How Much Is Inferred?
Measuring Inferential Distance In Datasets and Models

Anonymous ACL submission

Abstract

While much prior work involving the textual entailment relation between pairs of texts focuses on recognizing its existence, or generating texts that satisfy it, little attention has been given to measuring the semantic gap between the entailing and entailed texts. In this work, we introduce *inferential distance*, a discrete metric that quantifies the amount of newly inferred content introduced in a hypothesis relative to its premise. Based on quantifying lexically-novel "atomic" information units, our framework also includes an informative typology that classifies inferences into paraphrase, common-sense reasoning, and specialized knowledge. We further propose an LLM-based pipeline for automatically computing inferential distance and classifying inference types, and validate its reliability against human judgments. Through evaluation across Natural Language Inference and abstractive summarization benchmarks, we show that inferential distance serves as a powerful diagnostic tool that captures intrinsic dataset difficulty. Furthermore, it reveals a strong alignment between model performance and inferential complexity in entailment tasks, while in text generation settings models exhibit lower levels of abstraction compared to gold references.

1 Introduction

Textual entailment (Dagan et al., 2013) is the core semantic task of determining whether a source text (premise) logically entails a target sentence (hypothesis). In many Natural Language Processing (NLP) tasks, it is required to model the entailment relations between pairs of texts, either when classifying such relations, as in Natural Language Inference (NLI) and factual verification, or when generating texts that are entailed by a particular source context, as in summarization, question answering, and explanation generation. While most prior work has focused on modeling the sheer existence of the entailment relation between premise–hypothesis pairs, little attention has been given to measuring the amount of inference required to bridge the gap between such entailed pairs.

Premise (P): H-2A visa workers can't stay in the United States for more than one year.
Hypothesis (H): Workers who are on an H-2A visa can't stay in the <u>country</u> for <u>longer</u> than <u>12</u> <u>months</u>.
1. Novel words identified in the hypothesis are underscored.
2. Identifying inference atoms (S_{hi}) and source spans (S_{pi}). (S_{p1}) United States $\rightarrow$ (S_{h1}) the <u>country</u> (S_{p2}) more than $\rightarrow$ (S_{h2}) <u>longer</u> than (S_{p3}) one year $\rightarrow$ (S_{h3}) <u>12</u> <u>months</u> <u>*12*</u> <u>*months*</u> is atomic: it implies the meaning of *one year*. Resulting non-overlapping partition: Workers who are on an H-2A visa can't stay in [the <u>country</u>] for [<u>longer</u> than] [<u>12</u> <u>months</u>].
3. Inferential Distance: $D = 3$ (three inference atoms).

Table 1: Example of computing inferential distance (see Section 3.1 for the full definition and calculation).

In particular, in this paper we focus on quantifying the amount of inferences that are reflected by the presence of novel lexical items in the hypothesis, which are not present in the premise (beyond morphological inflections and derivations), while systematically identifying the underlying typology of those inferences.

For example, consider the premise P and the entailing hypotheses H in Table 1. H includes three units of novel inferred content: *the country*, *longer than* and *12 months*. In contrast, a hypothesis composed entirely of words already appearing in the premise (or their morphological variants) does not contain any newly inferred information.[1]

To quantify this gap, we introduce the metric of **inferential distance**. This metric captures the

[1] For example: *No more than a year can H-2A visa workers stay in the United States*.

number of distinct lexically-novel inferred information units introduced in a hypothesis beyond the premise, thereby directly reflecting the inferential gap between a premise and a hypothesis. Table 1 illustrates this computation, with details provided in Section 3.1. Further, to gain a more fine-grained understanding of the inferential gap, we also identify and classify three inference types, including paraphrase, common-sense reasoning, and specialized knowledge.

Next, we applied LLM-based models to determine inferential distance by identifying each newly inferred information unit and locating their entailing source spans. We validated the automated measure against human annotations on two datasets and found high agreement, confirming its reliability (see Appendix A for more details).

We then propose two important applications for this combined approach: **(1) Dataset Characterization**, where inferential distance and inference-type distributions within a dataset align closely with its inherent difficulty and models' ability to detect entailment, supporting the development of richer, more challenging benchmarks; and **(2) Model Behavior Analysis**, where examining model performance and generation behavior as a function of inferential distance and inference type enables a systematic identification of model strengths and weaknesses.

We demonstrate the utility of inferential distance through two representative applications: (a) Natural Language Inference (NLI) and (b) extreme summarization. For NLI, we analyze the LFQA, AggreFact-LLM-XSum, and SciFact datasets using GPT-3.5, GPT-4o, and GPT-o3-mini. At the dataset level, our analysis illustrates how our metric captures inherent dataset difficulty. At the model level, inferential distance is shown to correlate with confidence and performance: as distance increases, accuracy degrades, with models resolving surface-level paraphrasing more reliably than specialized knowledge.

For extreme summarization, we utilize the Gigaword and XSum datasets. At the dataset level, mapping the inferential relationships between source documents and gold references reveals the abstraction standards intrinsic to each benchmark. We find that gold summaries rely on rich commonsense and specialized-knowledge inferences, while larger inferential gaps consistently correspond to lower ROUGE performance scores. At the model level, language model generations fail to match this standard of abstraction, generating instead more surface-level paraphrases.

In summary, our main contributions are: (1) introducing inferential distance, a novel metric grounded in atomic information units and a typology of inference types; (2) providing and validating an LLM-based pipeline for automated metric computation and inference classification; and (3) demonstrating that our framework captures intrinsic differences in dataset difficulty and model behavior across NLI and summarization tasks.

2 Related Work

To the best of our knowledge, this work is the first to propose a relevant discrete measure for quantifying the amount of entailment inference exhibited in entailing premise–hypothesis pairs within the NLI and generation regimes. In this section, we survey prior works that address similar goals to ours, namely identifying the difficulty level of dataset instances, as well as prior works that address similar tasks to the main component of our approach, specifically the detection of novel information and atomic information.

Dataset and Instance Difficulty. Prior research evaluates NLI dataset and instance difficulty using either surface-level cues or indirect model dynamics. For instance, difficulty has been quantified via lexical or syntactic artifacts in the hypothesis alone (Gururangan et al., 2018) or through token-level information-theoretic measures (Ethayarajh et al., 2022). Other approaches infer difficulty indirectly by analyzing model behavior, such as the skill level required to predict an instance correctly (Lalor et al., 2019), training confidence patterns (Swayamdipta et al., 2020), or validation loss accumulation (Angelov et al., 2025). In contrast, we propose an intrinsic, model-agnostic semantic measure that explicitly quantifies the volume of lexically-novel information required to bridge the inferential gap between a premise and a hypothesis.

