LiTEx: A Linguistic Taxonomy of Explanations for Understanding Within-Label Variation in Natural Language Inference

Pingjun Hong* Beiduo Chen* Siyao Peng Marie-Catherine de Marneffe Barbara Plank
MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
Munich Center for Machine Learning (MCML), Germany FNRS, CENTRAL, UCLouvain, Belgium
Faculty of Computer Science and UniVie Doctoral School Computer Science, University of Vienna, Austria

TL;DR:

We introduce **LiTEx**, a linguistic taxonomy for categorizing NLI explanations, to analyze **within-label variation**. By annotating e-SNLI and validating the taxonomy, we show that LiTEx reveals how explanations relate to labels, and improves explanation generation to better match human reasoning.

INTRODUCTION

Highlight

Example A

Premise: A crowd is watching a group of men in suits with briefcases walk in formation down the street led by a woman holding a sign.

Hypothesis: The sign the woman is holding states that 'Freedom is free'.

Different highlights

Explanation 1: it doesn't tell you what the sign says.

Explanation 2: There's no explanation that the sign the woman is holding state that "Freedom is free".

Same explanation

Example B

Premise: A man in an Alaska sweatshirt stands behind a counter.

Hypothesis: The man is wearing a tank top.

Same highlight

Explanation 1: The man cannot simultaneously be wearing a sweatshirt and a tank top.

Explanation 2: A man in Alaska would typically not be wearing a tank top, as it is rather cold there most times of the year.

Different explanations

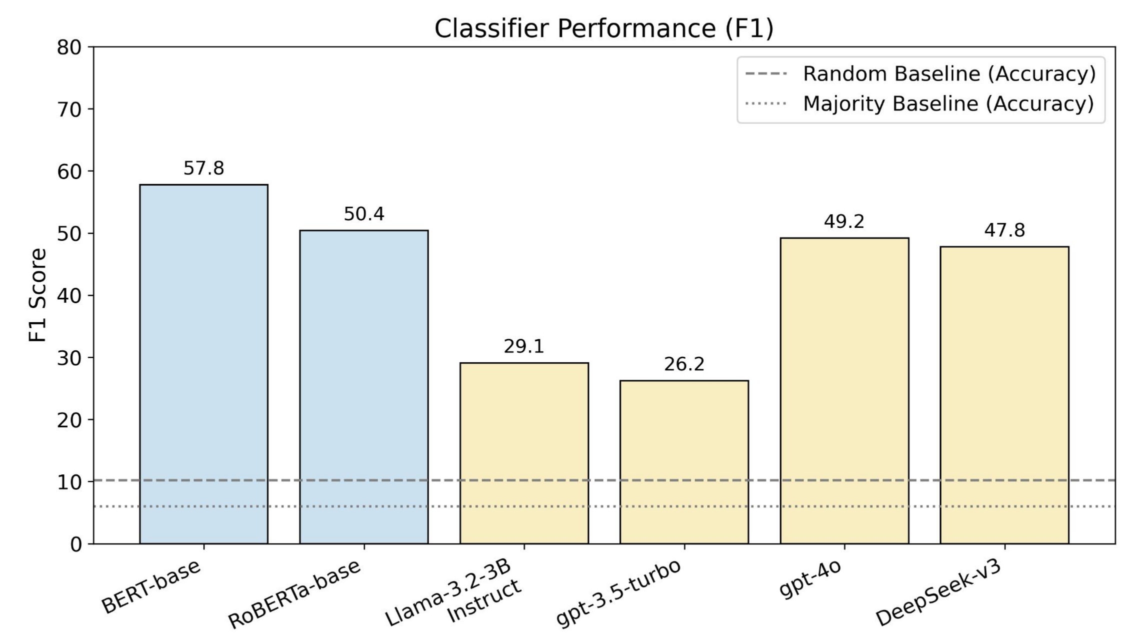
Taxonomy

Text-Based Reasoning: Coreference, Syntactic, Semantic, Pragmatic, Absence of Mention

Word-Knowledge Reasoning: Logic Conflict, Factual Knowledge, Inferential Knowledge

