LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Masterarbeit

im Studiengang Computerlinguistik

an der Ludwig- Maximilians- Universität München

Fakultät für Sprach- und Literaturwissenschaften

Within-Label Variation in Natural Language Inference: A Linguistic Taxonomy for Explanations and Its Impact on Model Interpretation of Label Decisions

vorgelegt von
Pingjun Hong

Betreuer:	M.Eng. Beiduo Chen
Prüfer:	Prof. Dr. Barbara Plank
Bearbeitungszeitraum:	11. März - 28. Juli 2025

Selbstständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig angefertigt, alle Zitate als solche kenntlich gemacht sowie alle benutzten Quellen und Hilfsmittel angegeben habe.

München, den 28. Juli 2025

.....
Pingjun Hong

Erklärung der verwendeten KI-Tools

Ich versichere, dass ich diese Arbeit eigenständig, ohne jede externe Unterstützung, außer den unten aufgeführten Ressourcen, angefertigt habe.

Purpose	**Section(s)**	**Tool**
Grammar check	Entire document	ChatGPT-4o
Translations	Abstract	DeepL, ChatGPT-4o
Code generation	annotation interfaces, explanation classification and generation codes	GitHub Copilot, ChatGPT-4o

Table 0.1: Tools used for different parts of the project

München, den 28. Juli 2025

.....
Pingjun Hong

Abstract

English

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. 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 usefulness 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.

As an initial step toward evaluating cross-dataset applicability, we apply the taxonomy to LiveNLI and VariErr, two NLI datasets with distinct annotation characteristics. We find that the taxonomy can be consistently applied and helps surface dataset-specific trends in explanation diversity and label alignment.

Deutsch

Es gibt zunehmend Belege für Human Label Variation (HLV) im Bereich der Natural Language Inference (NLI), bei der Annotatorinnen demselben Prämisse-Hypothese-Paar unterschiedliche Labels zuweisen. Eine bisher weitgehend übersehene Herausforderung stellt jedoch die Variation innerhalb eines Labels dar – Fälle, in denen sich Annotatorinnen zwar auf dasselbe Label einigen, jedoch unterschiedliche Begründungen dafür liefern. Mehrere NLI-Datensätze enthalten markierte Wörter als Erklärungen, doch dieselben Textstellen können aus unterschiedlichen Gründen hervorgehoben worden sein. Dies zeigen insbesondere Freitext-Erklärungen, die Einblicke in die Denkprozesse der Annotator*innen ermöglichen.

Um dieses Phänomen systematisch zu untersuchen und die Begründungen hinter NLI-Labels besser zu verstehen, führen wir LITEX ein – eine linguistisch fundierte Taxonomie zur Kategorisierung von Freitext-Erklärungen.

Auf Basis dieser Taxonomie annotieren wir einen Teil des e-SNLI-Datensatzes, überprüfen die Zuverlässigkeit der Taxonomie und analysieren deren Zusammenhang mit Labels, Hervorhebungen und Erklärungen im NLI.

Darüber hinaus evaluieren wir den Nutzen der Taxonomie für die automatische Generierung von Erklärungen. Wir zeigen, dass durch eine Generierung, die auf LITEX konditioniert ist, sprachlich menschenähnlichere Erklärungen entstehen als bei Modellen, die nur Labels oder Hervorhebungen berücksichtigen. Unser Ansatz erfasst somit nicht nur die Variation innerhalb eines Labels, sondern zeigt auch, wie eine taxonomiegesteuerte Erklärungs-Generierung die Lücke zwischen menschlichen und maschinellen Erklärungen effektiver überbrücken kann als bisherige Strategien.

Als ein erster Schritt zur Bewertung der Anwendbarkeit der Taxonomie über verschiedene Datensätze hinweg wenden wir sie auf LiveNLI und VariErr an – zwei NLI-Datensätze mit unterschiedlichen Annotationsmerkmalen. Unsere Analyse zeigt, dass die Taxonomie konsistent anwendbar ist und dabei hilft, datensatzspezifische Muster in der Vielfalt von Erklärungen und der Zuordnung zu NLI-Labels sichtbar zu machen.

Acknowledgment

At this point, I would like to sincerely thank everyone who supported me along the way, not just during the time I spent writing this thesis, but throughout my entire Master's journey.

First and foremost, I would like to thank my advisor, Beiduo Chen, for his continuous support, constructive feedback, and patient guidance throughout the entire thesis process. He not only introduced the initial idea for this work but also provided valuable insights that helped shape the direction and clarity of the research. I am also deeply grateful to Logan Peng, Prof. Marie-Catherine de Marneffe, and Prof. Barbara Plank for their thoughtful suggestions and encouragement. Their academic input helped transform a preliminary idea into a structured and well-developed research project. I especially appreciate their support in refining the scope of this thesis and contributing to the development of a publishable outcome.

Thanks are also due to Prof. Plank and the entire MaiNLP Lab, who welcomed and trusted an inexperienced graduate student. Being part of the lab, working on various research projects, and learning from the team opened the door to the academic world for me. It has been a privilege to work with such dedicated and inspiring people, people I sincerely hope to cross paths with again, as the German saying goes: „Man sieht sich immer zweimal im Leben."

Last but not least, I wish to thank my boyfriend Xiyan for his support throughout this journey. Your unwavering love has been a constant source of strength, especially during moments of self-doubt and stress. Thank you for standing by me, for being my best friend and life partner, for listening without hesitation, and for reminding me to keep going. Your presence has meant the world to me.

Contents

1 Introduction

1.1 Natural Language Inference

Natural Language Inference (NLI), also known as Recognizing Textual Entailment (RTE), is a fundamental task in natural language understanding that aims to determine whether a hypothesis can be inferred from a given premise. The origin of the task can be traced back to the PASCAL Recognizing Textual Entailment Challenge (Dagan et al., 2005). It aimed to define a general framework for semantic inference that could support various NLP applications, including question answering, information extraction, and summarization.

In the RTE framework, the task involves a pair of text fragments. One is the **Text** (T), which serves as the premise, and the other is the **Hypothesis** (H). The system must decide whether H logically follows from T — that is, whether T entails H. More concretely, textual entailment is formulated as a **directional semantic relation**: T entails H if, typically, a human reading T would conclude that H is most likely true. This practical definition of entailment relies on common human understanding of language and background knowledge. It is intentionally kept somewhat informal to make it applicable to a wide range of real-world scenarios (Condoravdi et al., 2003).

Although practical and intuitive, this informal definition of entailment has been widely discussed and revised over the years (Pavlick and Kwiatkowski, 2019). As noted by Dagan et al. (2005), the definition is initially "clearly not mature yet," prompting subsequent work to clarify the role of world knowledge and to distinguish between different types of inferences (e.g., entailment vs. implicature). For instance, Zaenen et al. (2005) advocate for more precise definitions to separate formal entailment from other pragmatic inferences, while Simons et al. (2011) argues that annotation tasks should remain "natural" to untrained annotators and reflect the inferences humans make in everyday communication.

Over time, the field has shifted toward favoring a more organic understanding of inference, emphasizing *inference* over *entailment*. This shift aligned with the empirical findings showing that semantic inferences are often uncertain, context-sensitive, and subject to inter-annotator disagreement (Poesio and Artstein, 2005; Simons et al., 2011; Pavlick and Callison-Burch, 2016). These findings suggest that NLI should account for the fact that people often interpret meaning differently, and that disagreement is a natural part of language understanding.

Reflecting this theoretical shift, the design of NLI tasks has also evolved, from binary entailment classification to more fine-grained categorizations of inference. The original RTE tasks used a binary classification format: entailment and non-entailment. Later work introduced a more fine-grained taxonomy of inference relations. A major milestone in this direction is the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015), which formalized NLI as a three-way classification problem: entailment, contradiction, and neutral. In this formulation, the neutral label is used when the premise and hypothesis are related, but the truth of the hypothesis cannot be determined from the premise alone. It typically captures cases of ambiguity or missing information and is often defined for crowdworkers as "neither" (Nie et al., 2020a) or "might be correct" (Bowman et al., 2015; Williams et al., 2018), serving as a catch-all category for relationships that do not clearly fit into entailment or contradiction. SNLI provides large-scale, high-quality human annotations derived from image captions. The merge of SNLI dataset enables the use of deep learning methods for sentence-level inference and also establishes NLI as a standard benchmark for testing semantic understanding for neural models. Table 1.1 illustrates example sentence pairs from the SNLI dataset, highlighting the three inference

categories: entailment, neutral, and contradiction, along with the individual annotator labels.

Premise:	A man inspects the uniform of a figure in some East Asian country.
Hypothesis:	The man is sleeping.
Label:	**contradiction** (C C C C C)
Premise:	An older and younger man smiling.
Hypothesis:	Two men are smiling and laughing at the cats playing on the floor.
Label:	**neutral** (N N E N N)
Premise:	A soccer game with multiple males playing.
Hypothesis:	Some men are playing a sport.
Label:	**entailment** (E E E E E)

Table 1.1: Examples from SNLI. Each example includes a premise, a hypothesis, and the final gold label (in color), which is determined by majority vote among five annotators. The labels in parentheses represent the individual annotations: E = entailment, N = neutral, C = contradiction.

Since the release of SNLI, a variety of NLI datasets have been introduced to capture different aspects of inference, including, for example MultiNLI (Williams et al., 2018), which expands genre coverage, and ANLI (Nie et al., 2020a), which increases reasoning complexity through adversarial annotation. However, even with these improvements, the task remains difficult due to its inherent reliance on subtle semantics and real-world knowledge.

1.2 Human Label Variation in NLI

NLI is a core task in NLP, requiring systems to determine whether a hypothesis can be inferred from a given premise. Traditionally, NLI has been treated as a classification problem with three mutually exclusive labels: *entailment*, *contradiction*, and *neutral*. These judgments are used as gold standards to train and evaluate computational models, under the assumption that there exists a unique, correct label for each premise-hypothesis pair. However, this single *ground truth* assumption has been increasingly challenged in recent work.

Recent research has demonstrated that Human Label Variation (HLV) - cases in which annotators assign different labels to the same premise-hypothesis pair - is not merely noise but a recurring and meaningful phenomenon in NLI tasks (Pavlick and Kwiatkowski, 2019; Nie et al., 2020a; Plank, 2022; Weber-Genzel et al., 2024). Far from being problematic, such disagreement can offer a valuable signal to models, encouraging them to capture subtler semantic distinctions and reflect more human-like reasoning (Dagan et al., 2005). In the context of NLI, such disagreements often arise from underspecified contexts, divergent pragmatic interpretations, or varying degrees of world knowledge among annotators (de Marneffe et al., 2012; Pavlick and Kwiatkowski, 2019; Mostafazadeh Davani et al., 2022). Rather than treating all such disagreements as annotation noise, recent work recognizes them as meaningful signals that reflect linguistic ambiguity, contextual nuance, and individual reasoning strategies (de Marneffe et al., 2012; Uma et al., 2022; Cabitza et al., 2023). As illustrated in Table 1.2, in the VariErr dataset, annotators were given the premise "Tommy brought the picture down with terrific force on his head". One annotator labeled this as *entailment*, explaining that "a picture hit Tommy's head with force," while another labeled it as *neutral*, noting that it's "ambiguous if Tommy hurt himself or someone else". Both explanations were validated as plausible. A third label, contradiction, was rejected because the explanation provided ("Tommy is not hurt but rather has bad strong emotion") was inconsistent with the premise. This example illustrates that not all disagreements reflect error — some represent legitimate variation grounded in plausible interpretations (Weber-Genzel et al., 2024).

NLI Item	
Premise	*As he stepped across the threshold, Tommy brought the picture down with terrific force on his head.*
Hypothesis	*Tommy hurt his head bringing the picture down.*
Annotator A	
Explanation_1	*A picture hit Tommy's head with force*
Label	**entailment**
Validation	✓
Annotator B	
Explanation_1	*Tommy is not hurt but rather bad strong emotion*
Label	**contradiction**
Validation	✗
Explanation_2	*Ambiguous if Tommy hurt himself or another guy*
Label	**neutral**
Validation	✓

Table 1.2: Example from the VariErr dataset illustrating that not all label disagreements indicate annotation error. While Annotator B's Contradiction label is rejected, both Entailment and Neutral are supported by plausible, validated explanations.

Moreover, disagreement in label-decision has been argued to be justified and informative in many cases, especially for tasks involving subjective or pragmatic interpretation (Sommerauer et al., 2020). Rather than deciding on a single truth, annotators may each hold a plausible view, especially in cases where multiple readings or rationals are supported by the linguistic or social context. This idea is also supported by recent work on *perspectivism*, which argues that data annotation should reflect the different beliefs and experiences of annotators, instead of forcing a single "correct" label (Cabitza et al., 2023). This view aligns with a growing movement in NLP to move beyond fixed, one-label-per-instance annotations and instead represent human judgments in more flexible ways, such as using label distributions or multiple perspectives.

In sum, HLV reveals that even seemingly simple classification tasks are often underpinned by complex interpretive processes. In NLI, acknowledging and modeling this variation is not only a way to improve alignment between human and machine reasoning but also an opportunity to rethink foundational assumptions about annotation, learning, and evaluation. As Plank (2022) emphasizes, HLV affects all stages of the machine learning pipeline — data, modeling, and evaluation — and must be addressed holistically to build more inclusive and trustworthy NLP systems.

1.3 Within-label Variation

However, while variation across labels has received considerable attention, comparatively less is known about within-label variation (Jiang et al., 2023) — cases in which annotators agree on the same inference label, yet offer different explanations or justifications for their decision. This phenomenon is particularly interesting because it reveals the plurality of valid reasoning strategies: even when agreement is reached on an outcome, the underlying paths to that decision may differ significantly. Understanding such variation is essential for developing interpretable systems and for capturing the richness of human inference beyond label selection.

One way to investigate within-label variation is through free-text explanations, which offer insight into the inferential processes underlying label decisions. Datasets such as e-SNLI (Camburu et al., 2018), VariErr (Weber-Genzel et al., 2024), and LiveNLI (Jiang et al., 2023) have expanded traditional NLI data by including human-written justifications,

thereby enriching the evidence available for analysis and training. Free-text explanations offer a rich perspective on reasoning variation. However, their open-ended form can be challenging to process systematically: they vary in form, length, and linguistic style, making it difficult to extract information that is directly useful for downstream analysis. As a result, structured formats are often used when collecting human explanations. One common mechanism is the use of highlights - subsets of input elements such as words, phrases, or sentences that provide insight into a prediction (Wiegreffe and Marasović, 2021; Tan, 2022). For instance, in the e-SNLI dataset, annotators are instructed to highlight the words they think are crucial for determining the inference label and to provide corresponding free-text explanations. 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.

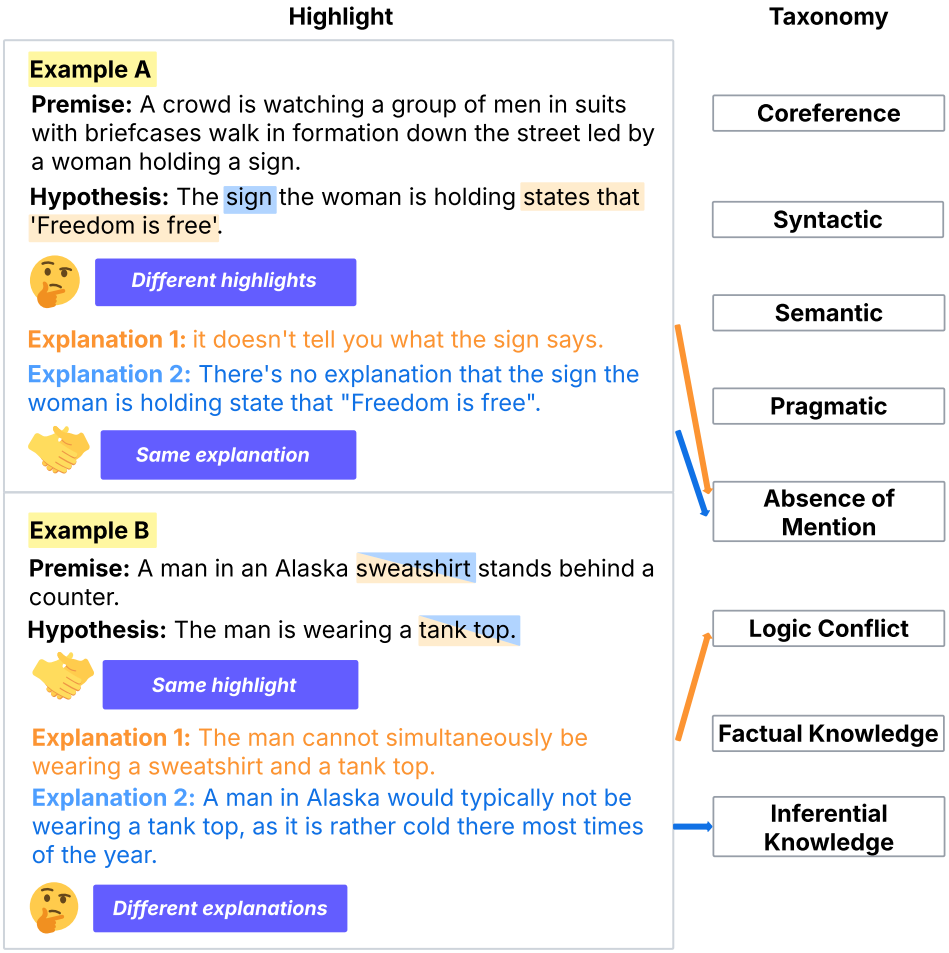

Figure 1.1: Our LITEX taxonomy reveals within-label variation not captured by highlights: the same highlights can yield different explanations (Example B), and vice versa (Example A).

Figure 1.1 presents two examples from the e-SNLI dataset that demonstrate the limits of relying solely on surface highlights. In Example A, the two explanations focus on the same core idea — that the premise does not clarify the content of the sign mentioned in the hypothesis — but differ in the spans highlighted to support this reasoning. One highlights the phrase "states that" while the other selects "sign", yet both explanations point to the same pragmatic gap: the absence of explicit supporting information in the premise. Conversely, Example B shows the opposite pattern: both annotators highlight the same words ("sweatshirt" and "tank top"), yet their explanations reflect distinct reasoning

types. One annotator appeals to logical inconsistency (a person cannot wear both a sweatshirt and a tank top simultaneously), while the other draws on inferential world knowledge (a person in Alaska would not typically wear a tank top due to the climate). Both examples demonstrate that highlights alone do not account for the full inferential intent behind an explanation and that within-label variation often emerges from deeper semantic or pragmatic distinctions.

1.4 Research Question and Thesis Structure

While prior works have researched disagreement across NLI labels, comparatively less is known about the variation that occurs *within* the same label, i.e., where annotators agree on an inference outcome (NLI label) but justify it using different reasoning paths. This raises our key research questions:

RQ1: Can a linguistically grounded taxonomy capture and structure the diversity of human reasoning underlying the same NLI label?

RQ2: Can the taxonomy improve the interpretability and alignment of explanations generated by large language models?

RQ3: Is the taxonomy applicable across diverse NLI datasets with different explanation styles and annotation protocols?

To address these questions, this thesis makes the following contributions:

1. We introduce LITEX, a LInguistic Taxonomy of EXplanations for understanding within-label variation in natural language inference. The taxonomy is designed to categorize free-text explanations into linguistically motivated reasoning types, such as paraphrastic, pragmatic, structural, and world-knowledge-based inference.

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, Large Language Models (LLMs) offer a scalable alternative for generating explanations in NLI (Chen et al., 2024a). We explore explanation generation using LLMs, which offer a scalable alternative to human annotation. Through the experiments, we demonstrate that taxonomy-based guidance provides a more effective signal for LLMs than highlight-based prompts.

4. We demonstrate the applicability of LITEX beyond e-SNLI by annotating two structurally distinct NLI datasets — LiveNLI and VariErr NLI. These datasets differ in their data collection protocols and explanation formats, allowing us to assess the consistency and flexibility of the taxonomy in capturing diverse reasoning styles across datasets.

To offer a high-level view of the thesis organization and how the core research questions are addressed, Figure 1.2 presents a structural overview of the thesis. The diagram summarizes the logical flow of the thesis: beginning with the motivation and background in Sections 1–2, which establish the challenge of within-label variation in NLI and motivate the need for a more structured understanding of explanations. Section 3 introduces the LITEX taxonomy, a linguistically grounded taxonomy for categorizing explanation types, and provides empirical validation through annotation studies. Section 4 evaluates the utility of the taxonomy in a generative setting, demonstrating how taxonomy-guided paradigms produce more human-aligned explanations compared to highlight-guided

paradigms and the baseline. Section 5 extends the analysis beyond the e-SNLI dataset to assess the generalizability of the taxonomy across structurally distinct datasets, LiveNLI and VariErr. Finally, Section 6 synthesizes insights from all previous chapters and revisits the three core research questions (RQ1–RQ3), discussing their implications for explanation modeling, NLI evaluation, and future research directions.

