ACL 2026 SAN DIEGO JULY 2-7

Agree, Disagree, Explain: Decomposing Human Label Variation in NLI through the Lens of Explanations

Pingjun Hong*, Beiduo Chen*, Siyao Peng, Marie-Catherine de Marneffe, Benjamin Roth, Barbara Plank

universität wien

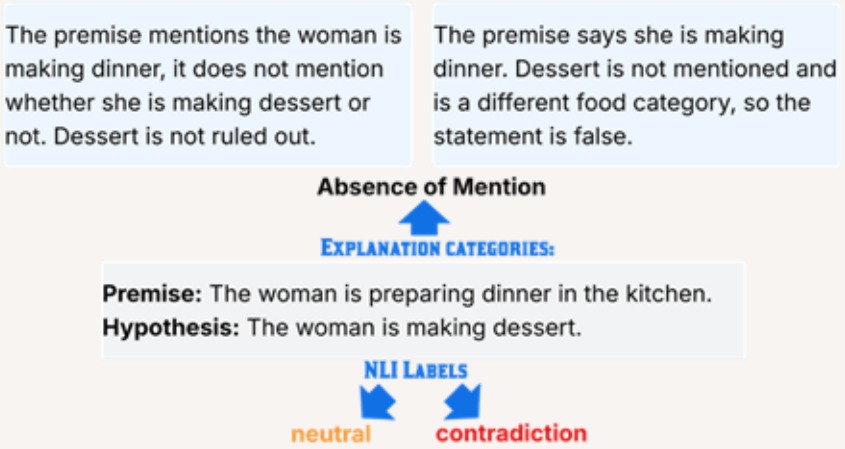
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

UCLouvain

mcml Munich Center for Machine Learning

Motivation: Labels Hide Reasoning

- NLI labels alone do not reveal why annotators agree or disagree.



- different NLI label, similar reasoning
- same NLI label, different reasoning

Research Questions

RQ1: Can reasoning categories reveal hidden agreement beyond labels?

RQ2: Do annotators exhibit stable reasoning preferences?

RQ3: What correlates better with explanation similarity: label agreement or reasoning-category agreement?

From Within-label Variation to Label Variation

- **Background: LiTEx Taxonomy**

LiTEx: Linguistically-informed Taxonomy of Explanations	
Text-Based Reasoning	**World-Knowledge Reasoning**
• Coreference • Semantic • Syntactic • Pragmatic • Absence of Mention • Logic Conflict	• Factual Knowledge • Inferential Knowledge

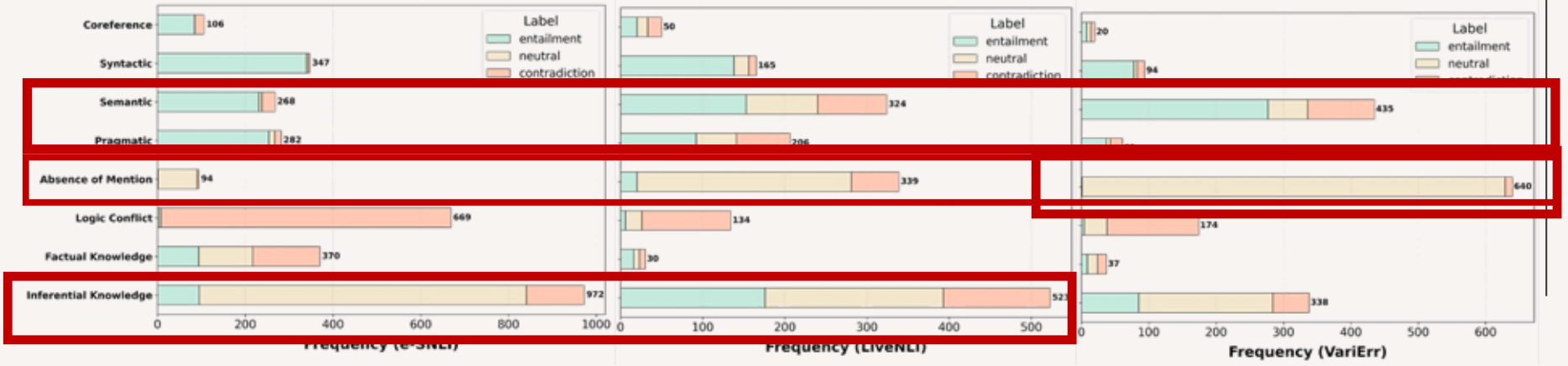
- Previous work: focus on within label variation (same label, different explanation types)
- *In this work, we extend the analysis to label variation.*