Novelty Detection Prior research detects text novelty to reduce redundancy in summarization (Ghosal et al., 2022), mitigate duplicate online content (Saikh et al., 2017; Ai et al., 2025), or evaluate model generalization (McCoy et al., 2021). Methodologically, redundancy-focused frameworks identify entirely unprecedented information using individual document pairs or multi-document setups (Ghosal et al., 2022; Saikh et al.,

2017), or using historical repositories (Ai et al., 2025). Targeting lexical novelty at the macro level, McCoy et al. (2021) flag text missing from a model's entire training corpus. In contrast, we focus strictly on entailing premise–hypothesis pairs. Instead of looking for completely new information or corpus-wide shifts, we evaluate how the hypothesis rephrases or abstracts content already anchored in the source text, using unseen single words —accounting for lemmas and morphological variants— to measure localized inferential complexity.

Atomic Information Units and Fact Decomposition. Prior works have used atomic facts for multiple goals, including factuality evaluation of generated text (Min et al., 2023; Gunjal and Durrett, 2024) and summarization content selection (Nenkova and Passonneau, 2004), yet granularity was often underspecified. Min et al. defined atomic facts as short statements that each contain one piece of information. Conversely, Min et al. (2023); Gunjal and Durrett (2024) introduce "molecular facts" claiming that over-isolating atomic facts strips the vital context needed for accurate verification and error localization. Nenkova and Passonneau loosely defined content units as clauses expressing the same meaning across multiple human summaries. We define atomic units specifically through inference, explicitly linking hypothesis units to source spans. This formalized approach achieves high inter-annotator agreement (Appendix A) and is easily automated. Unlike methods requiring external tools like AMR or QASem (Cattan et al., 2026), our framework is both fine-grained and self-contained, combining structured precision with surface-level simplicity.

3 Inferential Distance

To quantify the information gap between a premise and a hypothesis, we propose the *Inferential Distance* metric. Section 3.1 introduces the metric and formalizes our notion of an *novel inference atom*; Section 3.2 introduces the three inference types used in our analysis; and Section 3.3 describes the LLM-based measure implementation.

3.1 Metric Definition

Given an entailing premise-hypothesis pair, *inferential distance* measures the number of minimal lexically novel information units in the hypothesis, as defined below, each including at least one word that is absent from the premise. We refer to these as *novel inference atoms*. Table 1 provides a running example of this computation. In Step 2, the spans S_{h1}, S_{h2}, and S_{h3} are all distinct novel inference atoms.

An atomic information unit is not a well-defined term, and different previous works use different definitions depending on the task at hand, as pointed out in (Gunjal and Durrett, 2024). In this work, we focus on measuring atomic units in the context of inference, and identify how each novel inference atom was inferred—i.e., what source span S_p in the premise entails the new content. We define a novel inference atom as the minimal span of words in the hypothesis S_h that expresses a single coherent piece of new information to be evaluated for entailment.

To apply this definition systematically, we implement a **decomposition test**. A candidate hypothesis span S_h supported by a premise span S_p is considered atomic if and only if it cannot be broken down into smaller sub-spans (e.g., S_{h1} and S_{h2}) such that each sub-span is still independently entailed by the supporting premise context. For example, as shown in Table 1 (Step 2), the phrase *longer than 12 months* (entailed by *more than one year*) is **not** atomic. It can be successfully decomposed into *longer than* (entailed by *more than*) and *12 months* (entailed by *one year*), both of which independently satisfy entailment. Conversely, *12 months* cannot be further decomposed while remaining anchored to *one year* (or its sub-spans), rendering it a true inference atom. [2]

To apply this definition systematically, we implement a three-step automated procedure aligned with Table 1: **(1) Identify Novel Words:** We isolate hypothesis words (W_i) absent from the premise in surface form, lemma, or morphological derivatives, excluding stop words. **(2) Form Inference Atoms:** Each novel word is assigned to its minimal surrounding hypothesis span (S_{hj}) that constitutes a valid inference atom, and its entailing premise source span (S_{pj}) is identified. These spans form a strict partition over the novel words such that every highlighted word belongs to exactly one atom, though an atom may contain multiple novel words (e.g., 12 months). **(3) Compute Distance:** The **inferential distance** (D) is calculated as the total count of distinct novel inference atoms identified

[2] While theoretically more than a single valid annotation may exist, in practice one sensible annotation emerges, as reflected both in the observed data patterns and in the high inter-annotator agreement reported in Appendix A.

Paraphrase or Synonyms
For more than → Longer than
Political turmoil → Political crisis
A UN official said it is crucial → UN urges
Common Sense Reasoning
Prohibited from remaining in France → Can't stay in France
Heat wave → Hot weather
BDH won approval to buy the group → BDH is to take over the group
Specialized Knowledge
Fatwa → Islamic religious edict
Nets* → New Jersey
The first cloned mammal → Dolly the sheep

*Referring to the NBA franchise in New Jersey.

Table 2: Examples of inference types.

in the hypothesis.

Since the metric quantifies inferential relations between the premise and the hypothesis, it is defined only for premise-hypothesis pairs where the hypothesis is entailed by the premise. We also choose not to normalize by hypothesis length: while longer hypotheses may naturally introduce more new information, this variation reflects genuine differences in overall inference complexity within the NLI instance and is therefore intentionally preserved in the final distance.[3]

3.2 Inference Types

To allow for a deeper analysis, we categorize the type of inference required to derive each inference atom in the hypothesis from the premise. Inference type is a somewhat soft, qualitative label used to characterize the kind of knowledge needed for the inference. While prior work, such as the GLUE benchmark (Wang et al., 2019), includes a broad range of inference categories, we focus on three high-level and quite reliably distinguishable types: paraphrase or synonym, common sense reasoning, and specialized knowledge[4]. The categorization was validated by high inter-annotator agreement, as described in Appendix A. As shown in Sections 4 and 5, this three-way distinction sheds light on systematic differences in reasoning demands, dataset difficulty, model performance, and how model behavior compares to humans.

We define each of the three inference types below, and provide illustrative examples for each category in Table 2.

Paraphrase or Synonym. The information in the hypothesis span conveys the same meaning as the corresponding premise span within the context of this premise–hypothesis pair, only using different surface forms. No reasoning beyond linguistic equivalence is required.

Common Sense Reasoning. [5] The inference relies on general, everyday human experience knowledge that an average adult would intuitively possess. It does not require specialized or domain-specific understanding.

Specialized Knowledge. The inference depends on factual or specialized knowledge that goes beyond daily experience, such as information about geography, science, law, history, or technical domains.

