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*

EMNLP 2025 Suzhou, China 中国苏州

Introduction: Motivation

Within-label variation

- Annotators agree on the **same label** but provide **different explanations**.

Premise: This church choirs sings to the masses as they sing joyous songs from the book at a church.

Hypothesis: The church has cracks in the ceiling.

There is no indication that there are cracks in the ceiling of the church.

Not all churches have cracks in the ceiling.

Introduction: Research Questions

➢ **RQ1:**

What types of reasoning variation exist within the same NLI label, and how can they be systematically categorized?

➢ **RQ2:**

To what extent is the proposed taxonomy reliable and meaningful for capturing diverse types of human reasoning in NLI?

➢ **RQ3:**

What form of prompting best supports LLMs in generating diverse and accurate NLI explanations: taxonomy-based or highlight-based?

LiTEx: Linguistically-informed Taxonomy of NLI Reasoning

Taxonomy Categories

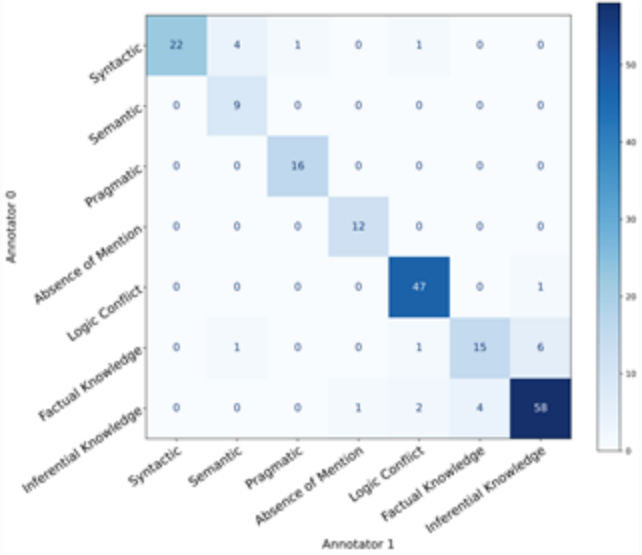
- LiTEx: Linguistically-informed Taxonomy of Explanations.
- Two high-level categories: **Text-Based (TB) Reasoning** and **World-Knowledge (WK) Reasoning**.

LiTEx	
Text-Based Reasoning (TB)	**World-Knowledge Reasoning (WK)**
• Coreference • Semantic • Syntactic • Pragmatic • Absence of Mention • Logic Conflict	• Factual Knowledge • Inferential Knowledge

LiTEx: Linguistically-informed Taxonomy of NLI Reasoning

Taxonomy Annotation and IAA

- **Main annotation dataset:** subset of the e-SNLI dataset (1,002 items) - 3,108 explanations.
- **IAA annotation dataset:** subset of the e-SNLI dataset (67 items) - 201 explanations.



- Cohen's kappa of 0.862

Taxonomy Categories	precision	recall	f1-score	support
Coreference	N/A	N/A	N/A	N/A
Syntactic	1.000	0.786	0.800	28
Semantic	0.643	1.000	0.783	9
Pragmatic	0.941	1.000	0.970	16
Absence of Mention	0.923	1.000	0.960	12
Logic Conflict	0.922	0.979	0.949	48
Factual Knowledge	0.789	0.652	0.714	23
Inferential Knowledge	0.892	0.892	0.892	65
accuracy		0.891		201
macro	0.873	0.901	0.878	201
weighted	0.897	0.891	0.889	201

