.34
Size	deepseek-r1	57	0.72	0.31	0.88	0.40
	gemma3	80	0.70	0.26	0.88	0.43
	qwen3	69	0.79	0.37	0.91	0.34

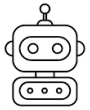
. [Giulianelli et al. 2023]

P: We were playing all sorts of sports, and you were not, so shut up and stop twitching,' the microbe's tone of voice changed, it was lower and more resounding.
H: We have played football, basketball, baseball and soccer and you haven't played anything so shut your mouth.

Same distribution between the 20 LLMs and the humans [0.8, 0.2, 0]

E - They stated they played all kinds of sports so it could have been football, basketball, baseball or soccer and it's true that one person did not play.

E - The premise shows a conversation where one party is asserting dominance by claiming superior sports participation. The microbe seems to be directly telling another party that they have played multiple sports while the other person has not. The hypothesis precisely mirrors the premise, listing the same sports (football, basketball, baseball and soccer) and using the same confrontational language of telling someone to shut up.

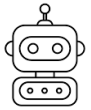


P: We were playing all sorts of sports, and you were not, so shut up and stop twitching,' the microbe's tone of voice changed, it was lower and more resounding.
H: We have played football, basketball, baseball and soccer and you haven't played anything so shut your mouth.

Same distribution between the 20 LLMs and the humans [0.8, 0.2, 0]

E - They stated they played all kinds of sports so it could have been football, basketball, baseball or soccer and it's true that one person did not play.

E - The premise shows a conversation where one party is asserting dominance by claiming superior sports participation. The microbe seems to be directly telling another party that they have played multiple sports while the other person has not. The hypothesis precisely mirrors the premise, listing the same sports (football, basketball, baseball and soccer) and using the same confrontational language of telling someone to shut up.

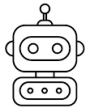


P: We were playing all sorts of sports, and you were not, so shut up and stop twitching,' the microbe's tone of voice changed, it was lower and more resounding.
H: We have played football, basketball, baseball and soccer and you haven't played anything so shut your mouth.

Same distribution between the 20 LLMs and the humans [0.8, 0.2, 0]

N - The premise mentions sports without saying which ones, the hypothesis mentions specific sports that might or might not have been played.

N - The premise states 'all sorts of sports' without specifying which sports were played, while the statement lists particular sports like football, basketball, baseball, and soccer. Since the hypothesis does not confirm or deny the inclusion of these specific sports, the hypothesis could be true if those sports were part of 'all sorts,' or false if other sports were played instead.



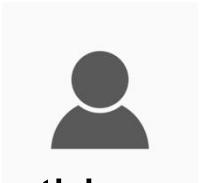
P: Do you want to see historic sights, tour museums and art galleries?
H: You wouldn't like to visit historic places, museums, and art galleries, would you?

[Groenendijk and Stokhof 1984]

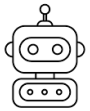
Different distributions between	LLMs	[0, 0.6, 0.4]
	humans	[0.5, 0.5, 0]

E - The premise and the hypothesis say basically the same thing, but one asks in a positive way and the other in a negative way.

E - The heart of both of these questions seems to match.



N - The premise is asking if someone would like to do these things, and there is no indication of whether or not they do.



N - The premise asks whether the person desires to experience historic sights, museums, and art galleries, but provides no information about their actual preference. The hypothesis asserts a negative preference with wouldn't like to visit and references equivalent attractions. Since the context only inquires without revealing the person's inclination, the hypothesis could be true if the person dislikes these activities or false if they enjoy them.

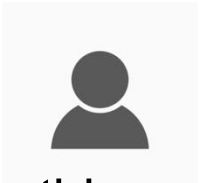
P: Do you want to see historic sights, tour museums and art galleries?
H: You wouldn't like to visit historic places, museums, and art galleries, would you?

[Groenendijk and Stokhof 1984]

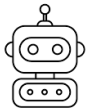
Different distributions between	LLMs	[0, 0.6, 0.4]
	humans	[0.5, 0.5, 0]

E - The premise and the hypothesis say basically the same thing, but one asks in a positive way and the other in a negative way.

E - The heart of both of these questions seems to match.



N - The premise is asking if someone would like to do these things, and there is no indication of whether or not they do.

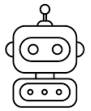


N - The premise asks whether the person desires to experience historic sights, museums, and art galleries, but provides no information about their actual preference. The hypothesis asserts a negative preference with wouldn't like to visit and references equivalent attractions. Since the context only inquires without revealing the person's inclination, the hypothesis could be true if the person dislikes these activities or false if they enjoy them.

P: Do you want to see historic sights, tour museums and art galleries?
H: You wouldn't like to visit historic places, museums, and art galleries, would you?

Different distribution between LLMs [0, 0.6, 0.4]
humans [0.5, 0.5, 0]

C - The premise asks if the person wants to see historic sights and tour museums and art galleries, which suggests the person has an interest in these activities. The hypothesis asserts the opposite—that the person wouldn't like to visit historic places, museums, and art galleries—which contradicts the interest expressed or implied in the premise. Therefore, if we assume the premise is true, the hypothesis is most likely false.



C - The premise expresses interest in visiting historic sights, museums, and art galleries, while the hypothesis suggests the opposite, implying disinterest. The key difference lies in the positive tone of the premise ('want to see') versus the negative tone of the hypothesis ('wouldn't like to visit'). This contradiction makes the hypothesis most likely false if the premise is true.

Wrapping up

Move away from a single label annotation

To provide robust natural language understanding, models need to distinguish items on which humans mostly agree from items leading to variation

In the current era of LLMs, we need explanations and a better way to evaluate these compared to humans'

Look at the data!

Thank you!

TACL action editor Anette Frank, the anonymous reviewers, Micha Elsner, Michael White, for insightful feedback, and Angélica Aviles Bosques for her help with the annotations.

github.com/njjiang/NLI_disagreement_taxonomy
github.com/njjiang/LiveNLI