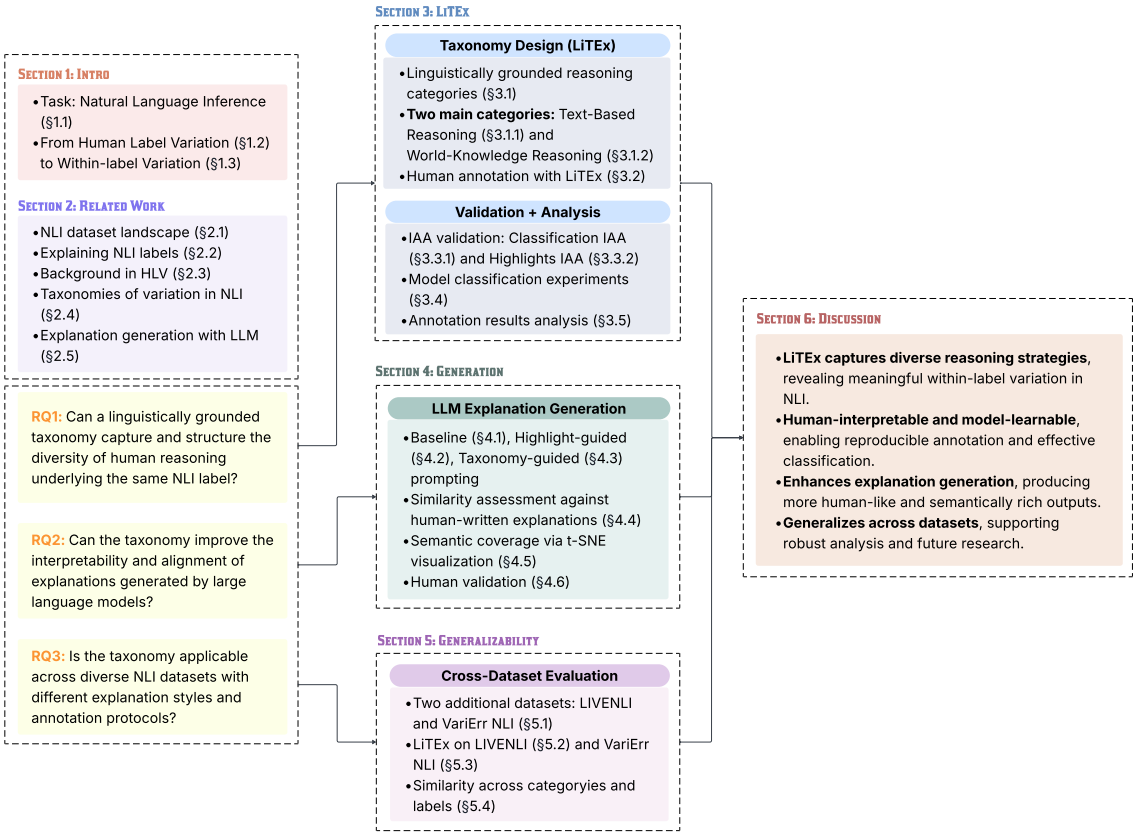

Figure 1.2: Structural overview of the thesis, including key components, experimental stages, and research questions.

The following overview summarizes the structure of the thesis and outlines how each section contributes to the overall research goals:

Section 2: Related Work surveys related work, laying the foundation for our investigation into explanation diversity. We begin with an introduction to the NLI task, followed by an overview of major NLI benchmarks and their evolution toward supporting interpretability. Section 2 then discusses existing works on collecting and analyzing explanations, including free-text explanations and supporting methods such as token-level highlights. We highlight recent work showing that variation in annotation is not just noise, but often reflects meaningful differences in interpretation. This includes studies on HLV and taxonomies that help categorize disagreement. Lastly, we also review how LLMs are now widely used to generate text-based explanations and how different prompting methods can affect the generation.

Section 3: LiTEx: Linguistically-informed Taxonomy of NLI Reasoning presents the central contribution of this thesis: LiTEx, a linguistically informed taxonomy for categorizing reasoning types in NLI explanations. The taxonomy distinguishes between **Text-Based** and **World-Knowledge** reasoning, and comprises eight fine-grained categories including *Coreference*, *Syntactic*, *Semantic*, *Pragmatic*, *Absence of Mention*, *Logic Conflict*, *Factual Knowledge*, and *Inferential Knowledge*. In this section, we describe the

taxonomy construction process, including category design, guiding questions, and illustrative examples for each category. This is followed by the annotation process description on e-SNLI explanations and inter-annotator agreement (IAA) studies to validate the taxonomy. We also conduct model classification experiments to assess whether the taxonomy is also machine-learnable. The final part of this section presents a comprehensive analysis of the taxonomy annotations, including the co-occurrence patterns between taxonomy categories and NLI labels, the similarity of explanations grouped by category diversity, and the relationship between highlighted text spans and reasoning types captured by the taxonomy.

Section 4: Taxonomy-Based and Highlight-Based Explanation Generation investigates the use of LITEX as a guidance signal for NLI explanation generation with LLMs. We compare three prompting paradigms: (1) a baseline providing only the premise, hypothesis, NLI label without further reasoning signal, (2) a highlight-based prompt, and (3) a taxonomy-conditioned prompt using LITEX categories. In addition to standard generation evaluation, comparing similarities between the generated explanations with human-written ones, we introduce a novel semantic coverage framework to evaluate how well LLM-generated explanations span the same reasoning space as human annotations. This involves t-SNE-based visualization and convex hull analysis to quantify the overlap and diversity of model outputs in embedding space.

Section 5: Assessing the Applicability of LiTEx Across NLI Benchmarks evaluates the generalizability of LITEX beyond the e-SNLI dataset. We apply our taxonomy to two additional NLI datasets: LiveNLI, containing 1,415 ecologically valid explanations (Jiang et al., 2023), and VariErr, which features fine-grained validation of 1,933 explanations (Weber-Genzel et al., 2024).

Section 6: Discussion concludes the thesis by summarizing the main findings and discussing what they mean for NLI evaluation, model training, and explanation generation. We reflect on the strengths and limitations of using a taxonomy to analyze free-text explanations and highlight open challenges, such as overlapping reasoning types and annotation ambiguity. We also suggest future directions, including evaluation setups that incorporate multiple annotator perspectives and better account for implicit reasoning, moving beyond surface-level explanation matching to consider the underlying inference strategies that different annotators may assume but not explicitly state.

2 Background and Related Work

2.1 Exploring the NLI Dataset Landscape

Natural Language Inference (NLI), also known as recognizing textual entailment (RTE), is a fundamental task in natural language understanding (NLU) that evaluates whether a **hypothesis** logically follows from, contradicts, or is neutral with respect to a given **premise**. Since its formalization in the PASCAL RTE Challenges (Dagan et al., 2005), NLI has evolved from small, rule-based datasets to large-scale benchmarks that test diverse reasoning capabilities, including lexical semantics, compositional structure, and real-world knowledge.

The introduction of the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015) marked a turning point by providing 570K human-annotated premise-hypothesis pairs, enabling data-driven approaches like neural networks to dominate NLI research. Its successor, MultiNLI (Williams et al., 2018), extended coverage to 10 genres of spoken and written English, introducing cross-genre generalization as a key challenge. These datasets played a pivotal role in driving advances in NLI modeling, serving as both a training ground and an evaluation benchmark. Early work leveraged their scale to develop the first effective neural sequential models, such as the ESIM (Enhanced Sequential Inference Model, Conneau et al. 2018a), which introduced attention mechanisms to capture fine-grained lexical and syntactic alignment between sentences. The release of larger and more diverse benchmarks like MultiNLI further pushed the field toward models capable of cross-domain generalization, culminating in the paradigm shift to pre-trained transformers like BERT (Devlin et al., 2019), whose bidirectional contextual representations achieved state-of-the-art performance on NLI tasks. However, these datasets also exposed critical limitations: studies like (Gururangan et al., 2018) demonstrated that many 'high-performing' models relied on superficial heuristics (e.g., lexical overlap for entailment, negation words for contradiction) rather than genuine semantic understanding, prompting the development of adversarial datasets and more rigorous evaluation protocols.

Subsequent datasets targeted limitations of general-purpose benchmarks. Adversarial NLI (ANLI) (Nie et al., 2020a) used iterative adversarial human-and-model-in-the-loop collection to create progressively harder rounds (R1–R3), exposing brittleness in state-of-the-art systems. ChaosNLI (Nie et al., 2020b) further decomposed ambiguity into semantic phenomena (e.g., lexical, syntactic) to diagnose fine-grained failures. Cross-lingual benchmarks like XNLI (Conneau et al., 2018b) (15 languages) and AmericasNLI (Ebrahimi et al., 2022) (indigenous languages) evaluated transfer learning, revealing gaps in low-resource language inference.

In addition to general-domain corpora, several datasets have been developed to evaluate NLI in specific domains. For example, SciNLI (Sadat and Caragea, 2022) is a corpus for NLI on scientific texts, designed to test a model's ability to reason over technical language and structured knowledge. Similarly, MedNLI (Romanov and Shivade, 2018) targets clinical inference based on medical records, and emphasized domain-specific terminology and implicit expert knowledge. These datasets highlight the challenge of transferring NLI models trained on general data to specialized contexts, where inference often depends on background expertise. Beyond single-domain focus, contemporary benchmarks also address discourse complexity: ContractNLI (Koreeda and Manning, 2021) escalates the task to multi-paragraph legal document analysis, whereas NLI-over-Context (Liu et al., 2020) extends inference to dialogue chains and narrative structures.

2.2 Explaining NLI Labels

Explanations play a crucial role in making NLI decisions interpretable, especially in applications where interpretability and error analysis are essential. 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. This variation has implications not only for downstream tasks but also for how models learn and generalize from supervision.

Among existing methods, token-level highlights (i.e., highlighting specific words or phrases in the input) serve as a proxy for explanations, guiding annotators to mark relevant spans that support their label choice (Wiegreffe and Marasović, 2021). 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).

In this work, we focus on analyzing explanations from e-SNLI, one of the most widely used datasets for supervised explanation generation. We leverage both free-text and token-level explanations to study how different reasoning strategies can underlie the same NLI label. Our goal is to go beyond simply collecting or generating explanations and instead use them to characterize variation in reasoning and better align model outputs with human judgments.

2.3 Analyzing Human Label Variation in NLI

Variation in NLP annotations is increasingly recognized not merely as annotation noise, but as a reflection of the inherent subjectivity and multiplicity of meaning in language. In many NLP tasks, such as sentiment analysis (Barnes et al., 2019; Majumdar et al., 2022; Hariri, 2024; Herrera-Poyatos et al., 2025), hate speech detection (Markov and Daelemans, 2022; Thapa et al., 2022), summarization (Gliwa et al., 2019; Kryscinski et al., 2019; Deas and McKeown, 2024), and natural language inference (Pavlick and Kwiatkowski, 2019; Nie et al., 2020a), multiple interpretations can be equally valid depending on context, perspective, or background knowledge. This phenomenon is often described as *human label variation* and has become a central topic in recent discussions on annotation, fairness, and evaluation (Poesio et al., 2018; Plank, 2022). Figure 2.1 (adapted from Plank, 2022) illustrates this notion, highlighting how disagreements between annotators can stem from genuine interpretative differences, rather than annotation mistakes.

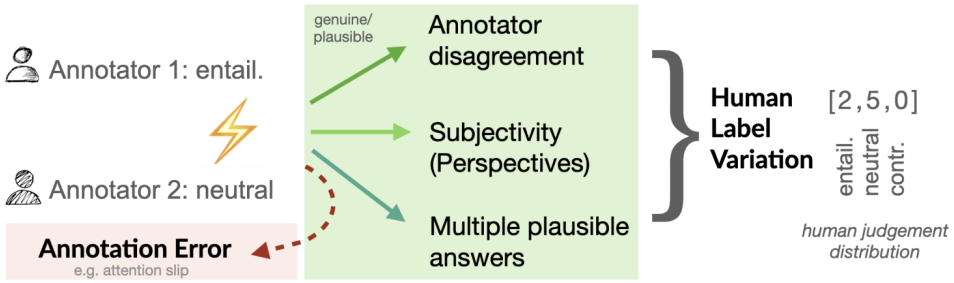

Figure 2.1: Illustration of human label variation, adapted from Plank (2022). Inherent disagreements between annotators may arise from genuine differences in interpretation, subjectivity, or the presence of multiple plausible answers — rather than annotation errors. This highlights that variation is often a reflection of language ambiguity rather than noise.

This variation has been studied from several complementary perspectives. The CrowdTruth framework (Aroyo and Welty, 2015) views annotation disagreement as a signal of ambigu-

ity, proposing that a range of reasonable interpretations can co-exist rather than converge on a single "ground truth". Similarly, Röttger et al. (2022) argued that annotators bring one or many beliefs to a task, which can influence their judgments, particularly in socio-political or subjective domains.

Beyond individual cognition, recent works also highlight the social dimensions of annotators, including demographics, cultural background, and lived experience. For instance, Sap et al. (2019, 2022) and Larimore et al. (2021) discussed how age, race, and gender can affect how annotators perceive toxicity or intent in language. Hershcovich et al. (2022) further emphasized the role of cultural framing in shaping what counts as offensive and acceptable language. They argued that NLP systems should take these cultural differences into account when being evaluated. Furthermore, the LeWiDi shared task (Leonardelli et al., 2023) promotes modeling NLP tasks without reconciling subjective differences, advocating for frameworks that reflect the diversity human judgments. Similarly, the MultiPICO corpus (Casola et al., 2024) introduces a multilingual, perspectivist dataset with demographic metadata, enabling the study of how sociocultural factors shape annotators' perceptions of irony and how LLMs can be evaluated accordingly. For the specific case of NLI, the CALI dataset (Huang and Yang, 2023) explicitly captures cross-lingual differences in inference by collecting labels and explanations from annotators in the U.S. and India, highlighting how cultural norms impact semantic interpretation.

These diverse interpretations are more broadly conceptualized under the umbrella of data perspectivism Cabitza et al. (2023); Wich et al. (2021); Sorensen et al. (2024), which promotes the idea that datasets should preserve and represent multiple human perspectives, rather than enforcing artificial consensus. This shift in perspective has implications not only for annotation and dataset design but also for how we evaluate and train NLP models, especially when deploying them in socially sensitive or high-stakes settings.

2.4 Formulating Taxonomies of Variation in NLI

Understanding how and why annotators arrive at different interpretations in NLI has led researchers to propose taxonomies that categorize the types of reasoning or inferences involved. Early work in this direction focused on classifying the types of knowledge or reasoning needed to bridge the gap between premise and hypothesis. For example, Sammons et al. (2010), Simons et al. (2011), and LoBue and Yates (2011) identified categories such as lexical relationship, coreference resolution, missing information, and temporal reasoning.

Later research shifted toward understanding annotation variation, especially when multiple annotators disagree or when inference is underspecified. Rather than treating such variation as noise, recent work views it as a reflection of underlying linguistic ambiguity or context-sensitive reasoning. Pavlick and Kwiatkowski (2019) argued that a significant portion of NLI disagreement is not due to annotation error. Jiang and de Marneffe (2022) responded to this by proposing a taxonomy that identifies characteristics of the items that can cause variation in annotation. Their taxonomy includes features such as vagueness, underspecificity, presupposition failure, and lexical polysemy—factors that frequently lead to disagreement or uncertainty. This perspective treats variation not as noise but as a reflection of linguistic complexity and context sensitivity.

Building on this, Jiang et al. (2023) shifted the focus from the NLI items. Rather than focusing solely on item properties, they analyzed the free-text justifications annotators wrote for their decisions by applying Jiang and de Marneffe (2022)'s taxonomy to the explanations. They uncovered diverse reasoning types such as pragmatic inference, paraphrasing, and co-hyponymy. Their study demonstrates that explanations offer a finer-grained view of agreement, revealing that even when labels are consistent, the reasoning behind them can vary substantially.

Our work builds on this direction by proposing a taxonomy of explanations themselves,

rather than of items or annotators. Specifically, we target cases where multiple explanations are given for the same NLI label, and aim to categorize the distinct reasoning strategies that can lead to that shared outcome. This focus on within-label variation complements prior work on inter-annotator disagreement by showing that even apparent agreement on labels can have diverse inferential paths underneath. Compared to Jiang et al. (2023), our taxonomy is thus grounded in the explanations. It also makes world knowledge in NLI reasoning more explicit.

2.5 Generating Explanations with LLMs

LLMs such as GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), and LLaMA (Meta, 2024) have demonstrated impressive capabilities in generating coherent and contextually appropriate natural language explanations. Recent studies explored the use of LLMs to generate natural language explanations across a range of NLP tasks - including question answering (Bohnet et al., 2023), commonsense reasoning (Rajani et al., 2019), and natural language inference (Chen et al., 2024a), aiming to improve transparency and support downstream analysis (Kunz and Kuhlmann, 2024).

One prominent strategy is chain-of-thought (CoT) prompting, which encourages models to reason step-by-step before arriving at a final answer. Li et al. (2024) proposed prompting LLMs to generate CoT explanations to improve the performance of small task-specific models. This research suggests that LLMs can be used not just to answer questions, but also to teach other models by showing how they arrived at their answers (Zhou et al., 2023; Wang et al., 2023).

Other studies have questioned whether LLMs can provide faithful self-explanations, i.e., justifications that reflect their internal reasoning. Huang et al. (2023) investigated whether LLMs could generate faithful self-explanations to justify their own predictions during inference. Chrysostomou and Aletras (2021) and Lampinen et al. (2022) highlighted the gap between fluency and faithfulness in explanation generation.

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. (2024a) 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.

3 LiTEx: Linguistically-informed Taxonomy of NLI Reasoning

3.1 Taxonomy Categories

To better capture and structure the human reasoning underlying the same NLI label through categorizing NLI explanations, we propose a taxonomy for NLI free-text explanations - LITEX: a Linguistically-informed Taxonomy of NLI Reasoning. Unlike prior work that focuses on modeling label-level disagreement, our goal is to explicitly classify the types of reasoning expressed in human-written explanations, especially those that lead to agreement on the same label despite diverse justifications for the decision.

Our taxonomy builds upon the work of Jiang and de Marneffe (2022), which identifies where NLI annotation disagreement stems from. To do so, Jiang and de Marneffe (2022) conducted a qualitative analysis of portions of the MNLI dataset (Williams et al., 2018), which they selected due to its diversity in genre and inference types, especially when compared to caption-based datasets that focus narrowly on visual scenes (e.g., SICK (Marelli et al., 2014), SNLI (Bowman et al., 2015)).

	Premise	**Hypothesis**
Uncertainty in Sentence Meaning		
[1] Lexical	Technological advances generally come in waves that crest and eventually subside.	Advances in electronics come in waves.
[2] Implicature	Today it is possible to buy cheap papyrus printed with gaudy Egyptian scenes in almost every souvenir shop in the country, but some of the most authentic are sold at The Pharaonic Village in Cairo where the papyrus is grown, processed, and hand-painted on site.	The Pharaonic Village in Cairo is the only place where one can buy authentic papyrus.
[3] Presupposition	What changed?	Nothing changed.
[4] Probabilistic Enrichment	It's absurd but I can't help it. Sir James nodded again.	Sir James thinks it's absurd.
[5] Imperfection	profit rather	Our profit has not been good.
Underspecification in Guidelines		
[6] Coreference	This was built 15 years earlier by Jahangir's wife, Nur Jahan, for her father, who served as Mughal Prime Minister.	Nur Jahan's husband Jahangir served as Mughal Prime Minister.
[7] Temporal Reference	However, co-requesters cannot approve additional co-requesters or restrict the timing of the release of the product after it is issued.	They cannot restrict timing of the release of the product.
[8] Interrogative Hypothesis	"How did you get it? " A chair was overturned.	"How did you get your hands on this object? "
Annotator Behavior		
[9] Accommodating Minimally Added Content	Indeed, 58 percent of Columbia/HCA's beds lie empty, compared with 35 percent of nonprofit beds.	58% of Columbia/HCA's beds are empty, said the report.
[10] High Overlap	Yet, in the mouths of the white townsfolk of Salisbury, N.C., it sounds convincing.	White townsfolk in Salisbury, N.C. think it sounds convincing.

Table 3.1: Taxonomy of potential sources of disagreement in NLI annotation, adapted from the framework introduced by Jiang and de Marneffe (2022). Each row illustrates a subtype with a premise–hypothesis pair exemplifying the source of interpretive variation.

