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, Germany FNRS, CENTAL, UCLouvain, Belgium
Faculty of Computer Science and UniVie Doctoral School Computer Science, University of Vienna, Austria
pingjun.hong@univie.ac.at, {beiduo.chen, siyao.peng, b.plank}@lmu.de,
marie-catherine.demarneffe@uclouvain.be

Abstract

There is increasing evidence of Human Label Variation (HLV) in Natural Language Inference (NLI), where annotators assign different labels to the same premise-hypothesis pair. However, *within-label variation*—cases where annotators agree on the same label but provide divergent reasoning—poses an additional and mostly overlooked challenge. Several NLI datasets contain highlighted words in the NLI item as explanations, but the same spans on the NLI item can be highlighted for different reasons, as evidenced by free-text explanations, which offer a window into annotators' reasoning. To systematically understand this problem and gain insight into the rationales behind NLI labels, we introduce LITEX, a linguistically-informed taxonomy for categorizing free-text explanations in English. Using this taxonomy, we annotate a subset of the e-SNLI dataset, validate the taxonomy's reliability, and analyze how it aligns with NLI labels, highlights, and explanations. We further assess the taxonomy's role in explanation generation, demonstrating that conditioning generation on LITEX yields explanations that are linguistically closer to human explanations than those generated using only labels or highlights. Our approach thus not only captures within-label variation but also shows how taxonomy-guided generation for reasoning can bridge the gap between human and model explanations more effectively than existing strategies.

1 Introduction

Natural Language Inference (NLI), a cornerstone task in Natural Language Processing (NLP), has inspired extensive research on human disagreement and model interpretability. A key focus of recent work has been Human Label Variation (HLV, Plank 2022) — cases in which annotators assign different labels to the same premise-hypothesis pair (Nie et al., 2020b; Jiang et al., 2023; Weber-Genzel et al., 2024). This variation has been acknowledged as a reflection of subjective judgment (Cabitza et al., 2023) and linguistic ambiguity (de Marneffe et al., 2012; Uma et al., 2022). Comparatively, the issue of *within-label variation* (Jiang et al., 2023) – cases where annotators agree on the same label, yet provide different explanations or rationales for their decision – has received less attention. Such variation reveals the plurality of valid reasoning strategies and highlights the richness of human inference beyond label selection.

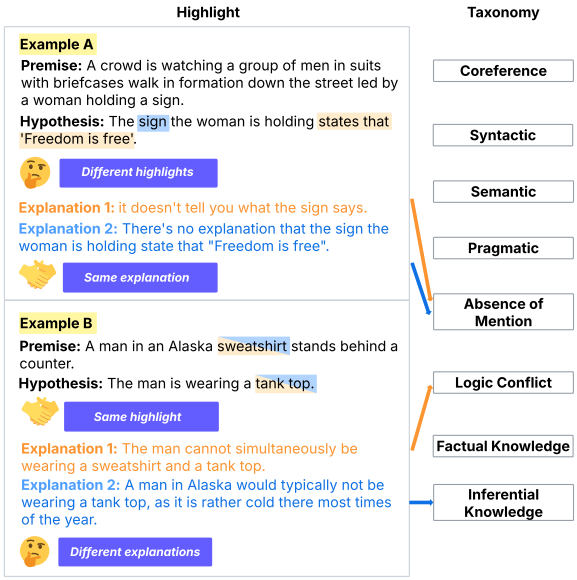

Figure 1: Our LITEX taxonomy reveals within-label variation not captured by highlights: the same highlights can yield different explanations (Example B), and vice versa (Example A).

Free-text explanations offer a rich perspective

[*] Equal contribution.
[†] Main work carried out while at LMU Munich.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34054–34074
November 4-9, 2025

on reasoning variation. However, their open-ended form makes it difficult to extract information that is directly useful for downstream analysis. As a result, structured formats are often used when collecting human explanations. Highlights are one such mechanism (Tan, 2022). Jiang et al. (2023) acknowledge that textual highlight spans alone are insufficient to capture deeper reasoning distinctions including within-label variation, especially when explanations focus on different parts of the input or rely on different assumptions. As illustrated in Figure 1, two explanations in Example B may share the same highlighted spans (here *sweatshirt* and *tank top*) but reflect different reasoning strategies (one annotator focuses on the fact that sweatshirt and tank top are not typically worn together, whereas the other says that one does not wear a tank top in Alaska); or conversely, different highlights may convey essentially the same explanation, as seen in Example A.

To address this gap, **(1)** we introduce LITEX, a LInguistic Taxonomy of EXplanations for understanding within-label variation in English natural language inference explanations. **(2)** We validate our taxonomy through human inter-annotator agreement and model-based classification. We further analyze its alignment with NLI labels and quantify within-label variation by examining category distribution and their similarity—demonstrating the taxonomy's ability to capture different types of explanations. **(3)** While human explanations are costly, LLMs offer a scalable alternative for generating explanations in NLI (Chen et al., 2025b). Through generation experiments, we demonstrate that taxonomy-based guidance provides a more effective signal for LLMs than highlight-based prompts.

2 Related Work

Explaining NLI Labels Explanations play a crucial role in making NLI decisions interpretable. As Tan (2022) highlights, explanations vary in form and quality, and improving their usefulness requires distinguishing between different explanation types and recognizing human limitations in producing them. Among existing methods, token-level highlights serve as a proxy for explanations, guiding annotators to mark relevant spans that support their label choice. Several NLI datasets provide such annotations (including free-text explanations also), collected either during labeling (e.g., LiveNLI (Jiang et al., 2023) and ANLI (Nie et al., 2020a)) or post-hoc (e.g., e-SNLI (Camburu et al., 2018)). Here, we focus on both types of explanations (free-text and highlights) from e-SNLI.

Taxonomies of Variation in NLI In the context of NLI, earlier taxonomies focused on categorizing the kind of inferences present in NLI items (Sammons et al., 2010; Simons et al., 2011; LoBue and Yates, 2011). Later work proposed a taxonomy that identifies characteristics of the items that can cause variation in annotation (Jiang and de Marneffe, 2022). Jiang et al. (2023) shifted the focus from the NLI items, collecting free-text explanations provided by the annotators themselves, applying Jiang and de Marneffe (2022)'s taxonomy to the explanations. Jayaweera and Dorr (2025) further argued for an ambiguity-aware NLI framework that detects ambiguous instances and classifies them using the taxonomy of Jiang and de Marneffe (2022).

Our work builds on this direction by proposing a taxonomy of explanations for instances that share the same NLI label, aiming to capture within-label variation in reasoning. Compared to Jiang et al. (2023), our taxonomy is thus grounded in the explanations. It also makes world knowledge in NLI reasoning explicit.

LLM-Based Explanation Generation Recent studies explored the use of LLMs to generate natural language explanations across a range of NLP tasks, aiming to improve transparency and support downstream analysis. Li et al. (2024) proposed prompting LLMs to generate chain-of-thought (CoT) explanations to improve the performance of small task-specific models. Chen et al. (2025a) further repurposed CoTs as a forward source of explanation-label pairs, applying discourse-guided segmentation to extract structured rationales. Huang et al. (2023) investigated whether LLMs could generate faithful self-explanations to justify their own predictions during inference.

In NLI, Jiang et al. (2023) employed GPT-3 to generate post-prediction explanations (predict-then-explain) and found this strategy to outperform CoT prompting. Chen et al. (2025b) showed that LLMs can effectively generate explanations to approximate human judgment distribution, offering a scalable and cost-efficient alternative to manual annotation. Building on this line of work, we use our proposed taxonomy to guide LLM prompting for more informative and human-aligned explanations.

Text-Based Reasoning (TB)		
Coreference	**Q:** **Check:**	*Does the explanation rely on resolving coreference between entities?* Determine whether the main entities in the premise and hypothesis refer to the same real-world referent, including via pronouns or phrases.
Syntactic	**Q:** **Check:**	*Does the explanation involve a change in sentence structure that preserves meaning?* Determine whether the premise and hypothesis differ in structure, such as active vs. passive, reordered arguments, or coordination/subordination, while preserving the same meaning.
Semantic	**Q:** **Check:**	*Does the explanation involve semantic similarity or substitution of key concepts?* Evaluate whether core words or expressions - including verbs, nouns, and adjectives - are semantically related between the premise and hypothesis. This includes synonymy, antonymy, lexical entailment, or category membership.
Pragmatic	**Q:** **Check:**	*Does the explanation rely on pragmatic cues like implicature or presupposition?* Look for meaning beyond the literal text - including implicature, presupposition, speaker intention, and conventional conversational meaning.
Absence of Mention	**Q:** **Check:**	*Does the explanation point out information not mentioned in the premise?* Check whether the hypothesis introduced information that is neither supported nor contradicted by the premise - i.e., it is not mentioned explicitly.
Logic Conflict	**Q:** **Check:**	*Does the explanation refer to logical constraints or conflict?* Evaluate whether the hypothesis interacts with the premise via logical structures, such as exclusivity, quantifiers ("only", "none"), or conditionals, which constrain or conflict with each other.
World Knowledge-Based Reasoning (WK)		
Factual Knowledge	**Q:** **Check:**	*Does the explanation rely on widely shared, intuitive facts acquired through everyday experience?* Determine whether the explanation invokes commonly known facts, such as physical properties or universal experiences, that are not stated in the premise.
Inferential Knowledge	**Q:** **Check:**	*Does the explanation rely on real-world norms, customs, or culturally grounded reasoning?* Determine whether the explanation requires reasoning based on general world knowledge, including cultural expectations, social norms, or typical causal inferences, that are not stated in the premise.

