Entailed Between the Lines: Incorporating Implication into NLI

Shreya Havaldar[*◇], **Hamidreza Alvari**[†], **John Palowitch**[†],
Mohammad Javad Hosseini[†], **Senaka Buthpitiya**[†], **Alex Fabrikant**[†]
◇University of Pennsylvania, †Google Deepmind
shreyah@seas.upenn.edu, {hamidrz, palowitch}@google.com

Abstract

Much of human communication depends on implication, conveying meaning beyond literal words to express a wider range of thoughts, intentions, and feelings. For models to better understand and facilitate human communication, they must be responsive to the text's implicit meaning. We focus on Natural Language Inference (NLI), a core tool for many language tasks, and find that state-of-the-art NLI models and datasets struggle to recognize a range of cases where entailment is implied, rather than explicit from the text. We formalize **implied entailment** as an extension of the NLI task and introduce the Implied NLI dataset (INLI) to help today's LLMs both recognize a broader variety of implied entailments and to distinguish between implicit and explicit entailment.[1] We show how LLMs fine-tuned on INLI understand implied entailment and can generalize this understanding across datasets and domains.

1 Introduction

Human communication is rich with implication. Emotions, social cues, insults, and a myriad of other messages are conveyed implicitly, often even more so than explicitly. In order for LLMs to reach human-level understanding of communication, they must be able to understand a text's implications.

Consider what a human reader learns from the following sentence: *After reading the ARR reviews, Kim had to go stuff himself with cheesecake.*

Beyond the explicit text that informs us that (a) Kim read ARR reviews, and (b) Kim subsequently was obliged to eat cheesecake, we also readily absorb a variety of implicitly entailed facts: (c) Kim found reading the reviews unpleasant, (d) Kim actually proceeded to eat cheesecake, and (e) Kim ate more than an ordinary amount of cheesecake.
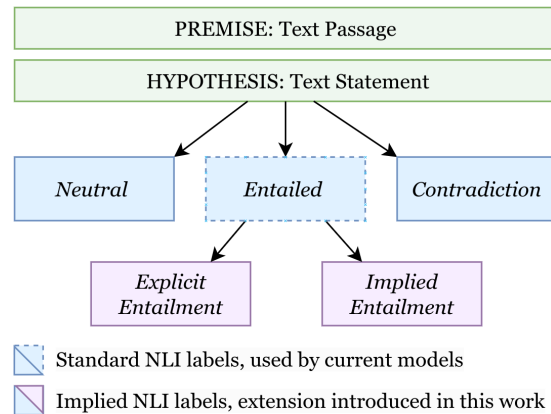

Figure 1: Extending NLI to introduce implication. We propose that entailment models should treat implications as entailed, while also distinguishing between explicit and implicit entailments.

Modern LLMs take on a number of roles and applications where understanding *implied entailments* like (c-e) is crucial. For instance, in the context of creative writing or mental health, it is vital to consider (c) to understand emotional states, often communicated implicitly (Gullestad and Killingmo, 2019). In the context of translation or cross-cultural communication, where patterns of implicit language use vary across languages and cultures, (Kim et al., 1998; Havaldar et al., 2023), it is important to understand the subtlety between (d) and (b), and to translate the figurative meaning of (e) rather than the verbatim hyperbole.

Prior work on implied language in NLP predominantly focuses on rigorously structured inputs – indirect answers to yes/no questions (George and Mamidi, 2020; Damgaard et al., 2021; Louis et al., 2020), pairwise entity selection (Hosseini et al., 2023), and logical consistency (Jeretic et al., 2020; Zheng et al., 2021). While useful for pragmatic understanding, these tasks require inputs that have a specific structure, and thus have limited applicability when it comes to LLM reasoning on natural, unstructured premises.

32274

Premise	Hypothesis	Entailed	Implicit	Label
Alex walked towards the checkout lane in the grocery store. The cashier asked if he had brought a bag for his groceries. With an angry grumble, Alex cursed himself for forgetting and told the clerk he needed to purchase a bag that evening.	Alex needs to purchase a bag for his groceries.	✓	✗	Explicit entailment
	The grocery store does not provide free bags.	✓	✓	Implied entailment
	The cashier at the checkout lane is middle-aged.	✗	–	Neutral
	Alex remembered to bring a bag to the grocery store.	✗	–	Contradiction

Table 1: Examples of the four labels in our implied entailment task. Unlike the standard entailment task, we draw the distinction between explicitly entailed and implicitly entailed hypotheses. Implications must be *inferable* (i.e. fully entailed in the premise) and *implicit* (i.e. not explicitly stated in the premise). Note that the definitions of neutral and contradictory remain the same as the standard task.