LiTEx: Linguistically-informed Taxonomy of NLI Reasoning

Model taxonomy classification

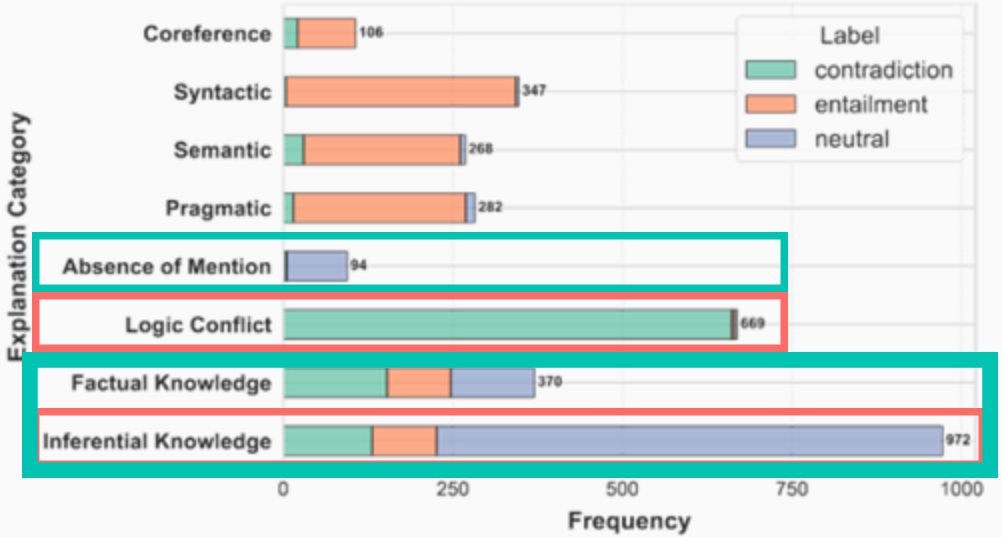
- **Fine-tuning:** BERT-base-uncased and RoBERTa-base.
- **Few-shot prompting:** Llama-3.2-3B-Instruct, GPT-3.5-turbo, GPT-4o and DeepSeek-v3

Classifiers	Acc	P	R	F1
Random Baseline	12.5	11.8	10.8	10.2
Majority Baseline	31.3	3.9	12.5	6.0
BERT-base	**70.2**	**60.5**	**57.9**	**57.8**
RoBERTa-base	68.9	48.4	53.4	50.4
Llama-3.2-3B-Instruct	35.7	44.0	35.7	29.1
gpt-3.5-turbo	30.5	31.7	30.5	26.2
gpt-4o	58.3	55.0	54.8	49.2
DeepSeek-v3	52.6	51.9	56.3	47.8

LiTEx: Linguistically-informed Taxonomy of NLI Reasoning

Taxonomy Analysis

- Co-occurrence of explanation categories and NLI labels.
- Distribution of LiTEx categories on the annotated explanations across NLI labels.



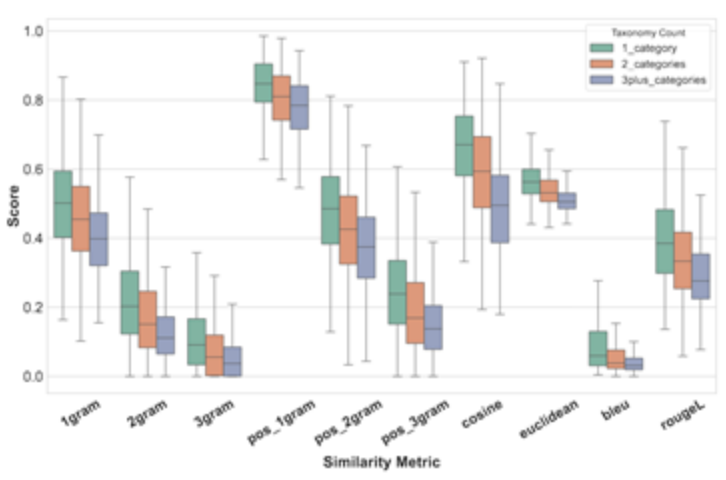
LiTEx: Linguistically-informed Taxonomy of NLI Reasoning

Taxonomy Analysis: Within-label Variation

- Distribution of NLI items that receive 1, 2, or ≥ 3 LiTEx categories on their explanations.

Category #	Entailment # (%)	Neutral # (%)	Contradiction # (%)	Total
1	76 (22.0)	171 (52.3)	142 (43.0)	389
2	179 (51.9)	139 (42.5)	156 (47.3)	474
≥ 3	90 (26.1)	17 (5.1)	32 (9.7)	139

- Explanation similarities grouped by number of LiTEx categories on an NLI item



Generating Explanations Using Taxonomy and Highlight

Model Generation Prompting Paradigms

- **Baseline** (NLI items and a label) vs. **Taxonomy-guided** prompting vs. Highlight-guided prompting.