Their proposed taxonomy organized sources of disagreement into three main categories: *Uncertainty in Sentence Meaning*, *Underspecification in Guidelines*, and *Annotator Be-*

havior, each comprising finer-grained subtypes such as *Lexical ambiguity*, *Presupposition*, and *Coreference*, which capture different interpretive challenges or annotation tendencies.

Table 3.1 (adapted from their paper) presents example premise-hypothesis pairs that illustrate each of the ten subtypes. For instance, the *Lexical* and *Implicature* rows exemplify how uncertainty on word meaning or pragmatics can lead to diverging judgments. Subtypes like *Coreference* and *Temporal Reference* highlight the role of under-specified relations, while *High Overlap* showcases how surface-level similarity can bias label choices (annotators tend to judge it as *Entailment* even if it is not strictly inferred from the premise). This taxonomy provides a valuable framework for understanding why annotators may disagree on the NLI label, but it was not designed to systematically categorize the explicit content of free-text explanations themselves.

Text-Based Reasoning (TB)		
Coreference	**Q:**	*Does the explanation rely on resolving coreference between entities?*
	Check:	Determine whether the main entities in the premise and hypothesis refer to the same real-world referent, including via pronouns or phrases.
Syntactic	**Q:**	*Does the explanation involve a change in sentence structure that preserves meaning?*
	Check:	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:**	*Does the explanation involve semantic similarity or substitution of key concepts?*
	Check:	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:**	*Does the explanation rely on pragmatic cues like implicature or presupposition?*
	Check:	Look for meaning beyond the literal text - including implicature, presupposition, speaker intention, and conventional conversational meaning.
Absence of Mention	**Q:**	*Does the explanation point out information not mentioned in the premise?*
	Check:	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:**	*Does the explanation refer to logical constraints or conflict?*
	Check:	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:**	*Does the explanation rely on widely shared, intuitive facts acquired through everyday experience?*
	Check:	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:**	*Does the explanation rely on real-world norms, customs, or culturally grounded reasoning?*
	Check:	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 3.2: Guiding questions and decision criteria for our LITEX taxonomy.

To create a linguistically grounded and explanation-focused categorization, we reinterpret and restructure previously identified disagreement types into a new taxonomy specifically designed for classifying free-text NLI explanations. Our proposed taxonomy, LITEX, focuses on explicitly stated reasoning in the explanations—capturing how annotators articulate their rationale rather than the implicit causes of label disagreement.

The taxonomy development process began with a close analysis of free-text explanations provided in the e-SNLI dataset (Camburu et al., 2018), alongside the disagreement categories introduced by Jiang and de Marneffe (2022). Through this exploration, we observed that explanations could be broadly grouped based on whether the reasoning relies purely on textual information from the premise–hypothesis pair or incorporates external world knowledge. This led us to define two high-level categories: **Text-Based (TB) Reasoning** and **World-Knowledge (WK) Reasoning**, as illustrated in Table 3.2.

For each explanation type, Table 3.2 provides:

- a **guiding question (Q)** to help annotators or models identify whether an explanation belongs to that category, and
- a **checklist-style description** that elaborates on what features or cues to look for during classification.

Under **TB Reasoning**, we retain the *Coreference* category from the taxonomy of Jiang and de Marneffe (2022), as referential interpretation is a common and linguistically grounded reasoning type. We further include broader linguistic phenomena by introducing the *Syntactic*, *Semantic*, and *Pragmatic* categories, where *Pragmatic* reasoning includes several pragmatic phenomena such as presupposition and implicature. In addition, our empirical inspection of e-SNLI revealed recurring patterns of reasoning based on textual inference. These include *Absence of Mention* (where the hypothesis introduces ungrounded content) and *Logic Conflict* (where a contradiction arises from structural or logical incompatibility in the textual content).

For **WK Reasoning**, we identified two key explanation types: *Factual Knowledge*, where the explanation cites a concrete or commonly accepted fact (e.g., "penguins cannot fly"), and *Inferential Knowledge*, where the annotator relies on broader commonsense or context-dependent inference (e.g., spatial or causal reasoning). These two subtypes differ in terms of specificity and generalizability of the external knowledge being used, and both are widely observed in the natural explanations from e-SNLI.

Together, this taxonomy captures a wide range of reasoning strategies in NLI explanation, grounded in linguistic theory and empirical observations from real-world data. It provides a foundation for analyzing and improving explanation generation systems by linking explanation form and content to specific reasoning mechanisms. The following two subsections provide detailed definitions of the eight subcategories in LITEX, accompanied by illustrative examples for each.

3.1.1 Text-Based (TB) Reasoning

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*. These six subcategories reflect how explanations leverage explicit textual elements. Below, we present the definitions of the LITEX categories along with corresponding illustrative examples. Most examples illustrated here (and in subsequent categories) tend to reflect Entailment cases, due to the underlying distribution of NLI labels across categories not being uniform. We refer readers to §3.5 for a more detailed discussion of label–category co-occurrence patterns.

Coreference This category captures explanations that rely on referential links between entities in the premise and hypothesis. It focuses on whether different expressions refer to the same real-world entity, such as pronouns, definite noun phrases, or names. Explanations in this category resolve ambiguity arising from coreference chains and often require entity tracking across the two sentences.

Illustrative Example 1:
Premise: The man in the black t-shirt is trying to throw something.
Hypothesis: The man is in a black shirt.
NLI Label: Entailment
Explanation: The man is in a black shirt refers to the man in the black t-shirt.

Illustrative Example 2:
Premise: A naked man rides a bike.
Hypothesis: A person biking.
NLI Label: Entailment
Explanation: The person biking in the hypothesis is the naked man.

Syntactic Syntactic reasoning involves transformations or structural differences between the premise and hypothesis that preserve semantic content. Examples include passive vs. active voice, fronting, coordination, or subordination. Explanations in this category assess whether such syntactic alternations retain the original meaning or lead to changes in entailment status. In the e-SNLI dataset, annotations in the *Syntactic* category are strongly associated with the entailment label.

Illustrative Example 1:
Premise: Two women walk down a sidewalk along a busy street in a downtown area.
Hypothesis: The women were walking downtown.
NLI 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.

Illustrative Example 2:
Premise: Bruce Springsteen, with one arm outstretched, is singing in the spotlight in a dark concert hall.
Hypothesis: Bruce Springsteen is a singer.
NLI Label: Entailment
Explanation: Springsteen is singing in a concert hall.

Semantic Explanations in this category are based on semantic similarity or substitution of key lexical items. This includes synonymy (e.g., "smart" vs. "intelligent"), antonymy, lexical entailment, or category relationships (e.g., "poodle" entails "dog"). Semantic reasoning identifies whether the overlap in meaning between the premise and hypothesis justifies entailment or contradiction.

Illustrative Example 1:
Premise: A man in a black tank top is wearing a red plaid hat.
Hypothesis: A man in a hat.
NLI Label: Entailment
Explanation: A red plaid hat is a specific type of hat.

Illustrative Example 2:
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.
NLI Label: Entailment
Explanation: Three men are people.

Pragmatic Pragmatic reasoning relies on meaning that goes beyond the literal content of the sentence. This includes implicature (e.g., "some" implying "not all"), presuppositions (e.g., "stopped" presupposing a prior action), and speaker intention. Explanations here involve interpreting conversational norms, context-driven inferences, or what is implied but not explicitly stated.

Illustrative Example 1:
Premise: A girl in a blue dress takes off her shoes and eats blue cotton candy.
Hypothesis: The girl is eating while barefoot.
NLI Label: Entailment
Explanation: If a girl takes off her shoes, then she becomes barefoot, and if she eats blue candy, then she is eating.

Illustrative Example 2:
Premise: A woman wearing bike shorts and a skirt is riding a bike and carrying a shoulder bag.
Hypothesis: A woman on a bike.
NLI Label: Entailment
Explanation: Woman riding a bike means she is on a bike.

Absence of Mention This category captures cases where the hypothesis introduces new information that is not supported or contradicted by the premise, but simply not mentioned. Such reasoning hinges on the principle that silence is not negation: the premise does not entail or contradict the hypothesis simply because the relevant detail is absent. This often results in an "neutral" label.

Illustrative Example 1:
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.
NLI 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.

Illustrative Example 2:
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.
NLI Label: Neutral
Explanation: It does not say anything about there being a lake.

Logic Conflict Logical reasoning involves evaluating formal relationships or constraints between the premise and hypothesis. This includes handling quantifiers ("all", "some", "none"), conditionals ("if...then..."), exclusivity, negation, or scalar terms. Explanations here analyze whether the two sentences are logically consistent.

Illustrative Example 1:
Premise: Five girls and two guys are crossing an overpass.
Hypothesis: The three men sit and talk about their lives.

NLI Label: Contradiction
Explanation: Three is not two.

Illustrative Example 2:
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.
NLI Label: Entailment
Explanation: The people described are standing outside, so naturally not inside the building.

3.1.2 World-Knowledge (WK) Reasoning

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). These two subcategories cover explanations that appeal to general knowledge, cultural norms, or learned associations not explicitly encoded:

Factual knowledge This category involves appeals to widely shared or encyclopedic facts about the world. The explanation assumes access to common truths—e.g., "Paris is the capital of France", "water freezes at 0°C", or "a dog is an animal". Inference depends on external knowledge not stated in either premise or hypothesis but considered universally accessible.

Illustrative Example 1:
Premise: Two people crossing by each other while kite surfing.
Hypothesis: The people are both males.
NLI Label: Neutral
Explanation: Not all people are males.

Illustrative Example 2:
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.
NLI Label: Neutral
Explanation: Not all houses have swimming pools.

Inferential knowledge Explanations in this category rely on typical social norms, expectations, or everyday reasoning. This includes commonsense inference (e.g., "If someone drops their phone, it might break"), as well as culturally grounded assumptions about what is likely or expected. Unlike factual knowledge, inferential knowledge is often probabilistic and context-sensitive.

Illustrative Example 1:
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.
NLI Label: Neutral
Explanation: A girl in a blue dress doesn't imply the girl is a flower girl at a wedding.

Illustrative Example 2:
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.
NLI 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.

As mentioned above, the guiding questions and decision criteria for each taxonomy category presented in Table 3.2 help annotators to identify the reasoning behind explanations. These diagnostic questions are designed to complement the formal definitions and serve as a practical aid for identifying the dominant reasoning type behind a given explanation. Together with the illustrative examples presented earlier, the table helps delineate the often subtle conceptual boundaries between categories, particularly in cases involving overlapping or ambiguous cues. 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*.

3.2 Taxonomy Annotation

To apply our taxonomy in practice and evaluate its applicability across a broad range of natural language inference phenomena, we conduct annotation on a subset of the e-SNLI dataset (Camburu et al., 2018). For convenience, we refer to the subset annotated according to LITEX as LITEX-SNLI throughout the paper.

Premise: An adult dressed in black holds a stick.
Hypothesis: An adult is walking away, empty-handed.
Label: contradiction
Explanation: Holds a stick implies using hands so it is not empty-handed.

Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her.
Hypothesis: A young mother is playing with her daughter in a swing.
Label: neutral
Explanation: Child does not imply daughter and woman does not imply mother.

Premise: A man in an orange vest leans over a pickup truck.
Hypothesis: A man is touching a truck.
Label: entailment
Explanation: Man leans over a pickup truck implies that he is touching it.

Figure 3.1: Examples from e-SNLI (Camburu et al., 2018). Annotators were given the premise, hypothesis, and label. They highlighted the words that they considered essential for the label and provided the explanations.

e-SNLI is an extension of the widely used SNLI dataset (Bowman et al., 2015). Each explanation in e-SNLI is accompanied by highlight annotations on the premise and hypothesis, indicating the textual spans that contributed to the labeling decisions and are relevant to the reasoning in the explanation. The e-SNLI annotation guidelines specify rules for highlighting words based on the NLI label. Annotators first assign a label (entailment, contradiction, or neutral) based on the relationship between the premise and hypothesis. They additionally highlight words in the premise and/or hypothesis that, in their view, support the explanation they provided for the label. The highlighting rules

are as follows: for entailment, at least one word must be highlighted in the premise; for contradiction, in both premise and hypothesis; and for neutral, only in the hypothesis. This makes e-SNLI a valuable resource not only for studying NLI classification but also for analyzing how humans localize reasoning evidence in natural language.

Figure 3.1 illustrates several representative examples from e-SNLI. These include diverse linguistic phenomena such as contradiction via physical state ("holds a stick" vs. "empty-handed"), referential ambiguity ("woman" vs. "mother/daughter"), and pragmatic entailment ("leans over" implying "touching"). The highlighted spans reflect the portions of the input that annotators deemed critical for labeling and explanation.

We select e-SNLI for our annotation task because it offers rich, crowd-sourced explanations written in unconstrained language, paired with high-quality highlight supervision. In addition, e-SNLI is particularly well-suited to our goal of studying within-label variation in reasoning. Each NLI item, defined by a premise, a hypothesis, and an NLI label, is accompanied by three distinct post-hoc free-text explanations written by different annotators. Each annotator also provides highlights within the premise and hypothesis, and their explanation is meant to justify the label based on these highlighted spans. These explanations often reflect diverse reasoning strategies, even though they support the same label. We randomly selected a subset (1,002 items) of the e-SNLI dataset for the annotation process using LITEX. We then 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. Consider the following annotated e-SNLI explanation with two categories from LITEX:

- **Premise**: A man wearing a red vest is walking past a black and green fence.
- **Hypothesis**: The man wearing the vest is sitting on the sofa.
- **NLI Label**: contradiction
- **Explanation**: A vest can be any color not just red. If a man is walking he is not sitting simultaneously.
- **Assigned Categories**: *Inferential Knowledge*, *Logic Conflict*

This explanation can be decomposed into two category-specific segments:

- **Inferential Knowledge**:
  "A vest can be any color not just red."
  NLI Label: contradiction
  This segment uses world knowledge to infer plausibility.
- **Logic Conflict**:
  "If a man is walking he is not sitting simultaneously."
  NLI Label: contradiction
  This segment highlights the logical conflict arising from simultaneous actions.

One annotator (the author of this thesis) is selected to conduct the annotation process. Prior to the annotation task, the annotator went through a dedicated training designed to ensure a clear understanding of the proposed LITEX. This training included the definitions of each explanation category, accompanied by a set of illustrative examples that aimed to capture the full range of phenomena covered by each category (see §3.1). These examples are selected from the original e-SNLI dataset to maintain consistency in domain and style. However, they are explicitly excluded from the subset of 1,002 instances used in our actual annotation process to avoid any risk of data leakage. The training phase emphasized category boundaries and edge cases, aiming to reduce ambiguity and maintain consistent

annotation decisions across varying linguistic scenarios. This 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.

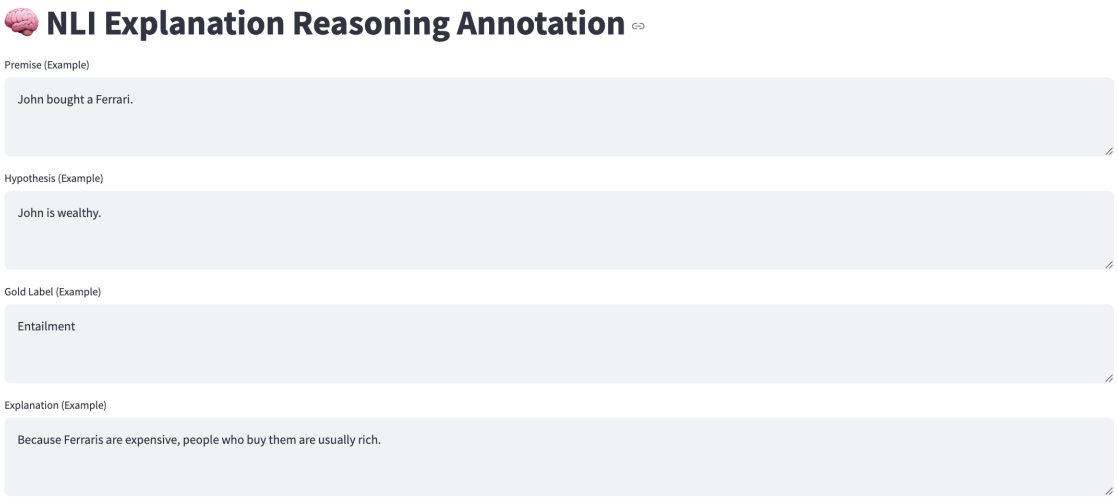

Figure 3.2: Annotation interface used in our reasoning category labeling process. Annotators are shown the premise, hypothesis, gold label, and explanation, and are asked to assign a reasoning category based on the taxonomy. The interface is implemented using **Streamlit**.

During the annotation process, we developed a simple and intuitive interface using **Streamlit** to facilitate efficient labeling. As shown in Figure 3.2, the annotator is presented with a premise–hypothesis pair, the corresponding gold NLI label, and a free-text explanation.

The annotator's task is to assign a reasoning category to each explanation by answering the set of guiding questions shown in Table 3.2, each corresponding to one of the eight taxonomy categories. This design encourages careful reflection on distinct reasoning types while reducing annotation ambiguity. An overview of the interface is shown in Figure 3.3.

Figure 3.3: Reasoning assessment interface used for explanation categorization. Each category presents a guiding question corresponding to one of the taxonomy's eight reasoning types, allowing annotators to assign the most appropriate category.

3.3 Inter-Annotator Agreement Analyses

To evaluate the reliability and reproducibility of our annotation framework, we conduct inter-annotator agreement (IAA) studies on two complementary tasks: reasoning category

classification and highlight span selection. The goal is to assess whether independent annotators can consistently apply our taxonomy to classify NLI explanations and whether they agree on which words or phrases support a given explanation. We report agreement scores and analyze both high-consensus and borderline cases to better understand the clarity and applicability of our annotation guidelines. The following two subsections present IAA results for explanation classification with taxonomy and highlight spans, respectively.

3.3.1 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, annotated 201 explanations from 67 extra e-SNLI items, using the proposed taxonomy. To ensure a fair and reliable evaluation, both annotators were given the same set of training materials and were blinded to each other's annotations during the labeling process.

The agreement between annotators is high, with a Cohen's κ score of 0.862, indicating strong consistency in applying the taxonomy. This level of agreement suggests that the reasoning categories are sufficiently well-defined to enable reproducible annotation across independent annotators.

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

Table 3.3: Inter-Annotator Agreement (IAA) classification report showing per-category precision, recall, and F1-scores on 201 explanations annotated using LITEX.

The full IAA classification report is shown in Table 3.3, including per-category precision, recall, F1-score, and support. Most categories demonstrate high agreement. For example, *Pragmatic* explanations achieve near-perfect performance (F1 = 0.970), reflecting a shared understanding of implicature- and presupposition-based reasoning. *Absence of Mention* (F1 = 0.960) and *Logic Conflict* (F1 = 0.949) also exhibit high agreement, indicating that both annotators could reliably identify explanations involving omitted information and logical inconsistency, respectively. These results affirm the robustness of these categories and suggest that they are operationalizable with minimal ambiguity.

On the other hand, some categories exhibit greater divergence between annotators. For instance, in the case of *Semantic* explanations, Annotator 1 labeled more instances as *Semantic* than Annotator 0, resulting in a relatively lower precision (0.643) when using Annotator 0's annotations as the gold standard. While recall remains perfect (1.0), this indicates a tendency of Annotator 1 to apply the *Semantic* label more liberally. This suggests that while annotators are generally inclusive in labeling semantic phenomena, there is occasional confusion with related categories—most notably *Pragmatic* or *Inferential Knowledge*, where meaning differences might stem from deeper contextual cues rather than pure lexical similarity. Similarly, *Factual Knowledge* exhibits lower recall (0.652), possibly due to the subtle boundary it shares with *Inferential Knowledge*, especially in

cases where commonsense and encyclopedic knowledge blur.

Notably, *Coreference* is absent from the classification reports. This is because none of the 201 randomly selected 201 explanations were labeled with LITEX by either annotator. Rather than indicating disagreement or mislabeling, this reflects the low empirical frequency of coreference-based reasoning in the e-SNLI dataset overall. While the category is conceptually important for linguistic inference, it is relatively rare in naturally occurring explanations, particularly in short, crowd-generated NLI explanations like those in e-SNLI.

Overall, macro- and weighted-average F1-scores (0.878 and 0.889, respectively) confirm that the taxonomy achieves high inter-annotator agreement across diverse explanation types, with only a few borderline cases requiring finer-grained disambiguation in future refinements. These results support the applicability and reliability of our taxonomy as a tool for categorizing reasoning in natural language inference.