NLI models (i.e. classifiers to decide if a hypothesis H is entailed by, contradicted by, or neutral with respect to a premise P) have been used for a variety of NLP applications – grounding, factuality, proposition segmentation, etc. (Stowe et al., 2022; Bhagavatula et al., 2020; Gao et al., 2023; Hosseini et al., 2024a). However, as we show in this paper, standard NLI datasets contain few implications. As a result, models trained on these benchmarks struggle to recognize implied entailments as entailments at all. Ensuring these cases are covered by entailment classifiers is important if LLMs are to understand nuances and subtleties in human communication.

Additionally, several applications of NLI, such as summarization evaluation or citation generation, would significantly benefit from a *distinction* between explicit and implied entailment, thus motivating our dataset's 4-way refinement to separate these entailment labels within the traditional 3-way NLI taxonomy.

In this paper, we introduce the **implied entailment** task and build a dataset on which NLI models can be fine-tuned to recognize implications as valid entailments. Concretely, when determining whether a given premise P entails a hypothesis H, we want NLI models to reason about whether H is *implicitly* or *explicitly* entailed in P (Figure 1). Specifically, our contributions are as follows:

1. We *formalize the implied entailment task*, distinguishing implied entailment from explicit entailment and introducing an NLI paradigm that reflects this distinction.

2. We build an *INLI – the implied NLI dataset*, with 10k premises mirroring real-world communication and 40k hypotheses that are implied, explicit, neutral, and contradictory.

3. We show that NLI models fine-tuned on INLI are able to recognize when a hypothesis is implied versus explicitly entailed, and that they *generalize across domains and datasets.*

2 Formalizing Implied Entailment

We start with the conventional 3-way distinction of NLI labels (Williams et al., 2018): Given a premise P and a hypothesis H, H is an entailed by P if it is definitely correct given P, contradicted by P if it is definitely incorrect given P, and neutral otherwise.

We then further refine entailment into two categories: explicit and implied. An implied entailment is a hypothesis that requires the reader to make additional cognitive deductions beyond the explicit language[2] used in the text:

- **Explicit entailment:** Follows directly from the text's lexical semantics (e.g. via synonymy and paraphrasing) and syntax (e.g. via pronominal co-reference, bridging, or other endophora).

- **Implied entailment:** Requires some sort of an additional cognitive step, such as logical reasoning, world knowledge, conversational pragmatics, or figurative language.

We use these definitions to create an evaluation scheme for human annotation of INLI in Section 4.3. Table 1 shows examples of hypotheses in our 4-way implied entailment task.

---

[2]Note that our definitions conflate the distinctions drawn in pragmatics/philosophy of literature between "explicatures" (Sperber and Wilson, 1986), "implicatures" (Grice, 1991), and "entailments" (Moldovan, 2019). In our framework, explicit entailment is equivalent to explicatures, while implied entailment combines implicatures and non-explicature entailments.

Dataset	% Implied
SNLI (Bowman et al., 2015)	9.33
MNLI (Williams et al., 2018)	3.68
ANLI (Nie et al., 2020)	15.66
WANLI (Liu et al., 2022)	5.48

Table 2: Quantifying implied entailments present in existing NLI benchmarks. Most NLI benchmarks contain a disproportionate amount of explicit entailments, highlighting the need for an implication-focused dataset.

3 Implied Language Understanding in Current Benchmarks & Models

To motivate the need for an implied entailment dataset, we first explore whether current NLI benchmarks already contain implicitly entailed hypotheses, and whether current NLI models already possess the ability to generalize to implications. Concretely, we answer the following questions:

1. To what extent do widely-used NLI benchmarks contain implied entailments?

2. How accurately can current NLI models infer implied entailments?

3.1 NLI Benchmarks Contain Few Implied Entailments

INLI (see Section 4) contains both implicit and explicit entailments. Using this data, we fine-tune a T5-XXL model (Raffel et al., 2020) to distinguish between implicit and explicit entailments, achieving 97.3% test accuracy. We then apply this model to entailments in existing NLI benchmarks – the Stanford NLI Corpus (Bowman et al., 2015), Multi-Genre NLI (Williams et al., 2018), Adversarial NLI (Nie et al., 2020), and Worker-and-AI NLI (Liu et al., 2022).