From Within-label Variation to Label Variation

- **Annotation on LiveNLI and VariErr**

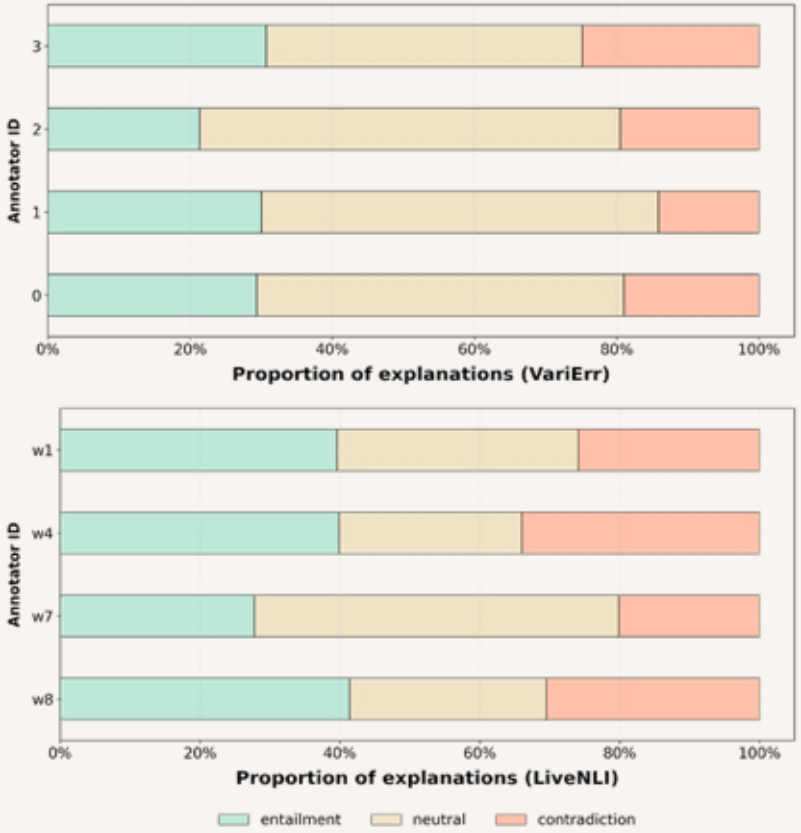
Datasets	Characteristic	#Explanations	IAA (κ)
LiveNLI	Ecologically valid explanations	1404	0.828
VariErr	Variation + annotation errors	1933	0.792