Prompt Type	Input Variant	LLMs Used
Taxonomy-Guided (Two-Stage)	Full taxonomy input	GPT-4o, DeepSeek-v3, Llama-3.3-70B
Taxonomy-Guided (End-to-End)	Full taxonomy input	GPT-4o, DeepSeek-v3, Llama-3.3-70B
Highlight-Guided (Human)	Indexed / In-text	GPT-4o, DeepSeek-v3, Llama-3.3-70B
Highlight-Guided (Model)	Indexed / In-text	GPT-4o, DeepSeek-v3, Llama-3.3-70B

Generating Explanations Using Taxonomy and Highlight

Model Generation Prompting Paradigms

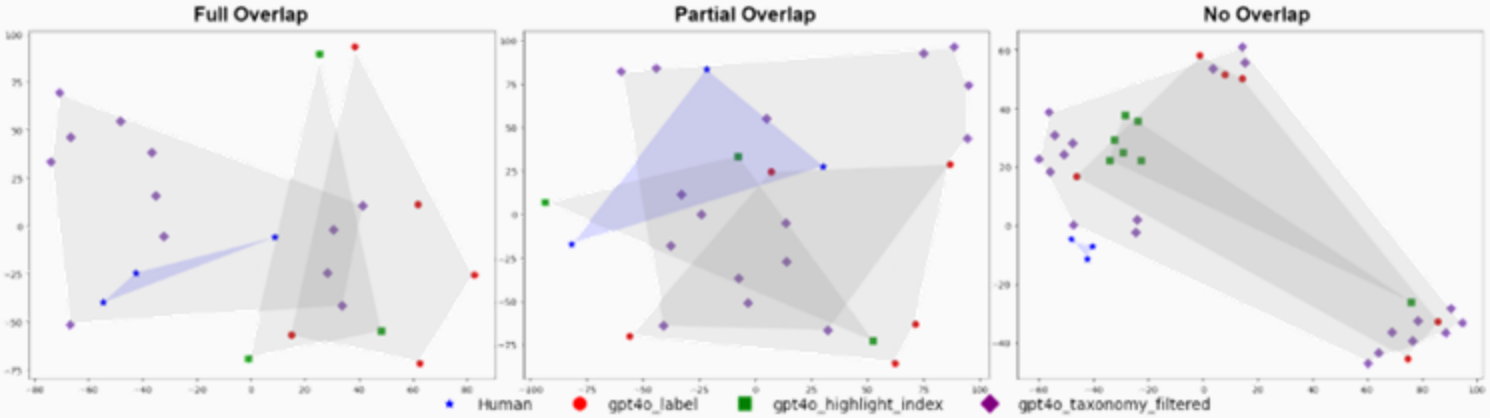
- **Baseline** vs. **Taxonomy-guided** vs. **Highlight-guided**.