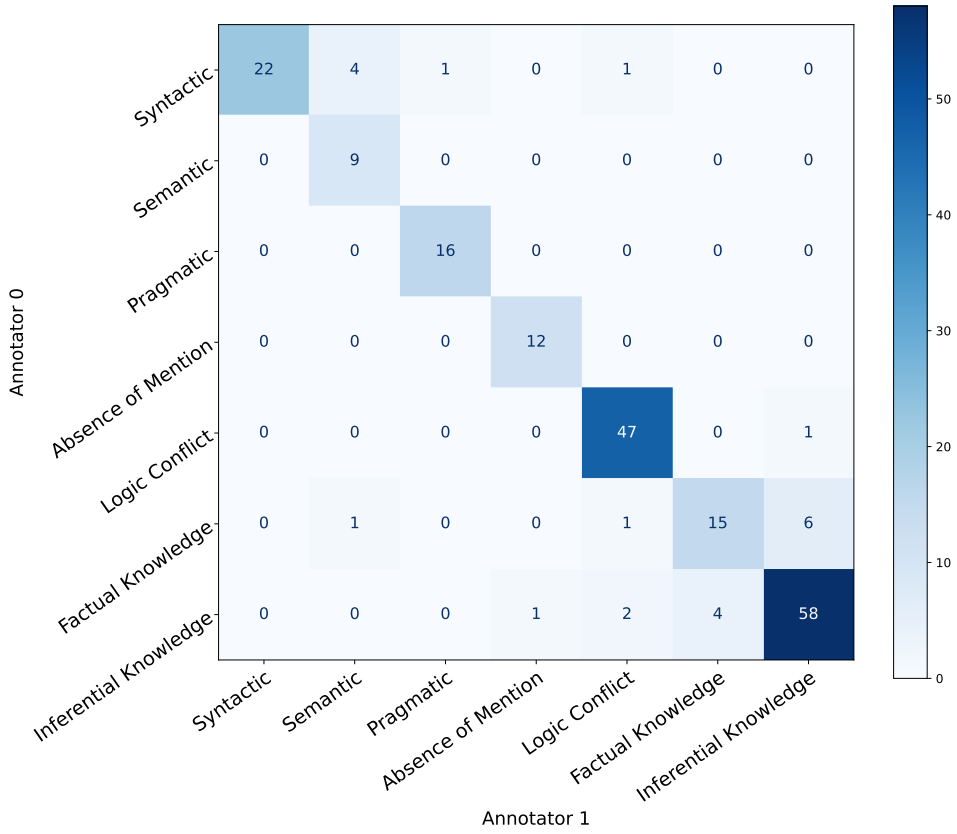

Figure 3.4: Inter-Annotator Confusion Matrix for Explanation Category Annotation.

Figure 3.4 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 *Logical Conflict* and *Inferential Knowledge*. Some confusion appears between semantically adjacent categories, such as *Factual Knowledge* vs. *Inferential Knowledge*, and *Semantic* vs. *Syntactic*. 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 (Figure 3.4). The following examples shed light on how subtle differences in reasoning can lead to divergent category assignments:

Case 1: 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.

Case 2: Inferential Knowledge 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.

Case 3: Inferential Knowledge vs. Factual Knowledge
Premise: A man reads the paper in a bar with green lighting.
Hypothesis: The man is reading the sportspage.
Explanation: The man could be reading something other than the sportspage.
Category (Annotator 0): Inferential Knowledge
Category (Annotator 1): Factual Knowledge

Analysis: Annotator 0 treats the issue as a reasoning gap, noting that the hypothesis assumes unstated information (a specific newspaper section) and this gap could not be filled with further inferential reasoning using world-knowledge. Annotator 1, however, sees this as a factual error, where the hypothesis makes an incorrect claim. This contrast shows how the same case can be interpreted either as a failure in logical inference or as a factual mismatch, highlighting how small wording differences influence classification.

Case 4: 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.

3.3.2 Highlights IAA

To understand whether human-generated highlights are consistent and reproducible, we conducted a highlight-level 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, following the same constraints as in e-SNLI. Each item included the premise, hypothesis, gold label, and 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.

3.4 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. In Table 3.4 the hyperparameter setup is listed. We use a relatively small batch size (8) and low learning rates (2e-5 and 3e-5) to promote stable convergence. No weight decay or warmup is applied, and all models are optimized using AdamW with linear learning rate decay. 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. We selected the best-performing checkpoint based on the highest macro-F1 on the dev set for final evaluation. To mitigate overfitting, we monitor performance after each epoch and apply early stopping when no improvement is observed.

As shown in Table 3.5, BERT-base consistently outperforms RoBERTa-base in terms of overall performance and shows more stable results across both data splits. At the category level, both models perform best on *Logic Conflict*, *Syntactic*, and *Inferential Knowledge*. These categories may have more consistent linguistic patterns that the models can learn. For example, negation markers or contrastive connectives in logical conflict, syntactic rephrasing (e.g., active/passive) in syntactic reasoning, or world knowledge cues (e.g.,"students" → "school") in inferential cases.

By contrast, *Coreference* receives consistently low scores, likely due to both the small number of annotated examples and its inherent complexity: coreference resolution requires discourse-level understanding rather than local lexical or syntactic cues. This makes it more challenging for models to learn without explicit supervision.

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

Across both models, the 50/0/50 split yields slightly better results than the 40/20/40 setting, likely due to the increased amount of training data, which helps mitigate sparsity in less frequent categories.

Explanation Category	**data split**	***roberta-base***			**bert-base**		
	(***train/dev/test***)	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Coreference	40/20/40	0.00	0.00	0.00	0.00	0.00	0.00
	50/0/50	1.00	0.04	0.07	0.00	0.00	0.00
Semantic	40/20/40	0.58	0.63	0.61	0.54	0.68	0.61
	50/0/50	0.57	0.68	0.62	0.54	0.64	0.59
Syntactic	40/20/40	0.64	0.74	0.68	0.61	0.77	0.68
	50/0/50	0.62	0.76	0.69	0.62	0.80	0.69
Pragmatic	40/20/40	0.53	0.74	0.62	0.57	0.65	0.61
	50/0/50	0.59	0.63	0.61	0.60	0.58	0.59
Absence of Mention	40/20/40	1.00	0.23	0.38	0.95	0.42	0.58
	50/0/50	0.93	0.52	0.67	0.96	0.41	0.57
Logic Conflict	40/20/40	0.81	0.83	0.82	0.78	0.87	0.82
	50/0/50	0.81	0.83	0.82	0.78	0.88	0.83
Factual Knowledge	40/20/40	0.61	0.51	0.55	0.57	0.50	0.53
	50/0/50	0.62	0.55	0.59	0.61	0.56	0.58
Inferential Knowledge	40/20/40	0.75	0.81	0.78	0.79	0.76	0.77
	50/0/50	0.79	0.82	0.80	0.80	0.79	0.80
Overall							
acuuracy	40/20/40	0.67			0.70		
	50/ 0/50	0.67			0.70		
maro avg	40/20/40	0.47	0.49	0.47	0.60	0.58	0.58
	50/0/50	0.48	0.53	0.50	0.61	0.58	0.58
weighted	40/20/40	0.61	0.67	0.64	0.68	0.70	0.68
	50/0/50	0.65	0.69	0.66	0.68	0.70	0.69

Table 3.5: RoBERTA and BERT fine-tuning results.

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 et al., 2024) and DeepSeek-v3 (DeepSeek-AI et al., 2025). We use a consistent prompting strategy across models, with all prompt templates detailed in Table 3.6. We experiment with zero-shot prompting (no training examples), one-shot prompting (a single annotated example as demonstration), and few-shot prompting (k = 2 examples per category) under six experimental settings:

1. **Instructions:** whether a general task description is provided (yes/no)
2. **Examples:** the number of annotated examples provided per category (0, 1, or 2)

This results in the following six settings, and we systematically evaluate all combinations of these two variables to assess their impact on model performance.

$$\text{Instruction} \in \{\text{no}, \text{yes}\}, \quad \text{Examples per category} \in \{0, 1, 2\}$$

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. Logical Structure 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. Logical Structure 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 3.6: Instruction prompts for LLMs as classifiers.

For few-shot settings, we selected either one or two representative examples from the training data for each of the eight categories to include in the prompt. The LLMs are

then instructed to classify each explanation by outputting the category index (1-8). We evaluate both classification accuracy and the distribution alignment between the LLM outputs and the annotated gold human label distributions, as reported in Table 3.7.

Classifiers	**Accuracy**	**Precision**		**Recall**		**F1**		***Invalid predictions***
		macro	*weighted*	*macro*	*weighted*	*macro*	*weighted*	
Llama-3.2-3B-Instruct								
baseline	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								
baseline	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								
baseline	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 3.7: Comparison of LLM-based classification results under varying prompting strategies and few-shot settings.

From the results, we observe that at the model level, GPT-4o consistently achieved the highest performance across most experimental configurations, with accuracy peaking at 0.594 in the *+ one example per category* setting. Its macro-F1 and weighted-F1 also reach 0.492 in the *+ instruction + one example per category* setting and 0.583 in the *+ one example per category* setting respectively, outperforming all other LLMs. Notably, GPT-4o maintains 100% valid predictions across all settings, indicating that it reliably follows the prompt format and consistently outputs a valid category index. DeepSeek-v3 also shows strong classification ability, with performance comparable to GPT-4o in several few-shot settings. For instance, in the *+ instruction + two example per category* setting, it achieves an accuracy of 0.526, a macro-F1 of 0.478, and a weighted-F1 of 0.518, all within range of the GPT-4o's results. DeepSeek-v3 likewise outputs valid predictions with near-perfect consistency. In contrast, GPT-3.5-turbo and Llama-3.2-3B-Instruct exhibit lower performance. Llama-3.2-3B-Instruct fails frequently to produce valid outputs. For example, in the *+ instruction* setting, 29.54% of the predictions are invalid, and even with two examples per category provided in prompt, the invalid rate remains at 8.91%. While GPT-3.5-turbo demonstrated better reliability, it still continues to exhibit occasional output formatting issues - e.g., 1.16% invalid predictions in the *+ instruction + two examples* setting.

Overall, these findings highlight that larger and more instruction-tuned LLMs such as GPT-4o and DeepSeek-v3 are more capable of taxonomy classification tasks, especially when supported by instructions including category descriptions and examples. However, smaller models such as Llama-3.2-3B-Instruct exhibit limitations, both in classification accuracy and output format reliability, which may restrict their applicability in explanation categorization tasks without further fine-tuning.

Table 3.8 provides an overview of the classification performance across models, alongside random and majority-class baseline comparisons. 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. The random baseline assumes uniform guessing across the eight classes, while the majority baseline always predicts the most frequent class. 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. However, their performance still falls short of fine-tuned models like BERT and RoBERTa, indicating that large models do not yet fully capture the fine-grained distinctions in our taxonomy without finetuning, which leaves room for future improvement.

Classifiers	Accuracy	Precision	Recall	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 3.8: 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. Precision, Recall, and F1 are at the macro-level.

3.5 Taxonomy Analysis

To understand the empirical behavior and utility of our proposed taxonomy, we conduct three complementary analyses. First, we examine how explanation categories co-occur with NLI labels, identifying systematic associations between certain reasoning types and inference outcomes. Second, we investigate within-label variation in e-SNLI by measuring the semantic diversity among explanations assigned the same label, using the taxonomy to characterize and quantify this variation. Finally, we analyze the relationship between taxonomy categories and the span length of human highlights, exploring whether different reasoning types correlate with distinct levels of lexical grounding. Together, these analyses provide insight into how the taxonomy organizes explanatory reasoning in NLI along both semantic and surface dimensions.

3.5.1 Co-occurrence of Explanation Categories and NLI Labels

Figure 3.5 plots the distribution of our explanation categories and their co-occurrence with NLI labels. Overall, we observe that different explanation categories show distinct distributions over NLI labels, suggesting that different types of reasoning are preferentially associated with specific label decisions.

As expected, the *Logic Conflict* category is strongly dominated by the contradiction label. This is consistent with the definition of the category, which captures explanations that point to logical inconsistency between the premise and hypothesis. The frequent use of logic-based reasoning in contradiction cases shows that identifying contradictions often relies on detecting structural conflicts.

Conversely, the *Syntactic*, *Semantic*, and *Pragmatic* categories are predominantly associated with the entailment label. These explanation types often reflect surface-level

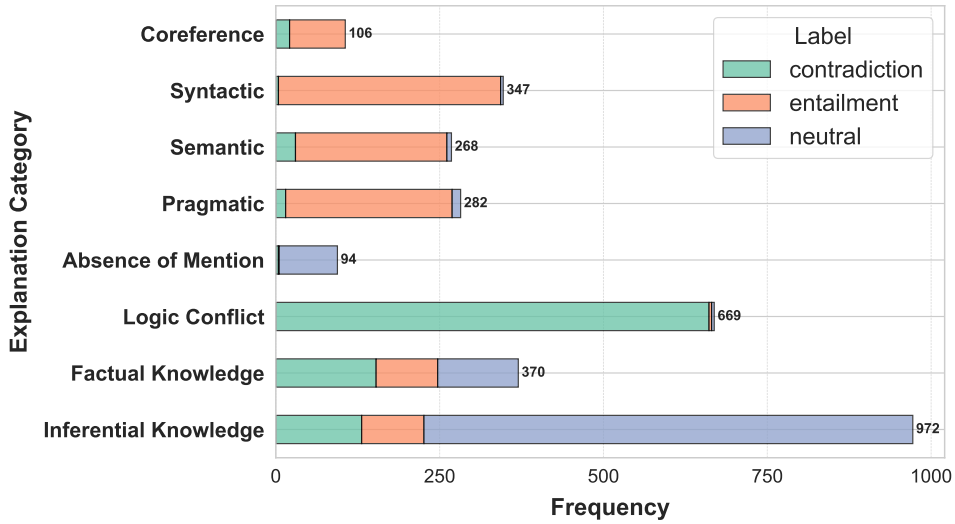

Figure 3.5: Distribution of LITEX categories on LITEX-SNLI explanations across NLI labels (n = 3,108).

alignment between the premise and hypothesis, such as syntactic-level paraphrase, synonymous expressions, or implicatures, which tend to support entailment judgments. The high frequency of these categories under entailment suggests that linguistically aligned sentence pairs often rely on these forms of reasoning to justify the positive label.

The *Factual Knowledge* and *Inferential Knowledge* categories appear more evenly distributed across all three NLI labels. This distribution reflects the flexible nature of world knowledge reasoning: factual or inferential cues can be used to support entailment (e.g., "Paris is in France"), to identify contradiction (e.g., when external knowledge contradicts the hypothesis), or to highlight neutrality due to missing links. The fact that these categories apply across all labels shows they play an important role in real-world NLI, not just in cases based on surface text.

Finally, *Absence of Mention* is strongly aligned with the neutral label. This is expected, as the neutral labels often rely on the identification of missing or unstated information - i.e., the hypothesis introduces content not explicitly entailed by the premise. This category reflects a key annotation strategy used to justify neutrality without asserting contradiction or entailment.

Together, these co-occurrence patterns not only validate the internal coherence of the taxonomy but also reveal how different forms of reasoning align with inferential outcomes, offering a richer perspective on what underlies label assignments in NLI datasets.

3.5.2 Within-label Variation

Table 3.9 gives the counts of our 1,002 NLI items for which the three (or more) explanations were annotated with 1, 2, or ≥ 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.

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

To quantify within-label variation more precisely, we compute pairwise similarity between the explanations for each NLI item using standard metrics, following Giulianelli et al. (2023) and Chen et al. (2024a). These include lexical metrics (1-gram, 2-gram, and 3-gram overlap), morphosyntactic metrics (part-of-speech n-gram overlap), and semantic similarity metrics (cosine similarity and Euclidean distance in sentence embedding space), as well as BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004), commonly used in natural language generation evaluation.

Figure 3.6 presents boxplots of these similarity scores, grouped by the number of LiTEx taxonomy categories assigned to the explanations of an NLI item (i.e., 1 category, 2 categories, or 3 or more categories). We observe a clear trend: as the number of assigned explanation categories increases, the average similarity between explanations tends to decrease. In particular, items whose explanations fall under a single taxonomy category (green boxes) generally show higher median similarity scores, suggesting that shared reasoning type contributes to more semantically aligned explanations. This supports our hypothesis that within-label variation is closely tied to underlying reasoning diversity. In contrast, items associated with multiple taxonomy categories (especially 3 or more) exhibit markedly lower median similarity and greater variance, suggesting more diverse and potentially conflicting reasoning strategies. These findings empirically validate the design of the taxonomy: when annotators select multiple reasoning categories for an NLI item, the explanations they provide are indeed more varied. This supports the taxonomy's utility in capturing within-label variation.

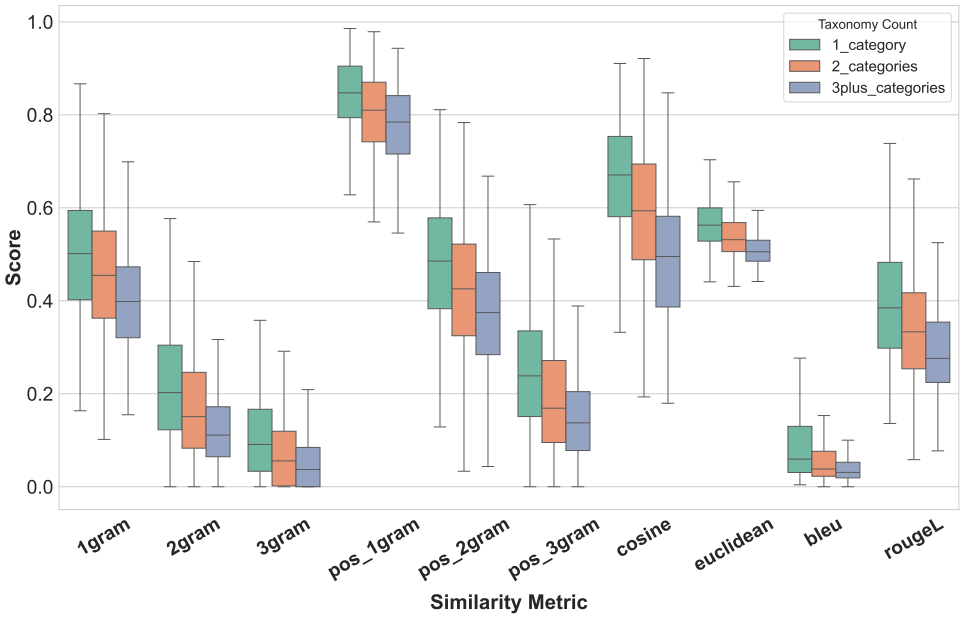

Figure 3.6: Boxplot of explanation similarities grouped by number of LITEX categories on an NLI item. Each boxplot illustrates the central tendency and dispersion of the data, facilitating comparison across conditions. The horizontal line inside each box indicates the median; whiskers denote the range of non-outlier data.

3.5.3 Highlight Length vs. Taxonomy Category

We analyze the highlight span lengths for different explanation categories in Figure 3.7. On average, premises and hypotheses contain 13.81 and 7.41 words.

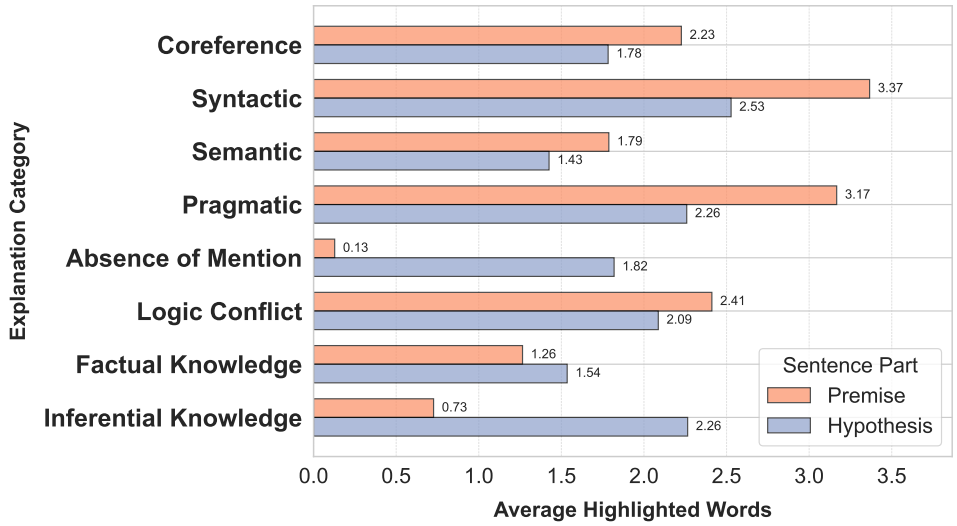

Figure 3.7: Average number of highlighted words in each premise-hypothesis pair across LITEX categories.