Table 1: Guiding questions and decision criteria for our LITEX taxonomy.

3 LITEX: Linguistically-informed Taxonomy of NLI Reasoning

To systematically capture the different types of reasoning strategies underlying within-label variation in NLI, we propose LITEX, a LInguistic Taxonomy of EXplanation classification, focusing strictly on the reasoning explicitly stated in the explanations.

3.1 Taxonomy Categories

LITEX organizes explanations into two broad categories based on their reliance on textual evidence or external knowledge, as shown in Table 1. This categorization builds on the work of Jiang and de Marneffe (2022).

The first broad category, *Text-Based (TB) Reasoning*, includes explanations that depend solely on *surface-level linguistic evidence found within the premise and hypothesis*, without appealing to world knowledge. Six subtypes are defined: *Coreference*, *Syntactic*, *Semantic*, *Pragmatic*, *Absence of Mention* and *Logic Conflict*.

The second category, *World-Knowledge (WK) Reasoning*, includes explanations that invoke background knowledge or domain-specific information beyond what is explicitly stated in the text. *Factual knowledge* refers to widely shared, intuitive facts acquired through everyday experience, such as *fire is hot*. *Inferential Knowledge* involves culturally or contextually grounded understanding, such as recognizing that *wearing white to a funeral is inappropriate* (a norm that varies across cultures) (Davis, 2017; Ilievski et al., 2021).

Table 1 presents guiding questions and decision criteria for each taxonomy category to help annotators identify the reasoning behind explanations. These questions, along with illustrative examples in Appendix A, clarify the conceptual boundaries between categories. For example, to distinguish between *Logic Conflict* and *Semantic*, consider the following two explanations: (a) *A man cannot be both tall and short at the same time* and (b) *Tall and short are not the same*. Explanation (a) reflects a logical inconsistency, pointing to the mutual exclusivity of properties, and thus labeled as *Logic Conflict*, whereas explanation (b) highlights lexical contrast or antonymy without explicit logical reasoning, and thus *Semantic*.

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

Table 2: Taxonomy classification results (%) on LITEX-SNLI. Fine-tuning methods are evaluated with a 50/50 data split; Prompt-based methods use taxonomy descriptions with two examples per category. P(recision), R(ecall), and F1 are at the macro-level.

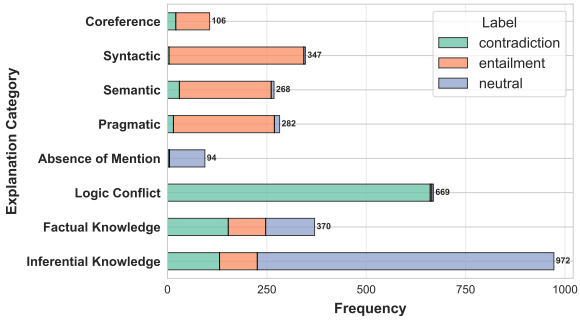

Figure 2: Distribution of LITEX categories on LITEX-SNLI explanations across NLI labels (n = 3,108).

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

Table 3: Distribution of NLI items that receive 1, 2, or >=3 LITEX categories on their explanations (n = 1,002).

3.2 Taxonomy Annotation

We randomly selected a subset (1,002 items) of the e-SNLI dataset, in which each item received three post-hoc human-written explanations accompanied by highlights. We conduct LITEX annotations on these explanations. To better capture distinct reasoning strategies, we manually segment the long explanations that potentially include multiple inferences into shorter ones. As a result, the original 3,006 explanations are expanded to 3,108. One trained annotator applied LITEX to these 3,108 explanations (and the associated premise, hypothesis, and NLI label are provided as context), labeling each with one of the eight categories.

3.3 Taxonomy Validation

To validate the consistency and generalizability of our LITEX taxonomy, we provide human inter-annotator agreement (IAA) and benchmark experiments on automatic explanation classification.

IAA We assess the consistency of our human annotations by calculating IAA on a subset of the e-SNLI explanations, separate from LITEX-SNLI used in our main experiments. Two annotators, the one from the initial phase and one newly recruited,[1] annotated 201 explanations from 67 extra e-SNLI items, using the proposed taxonomy. The agreement is high (Cohen's k of 0.862), suggesting that the taxonomy can be applied consistently between annotators. Appendix B presents the confusion matrix and representative examples of annotation disagreements.

Taxonomy Classification To validate the taxonomy and test its usefulness for automated classification, we fine-tuned two pre-trained language models, BERT-base-uncased (Devlin et al., 2019) and RoBERTa-base (Liu et al., 2019), to classify explanations in LITEX-SNLI to the annotated LITEX categories. We also few-shot prompt 4 generative AI models: Llama-3.2-3B-Instruct (Meta, 2024), GPT-3.5-turbo (Brown et al., 2020), GPT-4o (OpenAI, 2023) and DeepSeek-v3 (DeepSeek-AI et al., 2024); see Appendix C for details.

Table 2 gives the classification results. BERT-base and RoBERTa-base achieve strong results on this 8-way classification task, with macro-F1 scores of 57.8% and 50.4%, and accuracies of 70.2% and 68.9%, respectively. These results substantially surpass both a random baseline of 12.5% and a majority-class baseline of 31.3% (based on the dominant category, *Inferential Knowledge*), emphasizing the benefits of task-specific supervision. LLMs, when prompted with detailed taxonomy descriptions and illustrative examples, also perform better than random and majority-class baselines, further confirming our taxonomy's learnability.

In sum, the findings suggest that the proposed taxonomy is learnable, reinforcing its applicability for both annotation and LLM-based reasoning.

3.4 Taxonomy Analysis

Co-occurrence of Explanation Categories and NLI Labels Figure 2 plots the distribution of our explanation categories and their co-occurrence with NLI labels. We observe that different expla-

[1]Both are trained and paid according to national standards.

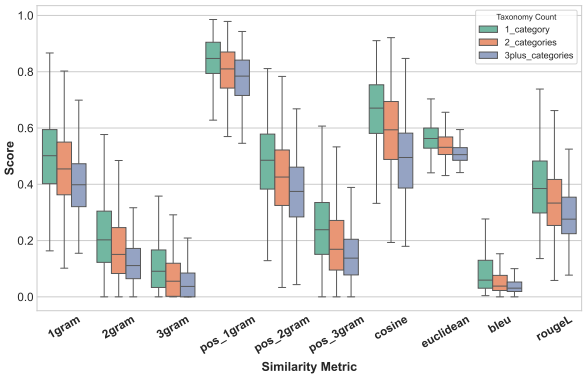

Figure 3: Boxplot of explanation similarities grouped by number of LITEX categories on an NLI item.

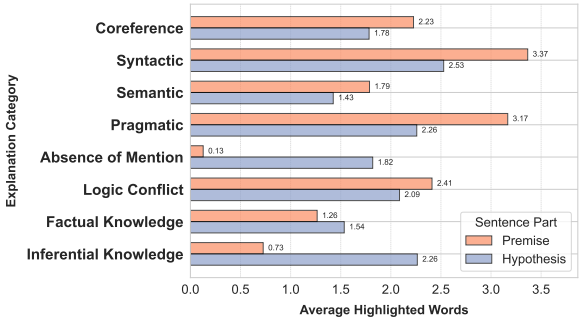

Figure 4: Average number of highlighted words in each premise-hypothesis pair across LITEX categories.

nation categories show distinct distributions over NLI labels. *Logic Conflict* is dominated by contradiction, because this category focuses on capturing logical inconsistency. *Syntactic*, *Semantic*, and *Pragmatic* are primarily associated with entailment, suggesting that these reasoning types tend to support alignment. *Factual Knowledge* and *Inferential Knowledge* are more evenly distributed across the labels, since world knowledge could be involved in different inference scenarios. Lastly, *Absence of Mention* aligns strongly with neutral, consistent with its reliance on unstated information.

Within-label Variation Table 3 gives the counts of our 1,002 NLI items for which the three (or more) explanations were annotated with 1, 2, or $\geq$ 3 LITEX categories (cf. §3.2 for explanation segmentation). These counts show that within-label variation is prevalent in e-SNLI, e.g., 613 out of 1,002 (61.2%) items received more than one taxonomy category across explanations.