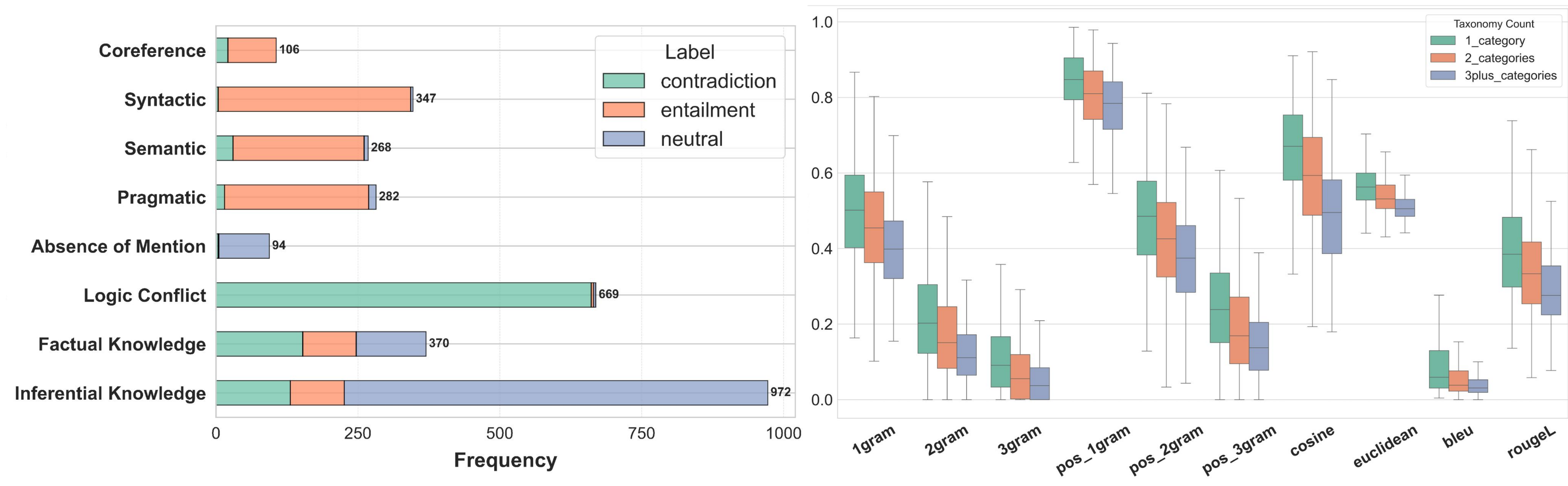
VALIDATION

- **IAA on subset:** Cohen's k of **0.862**
- **Model classification:**



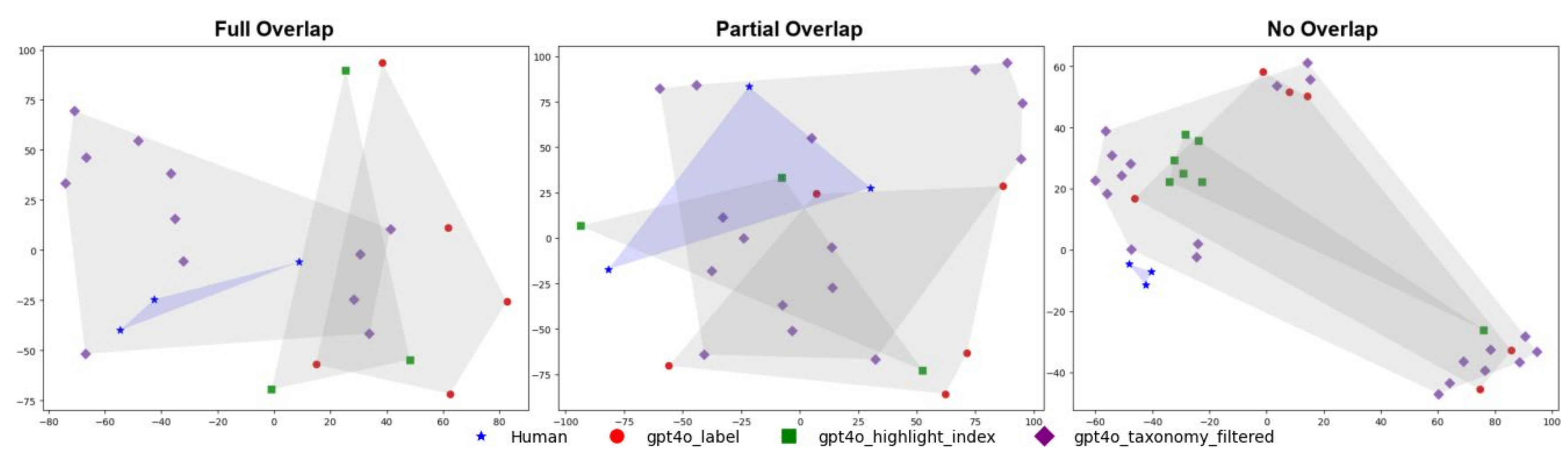
★ *Consistent taxonomy assignment across annotators*
★ *Learnable by fine-tuned and prompted models*

ANALYSIS



★ *Distribution of taxonomy categories across NLI labels reflects expected patterns.*
★ *Greater reasoning diversity corresponds to lower similarity.*