- **LiTEx Categories across NLI Labels**

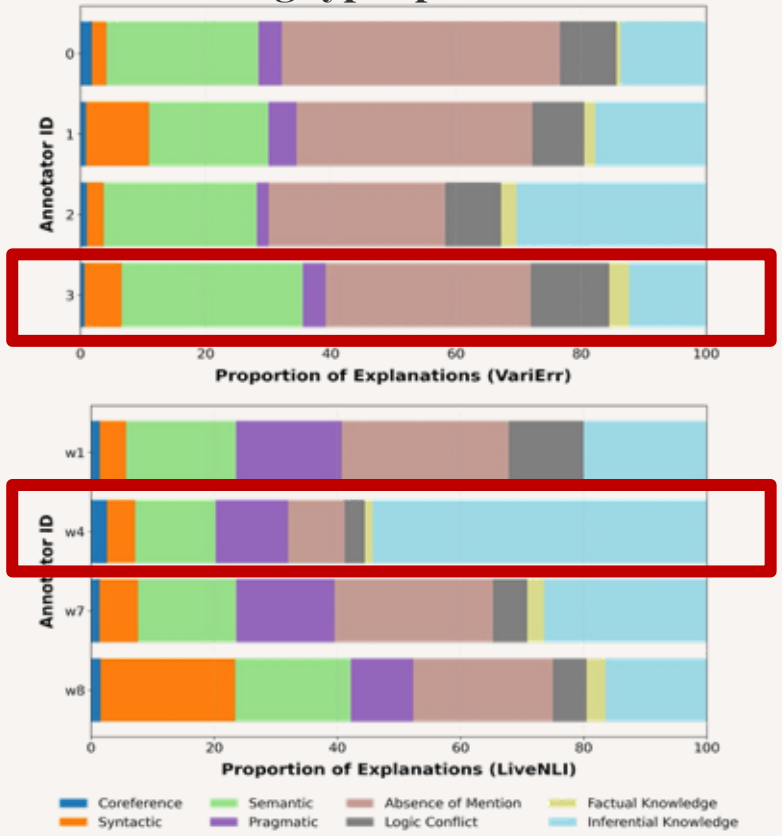


Label and Reasoning Preferences among Annotators

- Label preferences



- Reasoning types preferences



Measuring and Interpreting Agreement

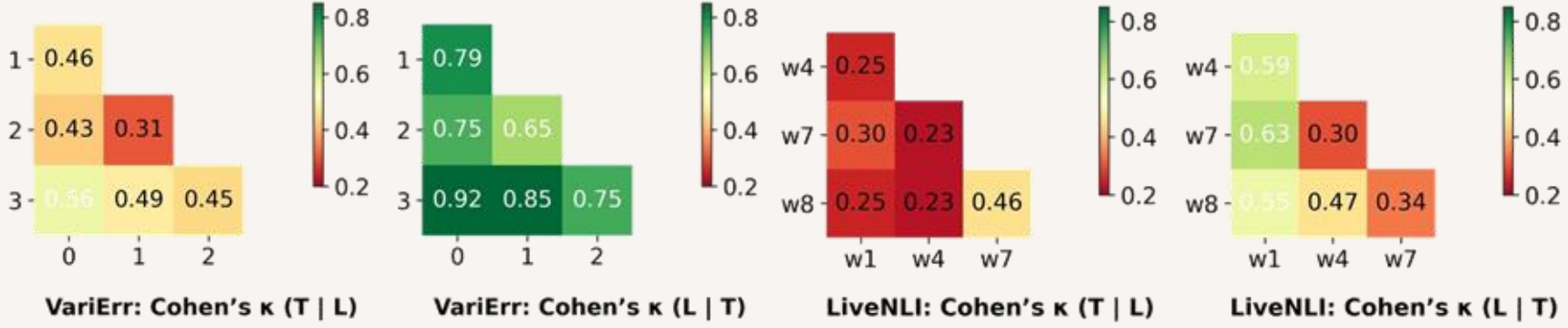
- Quantifying Annotator Agreement Beyond Labels

Agreement Class	Entropy	Support (%)	Category Agreement	token 1-gram	token 2-gram	POS 1-gram	POS 2-gram	cosine (%)	euclidean (%)
VariErr									
Full (4-0-0)	0.00	43.75	**0.76**	**35.05**	**11.53**	74.21	**35.06**	**52.87**	51.89
Partial (3-1-0)	0.81	28.95	0.60	34.72	10.62	**78.31**	34.85	52.81	**51.96**
Two Pairs (2-2-0)	1.00	23.36	0.56	30.80	8.50	73.47	31.23	49.22	51.02
Divergent (2-1-1)	1.50	3.95	0.50	32.02	9.96	70.37	31.50	48.21	50.91
LiveNLI									
Full (4-0-0)	0.00	21.74	**0.62**	40.31	10.26	**88.89**	41.96	**58.05**	**53.24**
Partial (3-1-0)	0.81	34.78	0.56	**40.44**	**11.95**	86.99	**44.02**	54.47	52.27
Two Pairs (2-2-0)	1.00	23.48	0.60	38.97	10.67	88.38	43.24	55.33	52.84
Divergent (2-1-1)	1.50	20.00	0.54	36.61	8.99	85.09	41.44	53.63	52.35

- Reasoning-category agreement better reflects explanation similarity than label agreement alone.

Measuring and Interpreting Agreement

- **Pairwise Annotator Agreement on Reasoning Category and Labeling**



- Same reasoning ↛ likely same label; Same label ↛ not necessarily same reasoning
- Divergence in the explanation reasoning categories is relatively more common than that in the labeling.

Measuring and Interpreting Agreement

Premise: The author began with a set of hunches or hypotheses about what can go wrong in agency management, and what would be evidence supporting—or contradicting—these hypotheses.

Hypothesis: The hunches provided by the author weren't realistic as it pertains to agency management.

[CHAOSNLI] [E,N,C]: [0.64, 0.06, 0.30]

Data	Ann.	NLI Label	Explanation	Exp. Category
LiveNLI	w1	Neutral	The context notes that the hunches were provided,but there is no information on their veracity or plausibility. Thus, the statement could be true or false, as it is not known whether they were realistic based on the provided information.	Absence of Mention
	w4	Contradiction	If an author is planning to write about a certain topic, they likely have enough knowledge to form informed opinions. Thus, it is much more likely that the statement is false, since these opinions would be at least somewhat realistic.	Inferential Knowledge
	w7	Contradiction	The author gave evidence to support the hunches, it is unlikely that the hunches were unrealistic.	Inferential Knowledge
	w8	Contradiction	It was not stated that the hunches were unrealistic.	Absence of Mention
VariErr	0	Neutral	It is not clear whether the hunches provided by the author were realistic or not.	Absence of Mention
	1	Neutral	It is not clear how realistic the hypotheses were.	Absence of Mention
	2	Neutral	The judgment of the hunches is not given in the context.	Absence of Mention
	3	Contradiction	The hunches could be realistic, as author provides potential evidence supporting these hypotheses.	Inferential Knowledge

TO SUM-UP

Key Findings:

- Labels alone are insufficient
- Reasoning categories reveal hidden structure
- Annotators show stable reasoning behaviors
- Explanation categories better capture semantic alignment

Future Directions:

- Multi-category explanations
- Annotator background modeling
- Explanation-aware evaluation
- Reasoning-aware dataset construction

THANK YOU!

Paper

Data