To quantify it further, we compute pairwise similarity of explanations for each NLI item using standard metrics, following Giulianelli et al. (2023) and Chen et al. (2025b). These include lexical (word n-gram overlap), morphosyntactic (POS n-gram overlap), and semantic similarity (cosine/Euclidean distance[2]), along with BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004). Figure 3 shows that explanation similarity decreases as the number of taxonomy categories increases, while explanations within the same category remain more similar, supporting the taxonomy's ability to capture within-label variation.

[2]We compute semantic similarity using sentence embeddings from Sentence-BERT (Reimers and Gurevych, 2019).

Highlights vs. Taxonomy We analyze highlight span lengths for different explanation categories in Figure 4. On average, premises and hypotheses contain 13.81 and 7.41 words. *Syntactic* explanations have the longest spans in both, reflecting sentence-level understanding. *Absence of Mention* highlights are minimal in premises but more in hypotheses, marking new mentions in the hypotheses. *Inferential* and *Factual Knowledge* rely on short spans in the premise, pointing to external knowledge needs. These observations demonstrate that the length of highlight spans and distribution vary systematically across reasoning types, offering evidence that different types of reasoning reveal distinct linguistic patterns in NLI explanations.

4 Explanation Generation using Taxonomy and Highlight

To investigate the interpretability and generalizability of our taxonomy, particularly in comparison to highlight approaches, we experiment on a practical usage: generating explanations with taxonomy or with highlight annotations. The goal is to generate, for a given NLI item and its label, multiple explanations that reflect different plausible reasoning paths. While collecting such varied human-authored explanations is expensive—and often infeasible to elicit from a single annotator—LLMs offer a scalable alternative (Chen et al., 2025b). We discuss various prompting paradigms (§4.1) and measure the similarities between LLM-generated and human explanations (§4.2).

4.1 Prompting Paradigms

We experiment with three prompting paradigms and evaluate our approach on three instruction-tuned LLMs: GPT-4o, DeepSeek-v3, and Llama-3.3-70B-Instruct, with full prompt templates presented in Appendix E.

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	
GPT4o *baseline*	0.291	0.117	0.049	0.882	0.488	0.226	0.556	0.524	0.051	0.272	24.995
highlight (indexed)	0.402	0.124	0.053	0.878	0.481	0.222	0.554	0.522	0.051	0.269	28.240
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**
highlight (indexed)	0.364	0.091	0.037	0.861	0.450	0.196	0.464	0.499	0.034	0.242	27.301
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
highlight (indexed)	0.317	0.065	0.024	0.807	0.408	0.173	0.367	0.478	0.031	0.199	24.987
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

Table 4: Similarity of LLM-generated explanations to human references.

Baseline The model only sees the NLI item (premise and hypothesis) and a label, and generates explanations based on this input.

Highlight-Guided Adding to the baseline inputs, we include highlight annotations of the premise and hypothesis—as indices (*indexed*) or tokens marked by surrounding ** in text (*in-text*). We ask the LLMs to first predict the highlighted tokens in the premise and hypothesis and subsequently generate relevant explanations. We report results in the *indexed* setup, as it yields marginally better average performance across metrics; see Appendix F for similar *in-text* results and when using e-SNLI highlights.

Taxonomy-Guided The model is provided with the taxonomy description (Table 1), one example for each of the eight reasoning categories, and the full taxonomy. We experiment with two prompting setups: *two-stage* and *end-to-end*. The *two-stage* setup separates classification and generation—first predicting the taxonomy label for a given NLI item, then generating explanations conditioned on it. The *end-to-end* approach performs both steps in a single prompt. This comparison addresses concerns that end-to-end generation may introduce a bias toward certain reasoning categories.

4.2 Model Generation Results

We evaluate similarities between LLM- and human-generated explanations using the same metrics as in §3.4. For each generated explanation, we evaluate it against the human-written references individually by computing all metrics. We then select the best-scoring reference for that explanation and retain its score. The score for each NLI item is then obtained by averaging over all its generated explanations. The final reported result is the average of these per-item scores across our entire dataset.

Table 4 reports our generation results. Notably, *end-to-end* taxonomy prompting performs best on GPT-4o and DeepSeek-v3, while *two-stage* prompting yields better performance on Llama 3.3. Across all models, taxonomy-guided generation achieves higher alignment with human explanations than both the baseline and highlight-based approaches. This is reflected in higher POS tag n-gram overlap, which captures morphosyntactic structural similarity, and in stronger semantic similarity metrics like Cosine. In contrast, highlight-guided explanations perform comparably or slightly worse than the baseline, and tend to have longer average lengths with lower lexical and semantic overlap with the references. This suggests that highlighting alone may not sufficiently inform the model to produce relevant explanations. It is also worth noting that the open-source Llama model performs on par with the closed-source GPT model.

While high similarity to human references is desirable, overly verbose content may indicate unnecessary redundancy (Holtzman et al., 2020). From Table 4, we observe that highlight-guided generations tend to produce longer explanations (e.g., 28.24 for GPT-4o and 30.42 for DeepSeek-v3) while yielding lower BLEU and ROUGE-L scores compared to both the baseline and taxonomy-guided variants. This indicates that the predicted highlights did not improve alignment with human-written explanations and may instead reflect redundancy. Rather, taxonomy-based methods result in higher similarity and more concise explanations. This effect, however, is driven by the taxonomy two-stage approach: while it produces notably

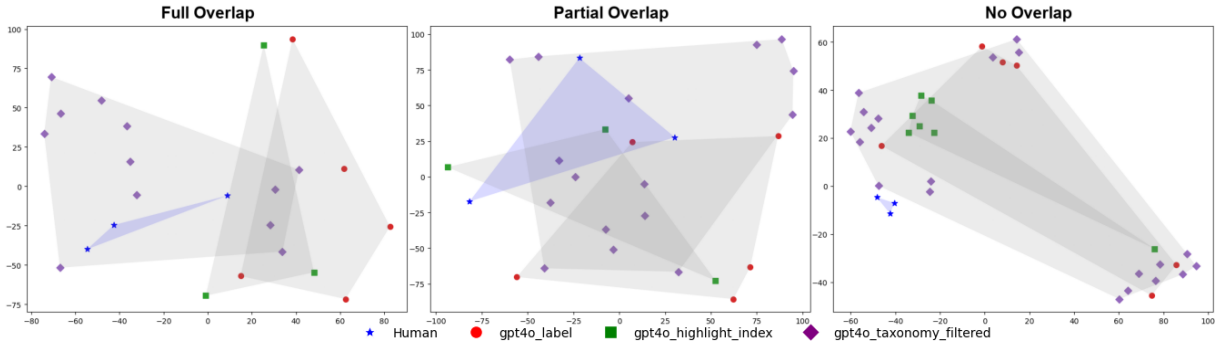

Figure 5: Representative t-SNE visualizations of explanation embeddings. The blue convex hull represents the span of human-written explanations, while the gray illustrates the spread of GPT4o-generated explanations.

shorter outputs, the end-to-end taxonomy variant often generates longer explanations, in some cases exceeding those of highlight-based methods for Llama.

4.3 Model Generation Validation

To assess the quality of model-generated NLI explanations, we conduct a round of human validation. Specifically, we evaluate 8,373 explanations produced by GPT-4o under the taxonomy *two-stage* prompting paradigm introduced in §4.1, as this setup yields a broader coverage of reasoning categories compared to the *end-to-end* variant, allowing for more comprehensive validation across the taxonomy. For each explanation, one trained annotator is provided with the premise, hypothesis, NLI label, the generated explanation, and the corresponding taxonomy category. The annotator is instructed to answer the following two binary questions:

1. NLI label consistency: Does the explanation fit the gold label? (Yes/No)

2. Taxonomy consistency: Does the explanation fit the taxonomy? (Yes/No)

The results show that overall 98.27% of the generated explanations align with the NLI label and 83.84% match the prompted taxonomy category. Some categories showed high alignment rates, such as *Syntactic* (94.88%) and *Absence of Mention* (92.47%), while others were more challenging, such as *Coreference* (57.25%) and *Logic Conflict* (63.05%). A detailed breakdown by taxonomy category, along with further discussion, is provided in Appendix G.

	Coverage		Area	
Mode	**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**

Table 5: Results on the semantic coverage of model explanations regarding human reference explanations.

5 Assessing Explanation Coverage: Human vs. LLM Outputs

Besides evaluating the similarity between human-written and LLM-generated explanations, the more fundamental question is *how much within-label variation can LLM-generated explanations capture.* Are LLMs too repetitive and only cover a subset of human explanations? Can LLMs unearth appropriate new explanations that are missing from a few human-written ones? This section presents our attempt to measure coverage in LLM-generated explanations. Given that LLMs are prompted to generate multiple explanations, we examine whether they can fully cover the semantic space of human explanations and potentially extend beyond it.