3.3 LLM-based Measure Implementation

To apply the inferential distance metric at scale, we used an automated annotation pipeline based on GPT-o3-mini. For each premise-hypothesis pair, we first used the SpaCy library to identify and highlight novel hypothesis words W_i, as defined in Section 3.1. Inference atoms were then extracted using a dedicated prompt of roughly 1,350 tokens (excluding the premise and hypothesis), which includes detailed task guidelines and three fully annotated examples. The model's output directly yields the inference atoms, whose count determines the inferential distance. The full prompt is provided in Appendix B.

Inference types were annotated separately using GPT-5[6] with a prompt of approximately 1,200 tokens; see Appendix C for the complete prompt.

We validated this pipeline against expert human annotations, establishing a ground truth with a 0.99 average NMI IAA (see Appendix A). The automated model strongly aligns with this baseline, scoring 0.865 NMI on Gigaword and 0.856 on LFQA, justifying its exclusive use in Sections 4 and 5.

[3] Code for computing inferential distance will be made publicly available on GitHub.

[4] corresponding to what is often referred to as *world knowledge*.

[5] Unrelated to a model's reasoning mechanisms, common sense reasoning here refers solely to the semantic inference type.

[6] `gpt-5` is the literal API model string we used.

Metric	LFQA	AggreFact-XSum	SciFact
Average Inferential Distance	1.03	1.78	2.56
Inference type distribution (% of Total Units)			
Paraphrase / Synonym	24.8%	10.5%	13.3%
Common Sense Reasoning	70.7%	55.5%	30.3%
Specialized Knowledge	4.5%	34.1%	56.4%
Model accuracy (Label = Entailed)			
GPT-3.5-turbo	90.2%	80.9%	73.8%
GPT-4o	97.0%	90.9%	72.8%
GPT-o3-mini	96.5%	78.9%	72.0%

Table 3: Results for NLI datasets. Model accuracy decreases as inferential distance increases.

3.4 Inferential Distance Use Cases

Inferential distance serves two complementary purposes: (1) characterizing the inferential complexity of datasets, and (2) analyzing model behavior with respect to inference complexity. We demonstrate these purposes through two illustrative case studies that showcase the analytical utility of the metric: analyzing inferential complexity in NLI datasets and models, and analyzing inferential complexity in summarization datasets and model-generated summaries. Our analyses are intended primarily as demonstrations of inferential distance as a diagnostic tool, rather than as broad conclusions about the datasets or models themselves.

In Section 4, we study the NLI use case by measuring inferential distance over three datasets and analyzing how entailment model performance varies with inferential complexity. In Section 5, we turn to summarization, using inferential distance to characterize dataset properties with respect to gold summaries and to analyze how model-generated summaries differ from gold summaries in their amount and type of inferred information.

4 Natural Language Inference

We next apply inferential distance to NLI dataset as a testbed for analyzing them and for examining how semantic inference complexity shapes model performance. Our goal is not to draw definitive conclusions about these datasets or models, but to illustrate the kinds of patterns that inferential distance can reveal.

4.1 NLI Datasets

We apply our approach to three NLI datasets spanning different domains: LFQA (Chen et al., 2025), containing LLM-generated answers to ELI5 questions (Fan et al., 2019) (we sampled 1,758 pairs); LLM-AggreFact-XSum, a subset of the LLM-AggreFact dataset (Tang et al., 2024a), comprising system-generated news summaries evaluated for factual consistency (766 pairs); and SciFact (Wadden et al., 2020), which contains expert-written atomic claims derived from the biomedical literature (1,392 pairs). Additional dataset details are provided in Appendix E.1.

4.2 NLI Dataset Inferential Distance Analysis

We use inferential distance to characterize the inherent inference complexity exhibited in the NLI datasets. For each entailment pair, we applied our inferential distance metric and classified the corresponding inference types to analyze the amount and nature of inferred information introduced by each dataset (See Section 3.3 for models used). Since the inferential distance metric is defined only for entailed pairs, we analyzed instances labeled as *entailed*.

Table 3 presents the average inferential distance for each dataset and the distribution of inference types. LFQA exhibits the lowest average distance, followed by LLM-AggreFact-XSum, while SciFact shows the highest. This variation indicates that SciFact contains substantially more newly inferred content than LFQA, making it more challenging for entailment recognition models, whereas the latter represents the least demanding setting under our metric.

Examining the inference type distribution reveals that LFQA not only introduces less new information but also tends to rely on more lexical and commonsense-based inferences, with the highest proportion of Paraphrase / Synonym inferences and the lowest proportion of specialized knowledge (4.5%). In contrast, SciFact exhibits the most semantically demanding inferences, with only 13.3% paraphrases and as much as 56.4% specialized-knowledge inferences, reflecting its emphasis on factual reasoning and domain-specific expertise. This disparity suggests that SciFact is inherently more challenging for models that struggle with specialized-knowledge reasoning, whereas LFQA is likely to yield higher performance across models due to its reliance on paraphrastic and everyday inference patterns, as analyzed below.

More broadly, these dataset-level insights demonstrate how inferential distance can be used both to characterize datasets and as a principled

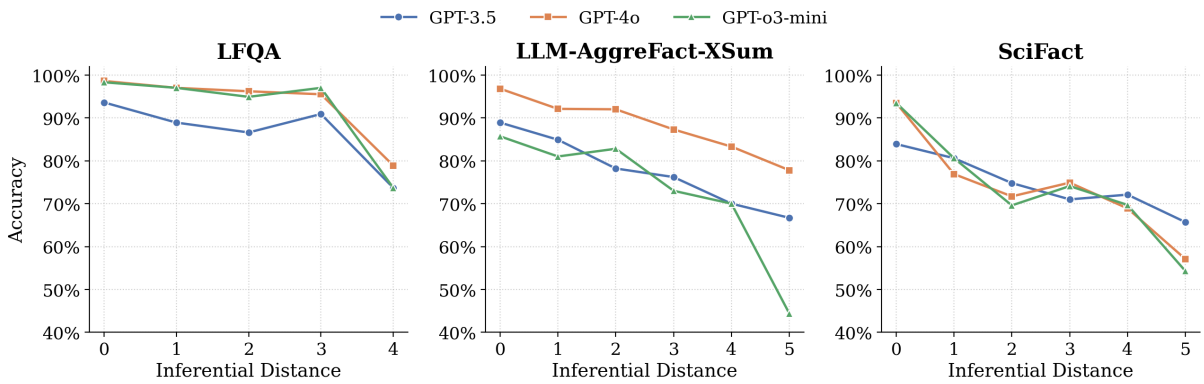

Figure 1: Effect of inferential distance on model accuracy across three NLI datasets. Accuracy consistently decreases as inferential distance increases.