Syntactic explanations show the longest average highlights in both the premise and the hypothesis, consistent with their need to holistic sentence-level understanding. *Pragmatic-level Inference* also shows relatively long highlight spans (premise $\approx$ 3.17 words, hypothesis $\approx$ 2.25 words), reflecting discourse-level reasoning across sentences (Lascarides and Asher, 1993). *Logic Conflict* explanations demonstrate moderate highlight lengths (premise $\approx$ 2.41 words, hypothesis $\approx$ 2.09 words), focusing on segments with logical relations. In contrast, *Semantic* explanations have shorter highlight spans (premise $\approx$ 1.79 words, hypothesis $\approx$ 1.43 words), targeting specific lexical items such as synonyms, antonyms, or semantic shifts. *Absence of Mention* explanations display little highlighting in premises but more in hypotheses, consistent with the task of marking unmatched information. Lastly, *Inferential Knowledge* (premise $\approx$ 0.73 words, hypothesis $\approx$ 2.26 words) and *Factual Knowledge* (premise $\approx$ 1.26 words, hypothesis $\approx$ 1.54 words)show short spans, reflecting reliance on external knowledge rather than textual cues.

These observations demonstrate that the length of highlight spans and distribution vary systematically across reasoning types, offering linguistic evidence that within-label variation reflects not just surface form differences but also deeper inference processes.

3.6 Interim Summary for Chapter 3

In this chapter, we introduce our proposed taxonomy LITEX for categorizing free-text explanations in NLI. In §3.1, we motivate the need for such a taxonomy based on patterns observed in existing free-text explanations. We then present the full taxonomy structure, which consists of two broad reasoning types - **Text-based Reasoning** and **World-Knowledge Reasoning** - along with eight fine-grained categories, each illustrated with examples.

In §3.2, we describe the annotation process on the e-SNLI dataset, outlining the annotation guidelines and procedures, including the training of annotators and implementation details.

To validate the reliability of our taxonomy, §3.3 reports IAA results on two levels: taxonomy-based explanation classification and explanation-supporting highlight spans. The high agreement scores suggest that the taxonomy is interpretable and consistently applicable across annotators.

In §3.4, we further evaluate the taxonomy's learnability by fine-tuning two pre-trained language models, BERT-base-uncased and RoBERTa-base, to classify free-text explanations into taxonomy categories. The models achieve strong performance, indicating that

the taxonomy is not only human-interpretable but also machine-learnable.

Finally, §3.5 presents three analyses based on our annotated data to better understand how the taxonomy captures reasoning variation. These include the co-occurrence of taxonomy categories with NLI labels, the presence and degree of within-label variation, and the relationship between explanation category and highlight span length. Together, these analyses demonstrate the descriptive value of the taxonomy in revealing the structure and diversity of inferential reasoning in NLI.

4 Generating Explanations Using Taxonomy and Highlight

In this chapter, we present our approach to generating free-text explanations for NLI using prompting strategies that incorporate both linguistic reasoning structures and input highlighting. Specifically, we compare three paradigms that provide varying levels of guidance to the model: a minimally guided baseline prompt, a highlight-guided approach that directs model attention to key input tokens, and our proposed taxonomy-guided prompting that incorporates explicit reasoning categories.

We introduce several prompting paradigms (§4.1–§4.3) that vary in their design and degree of structure: from minimally guided baseline prompts to highlight-guided inputs (both *human highlight* and *model highlight*), and finally to taxonomy-guided prompts that encode explicit reasoning categories. Each paradigm is implemented with multiple variants to assess fine-grained differences, such as indexed vs. in-text highlights and *two-stage* vs. *end-to-end* taxonomy prompting.

To evaluate explanation quality, we compare model-generated outputs to human-written references from the e-SNLI dataset using a suite of lexical, syntactic, and semantic similarity metrics (§4.4). This enables us to measure how well different prompting strategies align with human reasoning. We further analyze the semantic coverage of generated explanations using t-SNE visualizations and case studies (§4.5), offering insights into how well models capture within-label variation across NLI items.

Finally, we conduct a round of human validation (§4.6), where annotators assess whether the generated explanations are consistent with the gold NLI label and match the intended taxonomy category. These evaluations provide both quantitative and qualitative perspectives on the strengths and limitations of current LLMs in explanation generation and highlight the importance of structured prompting for eliciting interpretable and reasoning-grounded model behavior.

4.1 Baseline

To generate explanations for NLI examples, we conduct a series of generation experiments, prompting three instruction-tuned LLMs to generate explanations: GPT-4o, DeepSeek-v3, and Llama-3.3-70B-Instruct. The prompting paradigms are adapted from the instruction-based explanation generation setup proposed by Chen et al. (2024a). Each paradigm is designed to inject different forms of supervision or inductive bias into the generation process. 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.

The baseline prompting setup serves as our lower bound. The model only sees the NLI item (premise and hypothesis) and a label, and generates explanations based on this input. This setup tests how well an LLM can explain NLI reasoning without explicit structural guidelines.

Baseline Prompt:
```
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}
```

4.2 Generating Explanations Using Highlights

We explore a highlight-guided explanation generation setup, where token-level highlights in the premise and hypothesis are used to steer LLMs toward salient regions of the input. Our approach involves two sources of highlights:

- **Human highlights:** Taken directly from the e-SNLI dataset. These highlights were originally annotated by crowd workers as part of a three-part annotation process (label, explanation, highlight), and we use them here as guidance signals. We experiment with two ways of presenting them to the model: as raw token indices (*indexed* format) or by marking them directly in the input text (*in-text* format).

- **Model-predicted highlights:** Highlight spans are automatically generated using the same LLM that is later used for explanation generation (e.g., GPT-4o, DeepSeek-V3, or LLaMA 3.3-70B). For each NLI item, we first prompt the model with the premise, hypothesis, and gold label, asking it to select word indices in both the premise and hypothesis that are most relevant for justifying the inference. This setup ensures consistency across the two stages: the same model first identifies salient regions in the input and then uses those highlights to generate an explanation. In doing so, we simulate a fully automated pipeline where both focus selection and explanation are performed by the same LLM without human intervention.

For each source of highlights, we test two encoding strategies for integrating the highlighted information into the prompt:

- **Indexed:** Token indices of the highlighted words are explicitly included alongside the original sentence.

- **In-text:** The highlighted tokens are embedded directly in the text, marked with **.

These combinations yield four experimental conditions: *human highlight (indexed)*, *human highlight (in-text)*, *model highlight (indexed)*, and *model highlight (in-text)*.

Below, we describe the prompt formats used for model-based highlight generation, as well as the two explanation prompting variants that consume indexed or in-text highlights.

Highlight Generation: This prompt is used only in the model-based setting to predict relevant word indices in both the premise and hypothesis. Based on the provided NLI label, the LLM is instructed to select tokens that are critical for inference reasoning. Specific guidelines are given: e.g., for entailment, highlight at least one word in the premise; for contradiction, highlight words in both premise and hypothesis (as described for e-SNLI in §3.2). The predicted indices are later used as input for explanation generation.

Highlight Generation Prompt:

```
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:
...
```

Highlight Indexed: After obtaining the predicted highlights (i.e., token indices), these are provided explicitly to the model alongside the NLI input. The model is then prompted to generate explanations focusing on these indexed tokens.

General Instruction Prompt (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: In this variant, the same predicted highlights are inserted directly into the input sentences by surrounding the selected tokens with ** markers. The model receives the marked-up text and is instructed to focus on the highlighted parts when generating explanations. This approach keeps the input in a more natural text form while still guiding the model's attention.

General Instruction Prompt (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.
```

To clarify how the same highlight is presented under both variants, we provide an illustrative example below.

Example Highlight Encoding (for comparison):

- **Original sentence:** *The church has cracks in the ceiling.*
- **In-text (marked with **):** *The church has **cracks** **in** **the** **ceiling**.*
- **Indexed (token indices):** 3, 6, 4, 5

4.3 Generating Explanations Using Taxonomy

The model is provided with the taxonomy description (Table 3.2), one example for each of the eight reasoning categories, and the full taxonomy. We explore two overall prompting approaches that integrate the taxonomy into the generation process:

- **Classification + Two-Stage Generation:** The model first predicts the reasoning category (or categories) for a given NLI item (classification step), and then generates an explanation conditioned on the predicted category (generation step).
- **End-to-End Generation:** The model jointly identifies the relevant taxonomy categories and generates corresponding explanations in a single prompt.

This comparison is motivated by the hypothesis that end-to-end prompting, though efficient, may introduce bias toward certain salient categories. These templates are adapted and refined based on the approach of Chen et al. (2024b). 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.

In total, the taxonomy-guided setup includes three prompting variations corresponding to the two approaches:

Taxonomy Classification (Step 1 of Two-Stage): This prompt is used in the first stage of the two-stage setup. Given a premise, hypothesis, and label, the model is tasked with identifying all reasonable reasoning categories from the predefined LITEX that could support the inference. One example per category is provided to help the model ground its decisions.

Taxonomy Classification Prompt):

```
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.
```

Taxonomy-Guided Generation (Step 2 of Two-Stage): This prompting variation forms the second stage in the two-stage pipeline. After the taxonomy classification step, the predicted reasoning category (e.g., 1: *Coreference)* is selected and passed as part of the input to the generation prompt. The LLM is asked to generate all possible explanations that match the given taxonomy category, supported by a detailed category description and one illustrative example. This prompt design helps analyze how well LLMs can generate

category-specific explanations when provided with an explicit reasoning signal.

Taxonomy-Guided Generation Prompt:

```
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: In this variant, we merge reasoning category classification and explanation generation into a single prompt. The LLM first selects all relevant explanation categories from LITEX that could support the NLI label. It then generates corresponding explanations for each selected categories. The taxonomy definitions are included directly in the prompt. This approach tests whether LLMs can reason and generate in a more unified manner, but may also risk overfitting to frequent or more salient categories.