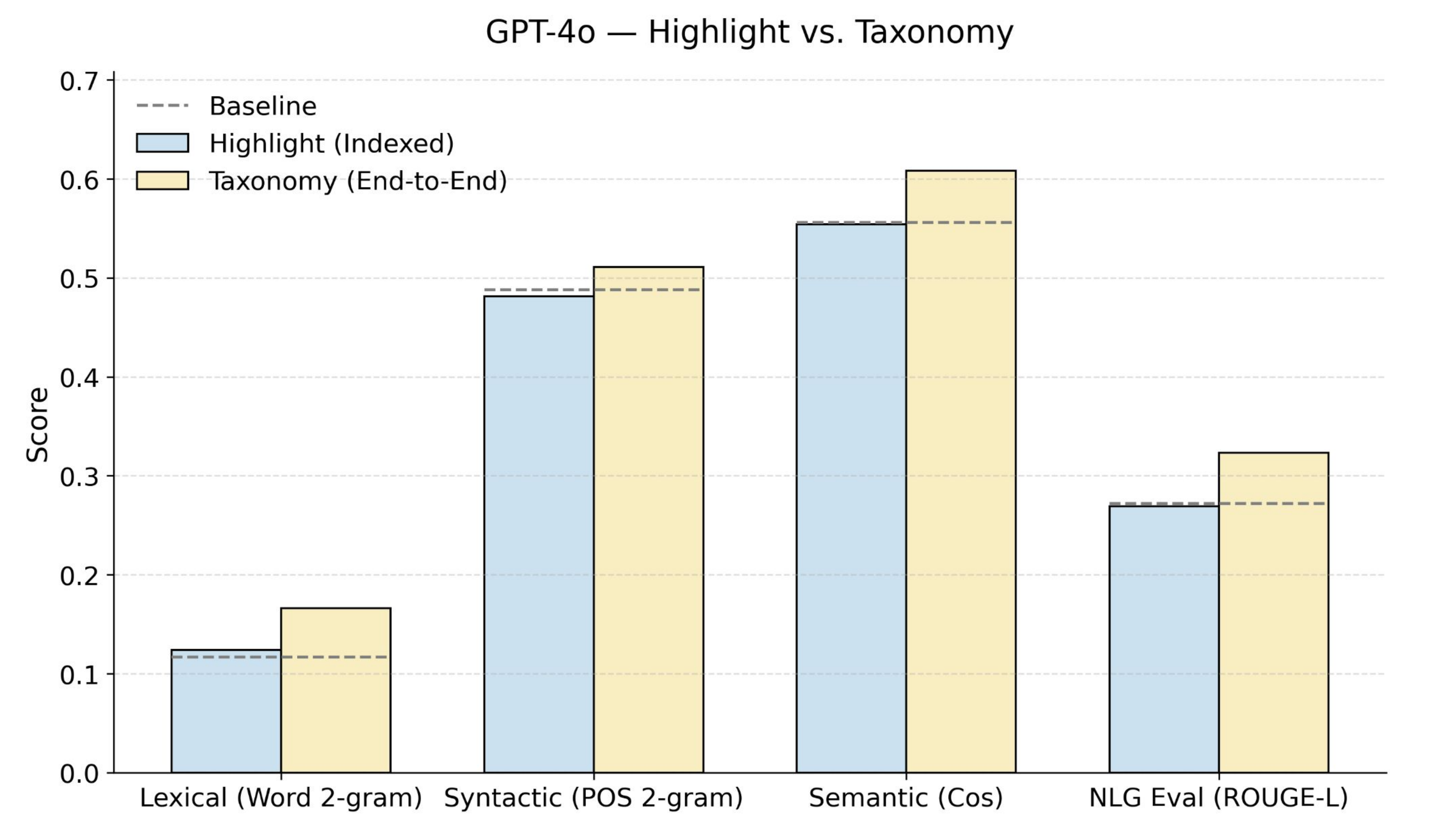
RESULTS



Mode (GPT4o)	Coverage		Area	
	Full	Partial	Rec	Prec
baseline	1.9	21.6	16.5	**5.7**
highlight-guided	1.1	13.5	10.0	4.7
taxonomy-guided	**10.7**	**56.1**	**49.3**	5.6

★ *Taxonomy-based prompting approaches consistently produce higher similarity scores*
★ *Taxonomy-guided outputs cover more of the human explanation space.*

- Models: **GPT4o**, DeepSeek-v3, Llama-3.3-70B
- Prompting Paradigms: Baseline, Highlight-guided, Taxonomy-guided



CONCLUSION

- Introduce **LiTEx**, a linguistic taxonomy designed to capture reasoning strategies in NLI explanations.
- **Taxonomy-guided generation** produces richer, more human-like explanations.
- **Enhanced e-SNLI** dataset with fine-grained taxonomy labels, offering a new resource.

RESOURCES

Paper Data

EMNLP 2025 Suzhou, China | 中国苏州
European Research Council Established by the European Commission

pingjun.hong@univie.ac.at
{beiduo.chen, siyao.peng, b.plank}@lmu.de,
marie-catherine.demarneffe@uclouvain.be