Figure 5 illustrates this semantic coverage for three representative instances from LITEX-SNLI. From left to right, the examples demonstrate: (1) full coverage, where the convex hull of model-generated explanations fully encloses the human explanation points; (2) partial coverage, where model generations cover some of the human reference points and (3) no coverage, where model outputs cover no human explanation point.

Premise:	*A person wearing a straw hat, standing outside working a steel apparatus with a pile of coconuts on the ground.*
Hypothesis:	*A person is near a pile of coconuts.*
Label:	Entailment

Source	**Explanations**
Human	- With a pile of coconuts implies near a pile of coconuts. [Semantic] - A person standing near a pile of coconuts would be standing outside. [Inferential Knowledge] - If the person is with the pile of coconuts on the ground, they are near them. [Inferential Knowledge]
Baseline	- There is no indication in the content that the person and the coconuts are separated by any significant distance. - The use of "near" in the statement aligns with the spatial arrangement described in the content, where the person and coconuts are part of the same setting. - The person is described as standing outside, suggesting they are near the objects in the scene.
Highlight	- The phrase "standing outside" in the content implies that the person is near the ground where the pile of coconuts is located, supporting the statement that the person is near the pile of coconuts. ["12,13,14,15,16", "3,4,5,6,7"] - The content describes "a pile of coconuts on the ground," which directly correlates to the statement's "a pile of coconuts," indicating the same pile is being referenced. ["6,7", "3"] - The person "working a steel apparatus" in the content might suggest proximity to the pile of coconuts since the work being done is likely related to the coconuts, thus supporting the statement that the person is near them. ["1,12,16,19", "1,3,7"]
Taxonomy	- Standing outside implies proximity to the objects on the ground, including the pile of coconuts. [Inferential Knowledge] - The word 'near' denotes closeness, which is implied by the person standing in the vicinity of the coconuts. [Semantic] - The hypothesis simplifies the structure of the premise by focusing on the person's proximity to the coconuts, which is implied by the premise. [Syntactic]

Table 6: Explanations from different generation strategies for one LITEX-SNLI item. For human explanations, annotator-assigned categories are in purple. Model-generated taxonomy categories and highlight indexes are in blue.

Proposed Measures We propose four measures, *full coverage*, *partial coverage*, *area precision*, and *area recall* to analyze the semantic space between model- and human-generated explanations using t-SNE visualizations and convex hull statistics (van der Maaten and Hinton, 2008).

We define *full coverage* as a binary condition: an NLI item is fully covered (yes) if all human explanation reference points fall within the convex hull spanned by the model explanations, and not covered (no) otherwise. Similarly, it is *partially covered* if at least one human reference point is within the model explanation space. *Full and partial coverage* computes the percentage of 1,002 LITEX-SNLI items whose explanations are fully or partially covered within the convex hull of the model explanations.

On the other hand, *area precision and recall* assess for each NLI item, the overlapping area between the space spanned by all reference explanations and that spanned by all model explanations. *Area precision* measures the ratio of the overlapping area over the area spanned by model explanations, and *area recall* over the area spanned by human explanations. We report the average of *area precision* and *area recall* over 1,002 instances.

Results Table 5 shows that taxonomy-guided explanation generation consistently achieves the highest full and partial coverage of reference explanation points. They also yield the highest average area recall and precision, in all test cases except the GPT4o baseline, indicating that the semantic space overlap between taxonomy-guided model explanations and human explanations is large.

In contrast, baseline and highlight-guided modes show much lower full and partial coverage and smaller overlap ratios. It indicates that the explanation spaces are less aligned with human explanations. Although highlight-guided outputs tend to form smaller and more concentrated explanation regions (as seen in their low area precision), this compactness does not mean their explanations are more meaningful. When guided by highlights, the model often fails to generate explanations that reflect the essential ideas expressed by humans. These results highlight that prompting using taxonomy-based guidance is more effective at generating human-aligned explanations in the embedding space.

We observe in Table 5 that GPT4o exhibits lower coverage compared to DeepSeek and Llama, which

can partly be attributed to the smaller number of explanations generated per instance. In our setup, the models are prompted to produce all possible explanations given an NLI instance, a label, and optionally a taxonomy category or relevant highlights. On average, GPT4o generates 3.59 explanations per example, while DeepSeek and Llama produce 5.90 and 6.14, respectively. This difference in output quantity naturally contributes to GPT4o's lower coverage.

Case Study Table 6 provides a concrete example (the leftmost case in Figure 5) where the human explanations are fully covered by the taxonomy-guided generation but only partially captured by label- and highlight-guided generations.

Human explanations focus on spatial proximity (*near*) and real-world expectations (i.e., coconuts being outdoors). The baseline and highlight-guided explanations also refer to the spatial proximity. However, the reasoning is less precise and often vague, lacking the structure seen in human explanations. Instead, taxonomy-guided generations are not only more coherent and concise, but also cover a broader range of reasoning types. In addition to producing outputs aligned with *Semantic* and *Inferential Knowledge*, they provide an additional *Syntactic*-labeled explanation, addressing the sentence simplification from premise to hypothesis.

However, while the taxonomy-generated explanation "standing outside implies proximity to the objects on the ground, including the pile of coconuts" captures the essence of the human-written "a person standing near a pile of coconuts would be standing outside," it is more abstract and less natural when expressing the casual contexts. All generated explanations, particularly highlight-guided ones, are also longer than the human-written ones, echoing the redundancy issue discussed in §4.2.

6 Conclusion

In this work, we introduce LITEX, a linguistically-informed taxonomy designed to capture different reasoning strategies behind NLI explanations, with a particular focus on within-label variation. The learnability evaluation shows that models, after fine-tuning or few-shot prompting, can effectively classify explanations into our taxonomy, demonstrating its practicality. We further demonstrate that taxonomy guidance consistently helps generation, resulting in model explanations that are semantically richer and closer to human explanations than baseline or highlight-based approaches.

Overall, our work bridges human reasoning strategies and model predictions in a structured way, providing a foundation for more interpretable NLI modeling. In addition, we enhance the e-SNLI dataset with fine-grained taxonomy categories for explanations, providing a resource to support future work. While our current evaluation focuses on a specific subset of NLI data, future work will extend this approach to broader variation-aware benchmarks such as ANLI (Nie et al., 2020a) and LiveNLI (Jiang et al., 2023). These extensions will enable a more comprehensive assessment of the taxonomy's generalizability across diverse inference settings. Annotations, generated explanations, and code will be released publicly upon publication.[3]

Limitations

While our taxonomy offers a structured and linguistically informed perspective to analyze different types of explanation in NLI, it has several limitations. First, the annotation process, though guided by detailed definitions, still involves subjective interpretation from a single annotator, especially for borderline categories such as *Factual Knowledge* versus *Inferential Knowledge*. This highlights the inherent subjectivity of explanation annotation, where different annotators may reasonably disagree on the most appropriate category. Second, our taxonomy focuses solely on explicit explanations provided in natural language. It does not account for the implicit reasoning process that may not be verbalized in text. This may limit the taxonomy's applicability to inferred or implied reasoning, especially when applying it to other NLI datasets without free-text explanations. Finally, our current experiments are conducted on the e-SNLI dataset, which may not represent the full spectrum of natural language inference.

Ethical considerations

We do not foresee any ethical concerns associated with this work. All analyses were conducted using publicly available datasets and models. No private or sensitive information was used. Additionally, we will release our code, prompts, and documentation to support transparency and reproducibility.

[3]Dataset and implementation are publicly available at https://github.com/mainlp/LiTEx for reproduction.

Acknowledgments

We thank the members of the MaiNLP lab for their insightful feedback on earlier drafts of this paper. We specifically appreciate the suggestions of Verena Blaschke, Andreas Säuberli, and Yang Janet Liu. Beiduo Chen acknowledges his membership in the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program. Marie-Catherine de Marneffe is a Research Associate of the Fonds de la Recherche Scientifique – FNRS. This research is supported by ERC Consolidator Grant DIALECT 101043235.

Use of AI Assistants The authors acknowledge the use of ChatGPT solely for assistance with grammar, punctuation, and vocabulary corrections, as well as for supporting coding tasks.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Beiduo Chen, Yang Janet Liu, Anna Korhonen, and Barbara Plank. 2025a. Threading the needle: Reweaving chain-of-thought reasoning to explain human label variation. *CoRR*, abs/2505.23368.

Beiduo Chen, Siyao Peng, Anna Korhonen, and Barbara Plank. 2025b. A rose by any other name: LLM-generated explanations are good proxies for human explanations to collect label distributions on NLI. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10777–10802, Vienna, Austria. Association for Computational Linguistics.

Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. 2024. "Seeing the big through the small": Can LLMs approximate human judgment distributions on NLI from a few explanations? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14396–14419, Miami, Florida, USA. Association for Computational Linguistics.