General Instruction Prompt (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)
```

4.4 Model Generation Results

4.4.1 Experimental Setups

We evaluate the quality of model-generated explanations by comparing them against human-written references from e-SNLI using a suite of automatic similarity metrics (described in §3.5.2). Each generated explanation is individually compared to the three gold

references associated with the same NLI item, and the best-matching reference is selected based on the metric score. These best scores are then averaged across all generated explanations for that item to obtain a single per-item score. The final reported score is the average across all NLI items in the dataset.

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

Table 4.1: Summary of prompt types and input variants used in explanation generation.

We experiment with multiple prompting strategies that incorporate different forms of inductive bias. Specifically, we explore:

- **Taxonomy-guided prompting**, with either a two-stage setup (classification then generation) or an end-to-end format;
- **Highlight-guided prompting**, where token-level supervision is provided either via human-annotated highlights (from e-SNLI) or model-predicted highlights, and encoded either as token indices (indexed) or inline markers (in-text).

The LLMs used for generation include GPT-4o, DeepSeek-v3, and Llama-3.3-70B-Instruct. Each prompt type is evaluated using all three models to ensure consistency across conditions. Table 4.1 summarizes the experimental configurations, showing the combination of prompt types, input variants, and LLMs used in our generation experiments.

4.4.2 General Results

For each generated explanation, we evaluate it against each of the three human references from e-SNLI and retain the score from the best-matching reference. We then average the scores over all generated explanations per NLI item. The final results reported in Table 4.2 are averages of these per-item scores over the full dataset.

4.4.3 Human Highlight vs. Model-Generated Highlight

We first compare the performance of explanations guided by human-annotated highlights from e-SNLI against those using model-predicted highlight spans. Across all models, model-predicted highlights consistently yield slightly better or comparable performance to human highlights in both lexical and semantic similarity metrics. For instance, in the case of GPT-4o, model highlight (indexed) achieves higher 1-gram (0.402 vs. 0.395), 2-gram (0.124 vs. 0.116), and 3-gram overlap (0.053 vs. 0.050), along with improved semantic similarity (cosine: 0.522 vs. 0.549) comparing to human highlight (indexed). These results suggest that LLMs are able to identify informative regions of the input with reasonable effectiveness, and that these model-generated highlights may help guide explanation generation. However, we note that no human evaluation has been conducted to directly assess the quality of these highlights relative to human-selected spans. One possible explanation is that model-predicted highlights are more internally consistent with the LLM's own generation mechanisms, leading to better alignment between attention and output content.

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

Table 4.2: Similarity of LLM-generated explanations to human references. Bold scores denote the best performance.

4.4.4 Highlight Indexed vs. Highlight In-Text

Across both human- and model-guided highlighting setups, we observe a consistent pattern: indexed highlighting performs comparably to in-text highlighting across most metrics, and outperforms it on certain dimensions. It is possibly due to the fact that the index format - where the model is directly told which word indices to focus on - provides more explicit and structured guidance, likely reducing ambiguity during explanation generation. For example, in GPT-4o, the model highlight (indexed) setting achieves higher n-gram and semantic scores than the in-text version (e.g., POS 3-gram overlap of 0.222 vs. 0.208; cosine similarity 0.554 vs. 0.555). This suggests that directly indexing tokens may more effectively constrain the model's focus, whereas marking tokens in-text introduces ambiguity due to varying context lengths and potential tokenization inconsistencies. However, the magnitude of this difference remains unclear, and further controlled evaluation would be needed to assess its statistical significance and practical impact.

4.4.5 Taxonomy Two-Stage vs. End-to-End

When comparing taxonomy-guided prompting strategies, we observe a nuanced tradeoff across models. On GPT-4o and DeepSeek-v3, the *end-to-end* prompting strategy performs best across nearly all metrics, achieving the highest n-gram overlap (e.g., 3-gram: 0.083 on GPT-4o), POS n-gram scores, and semantic similarity (e.g., cosine similarity: 0.608 on GPT-4o). In contrast, Llama-3.3-70B shows better performance under the *two-stage* setup, particularly for semantic alignment and lexical precision (e.g., ROUGE-L: 0.321 vs. 0.262).

This suggests that larger open-source models like Llama may benefit from explicit decomposition of tasks — first classifying the reasoning category and then generating explanations — while more instruction-tuned models like GPT-4o and DeepSeek-v3 can handle the reasoning integration more effectively in a unified *end-to-end* manner. Moreover, the *two-stage* approach may mitigate reasoning bias in open-source models, guiding them toward more faithful category-conditioned generation.

4.4.6 Baseline vs. Highlight-Guided vs. Taxonomy-Guided

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 similarity. 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 4.2, 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.

4.5 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.* In other words, are LLMs simply producing repetitive, template-like response that reflect only a narrow subset of human reasoning patterns? Or 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.

To this end, we propose a simple yet intuitive visualization-based analysis. As shown in Figure 4.1, we select three representative instances from LITEX-SNLI and visualize the embeddings of human and model-generated explanations in 2D space. The embedding representations are obtained using the all-distilroberta-v1 model, a widely used pretrained sentence encoder (Sanh et al., 2020). These embeddings are then projected into two dimensions using t-SNE, a non-linear dimensionality reduction technique commonly used for visualization of high-dimensional spaces.

Although t-SNE is primarily a visualization tool, prior work has leveraged it to study semantic spaces through geometric constructs such as convex hulls (Mimno and Lee, 2014). Following this line of work, we compute convex hulls around the projected points of both human-written and model-generated explanations to assess their relative coverage. The convex hull of a point set intuitively represents its semantic span in the projected space.

The figure illustrates three prototypical coverage patterns: (1) full coverage, where the convex hull of model-generated explanations fully encloses the human explanation points (blue stars); (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. These patterns offer a qualitative insight into the generative coverage and generalization capacity of LLMs with respect to human reasoning.

4.5.1 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

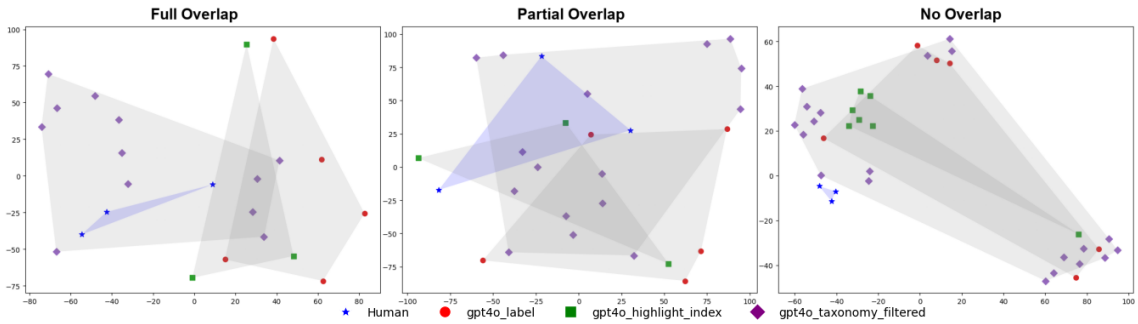

Figure 4.1: Representative t-SNE visualizations of explanation embeddings. The blue convex hull represents the span of human-written explanations, while the gray illustrates the spread of GPT4o-generated explanations.

t-SNE visualizations and convex hull statistics (van der Maaten and Hinton, 2008).

An NLI item is *fully covered* if all human explanation reference points are positioned within the convex hull spanned by the model explanations. 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.

4.5.2 Results

Table 4.3 shows that taxonomy-guided explanation generation consistently achieves the highest full and partial coverage of reference explanation points. For this analysis, we focus on two representative prompting setups: the *indexed* highlight format for model explanations and the *end-to-end* taxonomy-based prompting. These were selected based on earlier performance observations: *indexed* highlighting outperforms *in-text* for GPT-4o, and *end-to-end* generation yields better coverage than the *two-stage* variant for taxonomy-guided models. The selected setups also achieve the highest average area recall and precision across all test cases (except for the GPT-4o baseline), indicating substantial semantic overlap between model-generated and human-written explanations.

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

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.

4.5.3 Case Study

Table 4.4 provides a concrete example (the leftmost case in Figure 4.1) where the human explanations are fully covered by the taxonomy-guided generation but only partially captured by label- and highlight-guided generations.

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 4.4: 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.

The human explanations in this case revolve around two key reasoning strategies: spatial inference, grounded in the semantics of "near" and real-world contextual knowledge about coconuts typically being outdoors. These explanations are not only concise but

also closely aligned with the core entailment relation, linking physical location with co-occurrence plausibility. Notably, they include clear inferential cues such as "would be standing outside" and "implies near".

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. For example, phrases like "aligns with the spatial arrangement described in the content" remain vague and do not make explicit connections between specific premise and hypothesis elements.

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. This *Syntactic* explanation shows how the premise is shortened or simplified in the hypothesis. It is something people often understand naturally and it can also be seen as a reasonable explanation.

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. This points to a potential trade-off between coverage and naturalness when modeling explanations via structured supervision.

Finally, explanation length provides another lens into content density and redundancy. On average, human explanations are substantially shorter (16.3 tokens) than generated ones: 27.1 for the baseline model, 33.7 for highlight-guided, and 25.5 for taxonomy-guided. While taxonomy-guided explanations are closer in length to human ones, all model outputs remain longer, reinforcing the tendency of neural generation models to overgenerate or restate ideas. This pattern echoes the redundancy findings discussed in §4.4, suggesting that even with targeted supervision, generation systems may benefit from further constraints or post-editing mechanisms to align better with human communicative efficiency.

4.6 Model Generation Validation

To assess the quality of model-generated NLI explanations, we conduct a round of human validation. Specifically, we evaluate explanations produced by GPT-4o under the taxonomy *two-stage* prompting paradigm introduced in §4.3, 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. Validation is performed by a single trained annotator. For each explanation, the annotator is provided with the premise, hypothesis, NLI label, and 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)

To clarify the annotation criteria, we provide examples of failure cases for each dimension:

Label inconsistency:
Premise: An older woman tending to a garden.
Hypothesis: The lady is weeding her garden.
NLI Label: neutral
Explanation: "Older woman" might suggest more than one person, while "the lady" refers to one.

Incorrect: The explanation incorrectly focuses on referential ambiguity, implying a contradiction.

Taxonomy inconsistency:
Premise: A woman in a black jacket and a black and white striped skirt is watching a woman talking to a little boy sitting on concrete steps.
Hypothesis: A father looks at his wife and son.
Taxonomy Category: Logic Conflict
Explanation: The premise lacks any male figure who could be the father in the statement.
Incorrect: While the explanation correctly identifies a mismatch between the premise and hypothesis, the reasoning is not based on logical conflict (e.g., mutual exclusivity or contradiction). Instead, it relies on *Absence of Mention*, i.e., the hypothesis introduces new entities not entailed or contradicted by the premise.

We conduct human validation on a total of 8,373 model-generated explanations, covering 719 unique NLI instances. The aggregated results are as follows:

- **Validation Question 1 (label consistency)**: Yes: 8,228 (98.27%), No: 145 (1.73%)
- **Validation Question 2 (taxonomy consistency)**: Yes: 7,020 (83.84%), No: 1,353 (16.16%)

These results indicate that while the majority of model-generated explanations are correctly aligned with the provided NLI label, a notable proportion of explanations still mismatch with the intended taxonomy category. Specifically, around 16% of the generated explanations fail to conform to the assigned taxonomy category in the prompt. It suggests that controlling fine-grained reasoning structures remains a significant challenge for LLMs, even when provided with explicit category guidance.

These results raise important questions: *Are certain types of reasoning more difficult for models to express consistently?* And conversely, *are some taxonomy categories more naturally aligned with model generation behavior?* To answer these questions, we next break down the validation results by taxonomy category, examining how explanation-label and explanation-taxonomy alignment vary across different reasoning types as shown in Table 4.5.

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)
Logical Structure 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 4.5: Human validation results for model-generated explanations by taxonomy category. Q1: Whether the explanation support 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 §3.3.1, 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*.

Overall, these findings suggest that while LLMs can generate explanations that are coherent and generally aligned with the gold label, they sometimes fall short in accurately reflecting the specific type of reasoning required, particularly for discourse-sensitive or pragmatically complex categories such as *Coreference* and *Pragmatic*. The human validation results indicate high overall taxonomy consistency (over 80%), but also reveal that certain reasoning types remain more challenging for models to capture correctly. This highlights the potential benefit of more structured prompting or fine-grained supervision to guide models toward producing explanations that more faithfully represent the intended reasoning categories.

4.7 Interim Summary for Chapter 4

In this chapter, we investigate how our proposed taxonomy can support LLMs in generating high-quality free-text explanations for NLI. §4.1 to §4.3 introduce the design of our prompting paradigms, including the baseline setup, highlight-guided generation, and taxonomy-guided generation. We implement multiple variations within each setup, for example, *in-text* vs. *indexed* highlighting, and *two-stage* vs. *end-to-end* taxonomy prompting—to explore the effects of different guidance strategies on explanation quality.

In §4.4, we present the generation results from three LLMs under the proposed prompting strategies. This section focuses on evaluating the similarity between model-generated and human-written explanations using standard reference-based metrics. We compare performance across several dimensions, including human vs. model-predicted highlights, *indexed* vs. *in-text* highlights, and taxonomy *two-stage* vs. *end-to-end* prompts. We also assess performance across the three broader prompting categories: baseline, highlight-guided, and taxonomy-guided. Our results show that taxonomy-guided prompts consistently yield higher similarity scores, suggesting their effectiveness in aligning model outputs with human-written explanations.

§4.5 provides an explanation coverage analysis using t-SNE visualizations. This analysis evaluates how well LLM-generated explanations capture the semantic space of human explanations, thereby reflecting within-label variation. We supplement the quantitative

plots and coverage tables with qualitative case studies. The results show that taxonomy-guided generation tends to produce explanations with broader coverage and higher overlap with human-written explanations. Case studies further reveal that these explanations are often more coherent, informative, and diverse in reasoning types.

Finally, §4.6 presents a round of human validation to assess the quality of model-generated explanations. We evaluate the explanations on two dimensions: label consistency and taxonomy consistency. The results indicate that while LLMs generally produce coherent and label-aligned explanations, they occasionally fail to express the intended reasoning type accurately. This highlights the need for more structured prompting or training to guide models toward taxonomy-aligned reasoning.

5 Assessing the Applicability of LITEx Across NLI Benchmarks

To further evaluate the generalizability and applicability of our proposed taxonomy of explanations (LITEx) across diverse NLI benchmarks, we extend our annotation beyond the e-SNLI dataset. Specifically, we apply the LITEx taxonomy to two additional NLI datasets - **LiveNLI** (Jiang et al., 2023) and **VariErr** (Weber-Genzel et al., 2024) - which differ structurally and methodologically from e-SNLI.

We begin by introducing the two datasets and comparing their characteristics to e-SNLI (§5.1). We then present the annotation results in §5.2 and §5.3, focusing on analysis such as how LITEx explanation categories co-occur with NLI labels and how these patterns differ across datasets. In §5.4, we examine explanation similarity within fine-grained (category, label) groups to assess redundancy and internal consistency. In this subsection, we further analyze explanation variation at the level of individual NLI items, combining similarity metrics, visualization, and case studies to better understand what drives convergence or divergence in explanations within the same reasoning category.

5.1 Datasets: LiveNLI and VariErr

LiveNLI is a high-quality explanation dataset built upon a subset of the MNLI dataset (Williams et al., 2018), consisting of 122 NLI examples. Each example is annotated by at least 10 crowdworkers who independently assign one or more NLI labels (*true*, *either*, or *false*), highlight relevant spans in the premise and hypothesis, and provide a free-text explanation justifying their label choices (Jiang et al., 2023).

Inspired by previous work on explanation collection (Camburu et al., 2018; Wiegreffe and Marasović, 2021; Tan, 2022), LiveNLI adopts specific annotation guidelines to promote high-quality, interpretable explanations. Annotators were explicitly instructed to provide reasoning that adds new information or clarifies inference steps, rather than simply copying or paraphrasing the premise or hypothesis. They were also asked to refer to specific parts of the sentences and to highlight the words in the premise or hypothesis that were most relevant to their explanation. This design encourages explanations that are both informative and grounded in the input, making them more suitable for studying fine-grained reasoning and model alignment. Similar to e-SNLI, LiveNLI captures variation in human reasoning by collecting free-text explanations for NLI items. However, LiveNLI differs in several key ways. Firstly, it is ecologically valid; both the label and explanation come from the same annotator, preserving the natural reasoning process. Secondly, each instance may be associated with more than three explanations. Lastly, because LiveNLI includes multiple explanations across different NLI labels for the same instance, it enables the analysis of both within-label and cross-label explanation variation. This feature is something that e-SNLI does not support. The dataset thus provides a rich resource for studying ambiguity, interpretability, and the diversity of human reasoning in natural language inference. This characteristic of LiveNLI aligns well with our motivation for introducing LITEx, which aims to capture and model within-label variation in NLI.

VariErr is a complementary dataset focusing on variation and errors in English NLI. It contains 1,933 model-generated explanations for 500 re-annotated MNLI items, accompanied by 7,732 human validity judgments. Unlike LiveNLI, which builds on natural disagreement among annotators, VariErr includes both plausible variations (i.e., alternative valid hypotheses) and human annotation errors from the first round of labeling, with

the goal of analyzing where and why models fail (Weber-Genzel et al., 2024). Each item in VariErr is annotated in two consecutive rounds: first, the annotators provide a free-text explanation for the given NLI label; then, in a second step, they assess the validity of label-explanation pairs, either confirming the reasoning or identifying errors in the justification. Explanations are written in context, under relatively naturalistic settings, such as debugging model mistakes or interpreting near-miss predictions. Like LiveNLI, VariErr also allows for the presence of multiple plausible labels for a given item, especially in cases involving subtle ambiguity or underspecification.

What distinguishes VariErr is its explicit focus on diagnostic signal—the dataset includes both correct and incorrect examples, offering a contrastive perspective on explanation types associated with success versus failure in inference. The explanations have also undergone validation to ensure consistency with the intended label (either gold or majority), making them suitable for fine-grained annotation and reasoning analysis. To ensure reliability, we restrict our annotation to explanations that have been previously validated—i.e., those confirmed to be consistent with the self-validated or majority label.

Another key difference between LiveNLI and VariErr, as compared to e-SNLI, lies in the labeling structure of the datasets. While e-SNLI assigns a single gold label (i.e., entailment, contradiction, or neutral) to each NLI instance, both LiveNLI and VariErr may provide multiple plausible NLI labels per item. This shows that people can genuinely disagree on the correct label, which makes it harder to interpret and categorize explanations. Applying our taxonomy to these two datasets allows us to examine its usefulness in settings where explanations span across different NLI labels.

We follow the same annotation procedure described in Section §3.2, in which annotators are presented with a premise, hypothesis, NLI label, and a corresponding explanation. The annotation is conducted by one trained annotator whose task is to assign one or more appropriate categories from LITEX to the explanation, based on the reasoning type(s) explicitly evidenced in the text.

5.2 Annotation Results on LiveNLI

In this subsection, we present the annotation results on the LiveNLI dataset. A total of 1,404 [1] NLI explanations are annotated using LITEX. Among them:

- **1,234 explanations (87.9%)** are assigned a **single** explanation category.
- **167 explanations (11.9%)** are assigned with **two** explanation categories.
- **3 explanations (0.2%)** are assigned **three or more** categories.

This distribution suggests that while some explanations exhibit multiple taxonomy labels reasoning, the majority of explanations can still be characterized by a single dominant reasoning type.

Figure 5.1 visualizes the co-occurrence matrix of annotated explanation categories. The most frequently assigned categories were *Inferential Knowledge* (422 instances), *Semantic* (306 instances), and *Absence of Mention* (298 instances), followed by *Pragmatic* (194 instances), *Syntactic* (158 instances), and *Logic Conflict* (125 instances). Less frequent categories included *Coreference* (46 instances) and *Factual Knowledge* (28 instances), indicating that these reasoning types were less prominent or explicitly expressed in the dataset.

Co-occurrence patterns further reveal interactions between different types of reasoning. For example, *Absence of Mention* frequently co-occurred with *Inferential Knowledge* (35 instances), and *Inferential Knowledge* also showed notable overlap with *Semantic*(34 instances). Additionally, *Semantic* explanations often appeared alongside *Coreference* (19