Table 2 shows what percent of each of these benchmarks are classified as implied entailments by our fine-tuned model – most contain very few implied entailments, suggesting that the vast majority of entailed hypotheses in current benchmarks are explicit. The notable exception is ANLI, a dataset designed to adversarially distill exemplars that stump models. ANLI contains three subsets, or "rounds," with increasing difficulty, and we observe that the number of implied entailments increases as the rounds get harder.

This further suggests that modern NLI models struggle with understanding implication, as the hypotheses that challenge models most are implicit in nature. See Appendix A.1 for details on experiment setup and model training.

Training Dataset	Entailment Accuracy	
	Implied	Explicit
SNLI	0.500	0.943
MNLI	0.528	0.965
ANLI	0.714	0.983
WANLI	0.525	0.905

Table 3: Measuring how well models fine-tuned on existing NLI datasets can infer explicit entailments and implied entailments in INLI. Results suggest existing NLI datasets do not support generalization to implied entailments.

Human validation of Table 2. Training on implicit and explicit entailments from INLI and then testing on existing benchmarks requires manual validation to ensure our implicitness detection model can properly generalize. We verify that annotators agree with 92.0% of model outputs, with an inter-annotator agreement (Cohen's Kappa) of 0.768. Appendix A.1 contains additional details about annotation setup and results.

3.2 NLI Models Rarely Generalize to Implied Entailments

Despite the lack of implied entailments contained in NLI benchmarks, we also explore whether models trained on current datasets can successfully generalize and understand implications in INLI.

We fine-tune a T5-XXL model on each of the four NLI benchmarks, and then measure their accuracy in classifying implicitly and explicitly entailed hypotheses in INLI. Table 3 contains these results. Unsurprisingly, we see very high accuracy in inferring explicit entailments, but for implied entailments, most models perform around 50%, i.e. randomly guessing whether implied entailments are inferable or not.

The clear exception is the model trained on ANLI, suggesting that implied entailments in training data improves ability to infer implied entailments in other settings. However, there is still a significant performance gap between explicit vs. implied. Overall, Table 3 suggests a significant need for improvement in this task; see Appendix A.2 for experiment details.

Takeaways: NLI benchmarks predominantly contain explicit entailments; consequently, NLI models struggle to infer implied entailments. These findings highlight the need for a dataset that focuses on *implied entailment* across a variety of domains.

Dataset	Implicature Frame	Example
LUDWIG (George and Mamidi, 2020)	Question Indirect Answer Implied Meaning	Question: Would you like to go to a party tonight? Answer: I am too tired. Meaning: No
CIRCA (Louis et al., 2020)	Conversational Context Question Indirect Answer Implied Meaning	Context: Colleagues leaving work together on a Friday. Question: Do you want to hang out later? Answer: I could do with a stiff drink. Meaning: Yes
NORMBANK (Ziems et al., 2023)	Behavior Situational Context Implied Social Norm	Behavior: Play with your food Context: Have a food fight in a restaurant setting Social Norm: Taboo
SOCIALCHEM (Forbes et al., 2020)	Social Situation Implied Rule-of-Thumb	Situation: Telling my sister I would not donate a kidney to her. Rule-of-Thumb: You shouldn't expect someone to donate their organ to you.

Table 4: Datasets chosen for this work and their corresponding implicature frames. To ensure the implied entailments in INLI are of the highest quality, we transform existing implicature frames (i.e. data that contains an implicature) into the ⟨premise, implied entailment⟩ format required for NLI.
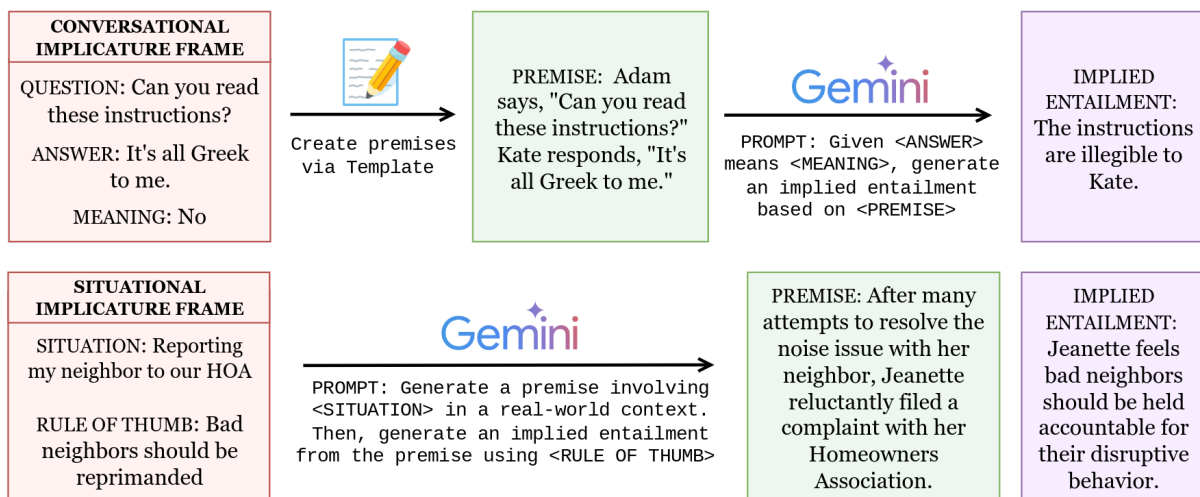

