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

TL;DR

We apply LiTEx to two NLI datasets and examine annotation variation through ***label agreement***, ***explanation similarity***, and ***taxonomy agreement***. Our analysis uncovers semantic ***alignment that cannot be captured by labels alone***, exposing richer individual differences in NLI annotation.

❖ Introduction

The premise mentions the woman is making dinner, it does not mention whether she is making dessert or not. Dessert is not ruled out.

The premise says she is making dinner. Dessert is not mentioned and is a different food category, so the statement is false.

Absence of Mention

Explanation categories:

Premise: The woman is preparing dinner in the kitchen.
Hypothesis: The woman is making dessert.

NLI Labels

neutral contradiction

The man cannot simutaneously be wearing a sweatshirt and a tank top.

A man in Alaska would typically not be wearing a tank top, as it is rather cold there most times of the year.

Logic Conflict **Inferential Knowledge**

Explanation categories

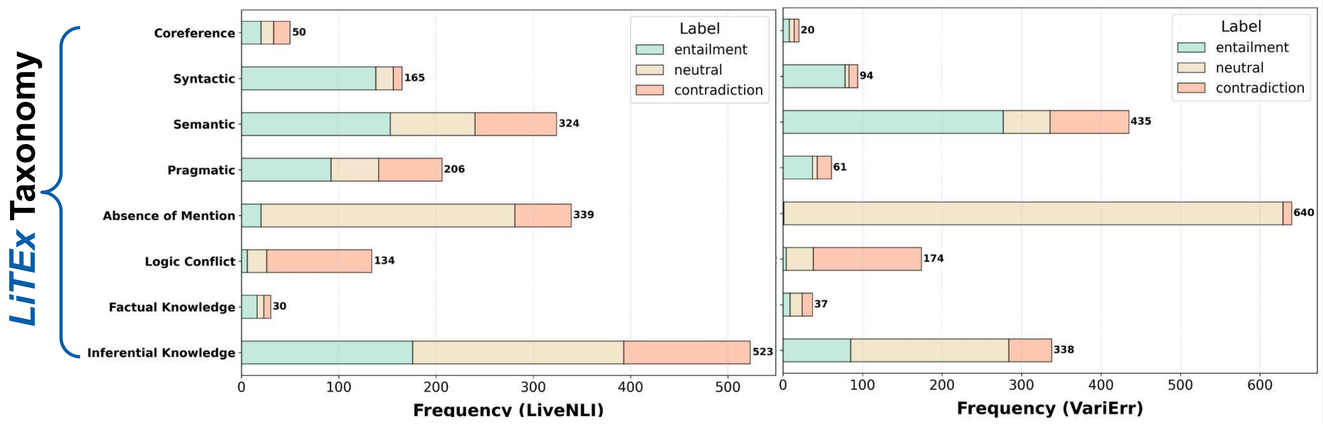
Premise: A man in an Alaska sweatshirt stands behind a counter.
Hypothesis: The man is wearing a tank top.