[1] We use the publicly released LiveNLI data, which contains 1,415 lines. After removing empty entries, 1,404 valid explanations remain.

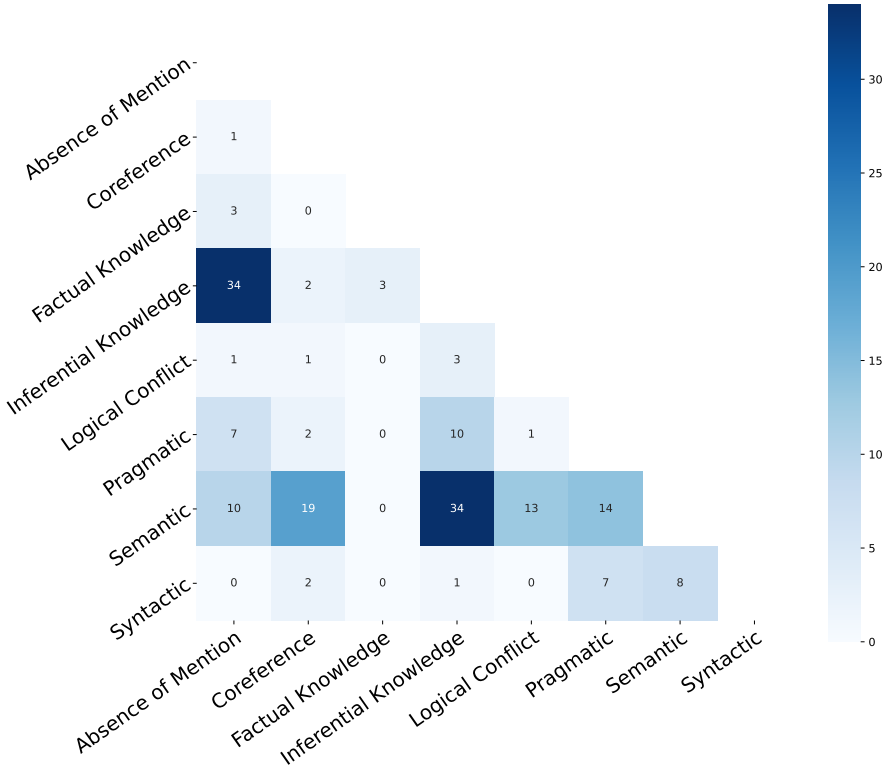

Figure 5.1: Co-occurrence Matrix of Explanation Categories in LiveNLI.

instances), suggesting that these categories are not mutually exclusive and can reflect layered reasoning processes in NLI explanations. Overall, these results demonstrate the suitability of the LITEx taxonomy for capturing diverse explanation types in LiveNLI, while also providing insight into how different forms of inference may co-occur within a single explanation.

Taxonomy-Guided Decomposition of Multi-Category Explanation To better leverage multi-category explanations, we adopt a taxonomy-guided decomposition strategy. Building on our annotation findings in §3.2, we split each explanation assigned multiple categories from LITEx into distinct sub-explanations, with each segment aligned to a single reasoning type. This approach is particularly suited to LiveNLI, where explanations tend to be longer in terms of word count and semantically richer than those in e-SNLI.

Consider the following annotated explanation with two categories from LITEx:

- **Premise**: In the summer, the Sultan's Pool, a vast outdoor amphitheatre, stages rock concerts or other big-name events.
- **Hypothesis**: Most rock concerts take place in the Sultan's Pool amphitheatre.
- **NLI Label**: true (entailment) — either (neutral)
- **Explanation**: The size of the amphitheatre suggests that it's easy to hold big events at Sultan's Pool, so the statement is probably true. However, nothing in the context limits big events from being held somewhere else, thus the statement may be true or false.
- **Assigned Categories**: *Inferential Knowledge*, *Absence of Mention*

This explanation can be decomposed into two category-specific segments:

- **Inferential Knowledge**:
  "The size of the amphitheatre suggests that it's easy to hold big events at Sultan's Pool, so the statement is probably true."
  NLI Label: *true (entailment)*
  This segment uses world knowledge to infer plausibility.

- **Absence of Mention**:
  "However, nothing in the context limits big events from being held somewhere else, thus the statement may be true or false."
  NLI Label: *either (neutral)*
  This segment highlights that the context does not explicitly restrict other alternatives.

This decomposition operation brings multiple benefits. Firstly, it increases the number of usable data instances without requiring new annotations, effectively augmenting the dataset. Secondly, by isolating individual reasoning strategies, the split explanations allow for finer-grained supervision of both model classification and model explanation generation. Lastly, the use of taxonomy ensures that each split is semantically coherent and anchored in linguistically motivated categories. This improves the interpretability of the resulting data and ensures more precise evaluation of model behavior across different types of reasoning. After splitting the multi-category explanations, the total number of explanations of the annotated LiveNLI dataset increases to 1,575.

Co-occurrence of Explanation Categories and NLI Labels To better understand how different reasoning strategies co-occur with NLI labels, we examine the distribution of annotated explanation categories in the LiveNLI dataset, as shown in Figure 5.2. The most frequently occurring categories are *Inferential Knowledge* and *Semantic*, while *Factual Knowledge* and *Coreference* are the least represented.

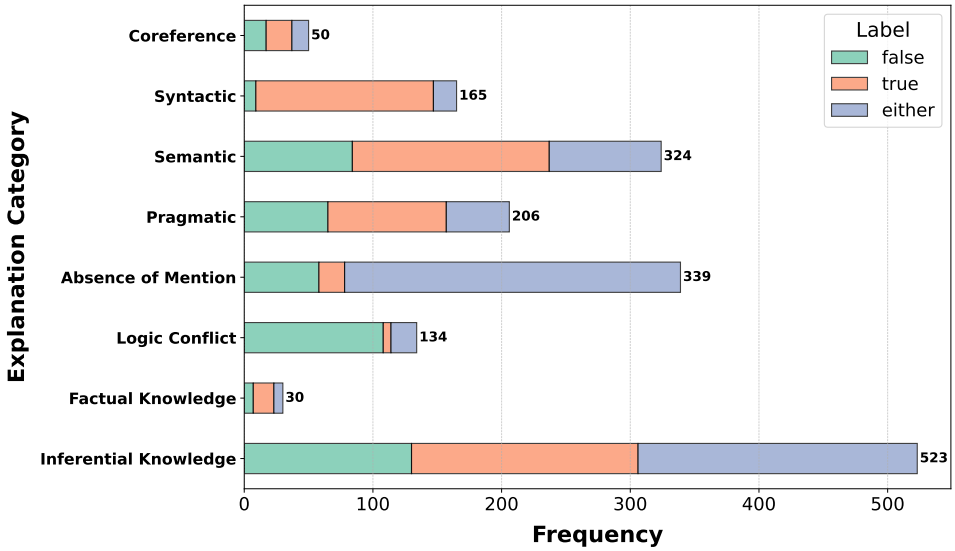

Figure 5.2: Distribution of LITEx categories on LiveNLI explanations across NLI labels (n = 1,575).

When comparing this distribution to that of the e-SNLI dataset illustrated in §3.5, we observe both similarities and notable differences. *Inferential Knowledge* remains the most dominant explanation category across both datasets, underscoring that many NLI explanations require bridging implicit gaps through background or world-informed knowledge. Similarly, *Coreference* appears infrequently in both corpora, suggesting that direct reasoning based on coreference resolution is relatively rare in both NLI datasets.

A key difference, however, lies in the higher proportion of *Semantic* explanations in LiveNLI. One likely reason for the higher proportion is that LiveNLI contains explanations that tend to be more detailed and elaborative. In the annotation guidelines of LiveNLI, annotators are asked to provide new information and refer to specific parts of the premise and the hypothesis, and also to avoid simply repeating the original sentences (Jiang et al., 2023). Therefore, annotators often explicitly describe the semantic relationships between the premise and hypothesis, leading to a higher number of explanations that fall into the *Semantic* category. In contrast, e-SNLI explanations are generally shorter, more direct, and often focus on the key inference without explicitly elaborating on semantic matches.

Another notable difference is the much greater dominance of the *Absence of Mention* category in LiveNLI compared to e-SNLI. This may be a side effect of LiveNLI's guideline emphasis on avoiding repetition and highlighting what is missing or underspecified in the input. As a result, annotators may be more inclined to frame their reasoning in terms of informational gaps, leading to a higher frequency of *Absence of Mention* explanations.

Conversely, the *Logic Conflict* category appears much less frequently in LiveNLI compared to e-SNLI. This may be due to the lower proportion of contradiction-labeled examples in LiveNLI (26.93%) compared to e-SNLI (32.79%). Since explanations in LiveNLI are sampled across all labels but are not uniformly distributed, the smaller number of contradiction instances may limit the model's ability to generalize well to this class. Additionally, annotators in LiveNLI may have preferred using softer reasoning strategies, especially when faced with ambiguous or underspecified sentence pairs, instead of explicitly identifying strict logical conflicts.

In addition to category frequency, the distribution of NLI labels within each explanation category also exhibits interesting patterns. Compared to e-SNLI, LiveNLI shows more balanced label distributions across *entailment (true)*, *neutral (either)*, and *contradiction (false)*, particularly in categories such as *Semantic*, *Pragmatic*, and *Inferential Knowledge*. This suggests that these reasoning types are not tightly coupled with a single label but are used across a wider range of inference scenarios. Notably, *Factual Knowledge* no longer shows a dominant association with the *contradiction (false)* label, unlike in e-SNLI, and *entailment (true)* is no longer the most dominant label in the *Semantic* and *Pragmatic* categories. These shifts indicate that LiveNLI supports more diverse reasoning–label associations, which may be due to its multi-annotator, multi-label structure and open-ended explanation prompts.

Finally, the co-occurrence patterns between explanation categories and NLI labels generally align with the taxonomy's intended semantics. For instance, *Logic Conflict* continues to strongly associate with the *false* label, while Absence of Mention predominantly co-occurs with the either label. These consistent patterns across datasets suggest that the LITEx taxonomy captures label-sensitive reasoning in a robust and interpretable manner, and that its structure remains applicable across different NLI formulations.

5.3 Annotation Results on VariErr

Co-occurrence of Explanation Categories and NLI Labels The stacked bar chart in Figure 5.3 presents the distribution of explanation categories in the VariErr dataset, annotated with the LITEx taxonomy, across the three NLI labels: *entailment*, *contradiction*, and *neutral*. The dataset contains 1,799 instances, each associated with a single explanation category.

The overall distribution of the LITEx explanation categories is notably skewed, with a few categories dominating. Most prominently, *Absence of Mention* is by far the most frequent category (n = 640), accounting for 35.6% of all explanations. This reflects a substantial number of *neutral* labels grounded in missing or unsupported information stated in the hypothesis. In contrast, categories such as *Pragmatic*, *Logic Conflict*, and *Inferential Knowledge* are comparatively underrepresented in VariErr. Specifically, *Pragmatic*

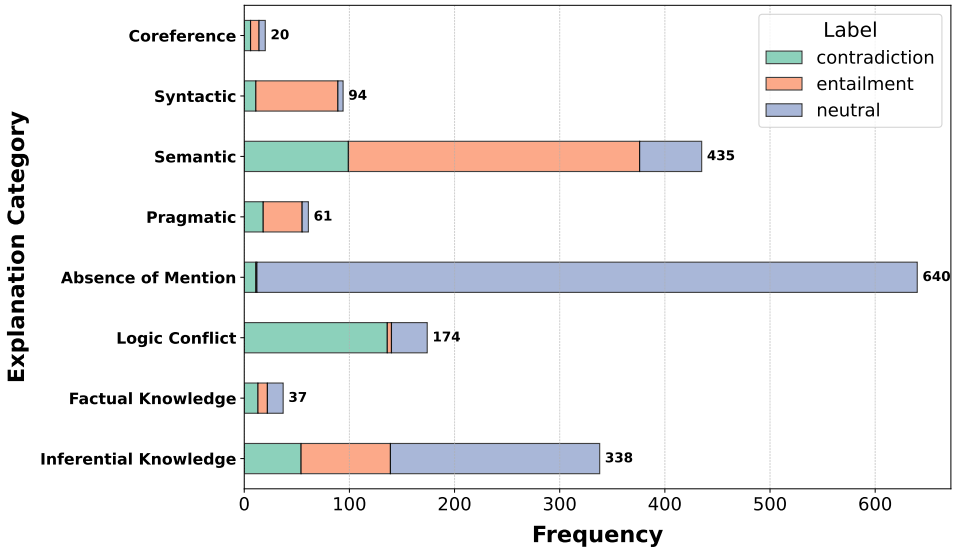

Figure 5.3: Distribution of LITEx categories on VariErr NLI explanations across NLI labels (n = 1,799).

explanations are rare (n = 61), despite their prevalence in datasets like LiveNLI, indicating a limited use of contextual or implicature-based reasoning in VariErr. *Logic Conflict* (n = 174) and *Inferential Knowledge* (n = 338) are also lower than expected given their central role in contradiction and entailment in other datasets. These observations suggest that VariErr explanations are more focused on surface-level mismatch (e.g., lexical absence) than on deeper pragmatic or inferential reasoning.

Semantic (n = 435) is the second most common category, covering 24.2% of the data, pointing to a strong reliance on lexical-level reasoning. The *Inferential Knowledge* category, while less dominant than in LiveNLI or e-SNLI, still makes up 18.2% of all explanations in VariErr, suggesting that real-world knowledge continues to play a meaningful—albeit reduced—role in NLI decisions.

In contrast, the *Coreference* (n = 20) and *Factual Knowledge* (n = 37) categories are clearly underrepresented. This is not unique to VariErr, similar patterns are also observed in our previous analysis of LiveNLI and e-SNLI, where these two categories appear much less frequently compared to others. One likely reason is that explanations relying purely on coreference (e.g., identifying entity links or pronoun resolution) or on stating factual background knowledge without additional reasoning are less commonly written down by annotators. Such reasoning may feel too "obvious" or "implicit" to be explicitly verbalized, or they may co-occur with more complex types (e.g., inferential or pragmatic), and thus get absorbed into broader categories during annotation. As a result, even though coreference and factual knowledge may support many NLI decisions, they are rarely identified as the main reasoning type.

Finally, the relationship between explanation categories and NLI labels in VariErr reveals several meaningful trends. The *Absence of Mention* category is almost exclusively linked to *neutral* labels, which aligns with its definition—these cases typically involve hypotheses that introduce information not explicitly entailed or contradicted by the premise. *Logic Conflict* is strongly associated with *contradiction*, reflecting how logical incompatibilities often signal incorrect hypotheses. These two patterns are consistent with observations from LiveNLI and e-SNLI. *Semantic* explanations are broadly distributed across *entailment*, *neutral*, and *contradiction*, but most frequent in entailment, indicating a preference for lexical reasoning to justify positive inferences. *Inferential Knowledge*, while more balanced across labels, shows a slight leaning toward *neutral*, indicating that incomplete or implicit reasoning may result in under-specification. Meanwhile, rare categories such as *Coreference*, *Factual Knowledge*, and *Pragmatic* do not show a clear pattern across

NLI labels, likely due to their low frequency, which limits strong conclusions about their alignment.

5.4 Explanation Similarity across Categories and Labels

To investigate whether our taxonomy can help identify redundant explanations and structure the space of explanatory variation, we analyze explanation similarity within fine-grained groups defined by both explanation category and gold NLI label. This approach is motivated by the fact that datasets like LiveNLI and VariErr exhibit both within-label and across-label variation: multiple explanations are provided for the same (premise, hypothesis) pair, potentially with differing gold labels.

5.4.1 Explanation Similarity across Reasoning Categories and Labels

Concretely, we group explanations by their assigned explanation category (from our taxonomy) and gold NLI label (*true*, *false*, *either* for LiveNLI; *entailment*, *contradiction*, *neutral* for VariErr), and compute pairwise similarity metrics within each group. If explanations grouped under the same category and label exhibit high semantic and structural similarity, this indicates that they may be redundant in content, and thus that the taxonomy is effective at partitioning the explanation space into semantically coherent clusters.

We compute a set of similarity metrics, selected from those introduced in §3.5.2, to capture both lexical and semantic overlap among explanations: cosine and Euclidean similarity of sentence embeddings (using a pretrained sentence encoder), unigram and bigram n-gram overlaps, and part-of-speech (POS) based n-gram overlaps. Groups with higher average similarity may indicate regions where explanation redundancy is higher and fewer examples may suffice, whereas low similarity suggests diverse or complementary reasoning patterns within the same label-category group.

Grouped Explanation Similarity Analysis in LiveNLI To better understand the distributional characteristics of NLI labels in the LiveNLI dataset, we first visualize the aggregated label probabilities for each item.

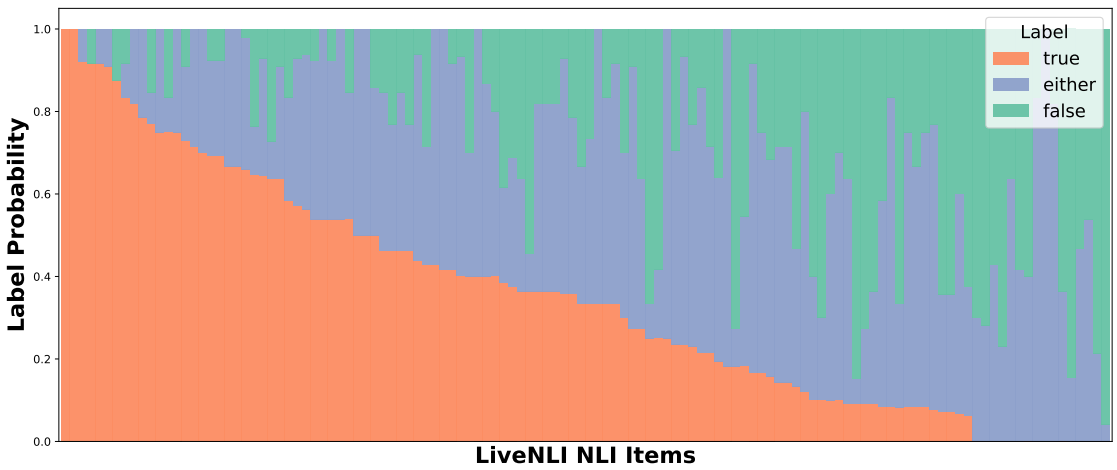

Figure 5.4: Normalized label distributions per NLI item in the LiveNLI dataset. Items are sorted by the proportion of *true* labels.

Figure 5.4 shows a stacked bar chart of normalized gold label distributions across all annotated items, capturing the proportion of responses labeled as *true*, *either*, or *false*. Items are sorted by the proportion of *true* labels to reveal general patterns in annotator agreement. As the figure illustrates, LiveNLI exhibits a diverse label distribution, with many

items showing ambiguity or disagreement among annotators, especially in cases where *either* dominates or co-occurs substantially with the other two labels. This motivates our subsequent analysis of how explanation similarity varies across explanation categories and NLI labels.

Category	Label	#Exp	Cosine	Euclid	2-gram	POS_2-gram
Coreference	true	20	0.484	0.501	0.120	0.433
	either	13	N/A	N/A	N/A	N/A
	false	17	0.535	0.514	0.128	0.433
Syntactic	true	138	0.539	0.522	0.114	0.431
	either	18	0.350	0.467	0.027	0.257
	false	9	0.626	0.536	0.154	0.385
Semantic	true	154	0.550	0.523	0.095	0.391
	either	87	0.520	0.513	0.080	0.408
	false	84	0.629	0.552	0.113	0.460
Pragmatic	true	93	0.537	0.518	0.104	0.463
	either	49	0.467	0.497	0.074	0.460
	false	65	0.619	0.541	0.106	0.445
Absence of Mention	true	21	0.647	0.545	0.159	0.476
	either	262	0.576	0.532	0.106	0.437
	false	58	0.561	0.532	0.103	0.414
Logic Conflict	true	6	N/A	N/A	N/A	N/A
	either	20	0.627	0.540	0.106	0.413
	false	108	0.605	0.542	0.130	0.466
Factual Knowledge	true	16	0.571	0.519	0.075	0.392
	either	7	0.620	0.534	N/A	0.162
	false	7	0.856	0.651	0.743	0.886
Inferential Knowledge	true	177	0.555	0.402	0.100	0.446
	either	218	0.529	0.519	0.090	0.454
	false	130	0.595	0.534	0.080	0.448

Table 5.1: Explanation similarity scores across categories and labels on the *LiveNLI* dataset annotated using LITEx. Green: top-3 values in each metric column; Red: bottom-3 values.

Table 5.1 reports the average explanation similarity scores across different combinations of explanation categories and NLI labels for the LiveNLI dataset. Specifically, for each *(category - label)* group, we first aggregate explanations by `pairID`, then compute all pairwise similarity scores between explanations under the same `pairID`. We then take the average similarity score per `pairID`, and finally compute the mean across all the pairs within the annotated LiveNLI dataset. This two-stage averaging ensures that each instance contributes equally, while capturing variation both within and across premise-hypothesis pairs. For interpretability, we highlight the top-2 values in green and the bottom-3 in red for each column. Notably, some *(category - label)* combination in Table 5.1 have their similarity scores marked as N/A - for example, *(Coreference - either)*, *(Logic Conflict - neutral)*, and *(Factual Knowledge - either)*. This occurs when no `pairID` within that group has more than one explanation. Since pairwise similarity requires at least two explanations per `pairID`, these groups do not satisfy the minimum condition for comparison and are therefore excluded from the results.

In Table 5.1, several *(category - label)* combinations exhibit high internal similarities, while others show notable variation with lower similarity scores. We can observe that the *(Factual Knowledge - false)* group yields the highest similarity scores across all four

metrics. A plausible explanation for this pattern is that for each premise-hypothesis pair, the factual knowledge or common fact required to support a contradiction decision tends to be consistent across annotators. In such cases, annotators tend to rely on similar implicit knowledge not explicitly stated in the text, which results in high semantic and structural consistency across their explanations.

Similarly, the *(Absence of Mention - true/false)* groups also achieve comparably high similarity scores. This suggests that annotators were able to identify and articulate the same missing information when explaining their reasoning, even if they may have differed in the NLI label they assigned. Notably, explanations in the *Absence of Mention* category often follow recurring phrasal patterns, such as "The statement is vague and does not clearly describe the situation", "The hypothesis is not supported by the premise" or "X in the hypothesis is not mentioned in the premise", which contributes to both lexical and syntactic overlap.

In contrast, several low-scoring combinations, especially those under the *either* label, highlight substantial divergence. In particular, the *(Syntactic - either)*, *(Semantic - either)*, and *(Pragmatic - either)* groups exhibit the lower explanation similarity across metrics. These categories inherently support more diverse reasoning styles, and annotators may focus on different surface forms, paraphrases, or wold alignments even for the same premise-hypothesis pair. This variation suggests that explanation generation in these contexts is more subjective and potentially more difficult to standardize.

Overall, the results on LiveNLI show that explanations within the same *(category - label)* group tend to be similar, especially when the underlying reasoning is anchored in clear factual knowledge or missing-information cues. Conversely, when the reasoning is more open-ended or ambiguous (as in the *either* cases), explanations diverge more significantly in both content and form. To examine whether explanation similarities within the same LITEX category and NLI label generalize beyond LiveNLI, we also computed the same similarity metrics on our VeriErr annotations.

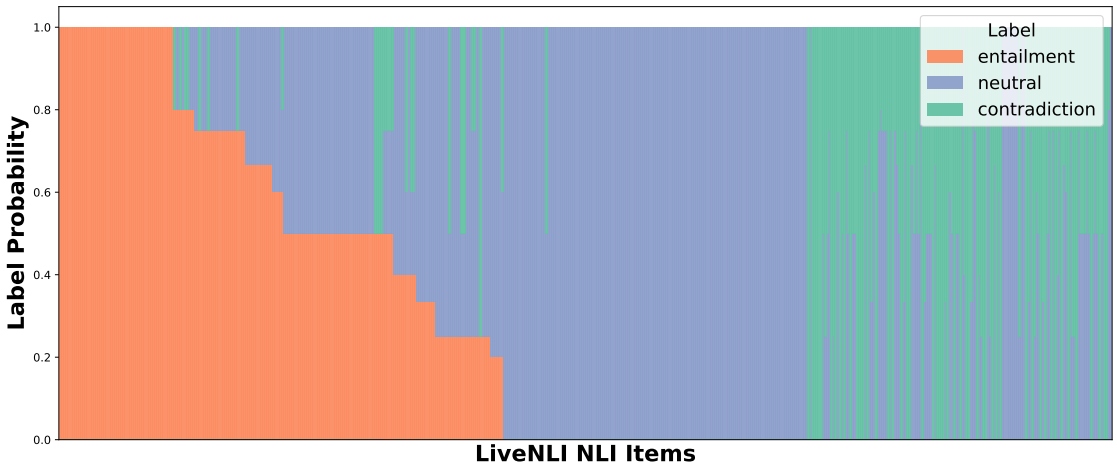

Figure 5.5: Normalized label distributions per NLI item in the VariErr dataset. Items are sorted by the proportion of *true* labels.

Grouped Explanation Similarity Analysis in VariErr Before analyzing explanation similarity on the VariErr dataset, we also first visualize the gold label distributions across annotated items to better understand the degree of label agreement. Figure 5.5 presents a stacked bar chart showing the normalized probability of each label—*entailment*, *neutral*, and *contradiction*—per item, sorted by the proportion of *entailment* annotations. Compared to LiveNLI, VariErr exhibits more concentrated distributions, with fewer examples displaying high levels of label ambiguity. Nonetheless, certain items still show substantial

variation in label assignment, indicating the presence of difficult or underspecified NLI cases. This visualization helps contextualize the similarity scores reported below, offering insight into how label ambiguity may interact with explanation variation.

Category	Label	#Exp	Cosine	Euclid	2-gram	POS_2-gram
Coreference	entailment	8	N/A	N/A	N/A	N/A
	neutral	6	0.666	0.550	0.194	0.387
	contradiction	6	0.509	0.502	N/A	0.085
Syntactic	entailment	78	0.417	0.492	0.082	0.175
	neutral	5	N/A	N/A	N/A	N/A
	contradiction	11	0.633	0.563	0.409	0.553
Semantic	entailment	277	0.608	0.544	0.157	0.377
	neutral	59	0.472	0.505	0.073	0.317
	contradiction	99	0.578	0.533	0.131	0.353
Pragmatic	entailment	37	0.645	0.553	0.163	0.479
	neutral	6	0.444	0.490	0.021	0.360
	contradiction	18	0.698	0.564	0.168	0.401
Absence of Mention	entailment	1	N/A	N/A	N/A	N/A
	neutral	628	0.531	0.524	0.120	0.377
	contradiction	11	0.637	0.540	0.160	0.240
Logic Conflict	entailment	4	0.529	0.508	N/A	0.154
	neutral	34	0.272	0.455	N/A	0.093
	contradiction	136	0.582	0.534	0.168	0.371
Factual Knowledge	entailment	9	0.592	0.525	0.095	0.238
	neutral	15	0.606	0.541	0.106	0.385
	contradiction	13	0.566	0.530	N/A	0.152
Inferential Knowledge	entailment	85	0.692	0.576	0.169	0.443
	neutral	199	0.607	0.538	0.107	0.425
	contradiction	54	0.684	0.566	0.130	0.381
Macro-Avg	—	—	0.582	0.530	0.141	0.362

Table 5.2: Explanation similarity scores across categories and labels on the *VariErr* dataset annotated using LITEX. Green: top-3 values in each metric column; Red: bottom-3 values.

Table 5.2 reports the corresponding explanation similarity scores on the VariErr dataset, using the same pairwise and per `pairID` averaging protocol as applied to LiveNLI. In the case of VariErr, the gold labels are standardized as *entailment*, *neutral*, and *contradiction*, following the NLI convention. Similar to before, the top-3 values in each metric column are highlighted in green, and the bottom-3 in red.

A closer comparison between LiveNLI and VariErr reveals how explanation similarity patterns vary not only by LITEX category and label, but also by dataset characteristics. One notable difference emerges in the *Inferential Knowledge* category: while similarity scores for this category remain moderate in LiveNLI, they rank among the top three across all metrics in VariErr. This suggests that, for NLI tasks, explanations invoking world knowledge may follow similar reasoning paths - especially when annotators share a common training background. Unlike LiveNLI, which is annotated by 48 native English speakers recruited via Surge AI (`surgehq.ai`) (Jiang et al., 2023), the VariErr explanations are produced by Master's students in Computational Linguistics (along with the first author of Weber-Genzel et al. (2024)), likely leading to more homogeneity in reasoning style and linguistic expression (Weber-Genzel et al., 2024). Another pattern observed in VariErr is the low similarity observed for the *(Logic Conflict - neutral)* combination, with the cosine similarity score dropping to 0.272. This suggests that in ambiguous cases, annotators explain logical conflict in varied ways, making this category especially diverse.

Despite differences in annotator pool and task formulation, both datasets reveal consistently low similarity in the *(Semantic - neutral)* and *(Pragmatic - neutral)* groups. These combinations appear to elicit a wider range of explanatory strategies, likely due to the open-ended nature of semantic reinterpretation or pragmatic inference, which permits multiple plausible justifications even for the same label. This reinforces our earlier findings on the inherently diverse reasoning styles within these categories.

In contrast, the *Coreference* category remains highly consistent across these two datasets, yielding high similarity scores across multiple labels. This further supports the notion that referential reasoning is often expressed through predictable linguistic structures (e.g., "X is referred to in the statement"), resulting in high surface-form redundancy. These findings highlight the robustness of our taxonomy across datasets and annotator populations, while also offering insight into which categories may benefit from targeted filtering or aggregation strategies in explanation modeling.

5.4.2 Intra-Item Explanation Similarity across Reasoning Types

To further examine the variation in explanation similarity within specific *(category - label)* combinations, we analyze the distribution of pairwise explanation similarity scores across different `pairID` (premise - hypothesis) groups.