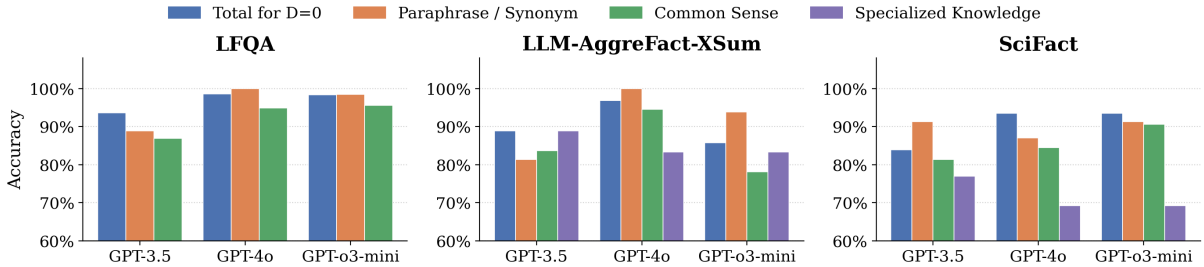

Figure 2: Impact of a single inferred atomic unit ($D = 1$) by inference type on model accuracy, relative to the $D = 0$ baseline. LFQA excludes Specialized Knowledge due to an insufficient number of examples. In LLM-AggreFact-XSum, Paraphrases or Synonym is based on a small subset of the data ($n = 16$).

tool when designing richer and more challenging benchmarks by deliberately controlling the amount and types of inferred information.

4.3 NLI Model Assessment

For each of the NLI datasets, we predicted whether each sentence pair was *entailed* or *not-entailed* using three ChatGPT models: GPT-3.5-turbo, GPT-4o, and the reasoning-enhanced GPT-o3-mini. The results are shown in Table 3.

As shown in the table, inferential distance is strongly aligned with models' ability to detect entailment: higher distances—reflecting greater semantic divergence between the premise and the hypothesis—consistently correspond to lower accuracy across all models. Somewhat unexpectedly, GPT-3.5-turbo outperforms the two larger models for SciFact, a finding we analyze below.

Figure 1 illustrates model accuracy per-instance as a function of inferential distance. For all models, accuracy is highest when the hypothesis contains no additional inferred content ($D = 0$), and declines steadily as inferential distance increases (up to small fluctuations). These results demonstrate that inferential complexity can substantially hinder model reasoning.

In a secondary analysis, we isolated the effect of each inference type by examining sentence pairs with exactly one inferred atomic unit ($D = 1$) relative to those with no newly inferred units ($D = 0$)[7]. Figure 2 shows model accuracy broken down by inference type. We analyze these effects across datasets, showing how this breakdown provides insight into both dataset characteristics and model behavior.

Across datasets, a consistent pattern emerges: paraphrastic inferences are easier to resolve, whereas inferences requiring common-sense reasoning or specialized knowledge confuse the models more often. This trend holds across nearly all models and datasets[8], reinforcing the intuition that lexical reformulation is generally less demanding for LLMs than deeper semantic inference.

In contrast, the effect of specialized-knowledge inferences is strongly dataset-dependent. In AggreFact-LLM-XSum, both GPT-3.5 and o3-mini exhibit minimal performance degradation on specialized-knowledge inferences relative to

[7] In LFQA, we excluded specialized-knowledge cases from this comparison due to their scarcity (only 4.5% of inferred units)

[8] Except for GPT-3.5 on AggreFact-LLM-XSum.

$D = 0$ (0% and 2.4% drops, respectively). We attribute this to the nature of specialized-knowledge cases in XSum, which often involve shallow entity enrichment (e.g., McCartney → Sir Paul McCartney) rather than substantive factual reasoning.

SciFact exhibits a markedly different pattern. Here, performance is highly sensitive to specialized-knowledge inferences: accuracy declines sharply for GPT-4o and GPT-o3-mini, while GPT-3.5 shows only a moderate decrease. This asymmetry helps explain GPT-3.5's superior overall performance on SciFact (Table 3), given that specialized-knowledge inferences constitute a majority of the dataset (56.4%).

Taken together, these findings highlight how inferential distance provides a unified lens for uncovering both dataset-level characteristics and models' varying abilities in tasks requiring inference.

5 Extreme Abstractive Summarization

We next apply inferential distance to abstractive summarization, under the standard assumption that the summary should remain fully entailed by the source document, with the goal of analyzing both dataset characteristics and generative model behavior. As in the NLI setting, our goal is not to draw definitive conclusions about the datasets or models, but to demonstrate the kinds of insights that inferential distance can provide.

5.1 Extreme Summarization Datasets

We first assess the Gigaword and XSum datasets by analyzing human-written gold summaries vs their source text, establishing how inferential distance and inference-type distributions are reflected in these datasets. We then examine how model-generated summaries compare to these gold summaries in terms of inferential complexity.

We evaluate the summarization setting using two popular benchmark datasets: Gigaword (Rush et al., 2015), which consists of article–headline pairs (we sampled 904 pairs), and XSum (Narayan et al., 2018)[9], where each BBC article is paired with its first sentence, used as a single-sentence summary (906 pairs). To ensure faithful summarization, we randomly sampled only entailing article–summary pairs from each dataset; Additional dataset and filtering details are provided in Appendix E.2.

[9]XSum gold summaries are well-documented to contain a high rate of factual discrepancies and out-of-context hallucinations (Maynez et al., 2020)

5.2 Summarization Dataset Assessment

For summarization datasets, inferential distance is measured between the source article and the reference summary, capturing the amount of inferred information required to produce this summary. The top row in Table 4 reports the resulting inferential distance statistics for each dataset. The XSum dataset exhibits nearly twice as much newly inferred information in its summaries ($D = 2.23$) compared to Gigaword ($D = 1.32$), which may be partly explained by the longer average length of XSum summaries (22 words) versus Gigaword's (8 words).

As shown in the Gold Distribution column of Table 5, Gigaword summaries tend to be more lexical in nature, containing 31.7% paraphrases compared to only 10% in XSum, and a smaller share of specialized-knowledge inferences (9.7% versus 27.8%). The proportion of common-sense reasoning inferences is relatively similar across both datasets.

Viewed through the lens of inferential distance, these findings indicate that XSum summaries are inherently richer and more abstractive than those in Gigaword, introducing substantially more inferred information. Consequently, XSum imposes a higher abstractive standard on text generation and is therefore expected to yield lower ROUGE scores by summarization models.

	Inf. Distance		**ROUGE-L** (↑)	
Summary Source	**Gigaword**	**XSum**	**Gigaword**	**XSum**
Gold	1.32	2.23	—	—
GPT-3.5	1.10	1.33	0.386	0.193
GPT-4o-mini	1.21	1.78	0.384	0.196

Table 4: Average inferential distance between source documents and summaries, and ROUGE-L scores for model generations against gold references. Evaluated models systematically generate lower distance than gold summaries, and higher gold distance correlates with lower ROUGE-L scores.