NLI labels

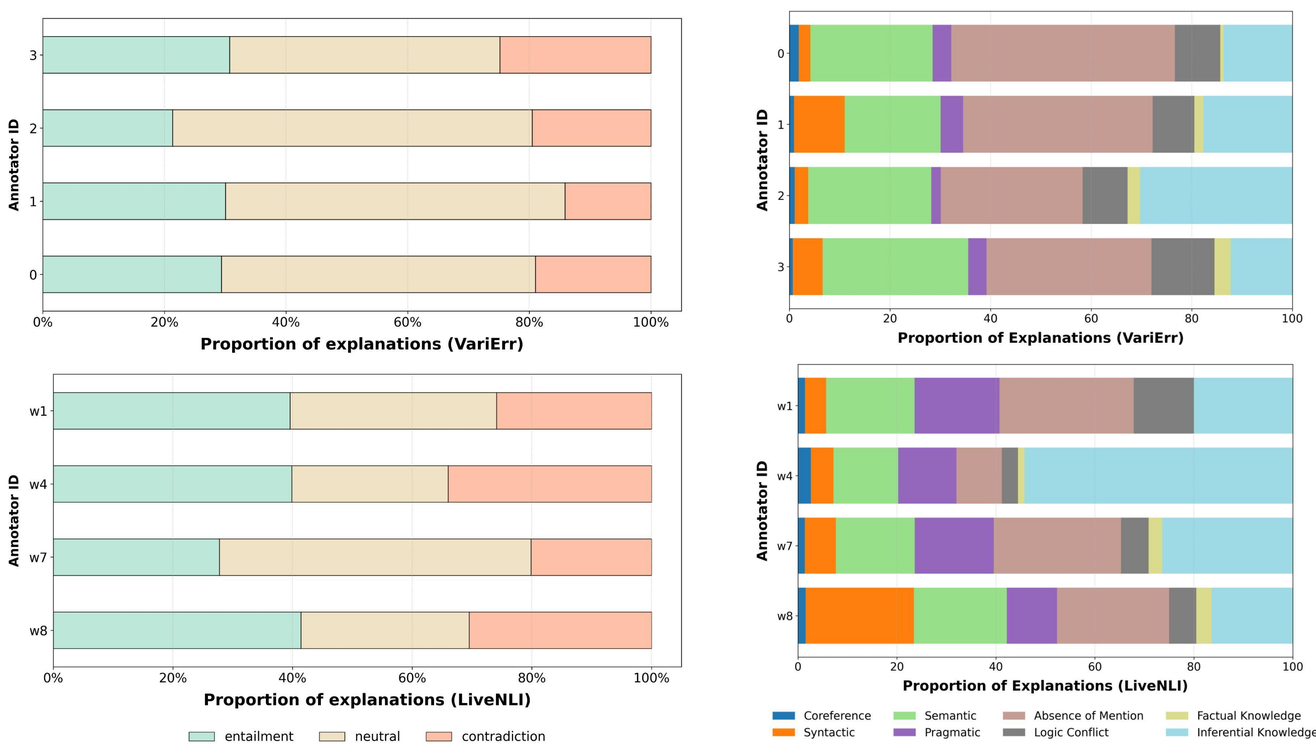
contradiction

- Annotators may diverge in NLI labels or explanation categories.
- *Distribution of NLI labels only is not enough.*

❖ LiTEx & Label Variation



❖ Label and Reasoning Preferences



❖ Measuring and Interpreting Agreement

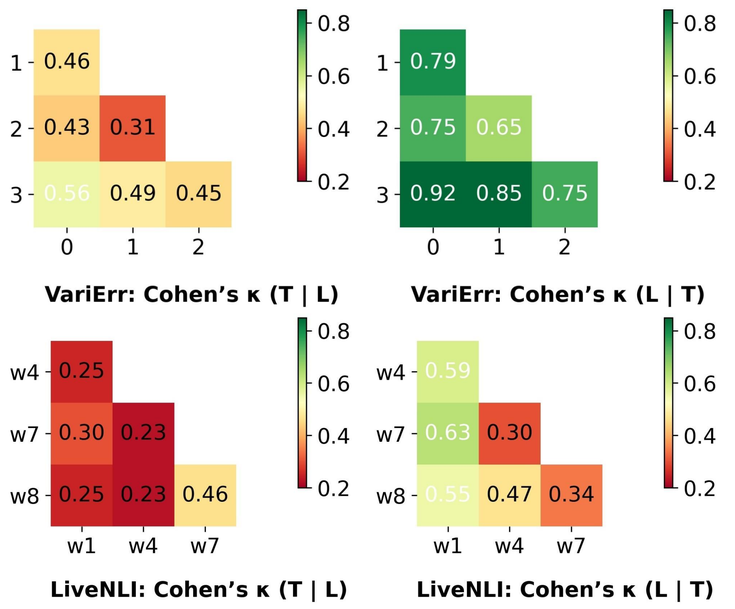
Label entropy, category agreement, and average explanation similarity

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

$$Jaccard(a,b) = \begin{cases} 1, if\ a = b \\ 0, otherwise \end{cases}$$

- Shared reasoning categories may better capture the semantic similarity of explanations than label agreement.

Pairwise Annotator Agreement



- Divergence in the explanation reasoning categories is relatively more common than that in the labeling.

❖ Conclusion

- **Beyond Labels:** Annotate two NLI datasets with LiTEx to study both within-label and label variation.
- **Beyond Distributions:** Combine NLI labels and explanation categories to uncover annotator-specific patterns.
- **Beyond Agreement:** Reasoning-category alignment better reflects explanation similarity than label agreement.

❖ Resources

Paper Data

{pingjun,hong, benjamin.roth}@univie.ac.at
{beiduo.chen, b.plank}@lmu.de, loganpeng1992@gmail.com
marie-catherine.demarneffe@uclouvain.be

ACL 2026 SAN DIEGO JULY 2-7