Ernest Davis. 2017. Logical formalizations of commonsense reasoning: A survey. *J. Artif. Int. Res.*, 59(1):651–723.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 80 others. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. 2023. Can large language models explain themselves? A study of LLM-generated self-explanations. *CoRR*, abs/2310.11207.

Filip Ilievski, Alessandro Oltramari, Kaixin Ma, Bin Zhang, Deborah L. McGuinness, and Pedro Szekely. 2021. Dimensions of commonsense knowledge. *Knowledge-Based Systems*, 229:107347.

Chathuri Jayaweera and Bonnie Dorr. 2025. From disagreement to understanding: The case for ambiguity detection in NLI. *CoRR*, abs/2507.15114.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine de Marneffe. 2023. Ecologically valid explanations for label variation in NLI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10622–10633. Association for Computational Linguistics.

Shiyang Li, Jianshu Chen, yelong shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2024. Explanations from large language models make small reasoners better. In *2nd Workshop on Sustainable AI*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 329–334, Portland, Oregon, USA. Association for Computational Linguistics.

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020a. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020b. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Mark Sammons, V.G.Vinod Vydiswaran, and Dan Roth. 2010. "Ask not what textual entailment can do for you...". In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1199–1208, Uppsala, Sweden. Association for Computational Linguistics.

Mandy Simons, Judith Tonhauser, David Beaver, and Craige Roberts. 2011. What projects and why. *Proceedings of SALT; Vol 20 (2010); 309-327*, 20.

Chenhao Tan. 2022. On the diversity and limits of human explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2173–2188, Seattle, United States. Association for Computational Linguistics.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

A Illustrative Examples of the Taxonomy

This section provides illustrative examples to clarify and exemplify our taxonomy. For each example, we present the premise, hypothesis, and human explanation as they appear in the original dataset, preserving all original text, including any typos or grammatical errors. In Table 7 and Table 8, two representative examples are listed for the two broad categories: *Text-Based (TB) Reasoning* and *World-Knowledge (WK) Reasoning*.

Coreference	
Premise:	The man in the black t-shirt is trying to throw something.
Hypothesis:	The man is in a black shirt.
Gold Label:	Entailment
Explanation:	The man is in a black shirt refers to the man in the black t-shirt.
Premise:	A naked man rides a bike.
Hypothesis:	A person biking.
Gold Label:	Entailment
Explanation:	The person biking in the hypothesis is the naked man.
Semantic	
Premise:	A man in a black tank top is wearing a red plaid hat.
Hypothesis:	A man in a hat.
Gold Label:	Entailment
Explanation:	A red plaid hat is a specific type of hat.
Premise:	Three man are carrying a red bag into a boat with another person and boat in the background.
Hypothesis:	Some people put something in a boat in a place with more than one boat.
Gold Label:	Entailment
Explanation:	Three men are people.
Syntactic	
Premise:	Two women walk down a sidewalk along a busy street in a downtown area.
Hypothesis:	The women were walking downtown.
Gold Label:	Entailment
Explanation:	The women were walking downtown is a rephrase of, Two women walk down a sidewalk along a busy street in a downtown area.
Premise:	Bruce Springsteen, with one arm outstretched, is singing in the spotlight in a dark concert hall.
Hypothesis:	Bruce Springsteen is a singer.
Gold Label:	Entailment
Explanation:	Springsteen is singing in a concert hall.
Pragmatic	
Premise:	A girl in a blue dress takes off her shoes and eats blue cotton candy.
Hypothesis:	The girl is eating while barefoot.
Gold Label:	Entailment
Explanation:	If a girl takes off her shoes, then she becomes barefoot, and if she eats blue candy, then she is eating.
Premise:	A woman wearing bike shorts and a skirt is riding a bike and carrying a shoulder bag.
Hypothesis:	A woman on a bike.
Gold Label:	Entailment
Explanation:	Woman riding a bike means she is on a bike
Absence of Mention	
Premise:	A person with a purple shirt is painting an image of a woman on a white wall.
Hypothesis:	A woman paints a portrait of a person.
Gold Label:	Neutral
Explanation:	A person with a purple shirt could be either a man or a woman. We can't assume the gender of the painter.
Premise:	A young man in a heavy brown winter coat stands in front of a blue railing with his arms spread.
Hypothesis:	The railing is in front of a frozen lake.
Gold Label:	Neutral
Explanation:	It does not say anything about there being a lake.
Logic Conflict	
Premise:	Five girls and two guys are crossing an overpass.
Hypothesis:	The three men sit and talk about their lives.
Gold Label:	Contradiction
Explanation:	Three is not two.
Premise:	Many people standing outside of a place talking to each other in front of a building that has a sign that says "HI-POINTE".
Hypothesis:	The group of people aren't inside of the building.
Gold Label:	Entailment
Explanation:	The people described are standing outside, so naturally not inside the building.

Table 7: Illustrative examples of the taxonomy (Text-Based Reasoning).

Factual Knowledge	
Premise:	Two people crossing by each other while kite surfing.
Hypothesis:	The people are both males.
Gold Label:	Neutral
Explanation:	Not all people are males.
Premise:	Here is a picture of people getting drunk at a house party.
Hypothesis:	Some people are by the side of a swimming pool party.
Gold Label:	Neutral
Explanation:	Not all houses have swimming pools.
Inferential Knowledge	
Premise:	A girl in a blue dress takes off her shoes and eats blue cotton candy.
Hypothesis:	The girl in a blue dress is a flower girl at a wedding.
Gold Label:	Neutral
Explanation:	A girl in a blue dress doesn't imply the girl is a flower girl at a wedding.
Premise:	A person dressed in a dress with flowers and a stuffed bee attached to it, is pushing a baby stroller down the street.
Hypothesis:	An old lady pushing a stroller down a busy street.
Gold Label:	Neutral
Explanation:	A person in a dress of a particular type need neither be old nor female. A street need not be considered busy if only one person is pushing a stroller down it.

Table 8: Illustrative examples of the taxonomy (World Knowledge-Based Reasoning).

B Taxonomy Validation: IAA Classification Report

Figure 6 presents the inter-annotator confusion matrix for explanation category annotation, used to validate the proposed taxonomy. Overall, we observe strong agreement across most categories, with especially high consistency in categories such as *Logic Conflict* and *Inferential Knowledge*. Some confusion appears between semantically adjacent categories, such as *Factual Knowledge* vs. *Inferential Knowledge*, and *Semantic* vs. *Syntactic*.

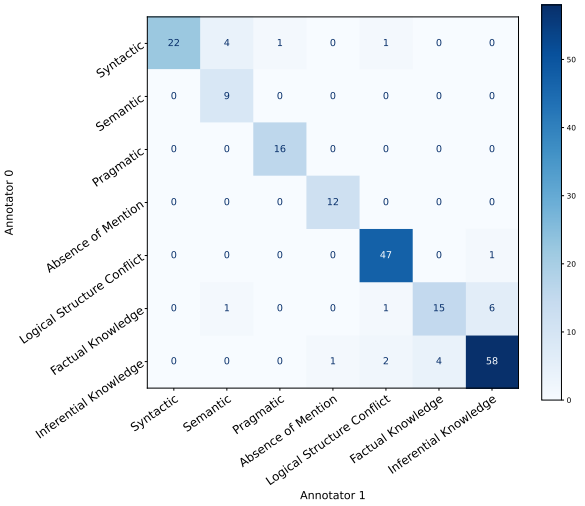

Figure 6: Inter-Annotator Confusion Matrix for Explanation Category Annotation.

To better understand the nature of inter-annotator disagreement in our taxonomy-based labeling, we present a qualitative analysis of several items with mismatched labels in the confusion matrix. The following examples shed light on how subtle differences in reasoning can lead to divergent category assignments:

Factual Knowledge vs. Logic Conflict
Premise: An old man with a package poses in front of an advertisement.
Hypothesis: A man walks by an ad.
Explanation: Poses is different from walks.
Category (Annotator 0): Factual Knowledge
Category (Annotator 1): Logic Conflict
Analysis: Annotator 0 likely interprets the explanation as highlighting a factual discrepancy in the physical action "posing" vs."walking"), treating this as a knowledge-based distinction about what the person is doing. Annotator 1, on the other hand, may view the same contrast as introducing a logical inconsistency in the event semantics—i.e., the man cannot be simultaneously posing and walking, which reflects a conflict in entailment assumptions. This illustrates how borderline cases between fact-based knowledge and event logic can be interpreted differently, especially when both literal and inferential mismatches are present.

Inferential vs. Factual Knowledge
Premise: A young family enjoys feeling ocean waves lap at their feet.
Hypothesis: A young man and woman take their child to the beach for the first time.
Explanation: The young family does not mean that they have a child at the beach.
Category (Annotator 0): Inferential Knowledge
Category (Annotator 1): Factual Knowledge
Analysis: Annotator 0 interprets the inference from "young family" to "having a child present" as requiring reasoning with world knowledge about family structures. In contrast, Annotator 1 views this as an incorrect factual claim, where the hypothesis wrongly assumes a child is present. This disagreement highlights the challenge of distinguishing between inferential reasoning and factual correction, indicating a need for clearer taxonomy boundaries.