5.3 Summarization Model Assessment

We used GPT-3.5-turbo and GPT-4o-mini to generate summaries for the articles in the Gigaword and XSum datasets.

All models were prompted with a standardized instruction to produce concise, single-sentence summaries of comparable length to the gold references (22 words for XSum and 8 words for

Inference Type	Gold %	%Δ GPT-3.5	%Δ GPT-4o-mini
Gigaword			
Para. / Syn.	31.7%	+28.1%	+22.0%
Common Sense	58.7%	–31.6%	–15.6%
Spec. Knowl.	9.7%	–72.2%	–65.2%
XSum			
Para. / Syn.	10.0%	+66.3%	+115.6%
Common Sense	62.2%	–34.8%	–10.2%
Spec. Knowl.	27.8%	–87.9%	–87.2%

Table 5: Relative change (%Δ) in inference type compared to gold. Positive/negative values indicate increase/decrease.

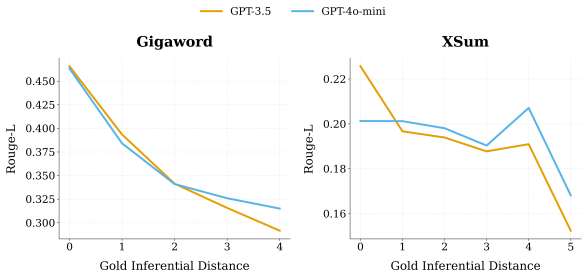

Figure 3: Rouge-L scores between model-generated and gold summaries as a function of the gold summary's inferential distance, for Gigaword (left) and XSum (right). Larger inferential distance corresponds to lower scores, leading to lower alignment with the reference.

Gigaword) (The full prompt is provided in Appendix F.1).

We apply our inferential distance metric to summaries generated by GPT-3.5-turbo and GPT-4o-mini, and report the resulting inferential distance and inference-type analyses in Tables 4 and 5, respectively. In both datasets, the gold summaries contain noticeably more inferred information than those generated by GPT-3.5-turbo and GPT-4o-mini. For Gigaword, gold summaries show a slightly higher inferential distance (1.32) than GPT-3.5-turbo (1.10) and GPT-4o-mini (1.21). In XSum, the difference is larger, with human summaries averaging 2.23 versus 1.33 and 1.78 for the two models, respectively. Our metric reveals that, while both models tend to produce summaries containing less inferred information than the gold summaries, GPT-4o-mini exhibits inferential distances closer to the gold summaries.

To further probe this relationship, we analyze how the inferential distance between the source context and the gold summary is related to model-generated summaries. To this end, we compute ROUGE-L similarity between the model outputs and the gold summaries. As shown in Figure 3, gold summaries with higher inferential distance consistently correspond to lower ROUGE-L model scores, indicating that as the semantic gap between the source and gold summary widens, model-generated summaries are less likely to exhibit the same lexical transformations as those in the gold summaries. This inverse relationship is also evident at the dataset level in Table 4, where XSum's higher gold distance corresponds, as predicted, to lower average ROUGE-L scores across models.

Turning to inference type distributions in Table 5, GPT-3.5-turbo exhibits a strong bias toward paraphrases and synonyms relative to human summaries (+28.1% in Gigaword, +66.3% in XSum), alongside substantially fewer common-sense (-31.6%, -34.8%) and specialized-knowledge inferences (-72.2%, -87.9%). In line with the higher inferential distance observed earlier, GPT-4o-mini more closely matches human patterns of common-sense reasoning in both datasets, but shows only limited improvement in specialized-knowledge inferences.

Overall, although the models can exceed human proportions of paraphrasing and while larger models almost match some common-sense inferences, they consistently struggle to produce specialized knowledge inferences.

6 Conclusions

In this paper, we introduced Inferential Distance, a novel metric for quantifying the semantic gap in entailing text pairs in terms of lexically-novel atomic information units. Combined with a three-way typology of inference types, our framework reveals systematic patterns in both datasets and model behavior. Empirically, increasing inferential distance is consistently aligned with degradation in models' ability to detect entailment, while inference limitations prevail particularly with respect to specialized-knowledge reasoning. In text generation, we identify an abstraction boundary where models exhibit substantially fewer common-sense and specialized-knowledge inferences than gold references. These findings illustrate the potential utility of our metric for dataset and model developers, across settings in which the entailment relation is involved. We prospect additional future utility of our measure, such as for targeted fine-tuning, prompt design, and robustness-oriented training strategies (Liu et al., 2019) and confidence calibration.

Limitations

While the inferential distance metric provides a structured framework for quantifying the semantic gap in entailment pairs, several limitations remain. First, our current validation and empirical evaluation rely exclusively on English-language benchmarks (LFQA, LLM-AggreFact, SciFact, Gigaword, and XSum). The applicability of our linguistic taxonomy—particularly regarding morphological variants and paraphrastic equivalence—has not yet been tested on morphologically rich or low-resource languages.

Second, the categorization of inference types (Paraphrase, Common Sense, and Specialized Knowledge) carries an inherent degree of subjectivity. Although our expert human annotations demonstrate high inter-annotator agreement (Appendix A), boundary lines between daily commonsense reasoning and domain-specific specialized knowledge can blur depending on the target annotator's background.

Finally, our metric assumes that the underlying hypothesis is fully entailed by the premise. In broader, unmoderated generation contexts where models produce factual hallucinations or complete non-sequiturs, the current alignment mechanism would require an upstream filtering pipeline to guarantee valid distance calculations.

References

Bo Ai and 1 others. 2025. NovAScore: A new automated metric for evaluating document level novelty. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, Abu Dhabi, UAE.

Aleksandar Angelov, Tsegaye Misikir Tashu, and Matias Valdenegro-Toro. 2025. Difficulty estimation in natural language tasks with action scores. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 351–364, Albuquerque, New Mexico. Association for Computational Linguistics.

Arie Cattan, Paul Roit, Shiyue Zhang, David Wan, Roee Aharoni, Idan Szpektor, Mohit Bansal, and Ido Dagan. 2026. Localizing factual inconsistencies in attributable text generation. *Transactions of the Association for Computational Linguistics*, 14:100–123.

Hung-Ting Chen, Fangyuan Xu, Shane Arora, and Eunsol Choi. 2025. Understanding retrieval augmentation for long-form question answering. *arXiv*.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Tirthankar Ghosal, Tanik Saikh, Tameesh Biswas, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Novelty detection: A perspective from natural language processing. *Computational Linguistics*, 48(1):77–117.