In Figure 5.6 and Figure 5.7, we present grouped boxplots of similarity metrics (cosine similarity, euclidean similarity, bigram overlap, and POS bigram overlap) for four representative combinations in LiveNLI and VariErr. We highlight the two combinations with the highest within-group similarity (in green) and the two with the lowest similarity (in orange).

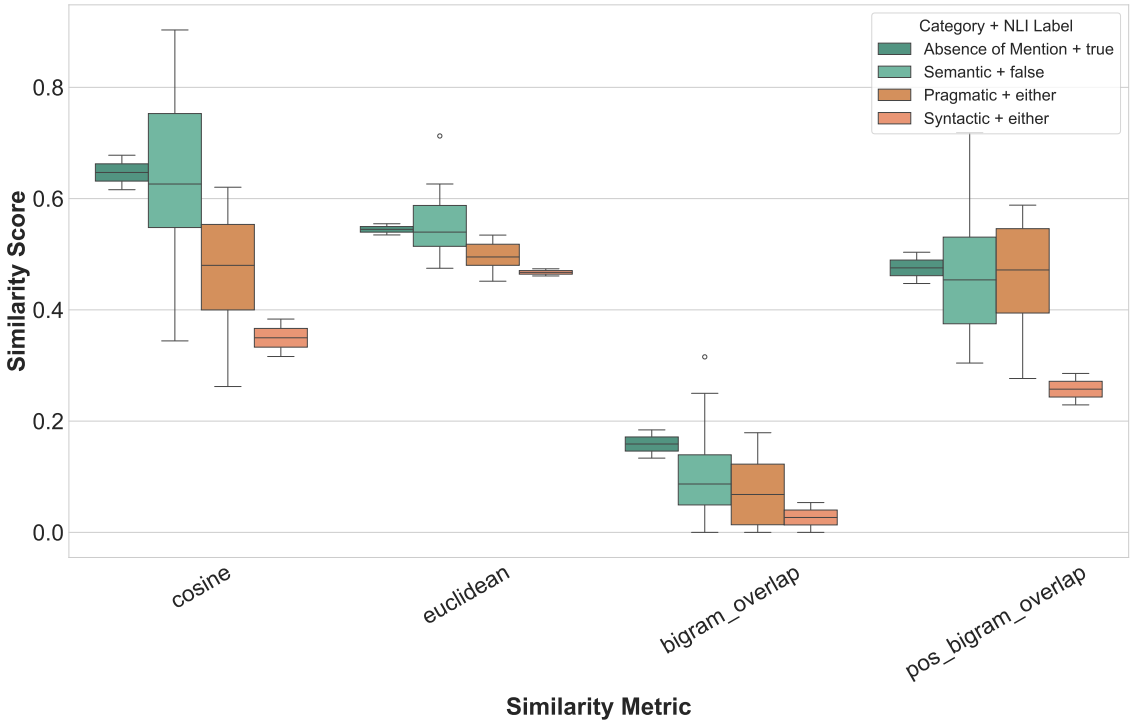

Figure 5.6: Grouped boxplot of explanation similarity metrics on the LiveNLI dataset. The green bars correspond to the highest within-group similarity, while the orange bars represent the lowest-similarity combinations.

In Figure 5.6, we observe that in LiveNLI the *(Absence of Mention - true)* combination exhibits the highlight within-group similarity across multiple metrics, especially in terms of bigram overlap. This suggests that when annotators agree that a hypothesis is likely but

not explicitly mentioned in the premise, they tend to rely on similar reasoning rationale. Typically, the annotators refer to the lack of relevant information or the context not being definitive enough in the provided free-text explanation. The explanations converge in both content and linguistic form, leading to high similarity scores. On the other hand, the *(Semantic - false)* and *(Syntactic - either)* combinations show the lowest similarity. These categories are more ambiguous by nature. For *(Syntactic - either)*, annotators may interpret minor syntactic shifts differently or focus on syntactic paraphrasing, resulting in varied explanations. For *(Semantic - false)*, semantic mismatch can be interpreted in numerous ways. As such, explanations could diverge both in content and structure.

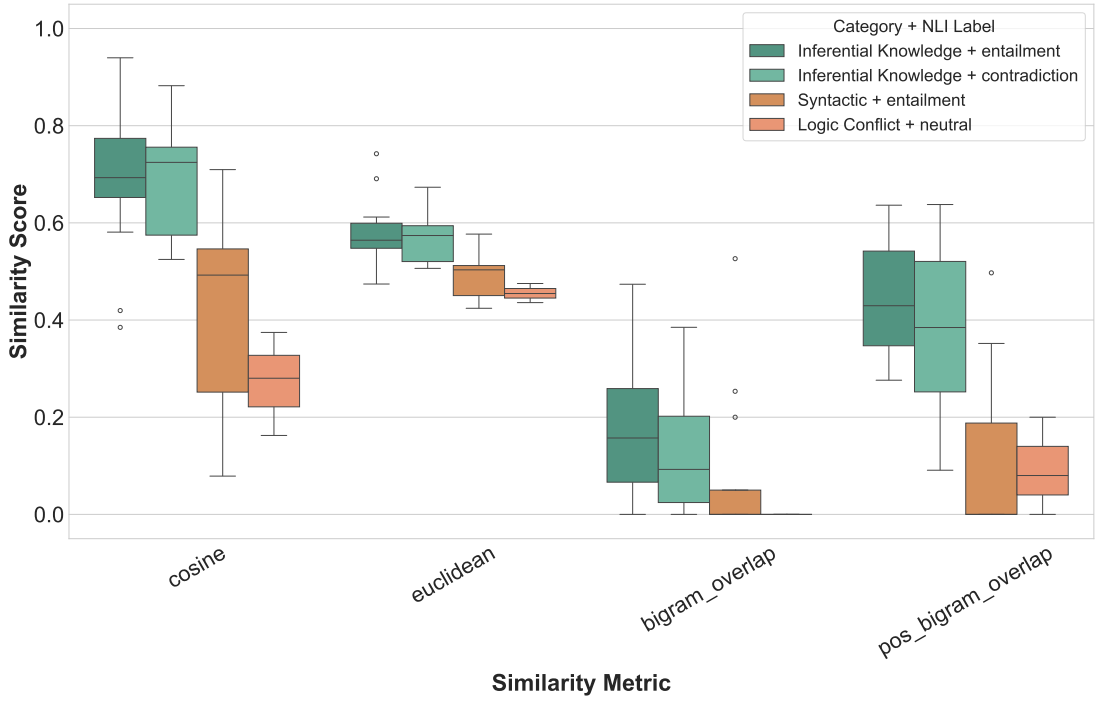

Figure 5.7: Grouped boxplot of explanation similarity metrics on the VariErr dataset. The green bars correspond to the highest within-group similarity, while the orange bars represent the lowest-similarity combinations.

In Figure 5.7, the *(Inferential Knowledge - entailment)* combination shows the highest similarity in VariErr. This suggests that for inferentially entailed hypotheses using external world knowledge, annotators tend to share a common inferential pathway - for instance, recognizing paraphrased sentiment or implicit causality - which leads them to provide consistent and lexically aligned explanations. Conversely, the *(Logic Conflict - neutral)* and *(Syntactic - entailment)* score the lowest. In *(Logic Conflict - neutral)*, the low similarity suggests that annotators may detect different types of logical inconsistencies or ambiguities, or disagree on whether there is conflict at all. For *(Syntactic - entailment)*, the low similarity likely stems from surface-level rephrasing or syntactic alternations that require different types of interpretation.

5.4.3 Case Study: Interpreting Explanation Similarity Extremes

To better understand what contributes to high or low explanation similarity within a given *(category - label)* group, we qualitatively examine the top and bottom-scoring example pairs selected from the LiveNLI and VariErr datasets (see Table 5.3).

Category	Label	Explanations
		Dataset: LiveNLI
Absence of Mention	true	**Highest Similarity** **Ex_1**: The context is a question about whether or not Indian food is kosher. The statement says Indian food are not Kosher. This could be true or false but the context does not provide any more information for us to assess. **Ex_2**: The context asks a question, and the statement answers it. The statement can be either likely true or likely false depending on whether the answer is factually correct or not. **Lowest Similarity** **Ex_1**: I do not see the connection between the party-line inclusive party and these two sentences, or abortion, to give a solid answer. **Ex_2**: Republican politician did invoke the party-line on abortions but it's not clear if it's because their constituents insist on it. **Ex_3**: The context statement is ambiguous – I can't tell what "the party-line inclusive party" means. But the statement that politicians rely on party-line seems to be true based on the context, but what I can't determine is that constituents insist on it. There is no evidence from context that constituents have an opinion, so the statement could be true or false.
Syntactic	either	**Highest Similarity** **Ex_1**: As the statement merely rephrases the context into a question with entirely different meaning, you cannot infer anything from it. **Ex_2**: Both questions are very similar, while one is asking in the past tense and the other is in the present tense so it could be true or false. **Lowest Similarity** **Ex_1**: Both are asking the same question of whether or not you want to see historic sights and tour museums and art galleries. **Ex_2**: These two are questions that technically ask the same thing. 'Would you' and 'would you not' are in essence the same but they could be understood in different ways.
		Dataset: VariErr
Inferential Knowledge	entailment	**Highest Similarity** **Ex_1**: Making violence beautiful is a positive way of looking at violence. **Ex_2**: People made violence beautiful is a way to look at it positively. **Lowest Similarity** **Ex_1**: Publishing a Notice is a way HCFA tried to keep everyone informed. **Ex_2**: They published the rules. So theoretically everyone can see it.
Logic Conflict	neutral	**Highest Similarity** **Ex_1**: The statement is irrelevant to what is discussed in the context. I think the topics are different. **Ex_2**: The statements seem to be completely unrelated, with conflicts in information. **Ex_3**: Not relevant information, different topics. **Lowest Similarity** **Ex_1**: The context talks only about a question on Monday, not Tuesday. **Ex_2**: Irrelevant information provided. **Ex_2**: It was created to hide his true intention to manipulate.

Table 5.3: Examples of explanation pairs corresponding to the highest and lowest within-group similarity in four *(Category, Label)* combinations from the LiveNLI and VariErr datasets. For each group, we show the pair with the highest and the lowest similarity based on multiple metrics. These examples help qualitatively verify the aggregate trends observed in Figure 5.6 and Figure 5.7.

For LiveNLI, the combination *(Absence of Mention, true)* achieves the highest within-group similarity score. The two most similar explanations in this group both emphasize that the hypothesis introduces new information (e.g., whether Indian food is kosher) that cannot be verified from the given premise. Despite minor differences in wording, they follow a consistent reasoning pattern: the absence of relevant information in the premise leads to a neutral or indeterminate judgment. Lexical overlap is also notably high—terms like "context", "question", and statement appear in both explanations, contributing to higher bigram overlap scores.

In contrast, explanations with low similarity in this group diverge in how they interpret what is ambiguous or missing. Some focus on the lack of connection to political context (e.g., abortion), while others highlight uncertainty around specific phrases or terms. This suggests that although the overall reasoning strategy — pointing out ambiguity or missing information — is shared, annotators may differ in what exactly they perceive as absent, leading to greater variation in explanation content.

The combination *(Syntactic, either)* exhibits the lowest within-group similarity scores, as shown in Figure 5.6. In this group, explanations associated with the same item (pairID) vary notably in both phrasing and lexical choice. The two explanations with the highest similarity consistently recognize that the hypothesis is essentially a syntactic reformulation of the premise, differing only in tense or surface form. Despite minor wording differences, they converge on a shared interpretation: that syntactic variation does not imply a semantic shift. This consistency reflects aligned annotator reasoning.

In contrast, the lowest similarity pair diverges more in both expression and focus. One explanation views the questions as nearly identical, while the other emphasizes possible variation in meaning, suggesting interpretive ambiguity. This explanation also makes direct reference to specific token-level distinctions (e.g., "would you" vs. "would you not"), which is absent in the other. The discrepancy indicates that even when explanations fall under the same reasoning category, annotators may differ in how explicitly they tie their judgments to surface-level features.

For the VariErr dataset, the combination *(Inferential Knowledge - entailment)* yields the highest within-group similarity scores. The two explanations with the top-scoring pair both have a shared reasoning pattern: the hypothesis presents a positive framing of violence, and this interpretation aligns with how the premise is worded. Despite slight lexical variation, e.g., one uses a "positive way of looking at violence", the other "a way to look at it positively", both explanations clearly identify and justify the inference as one of alignment in sentiment or framing. Both explanations are concise, focused, and exhibit minimal surface-level variation, contributing to their high similarity.

In contrast, while the lowest similarity explanations in this combination point to actions by an institution (e.g., "HCFA"), their phrasing diverges significantly. One explanation introduces additional context (e.g., "purpose of the notice"), while the other states a factual outcome without elaboration. This results in lower lexical and semantic alignment, despite falling under the same reasoning category.

The combination *(Logic Conflict - neutral)* ranks among the lowest in similarity scores across the VariErr dataset. In the lowest similarity examples, one explanation focuses on a temporal mismatch (e.g., "Monday, not Tuesday"), while the other two explanations provide vague comments (e.g., "irrelevant information" or "intent to manipulate"). These explanations diverge both in surface form and in how the logical conflict is framed, ranging from specific factual contradiction to high-level explanations, leading to low semantic and token-level similarity.

Conversely, the highest similarity explanations share a consistent theme: they all point out that the hypothesis is irrelevant or mismatched with the premise. Terms such as "statement", "different", "conflict", and "irrelevant" recur across the three explanations, reinforcing a shared interpretation of logical disconnection. This convergence in both word choice and interpretive focus contributes to high overlap scores across metrics.

This case study confirms that the similarity scores effectively reflect the degree of semantic overlap among explanations: examples labeled as having high similarity indeed exhibit consistent reasoning and phrasing upon closer inspection, while low-scoring examples often diverge in both content and surface form. Within a given *(category, label)* combination, high-similarity explanation sets tend to focus on a shared reasoning, even if individual explanations vary slightly in wording. In contrast, low-similarity groups often differ in how annotators frame their reasoning or interpret ambiguous elements in the input.

These observations suggest that combining the NLI label and explanation category can serve as a useful signal for identifying explanation clusters with redundancy under the same premise – hypothesis pair. That is, explanations with the same *(category, label)* are more likely to share underlying reasoning patterns. However, further analysis is needed to formalize this connection and determine practical thresholds for defining and controlling redundancy in explanation datasets.

5.5 Interim Summary for Chapter 5

In this chapter, we extend our proposed taxonomy of explanations (LITEX) beyond the e-SNLI dataset to evaluate its generalizability and applicability in structurally and methodologically diverse NLI settings. We apply the taxonomy to two complementary datasets, LiveNLI and VariErr, which differ from e-SNLI in their annotation design, label distributions, and explanation styles.

In §5.1, we introduce and compare the two datasets in relation to e-SNLI, highlighting their unique characteristics and how they complement our analysis.

In §5.2 and §5.3, we present the annotation results on LiveNLI and VariErr using the LITEX explanation categories. Specifically, we analyze the co-occurrence patterns between explanation categories and NLI labels, revealing both similarities and differences compared to e-SNLI.

In §5.4, we investigate whether our taxonomy helps identify redundant explanations and structure the space of explanatory variation. To this end, we compute explanation similarity within fine-grained groups defined by both explanation category and gold NLI label. The results show that explanation similarity varies systematically across different *(category – label)* combinations. Furthermore, in §5.4.2 we further examine variation in explanation similarity at the level of individual NLI items. We analyze the distribution of pairwise similarity scores across different `pairID` (premise – hypothesis) groups, visualize representative patterns using boxplots, and include a qualitative case study of the highest- and lowest-scoring explanation pairs from both LiveNLI and VariErr. This analysis confirms that high-scoring examples reflect consistent reasoning and expression, while low-scoring examples exhibit more diverse and divergent explanation strategies.

6 Discussion

6.1 Conclusion

This work introduces LITEX, a linguistically-informed taxonomy designed to capture different reasoning strategies behind NLI explanations, with a particular focus on within-label variation. Through a human annotation process, model-based classification, and explanation generation experiments, we demonstrate the value of this taxonomy in structuring, interpreting, and guiding NLI explanation tasks. The learnability evaluation by fine-tuned models and LLMs shows that models, after fine-tuning or few-shot prompting, can effectively classify explanations into our taxonomy, demonstrating its practicality.

Through large-scale annotation on the e-SNLI dataset, we demonstrate that LITEX effectively exposes within-label variation and aligns with the reasoning types defined in our taxonomy. Validation via IAA and model-based classification further confirms that the taxonomy is both reproducible across annotators and also learnable by models, underscoring its utility as a classification tool for NLI explanation analysis.

Beyond annotation, we showed that taxonomy-guided generation produces explanations that are semantically richer and more human-like than baseline or highlight-based approaches. Human evaluation of generated explanations revealed that with appropriate prompting techniques, models can produce valid and label-aligned rationales at scale. To assess generalizability, we applied LITEX to two additional NLI datasets — LIVENLI and VariErr — where it successfully facilitated explanation categorization, confirming its adaptability to diverse NLI tasks.

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.

6.2 Future Work

While this thesis provides an initial step towards structuring and analyzing explanations through a linguistically grounded taxonomy LITEX, several avenues remain open for future exploration.

Expanding the taxonomy to other tasks and datasets Our current taxonomy is designed and tested primarily on NLI data. A natural extension would be to investigate whether LITEX categories generalize to other explanation-based tasks, such as commonsense QA, fact verification, or multi-hop reasoning. Applying the taxonomy to these settings may reveal new categories or lead to refinement of the current framework.

Fine-tuning models for taxonomy-aware generation While this thesis uses prompting to guide LLMs toward specific explanation types, future work could involve fine-tuning smaller, efficient models to generate explanations aligned with the LITEX taxonomy. This would benefit settings where controllability, interpretability, or explanation diversity is important.

Measuring explanation usefulness and diversity Beyond similarity and lexical overlap, future work could develop metrics for measuring explanation usefulness to end users. This

would involve conducting user studies or building evaluation datasets where explanations are rated for clarity or insightfulness value.

Investigating implicit reasoning strategy The current taxonomy focuses exclusively on explicitly stated reasoning in written free-text explanations. However, annotators may rely on additional implicit reasoning types when arriving at their NLI label decisions, which are often left unstated. Future work could explore ways to uncover and model these hidden steps. For example, through think-aloud protocols (Lewis, 1982) or multi-step annotation formats. Defining a more detailed annotation protocol that captures both explicit and implicit reasoning, possibly in combination with annotator background analysis (e.g., expertise, linguistic training), may provide a richer understanding of the reasoning process behind NLI judgments.

References

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Jeremy Barnes, Lilja Øvrelid, and Erik Velldal. 2019. Sentiment analysis is not solved! assessing and probing sentiment classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 12–23, Florence, Italy. Association for Computational Linguistics.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed question answering: Evaluation and modeling for attributed large language models.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 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.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Beiduo Chen, Siyao Peng, Anna Korhonen, and Barbara Plank. 2024a. A rose by any other name: LLM-generated explanations are good proxies for human explanations to collect label distributions on NLI.

Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. 2024b. "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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

George Chrysostomou and Nikolaos Aletras. 2021. Improving the faithfulness of attention-based explanations with task-specific information for text classification.

Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*, pages 38–45.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2018a. Supervised learning of universal sentence representations from natural language inference data.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. pages 177–190.

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.

Nicholas Deas and Kathleen McKeown. 2024. Summarization of opinionated political documents with varied perspectives.

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, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu

Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. Deepseek-v3 technical report.

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.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. 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.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data.

Walid Hariri. 2024. Sentiment analysis of citations in scientific articles using chatgpt: Identifying potential biases and conflicts of interest.

David Herrera-Poyatos, Carlos Peláez-González, Cristina Zuheros, Andrés Herrera-Poyatos, Virilo Tejedor, Francisco Herrera, and Rosana Montes. 2025. An overview of model uncertainty and variability in llm-based sentiment analysis. challenges, mitigation strategies and the role of explainability.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. 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*.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

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.

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

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.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

Jenny Kunz and Marco Kuhlmann. 2024. Properties and challenges of llm-generated explanations.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Alex Lascarides and Nicholas Asher. 1993. A semantics and pragmatics for the pluperfect. In *Sixth Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, The Netherlands. Association for Computational Linguistics.

Elisa Leonardelli, Alexandra Uma, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, and Massimo Poesio. 2023. Semeval-2023 task 11: Learning with disagreements (lewidi).

C. Lewis. 1982. *Using the "thinking Aloud" Method in Cognitive Interface Design*. Research report. IBM Thomas J. Watson Research Division.

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.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2020. Natural language inference in context – investigating contextual reasoning over long texts.

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.

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.

Srijita Majumdar, Debabrata Datta, Arpan Deyasi, Soumen Mukherjee, Arup Bhattacharjee, and Anal Acharya. 2022. Sarcasm analysis and mood retention using nlp techniques. *International Journal of Information Retrieval Research*, 12:23.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ilia Markov and Walter Daelemans. 2022. The role of context in detecting the target of hate speech. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 37–42, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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

David Mimno and Moontae Lee. 2014. Low-dimensional embeddings for interpretable anchor-based topic inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1319–1328, Doha, Qatar. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

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, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai

Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 technical report.

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.

Ellie Pavlick and Chris Callison-Burch. 2016. Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173, Berlin, Germany. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

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.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan. Association for Computational Linguistics.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New Orleans, Louisiana. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Mobashir Sadat and Cornelia Caragea. 2022. Scinli: A corpus for natural language inference on scientific text.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. 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.

Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2020. Would you describe a leopard as yellow? evaluating crowd-annotations with justified and informative disagreement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4798–4809, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A roadmap to pluralistic alignment.

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.

Surendrabikram Thapa, Aditya Shah, Farhan Jafri, Usman Naseem, and Imran Razzak. 2022. A multi-modal dataset for hate speech detection on social media: Case-study of russia-Ukraine conflict. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 1–6, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

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.

Maximilian Wich, Christian Widmer, Gerhard Hagerer, and Georg Groh. 2021. Investigating annotator bias in abusive language datasets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1515–1525, Held Online. INCOMA Ltd.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Annie Zaenen, Lauri Karttunen, and Richard Crouch. 2005. Local textual inference: Can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 31–36, Ann Arbor, Michigan. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models.

List of Figures

List of Tables

Submitted Software and Data Files

The materials accompanying this master thesis are made available under `https://github.com/PingjunHong/masterthesis`. This include:

- The original LaTeX source files of this thesis
- The compiled PDF version of the thesis
- `data/` – Annotated datasets used in the experiments
- `annotation_interface/` – Web-based interface for manual annotation
- `iaa/` – Inter-annotator agreement (IAA) results
- `classification/` – Python scripts for explanation classification and corresponding outputs
- `generation/` – LLM-generated explanations and Python scripts for generation
- `human_validation/` – Results of human validation of model-generated explanations
- `similarity_analysis/` – Script and results for analyzing the similarity of generated explanations
- `images/` – Figures and diagrams included in this thesis
- `README.md` – Overview of the repository structure and instructions for reproducing results
- `requirements.txt` – List of required Python packages and dependencies