Mode	Word n-gram			POS n-gram			Semantic		NLG Eval		Avg_len
	1-gram	2-gram	3-gram	1-gram	2-gram	3-gram	Cos.	Euc.	BLEU	ROUGE-L	
GPT-4o											
baseline	0.291	0.117	0.049	0.882	0.488	0.226	0.556	0.524	0.051	0.272	24.995
human highlight (indexed)	0.395	0.116	0.050	0.882	0.478	0.219	0.549	0.521	0.047	0.264	30.771
human highlight (in-text)	0.367	0.085	0.031	0.873	0.442	0.187	0.519	0.511	0.034	0.269	28.606
model highlight (indexed)	0.402	0.124	0.053	0.878	0.481	0.222	0.554	0.522	0.051	0.269	28.240
model highlight (in-text)	0.380	0.109	0.044	0.888	0.468	0.208	0.555	0.523	0.044	0.270	28.160
taxonomy (two-stage)	0.418	0.128	0.071	0.886	0.495	0.242	0.593	0.537	0.071	0.314	19.991
taxonomy (end-to-end)	**0.437**	**0.166**	**0.083**	**0.898**	**0.511**	**0.255**	**0.608**	**0.540**	**0.074**	**0.323**	26.672
DeepSeek-v3											
baseline	0.369	0.087	0.034	0.847	0.449	0.195	0.428	0.490	0.042	0.245	20.288
human highlight (indexed)	0.358	0.084	0.033	0.864	0.436	0.184	0.463	0.498	0.035	0.243	29.293
human highlight (in-text)	0.362	0.091	0.033	0.885	0.449	0.191	0.551	0.522	0.036	0.261	28.527
model highlight (indexed)	0.364	0.091	0.037	0.861	0.450	0.196	0.464	0.499	0.034	0.242	27.301
model highlight (in-text)	0.341	0.073	0.026	0.869	0.422	0.171	0.447	0.457	0.030	0.248	31.328
taxonomy (two stage)	0.391	0.122	0.055	0.884	0.475	0.219	0.544	0.522	0.057	0.293	20.894
taxonomy (end-to-end)	**0.404**	**0.140**	**0.067**	**0.897**	**0.486**	**0.233**	**0.556**	**0.528**	**0.063**	**0.306**	25.960
Llama-3.3-70B											
baseline	0.392	0.106	0.044	0.863	0.478	0.224	0.466	0.496	0.046	0.250	27.148
human highlight (indexed)	0.362	0.082	0.031	0.859	0.446	0.194	0.453	0.484	0.035	0.228	29.912
human highlight (in-text)	0.348	0.059	0.019	0.875	0.415	0.165	0.499	0.505	0.024	0.270	34.827
model highlight (indexed)	0.317	0.065	0.024	0.807	0.408	0.173	0.367	0.478	0.031	0.199	24.987
model highlight (in-text)	0.300	0.047	0.014	0.831	0.385	0.150	0.400	0.486	0.021	0.227	29.763
taxonomy (two-stage)	**0.444**	**0.167**	**0.082**	0.889	**0.512**	**0.256**	**0.609**	**0.541**	**0.078**	**0.321**	22.340
taxonomy (end-to-end)	0.383	0.110	0.048	**0.896**	0.499	0.232	0.505	0.510	0.047	0.262	28.870

Generating Explanations Using Taxonomy and Highlight

Results: Assessing explanation coverage

- Representative t-SNE visualizations of explanation embeddings
- **Blue convex hull**: span of human-written explanations
- **Gray convex hull (purple points)**: GPT4o-generated explanations



Generating Explanations Using Taxonomy and Highlight

Results: Semantic coverage of model explanations regarding human reference explanations

Mode	Coverage		Area	
	Full	Partial	Rec	Prec
GPT4o *baseline*	1.9	21.6	16.5	**5.7**
highlight (indexed)	1.1	13.5	10.0	4.7
taxonomy (end-to-end)	**10.7**	**56.1**	**49.3**	5.6
DeepSeek-v3 *baseline*	4.0	20.5	17.5	2.7
highlight (indexed)	2.3	14.9	12.5	2.9
taxonomy (end-to-end)	**17.8**	**61.8**	**54.7**	**3.8**
Llama-3.3-70B *baseline*	1.7	15.4	12.2	2.9
highlight (indexed)	0.5	8.2	6.5	2.5
taxonomy (end-to-end)	**16.7**	**65.2**	**59.8**	**5.7**

Generating Explanations Using Taxonomy and Highlight

Model Generation Validation: Human validation on explanations produced by GPT-4o (719 items - 8,373 explanations)

- **NLI label consistency**: : Does the explanation fit the gold label? (Yes/No)
- **Taxonomy consistency**: **:** Does the explanation fit the taxonomy? (Yes/No)

Q1: *Yes (98.27%), No (1.73 %)*

Q2: *Yes (83.84%), No: (16.16 %)*

Discussion

Conclusion:

- A linguistically-driven taxonomy to capture NLI reasoning diversity.
- Taxonomy-guided generation produces richer, more human-like explanations.
- Enhanced the e-SNLI dataset with fine-grained taxonomy labels, offering a new resource.

Future Work:

- Extend to broader NLI benchmarks to evaluate taxonomy generalizability across diverse settings.
- Evaluate the generated explanations more deeply, including their faithfulness to model reasoning and usefulness for end users.

Thank you!

Paper

Data