Anisha Gunjal and Greg Durrett. 2024. Molecular facts: Desiderata for decontextualization in LLM fact verification. *arXiv preprint arXiv:2406.20079*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

John P. Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. 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 4249–4259, Hong Kong, China. Association for Computational Linguistics.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019. Inoculation by fine-tuning: A method for analyzing challenge datasets. 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 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2021. How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN. *arXiv preprint arXiv:2111.09509*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Tanik Saikh, Tirthankar Ghosal, Asif Ekbal, and Pushpak Bhattacharyya. 2017. Document level novelty detection : Textual entailment lends a helping hand. In *Proceedings of the 14th International Conference on Natural Language Processing*, pages 131–140, Kolkata, India. NLP Association of India (NLPAI).

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. Minicheck: Efficient fact-checking of llms on grounding documents. *arXiv preprint arXiv:2404.10774*.

Xiangru Tang and 1 others. 2024b. Llm-aggrefact: Benchmarking large language models on aggregated factuality. *Preprint*, arXiv:2403.12345. To appear.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. SciFact: Fact verification in scientific literature. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of ICLR*.

A Human Annotation

We validated our automated annotation method through expert human annotation. The expert annotator followed the guidelines outlined in Appendix D and was instructed to measure inferential distance accordingly. The human-annotator guidelines were closely aligned with the prompt used for automated annotation. They instructed the annotator to identify atomic information units based on the novel words T_i in the hypothesis, as well as to annotate the corresponding inference type. We randomly sampled 130 entailing sentence pairs from the Gigaword dataset and 122 from LFQA. Pairs identified as non-entailing or lacking novel tokens in the hypothesis were excluded. As in the automated setting, novel tokens T_i were explicitly highlighted for the annotator.

We then applied our automatic annotation method to these human-annotated examples and measured agreement separately for inferential distance and inference type.

For inferential distance, we treat the annotation task as clustering novel tokens, where the distance corresponds to the total number of clusters. Agreement between the automated and expert annotations was measured using the Normalized Mutual Information (NMI) metric. The average NMI scores for the Gigaword and LFQA datasets were 0.865 and 0.856, respectively, demonstrating strong alignment and thus confirming the reliability of our automated inferential distance annotation.

Because the premise and hypothesis spans can differ even when their corresponding clusters are semantically similar, we evaluated inference type agreement separately. We prompted the model (GPT-5) to predict the inference type given the expert annotator's spans S_p and S_h, and computed the micro-F1 across the three inference types. The resulting F1 scores were 80.1% for Gigaword and 78.4% for LFQA. The moderate level of agreement stems from the inherent ambiguity of the inference type taxonomy and the partial overlap between categories. Despite this, the agreement remains sufficiently robust to support the qualitative and quantitative analyses in our methodology.

To further validate the expert annotations, we

measured inter-annotator agreement on a subset of 48 examples that overlapped with the expert-labeled set (22 from LFQA and 26 from Gigaword). Agreement on inferential distance was near-perfect, with an average NMI of 0.99. For inference type annotation, agreement was similarly high, with a micro-averaged F1 of 94.1% and a macro-averaged F1 of 95.8%. These results indicate a very high level of consistency across annotators, providing strong evidence that the annotations can be reliably treated as ground truth.

Recruitment and Training We hired and trained student annotators for the validation task. They attended a training session covering task definitions, minimal atomic inference spans, and our three inference types, as outlined in Appendix D. Annotators completed a calibration phase on pilot examples prior to the final evaluation to ensure alignment with guidelines.

Compensation and Ethics All student annotators were financially compensated through standard institutional student hourly wages, meeting or exceeding local minimum wage requirements. The task involved only semantic analysis of public datasets; no personally identifiable information (PII) was collected.

B Prompt for Inferential Distance

```
Your job is to be an annotator for NLP task. I'll define a task, and then I'll give you a few examples for the task.

===

Task description Textual entailment is the task of determining whether a sentence B is entailed (can be induced) from another sentence A. Both sentences usually contain multiple pieces of information. In this task, we would like to understand which part of A entails which part of B. For example, A (premise): Italy faced power cuts due to a prolonged heat wave affecting hydro-electric stations B (hypothesis): Hot weather threatens Italy with power cuts "a prolonged heat-wave" entails "hot weather", and "Italy faced power cuts due to" entails "threatens Italy with power cuts"

In this task, we want to find (1) the new unit(s) of information in the hypothesis (i.e., spans of tokens that appear in the hypothesis but not in the premise) and (2) mark which part of the premise entails each new unit of information (a.k.a. entailment explanation). Each unit in the hypothesis must be atomic, i.e., can't be broken down into smaller units while keeping entailment relations with the premise. We allow non-consecutive spans.

Atomic span

To check if a span is an atomic unit: let Sh the span in the hypothesis and Sp be the span in the premise, can you further decompose the Sh (and thus the Sp if needed) into smaller units of information, i.e. Sh1 and Sh2 (and Sp1 and Sp2 if required) and get more spans where each one is still entailed by Sp (or by Sp1 or Sp2) and makes sense. Try to make the span of the atomic units as minimal as possible so it still explains the new information.

Example

Premise: H-2A workers are legally prohibited from remaining in the United States for more than one year. Hypothesis: Workers who are in the United States on an H-2A visa can't stay in the country for longer than 12 months. For example: (Sp) united states → (Sh) the country: atomic (Sp) more than one year → (Sh) longer than 12 months: not atomic (Sp1) more than → (Sh1) longer than (Sp2) one year → (Sh2) 12 months

Guidelines for the task

1. Identify New Information a. Check each token in the hypothesis. If it does not appear in the premise in the same or a lemmatized form, it is a candidate for a "new" token.

b. Morphological Variants: - If a token in the hypothesis (e.g., bombing) is simply a morphological or closely related form of a token in the premise (e.g., bombed-out), and no new information or nuance is introduced (both refer to the same event/root concept), do not treat it as new. - Only consider it new if the change in form alters or adds meaning in a way not conveyed by the premise.

c. Group new tokens into atomic units—the smallest possible spans that: - Introduce a distinct, standalone piece of information. - Cannot be further broken down without losing or distorting the newly introduced meaning.

2. Examples of Atomic Unit Boundaries

If two or more new tokens together form a single cohesive idea (e.g., "goes on tour" means something different than "goes on" + "tour" individually), keep them as one unit.

If a new concept can be split into smaller meaningful bits that each correspond to different information in the premise or to different inference links, split them (e.g., "more than one year" » split into "more than" and "one year" if they map separately to "longer than" and "12 months").
```

4. Annotation Output Format

For each atomic unit identified in the hypothesis, output a line of the form: [premise span that entails the hypothesis atomic unit] » [atomic unit in hypothesis] – Highlight new words in bold (e.g., prayer leader).