Syntactic vs. Semantic
Premise: Two children are laying on a rug with some wooden bricks laid out in a square between them.
Hypothesis: Two children are on a rug.
Explanation: To say the children are "laying on" a rug is rephrasing "on" a rug.
Category (Annotator 0): Syntactic
Category (Annotator 1): Semantic
Analysis: Annotator 0 classifies the change from "laying on" to "on" as a simple syntactic variation, treating it as a surface-level rewording. In contrast, Annotator 1 interprets this shift as semantically meaningful, possibly inferring that "laying on" conveys posture or state, thus labeling it as a Semantic shift. This disagreement illustrates a key challenge in NLI: distinguishing between purely syntactic paraphrases and cases where subtle wording changes alter meaning. Such distinctions become especially nuanced when modifications involve minor phrasing differences.

C Taxonomy Validation: LM and LLM Classification

In Table 10 the hyperparameter setup of fine-tuning BERT and RoBERTa is listed. We follow a standard supervised classification pipeline, where the model takes as input the concatenated premise, hypothesis, label, and explanation, and predicts the correct explanation category among eight categories. For validation, we measured both the classification accuracy and the macro-F1 score across the explanation categories, as shown in Table **??**. We selected the best-performing checkpoint based on the highest macro-F1 on the dev set for final evaluation.

We design a set of experiments to assess the ability of LLMs to classify NLI explanations into one of eight fine-grained categories (introduced in Section 3). Our evaluation covers zero-shot prompting (no training examples), one-shot prompting (a single annotated example), and few-shot prompting (two examples per category). A consistent prompting strategy is applied across models, with all templates provided in Table 11.

Hyperparameter	**BERT**	**RoBERTa**
Learning Rate Decay	Linear	Linear
Weight Decay	0.0	0.0
Optimizer	AdamW	AdamW
Adam ϵ	1e-8	1e-8
Adam β_1	0.9	0.9
Adam β_2	0.999	0.999
Warmup Ratio	0%	0%
Learning Rate	2e-5	3e-5
Batch Size	8	8
Num Epoch	4	3

Table 10: Hyperparameter used for fine-tuning BERT and RoBERTa models.

Specifically, we experiment with Llama-3.2-3B-Instruct (Meta, 2024), GPT-3.5-turbo (Brown et al., 2020), GPT-4o (OpenAI, 2023), and DeepSeek-v3 (DeepSeek-AI et al., 2024), under six experimental configurations:

1. without instruction and without examples
2. with general task instruction but no examples
3. with one example per category
4. with two representative examples per category
5. with instruction plus one example per category
6. with instruction plus two examples per category

For the few-shot settings, one or two representative examples from the training set were selected for each of the eight categories and incorporated into the prompt. The LLMs were instructed to output the category index (1–8) for each explanation. We evaluate both classification accuracy and the distributional alignment between LLM predictions and the gold human label distributions, as reported in Table 12.

From the results, we observe that GPT-4o consistently delivers the strongest performance across most experimental configurations, achieving its highest accuracy of 0.594 in the + *one example per category* setting. Its macro-F1 reaches 0.492 in the + *instruction* + *one example per category* setting, while its weighted-F1 peaks at 0.583 in the + *one example per category* setting, both surpassing all other LLMs. To gain deeper insight, we further analyze the confusion cases under the + one example per category setting, focusing on GPT-4o as a representative model.

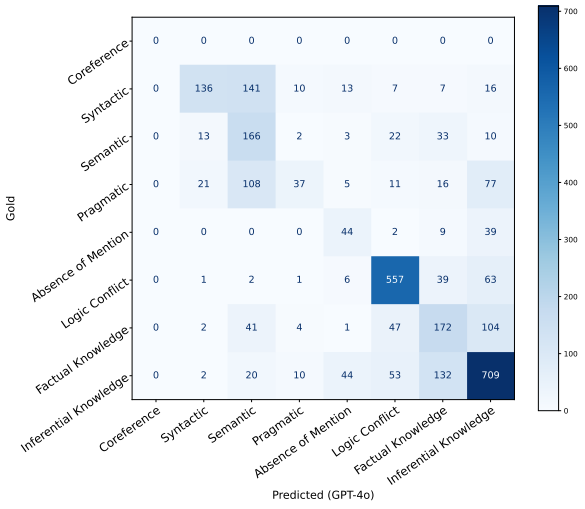

Figure 7: Confusion matrix of GPT-4o on the NLI explanation classification task using the + *one example per category* prompting style.

Figure 7 provides a detailed view of the error patterns in GPT-4o's predictions. We observe that *Syntactic* and *Semantic* are frequently confused, indicating that the model has difficulty capturing the fine-grained distinction between structural and meaning-oriented reasoning. Similarly, a considerable number of *Factual Knowledge* instances are mislabeled as *Inferential Knowledge*, suggesting that GPT-4o often fails to separate lexical associations from broader factual inferences. To further illustrate these confusions, consider the following examples:

Syntactic vs. Semantic

Mode	General Instruction Prompt
without instruction and example	**"role": "user", "content":** You are an expert in solving Natural Language Inference tasks. Your task is to classify the following explanations into one of the categories listed below. Each category reflects a specific type of inference in the explanation between the premise and hypothesis. Here are the categories: 1. Coreference 2. Syntactic 3. Semantic 4. Pragmatic 5. Absence of Mention 6. Logic Conflict 7. Factual Knowledge 8. Inferential Knowledge
+ *instruction*	**"role": "user", "content":** You are an expert in solving Natural Language Inference tasks. Your task is to classify the following explanations into one of the categories listed below. Each category reflects a specific type of inference in the explanation between the premise and hypothesis. Here are the categories: 1. Coreference - The explanation resolves references (e.g., pronouns or demonstratives) across premise and hypothesis. 2. Syntactic - Based on structural rephrasing with the same meaning (e.g., syntactic alternation, coordination, subordination). If the explanation itself is the rephrasing of the premise or hypothesis, it should be included in this category. 3. Semantic - Based on word meaning (e.g., synonyms, antonyms, negation). 4. Pragmatic - This category would capture inferences that arise from logical implications embedded in the structure or semantics of the text itself, without relying on external context or background knowledge. 5. Absence of Mention - Lack of supporting evidence, the hypothesis introduces information that is not supported, not entailed, or not mentioned in the premise, but could be true. 6. Logic Conflict - Structural logical exclusivity (e.g., either-or, at most, only, must), quantifier conflict, temporal conflict, location conflict, gender conflict etc. 7. Factual Knowledge - Explanation relies on common sense, background, or domain-specific facts. No further reasoning involved. 8. Inferential Knowledge - Requires real-world causal, probabilistic reasoning or unstated but assumed information. Respond **only with the number (1–8)** corresponding to the most appropriate category.

Table 11: Instruction prompts for LLMs as classifiers.

Premise: A man in a black shirt overlooking bike maintenance.
Hypothesis: A man watches bike repairs.
Explanation: A man is watching the bike maintenance which is repairs.
Category (Human): Syntactic
Category (GPT-4o): Semantic
Analysis: Human annotators classify this case as *Syntactic*, since the paraphrase between "maintenance" and "repairs" is treated as a surface-level syntactic variation. In contrast, GPT-4o labels it as *Semantic*, interpreting the paraphrase primarily as a meaning equivalence rather than a structural rewording.

Inferential vs. Factual Knowledge
Premise: A blond-haired doctor and her African American assistant looking through new medical manuals.
Hypothesis: A doctor is studying.
Explanation: Answer: Just because the doctor is studying it doesn't mean he is reading medical manuals.
Category (Human): Inferential Knowledge
Category (GPT-4o): Factual Knowledge
Analysis: Human annotators label this case as *Inferential Knowledge*, since the reasoning requires recognizing the pragmatic gap between "studying" in general and "studying medical manuals" in particular. GPT-4o, however, classifies it as *Factual Knowledge*, suggesting that it grounds the judgment in the surface facts of the premise rather than modeling the inference beyond what is explicitly mentioned.