Examples

Consider the following sentence pairs (new hypothesis tokens are underlined) below.

Example 1

Premise: H-2A workers are legally prohibited from remaining in the United States for more than one year. Hypothesis: Workers who are in the United States on an H-2A visa can't stay in the country for longer than 12 months.

Annotation output:

H-2A workers » H-2A visa Prohibited from remaining » can't stay United States » the country For more than » longer than One year » 12 months

Example 2

Premise: Italy faced power cuts due to a prolonged heat wave affecting hydro-electric stations Hypothesis: Hot weather threatens Italy with power cuts.

Annotation output:

heat wave » hot weather Italy faced power cuts due to » threatens Italy with power cuts

Example 3

Premise: The city is often referred to simply as Penang. Hypothesis: Penang is another name for the city.

Annotation output:

referred to as » another name Thus, the only new atomic entailed unit is 'another name'.

Output guidelines:

Make sure your answer is a JSON containing a list of dictionaries of all atomic units. For example: [{ "premise_span": "H-2A workers", "hypothesis_span": "H-2A visa" }, { "premise_span": "Prohibited from remaining", "hypothesis_span": "can't stay" }]

Mandatory Coverage: Every word originally marked with ** in the hypothesis must appear in exactly one hypothesis_span atomic unit. Do not omit any bold-marked word, and ensure that each bold-marked word is included only once across all annotated atomic units.

In your response in the hypothesis_span, keep the bold (**word**) for words that are originally highlighted in the hypothesis.

===

Now, please annotate the sentence pair below.

Premise: {{premise}} Hypothesis: {{hypothesis}}

C Prompt for Inference Type

Your job is to be an annotator for an NLP task. I'll define a task, and then I'll give you a few examples for it.

===

Task description:

Textual entailment is the task of determining whether a sentence B is entailed (can be induced) from another sentence A. Both sentences usually contain multiple pieces of information. In this task, we want to understand which part of A entails which part of B, and *how*.

For example:

A (premise): Italy faced power cuts due to a prolonged heat wave affecting hydro-electric stations B (hypothesis): Hot weather threatens Italy with power cuts

"a prolonged heat-wave" entails "hot weather" [Inference Type: Common Sense Reasoning] and "Italy faced power cuts due to" entails "threatens Italy with power cuts" [Inference Type: Common Sense Reasoning]

—

In this task, identify the new units of information in the hypothesis (i.e., spans that appear in the hypothesis but not in the premise), and for each, indicate which part of the premise entails it (the entailment explanation). Then categorize the correct Inference Type.

—

Inference Types to choose from:

Paraphrase or Synonyms Both terms convey essentially the same situation, event, or state, even if there are small differences in focus, tense, or mechanics. No outside knowledge or reasoning is needed to see the equivalence. Use this when both spans express the same idea directly — each entails the other. If it only works one way or needs reasoning about implications or effects → Not Paraphrase. Abbreviations (e.g. United States of America » USA) also count as Paraphrase or Synonyms.

Examples: quick response » fast reply happy » glad ahead of the deadline » before the deadline response » replying launching » beginning prime minister »

pm Not a Paraphrase: to buy » to take over

—

Common Sense Reasoning Based on universally shared, everyday human experience. If an average adult would know the fact without study, it's common sense. Use this when the two spans are not direct synonyms and understanding requires reasoning about effects or implications. Applies when the entailment goes one way — A implies B, but not vice versa. Hypernym-hyponym relations (e.g. dog » animal) also count as Common Sense Reasoning.

Check yourself: Would a typical adult know this without reading Wikipedia? If yes, it's Common Sense Reasoning.

Examples: he was naked » he didn't wear anything the crowd cheered » something good happened in the show he bought cold medicine and stayed home » he is feeling unwell it was dark outside » it was after sunset he is putting the key in the ignition » he is starting his car it rained on the ground last night » the ground is probably wet she put on a jacket » it was cold outside boat » vehicle begin distributing » release Not Common Sense: closed up » edge up (domain-specific stock market phrasing)

—

Specialized Knowledge Requires factual or specialized knowledge (geography, science, law, history, etc.) that most adults don't automatically know. Ask: (1) Could someone without special education know this from daily life? (2) Does it require a specific fact or definition? If (1)=yes and (2)=no → Common Sense. Otherwise → specialized Knowledge.

Examples: The cleric issued a fatwa » The cleric issued an Islamic religious edict They went to see the Nets » They visited New Jersey Scientists created the first cloned mammal » Scientists created Dolly the sheep

—

=== ***Examples — Consider the following sentence pairs:***

Example 1 Premise: H-2A workers are legally prohibited from remaining in the United States for more than one year. Hypothesis: Workers who are in the United States on an H-2A visa can't stay in the country for longer than 12 months.

Annotation output: H-2A workers » H-2A visa [Specialized Knowledge] Prohibited from remaining » can't stay [Common Sense Reasoning] United States » the country [Specialized Knowledge] For more than » longer than [Paraphrase or Synonyms] One year » 12 months [Paraphrase or Synonyms]

—

Example 2 Premise: Italy faced power cuts due to a prolonged heat wave affecting hydro-electric stations. Hypothesis: Hot weather threatens Italy with power cuts.

Annotation output: heat wave » hot weather [Common Sense Reasoning] Italy faced power cuts due to » threatens Italy with power cuts [Common Sense Reasoning]

—

Example 3 Premise: The city is often referred to simply as Penang. Hypothesis: Penang is another name for the city.

Annotation output: referred to as » another name [Common Sense Reasoning]

—

Please categorize the correct Inference Type for the sentence pair below. Your output must be only one of the options above (no additional text).

Premise: premise Hypothesis: hypothesis Inference to be categorized: {{_source}} » {{_destination}}

D Annotation Instructions

D.1 Task Description

Textual entailment is the task of determining whether a sentence B is entailed (can be induced) from another sentence A. Both sentences usually contain multiple pieces of information. In this task, we would like to understand which part of A entails which part of B.

For example,

> **A (premise):** Italy faced power cuts due to a prolonged heat wave affecting hydro-electric stations
> **B (hypothesis):** Hot weather threatens Italy with power cuts

"a prolonged heat-wave" entails "hot weather", and "Italy faced power cuts due to" entails "threatens Italy with power cuts".

In this task, we want to find (1) the new unit(s) of information in the hypothesis (i.e., spans of tokens that appear in the hypothesis but not in the premise) and (2) mark which part of the premise entails each new unit of information (a.k.a. entailment explanation). Each unit in the hypothesis must be atomic, i.e., can't be broken down into smaller units while keeping entailment relations with the premise. We allow non-consecutive spans.

D.2 Atomic Span

A rule of thumb to check if a span is an atomic unit: let S_p be the span in the premise and S_h the span in the hypothesis, can you further decompose the S_h (and thus the S_p if needed) into smaller units of information, i.e., S_{h1} and S_{h2} (and S_{p1} and S_{p2} if required) and get more spans where each one is still entailed by S_p (or by S_{p1} or S_{p2}) and makes sense.

Try to make the span of the atomic units as minimal as possible so it still explains the new information.

D.2.1 Example

> **Premise:** H-2A workers are legally prohibited from remaining in the United States for more than one year.
> **Hypothesis:** Workers who are in the United States on an H-2A visa can't stay in the country for longer than 12 months.

For example:

- (S_p) united states $\rightarrow$ (S_h) the country: atomic
- (S_p) more than one year $\rightarrow$ (S_h) longer than 12 months: not atomic
- (S_{p1}) more than $\rightarrow$ (S_{h1}) longer than
- (S_{p2}) one year $\rightarrow$ (S_{h2}) 12 months

D.3 Guidelines for the Task

D.3.1 1. Identify New Information

a. Check each token in the hypothesis. If it does not appear in the premise in the same or a lemmatized form, it is a candidate for a "new" token.

b. Morphological Variants:

   - If a token in the hypothesis (e.g., bombing) is simply a morphological or closely related form of a token in the premise (e.g., bombed-out), and no new information or nuance is introduced (both refer to the same event/root concept), do not treat it as new.
   - Only consider it new if the change in form alters or adds meaning in a way not conveyed by the premise.

c. Group new tokens into atomic units—the smallest possible spans that:

   - Introduce a distinct, standalone piece of information.
   - Cannot be further broken down without losing or distorting the newly introduced meaning.

D.3.2 2. Assign an Entailment Relation

For each atomic unit in the hypothesis, find the span in the premise that entails (or logically connects to) this new piece of information.

Choose the appropriate entailment type(s).

D.4 Inference Types to Choose From

Paraphrase or Synonyms Both terms convey essentially the same situation, event, or state, even if there are small differences in focus, tense, or the mechanics of how it happens. No outside factual knowledge or reasoning is needed to see the equivalence. Same surface meaning or using metaphor, but no reasoning is needed. Use this when both spans express the same idea directly, such that each entails the other.

Test yourself: Ask: "If I only had sentence A, could I replace its span with sentence B's span (and vice versa) without changing the meaning?"

If yes in both directions $\rightarrow$ Paraphrase or Synonyms.
If it only works one way or needs reasoning about implications, intent, or effects $\rightarrow$ Not Paraphrase.

We also consider any abbreviations (e.g. United States of America $\rightarrow$ USA) as Paraphrase or Synonyms.

Examples for Paraphrase or Synonyms:

- quick response $\rightarrow$ fast reply
- happy $\rightarrow$ glad
- ahead of the deadline $\rightarrow$ before the deadline
- response $\rightarrow$ replying
- launching $\rightarrow$ beginning
- prime minister $\rightarrow$ pm [abbreviation is considered as Paraphrase or Synonyms]

Not a Paraphrase or Synonyms example: to buy $\rightarrow$ to take over [aren't bi-directional entailment. take over doesn't mean buy, though buy might mean takeover].

Common Sense Reasoning Common Sense Reasoning is based on universally shared, everyday human experience. If an average adult (anywhere in the world, without formal study) would immediately know the fact, it is common sense. Common sense reasoning does not require deep understanding in the domain, and no need to use Wikipedia or books to know the answer.

Use this label only when the two spans are not direct synonyms, and understanding the entailment requires reasoning about effects, implications, or unstated links. That is, if you must ask "why/how does A lead to B?" instead of simply recognizing two ways of saying the same thing. Use this when the entailment only goes one way.

We also consider any hypernym-hyponym (e.g. dog → animal) as Common Sense Reasoning.

Examples:

- he was naked → he didn't wear anything
- the crowd cheered → something good happened in the show
- he bought cold medicine and stayed home → he is feeling unwell
- it was dark outside → it was after sunset
- he is putting the key in the ignition → he is starting his car
- it rained on the ground last night → the ground is probably wet
- she put on a jacket → it was cold outside
- boat → vehicle (hypernym is considered as Common Sense Reasoning)
- begin distributing → release

Example of NOT Common Sense Reasoning: "closed up" → "edge up" (both are stock-market shorthand for the same outcome, just phrased differently).

Specialized Knowledge Specialized knowledge requires specific facts beyond common sense. Requires factual or specialized knowledge (geography, science, law, history, technical terms, etc.) that most adults don't automatically know.

Before labeling as Specialized Knowledge, ask yourself:

1. Could someone with no special education know this from everyday life?
2. Does it require recalling a specific location, technical definition, or published fact?

If (1) = yes and (2) = no → Common Sense Reasoning.

Examples:

- Fatwa → Islamic religious edict
- They went to see the Nets → They visited New Jersey
- The first cloned mammal → Dolly the sheep

E Datasets

E.1 NLI Datasets

LFQA (Chen et al., 2025) contains LLM-generated answers to ELI5 ("Explain Like I'm Five") questions (Fan et al., 2019), evaluated against retrieved documents. Human annotators labeled each sentence for factual support. Following MiniCheck (Tang et al., 2024a), we collapsed fine-grained labels into a binary *supported* / *unsupported* decision and constructed a dataset of 1,758 examples, balanced across the two labels.

LLM-AggreFact-XSum (Tang et al., 2024b) is part of the LLM-AGGREFACT benchmark and consists of complete system-generated abstractive summary sentences expert-annotated for factual consistency with their source articles. We used 766 binary labels (*supported*/*unsupported*) and balanced the dataset across the two classes, as in LFQA.

SciFact (Wadden et al., 2020) comprises expert-written scientific claims paired with literature abstracts and annotated by domain experts for entailment. Each pair links a scientific claim to a supporting or refuting evidence sentence. For our experiments, we used both the train and dev sets, comprising 1,392 pairs.

E.2 Extreme Summarization Datasets

Since headlines and summaries in both datasets often contain information that is not explicitly stated in the source document, we use GPT o3-mini to predict whether each summary is logically derivable from its context and filter out non-entailing cases.

Gigaword Corpus (Rush et al., 2015) contains about 3.8M article–headline pairs. We randomly sampled 904 pairs, using the article sentence as the premise and the headline as the hypothesis.

XSum (Narayan et al., 2018) contains 226,711 BBC articles, each paired with a single-sentence abstractive summary. For our analysis, we randomly sampled 906 article–summary pairs.

F Summarization Prompts

F.1 Standard Prompt

```
Summarize this news article in only 1 sentences, not exceeding
[target_summary_length] words, keep only the most important details.
News Article: [prompt_article]
```