In contrast, *Absence of Mention* is classified with high reliability, as reflected by the strong diagonal concentration in its row. These observations highlight that GPT-4o is more robust when reasoning relies on explicit absence cues, while it struggles

Classifiers	Accuracy	Precision		Recall		F1		*Invalid predictions*
		macro	*weighted*	macro	*weighted*	macro	*weighted*	
Llama-3.2-3B-Instruct	0.357	0.440	0.581	0.373	0.357	0.291	0.310	0 (0.00%)
+ instruction	0.229	0.379	0.465	0.281	0.229	0.227	0.256	918 (29.54%)
+ one example per category	0.340	0.393	0.540	0.343	0.340	0.255	0.293	23 (0.74%)
+ two example per category	0.160	0.243	0.302	0.252	0.160	0.139	0.163	277 (8.91%)
+ instruction + one example per category	0.357	0.440	0.581	0.272	0.357	0.291	0.310	0 (0.00%)
+ instruction + two example per category	0.538	0.484	0.591	0.402	0.538	0.397	0.522	0 (0.00%)
gpt-3.5-turbo	0.289	0.264	0.351	0.286	0.289	0.239	0.279	0 (0.00%)
+ instruction	0.366	0.314	0.431	0.357	0.366	0.295	0.336	0 (0.00%)
+ one example per category	0.175	0.162	0.244	0.155	0.175	0.139	0.182	28 (0.90%)
+ two example per category	0.297	0.281	0.403	0.265	0.297	0.237	0.308	1 (0.03%)
+ instruction + one example per category	0.274	0.286	0.393	0.264	0.274	0.236	0.290	36 (1.16%)
+ instruction + two example per category	0.305	0.317	0.420	0.301	0.305	0.262	0.303	8 (0.26%)
gpt-4o	0.433	0.402	0.495	0.409	0.433	0.321	0.411	0 (0.00%)
+ *instruction*	0.410	0.465	0.536	0.438	0.410	0.357	0.404	0 (0.00%)
+ *one example per category*	**0.594**	0.530	0.619	0.486	**0.594**	0.476	**0.583**	0 (0.00%)
+ *two example per category*	0.589	**0.545**	0.631	0.532	0.589	0.491	0.579	0 (0.00%)
+ *instruction + one example per category*	0.583	0.550	0.643	0.548	0.583	0.491	0.578	0 (0.00%)
+ *instruction + two example per category*	0.574	0.541	**0.648**	0.552	0.574	**0.492**	0.573	0 (0.00%)
DeepSeek-v3	0.340	0.306	0.409	0.389	0.340	0.268	0.312	1 (0.03%)
+ instruction	0.422	0.423	0.508	0.480	0.422	0.369	0.388	0 (0.00%)
+ one example per category	0.540	0.483	0.592	0.514	0.540	0.461	0.529	0 (0.00%)
+ two example per category	0.560	0.498	0.611	0.520	0.560	0.475	0.552	0 (0.00%)
+ instruction + one example per category	0.495	0.504	0.603	0.544	0.495	0.453	0.474	0 (0.00%)
+ instruction + two example per category	0.526	0.519	0.626	**0.563**	0.526	0.478	0.515	0 (0.00%)

Table 12: LLM as classifiers results.

when categories require distinguishing subtle linguistic or knowledge-based inferences.

To further assess the impact of supervised adaptation, we finetune Llama-3.2-3B-Instruct using LoRA (Hu et al., 2022), a parameter-efficient fine-tuning method. We adopt a 50/50 train-test split based on pairID. Fine-tuning is conducted using SFTTrainer with standard causal language modeling objectives and a maximum input length of 512 tokens. The LoRA configuration used is displayed in Table 13.

Hyperparameter	Value
Model	Llama-3.2-3B-Instruct
Gradient Accumulation	4
Max Sequence Length	512
Warmup Steps	50
Scheduler	Cosine
Learning Rate	2e-4
Batch Size	4
Num Epoch	3
Trainer	SFTTrainer (TRL)

Table 13: Training hyperparameters used for LoRA fine-tuning on Llama-3.2-3B. LoRA settings: $r = 8$, $\alpha = 16$, dropout = 0.05.

The fine-tuned Llama-3.2-3B model achieves an accuracy of 0.509 and a macro-F1 score of 0.302 on the test set. Detailed per-category results are presented in Table 14.

Explanation Category	Precision	Recall	F1
Coreference	0.429	0.052	0.092
Semantic	0.250	0.489	0.331
Syntactic	0.548	0.182	0.273
Pragmatic	0.273	0.200	0.231
Absence of Mention	0.000	0.000	0.000
Logic Conflict	0.735	0.758	0.746
Factual Knowledge	0.138	0.041	0.064
Inferential Knowledge	0.562	0.861	0.680
Summary			
accuracy		0.509	
F1 Score (macro)		0.302	

Table 14: LoRA fine-tuning results using Llama-3.2-3B-Instruct on the explanation categorization task.

While zero-shot prompting offers a lightweight baseline, these results suggest that parameter-efficient fine-tuning can boost performance in structured reasoning categories such as *Logical Conflict* and *Inferential Knowledge*. However, performance remains limited in categories such as *Factual Knowledge*, which require external world knowledge, and *Absence of Mention*, where low performance may be attributed to the small number of training examples.

We accessed GPT-3.5 and GPT-4o via OpenAI's hosted API and DeepSeek-V3 via DeepSeek's hosted API. Experiments with Llama-3.2-3B-

Instruct were run on a single NVIDIA A100 GPU.

D Human Highlight IAA

To understand whether human-generated highlights are consistent and reproducible, we conducted a highlight-level inter-annotator agreement (IAA) study on 201 items from the e-SNLI dataset. Two annotators were asked to highlight the parts of the premise and hypothesis that support the given explanation. Each item included the premise, hypothesis, gold label and the explanation.

We measured agreement using Intersection over Union (IoU). The results are as follows:

- **Annotator 1 vs Annotator 2**: 0.889
- **Annotator 1 vs e-SNLI Highlight**: 0.659
- **Annotator 2 vs e-SNLI Highlight**: 0.712

These results show that the two annotators had high agreement with each other, suggesting that the highlighting task is fairly consistent when done by different people. However, their agreement with the original e-SNLI highlights is lower, which means there are some differences in how people choose text spans, even when they agree on the explanation. This may be partially attributed to differences in annotation setup: in e-SNLI, the same annotator provided the NLI label, explanation, and highlight jointly, whereas in our IAA study, annotators re-annotated highlights for a given explanation under fixed label and span constraints. Although we adopted the same span-level constraints as e-SNLI (e.g., highlighting only the hypothesis for neutral items), our task required linking highlights to prewritten explanations rather than authoring them jointly, introducing a structural difference that may affect highlight choices.

E Prompting Templates for Generating Model Explanations

For the generation experiments, we prompt three LLMs to generate NLI explanations: GPT-4o, DeepSeek-V3, and Llama-3.3-70B-Instruct. We accessed GPT-4o via OpenAI's hosted API and DeepSeek-V3 via DeepSeek's hosted API. The generation experiments using Llama-3.3-70B-Instruct were conducted on two NVIDIA A100 GPUs.

Table 15 presents the prompt templates used to generate NLI explanations from LLMs. These templates are adapted and refined based on the approach of Chen et al. (2024). For LLMs that imply a "system" role within their chat format, the "system" role content is unset to maintain alignment with the design choices applied to other LLMs.

Mode	General Instruction Prompt
baseline	You are an expert in Natural Language Inference (NLI). Please list all possible explanations for why the following statement is {gold_label} given the content below without introductory phrases. Context: {premise}, Statement: {hypothesis}
highlight indexed	You are an expert in Natural Language Inference (NLI). Your task is to generate possible explanations for why the following statement is {gold_label}, focusing on the highlighted parts of the sentences. Context: {premise}, Highlighted word indices in Context: {highlighted_1} Statement: {hypothesis}, Highlighted word indices in Statement: {highlighted_2} Please list all possible explanations without introductory phrases.
highlight in-text	You are an expert in Natural Language Inference (NLI). Your task is to generate possible explanations for why the following statement is {gold_label}, focusing on the highlighted parts of the sentences. Highlighted parts are marked in "**". Context: {marked_premise} Statement: {marked_hypothesis} Please list all possible explanations without introductory phrases.
highlight generation	You are an expert in NLI. Based on the label 'gold_label', highlight relevant word indices in the premise and hypothesis. Highlighting rules: For entailment: highlight at least one word in the premise. For contradiction: highlight at least one word in both the premise and the hypothesis. For neutral: highlight only in the hypothesis. Premise: {premise}, Hypothesis: {hypothesis}, Label: {gold_label} Please list **3** possible highlights using word index in the sentence without introductory phrases. Answer using word indices **starting from 0** and include punctuation marks as tokens (count them). Respond strictly this format: Highlight 1: Premise_Highlighted: [Your chosen index(es) here] Hypothesis_Highlighted: [Your chosen index(es) here] Highlight 2: ...
taxonomy (two-stage)	You are an expert in Natural Language Inference (NLI). Given the following taxonomy with description and one example, generate as many possible explanations as you can that specifically match the reasoning type described below. The explanation is for why the following statement is {gold_label}, given the content. The explanation category for generation is: {taxonomy_idx}: {description} Here is an example: Premise: {few_shot['premise']}, Hypothesis: {few_shot['hypothesis']} Label: {few_shot['gold_label']}, Explanation: {few_shot['explanation']} Now, consider the following premise and hypothesis: Context: {premise} Statement: {hypothesis} Please list all possible explanations for the given category without introductory phrases.
taxonomy end-to-end	You are an expert in Natural Language Inference (NLI). Your task is to examine the relationship between the following content and statement under the given gold label, and: First, identify all categories for explanations from the list below (you may choose more than one) that could reasonably support the label. Second, for each selected category, generate all possible explanations that reflect that type. The explanation categories are: {taxonomy_idx}: {description} Context: {premise}, Statement: {hypothesis}, Label: {gold_label} Please list all possible explanations without introductory phrases for all the chosen categories. Start directly with the category number and explanation, following the strict format below: 1. Coreference: - [Your explanation(s) here] ... (continue for all reasonable categories)
taxonomy two-stage (classification)	You are an expert in Natural Language Inference (NLI). Your task is to identify all applicable reasoning categories for explanations from the list below that could reasonably support the label. Please choose at least one category and multiple categories may apply. One example for each category is listed as below: {examples_text} Given the following premise and hypothesis, identify the applicable explanation categories: Premise: {premise}, Hypothesis: {hypothesis}, Label: {gold_label} Respond only with the numbers corresponding to the applicable categories, separated by commas, and no additional explanation.

Table 15: Instruction prompts for LLMs to generate NLI explanations (all prompts are issued as user messages in the chat format).

F Additional Generation Results

Table 16 presents the full evaluation results of our explanation generation experiments, covering two highlight formats (indexed vs. in-text) and both human-provided and model-generated highlights.

Human Highlights vs. Model Generated Highlights Overall, model highlights achieve comparable performance to human highlights across most lexical and semantic metrics, with slight improvements in certain surface-level features (e.g., BLEU, ROUGE-L). However, these gains are often marginal. Notably, models like Llama-3.3-70B show a larger drop in similarity metrics when using model-generated highlights, indicating that automatic highlight classification may not always align with human judgment.

Indexed vs. In-text We compare the indexed and in-text variants of human and model highlights to assess whether highlight format affects similarity scores. Across all three models, the performance differences between the two formats are generally minor with the indexed variant performing slightly better. For instance, GPT-4o yields similar scores in both settings (e.g., cosine: 0.549 vs. 0.519 for human highlights; 0.554 vs. 0.555 for model highlights). The same trend holds for DeepSeek-v3 and Llama-3.3-70B, where average performance differences across metrics remain negligible.

G Human Validation of Model-Generated Explanations

Table 17 reports human validation results for model-generated explanations, broken down by taxonomy category. This analysis helps us better examine how explanation faithfulness and taxonomy alignment vary across different reasoning types.

Taxonomy	Q1 Yes (%)	Q1 No (%)	Q2 Yes (%)	Q2 No (%)
Coreference	269 (97.46)	7 (2.54)	158 (57.25)	118 (42.75)
Syntactic	780 (99.87)	1 (0.13)	741 (94.88)	40 (5.12)
Semantic	1716 (95.12)	88 (4.88)	1273 (70.57)	531 (29.43)
Pragmatic	131 (99.24)	1 (0.76)	109 (82.58)	23 (17.42)
Absence of Mention	3794 (99.16)	32 (0.84)	3538 (92.47)	288 (7.53)
Logic Conflict	428 (98.85)	5 (1.15)	273 (63.05)	160 (36.95)
Factual Knowledge	949 (99.16)	8 (0.84)	789 (82.45)	168 (17.55)
Inferential Knowledge	161 (98.17)	3 (1.83)	139 (84.76)	25 (15.24)

Table 17: Human validation results for model-generated explanations by taxonomy category. Q1: Whether the explanation supports the gold label. Q2: Whether the explanation matches the assigned taxonomy.

Across all categories, validation question 1 — evaluating whether the explanation supports the gold NLI label — yields consistently high agreement, with most categories exceeding 98% "Yes" responses. This indicates that the generated explanations are largely faithful to the NLI decision, regardless of the reasoning type. In contrast, validation question 2 — assessing whether the explanation aligns with the specified taxonomy — shows greater variation across categories. Categories such as *Syntactic* and *Absence of Mention* achieve the highest taxonomy agreement, with 94.88% and 92.47% of explanations remaining consistent with their respective reasoning types. These categories tend to involve explicit cues, which may be easier for LLMs to identify and replicate during generation. For example, explanations like "A is a rephrase of B" or "A in the premise is rephrased in the hypothesis" are common and prototypical forms of the *Syntactic* category. Similarly, in the *Absence of Mention* category, model outputs often include patterns such as "The premise discusses A but does not mention B" or "A is absent from the premise", which directly map onto the intended reasoning structure and are relatively easy to pattern-match.

In contrast, categories like *Coreference* (57.25%) and *Logic Conflict* (63.05%) show significantly lower alignment with the taxonomy categories. These types require discourse-level understanding or implicit logical inference, such as tracking entity references across clauses or identifying contradictions in different logical forms (temporal contradiction, location contradiction, gender conflict, etc.). Such reasoning is more abstract and difficult to control through prompting, which likely explains the increased rate of taxonomy mismatches.

Categories such as *Semantic*, *Factual Knowledge*, and *Inferential Knowledge* fall in an intermediate range (70–85%), likely due to their broader and more flexible definitions. For instance, semantic reasoning can often overlap with world knowledge or pragmatic cues, making it harder for models (and annotators) to sharply distinguish the boundaries of the category. This pattern is consistent with our IAA findings reported in §3.3, where we observed lower precision for *Semantic* (0.643) and lower recall for *Factual Knowledge* (0.652). These results point to potential ambiguities in distinguishing these categories from others, particularly from *Inferential Knowledge*.

Mode	Cosine	Euclidean	1gram		2gram		3gram		BLEU	ROUGE-L	Avg_len
			Word	POS	Word	POS	Word	POS			
GPT4o											
baseline	0.556	0.524	0.291	0.882	0.117	0.488	0.049	0.226	0.051	0.272	24.995
human highlight (indexed)	0.549	0.521	0.395	0.882	0.116	0.478	0.050	0.219	0.047	0.264	30.771
human highlight (in-text)	0.519	0.511	0.367	0.873	0.085	0.442	0.031	0.187	0.034	0.269	28.606
model highlight (indexed)	0.554	0.522	0.402	0.878	0.124	0.481	0.053	0.222	0.051	0.269	28.240
model highlight (in-text)	0.555	0.523	0.380	0.888	0.109	0.468	0.044	0.208	0.044	0.270	28.160
model taxonomy (two-stage)	0.593	0.537	0.418	0.886	0.128	0.495	0.071	0.242	0.071	0.314	19.991
model taxonomy (end-to-end)	**0.608**	**0.540**	**0.437**	**0.898**	**0.166**	**0.511**	**0.083**	**0.255**	**0.074**	**0.323**	26.672
DeepSeek-v3											
baseline	0.428	0.490	0.369	0.847	0.087	0.449	0.034	0.195	0.042	0.245	20.288
human highlight (indexed)	0.463	0.498	0.358	0.864	0.084	0.436	0.033	0.184	0.035	0.243	29.293
human highlight (in-text)	0.551	0.522	0.362	0.885	0.091	0.449	0.033	0.191	0.036	0.261	28.527
model highlight (indexed)	0.464	0.499	0.364	0.861	0.091	0.450	0.037	0.196	0.034	0.242	27.301
model highlight (in-text)	0.447	0.457	0.341	0.869	0.073	0.422	0.026	0.171	0.030	0.248	31.328
model taxonomy (two stage)	0.544	0.522	0.391	0.884	0.122	0.475	0.055	0.219	0.057	0.293	20.894
model taxonomy (end-to-end)	**0.556**	**0.528**	**0.404**	**0.897**	**0.140**	**0.486**	**0.067**	**0.233**	**0.063**	**0.306**	25.960
Llama-3.3-70B											
baseline	0.466	0.496	0.392	0.863	0.106	0.478	0.044	0.224	0.046	0.250	27.148
human highlight (indexed)	0.453	0.484	0.362	0.859	0.082	0.446	0.031	0.194	0.035	0.228	29.912
human highlight (in-text)	0.499	0.505	0.348	0.875	0.059	0.415	0.019	0.165	0.024	0.270	34.827
model highlight (indexed)	0.367	0.478	0.317	0.807	0.065	0.408	0.024	0.173	0.031	0.199	24.987
model highlight (in-text)	0.400	0.486	0.300	0.831	0.047	0.385	0.014	0.150	0.021	0.227	29.763
model taxonomy (two-stage)	**0.609**	**0.541**	**0.444**	0.889	**0.167**	**0.512**	**0.082**	**0.256**	**0.078**	0.321	22.340
model taxonomy (end-to-end)	0.505	0.510	0.383	**0.896**	0.110	0.499	0.048	0.232	0.047	0.262	28.870

Table 16: Full evaluation results for LLM-generated explanations (lexical, morphosyntactic, semantic, and summarization levels).