Figure 2: Stage 1: Implicature Augmentation. Given existing data that contains an implicature (e.g. the answer to an indirect question, a widely-accepted social norm, etc.), we augment the data into a premise and implicitly entailed hypothesis. In this figure we show augmented examples from LUDWIG and SOCIALCHEM.

4 The Implied NLI Dataset (INLI)

To construct INLI, we introduce a novel procedure to leverage existing datasets that exemplify language pragmatics.

Specifically, given a structured example of implied meaning in a social context (see Table 4), we prompt `Gemini-Pro` to derive a ⟨premise, implied entailment⟩ pair that can be used to train NLI models. This strategy is more efficient than traditional human-generation strategies of traditional NLI datasets (Bowman et al., 2015; Nie et al., 2020) in the following three areas:

1. **Generation cost:** We sidestep the need to train annotators to generate implied hypotheses from premises.
2. **Quality:** We derive implied entailments from

pre-existing, human-vetted datasets, offering a high-quality foundation for INLI.

3. **Reproducibility:** Our procedure is easily reproducible for researchers who want to expand or generalize INLI.

Our pipeline to build INLI has two main components: (1) Augmenting Implicature Frames, and (2) Alternative Hypothesis Generation. We discuss each component in detail below.

4.1 Augmenting Implicature Frames

Given understanding implication is a complex and subjective task, we utilize past datasets that contain *implicature frames*: structured templates that express an implied meaning arising from a conversational or behavioral setting. Table 4 gives exam-
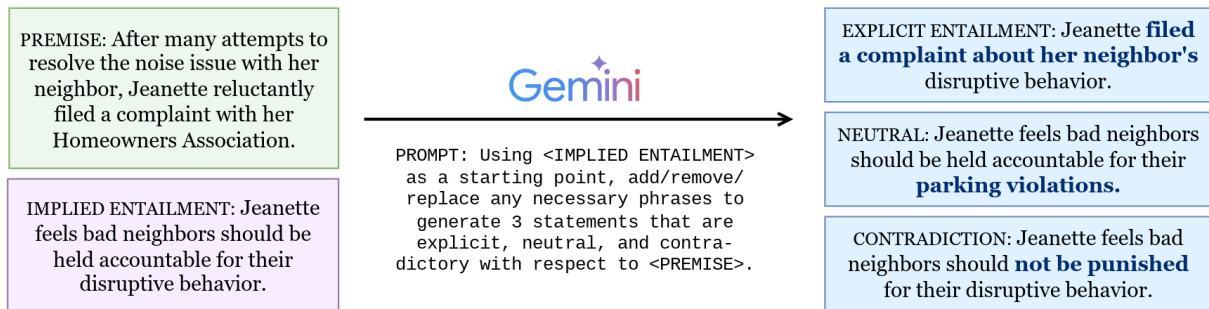
Figure 3: Stage 2: Alternative Hypothesis Generation. Given premises and their corresponding implied entailments, we generate three additional hypotheses (explicit, neutral, and contradictory) by subtly modifying the implied entailment. This ensures we build a more challenging dataset for the implied entailment task.

ples of the four datasets we use in this work and their corresponding implicature frames. Our goal is to convert such implicature frames into ⟨premise, hypothesis⟩ pairs that can be used to train entailment models, where the hypothesis expresses the implied meaning present in the original frame. We select 10k frames from these datasets and few-shot prompt `Gemini-Pro` to perform this conversion.

Specifically, we design an augmentation framework that takes in an implicature frame, like those in Table 4, and extracts a corresponding premise and implied entailment. This process is different for conversational implicatures (LUDWIG, CIRCA) and situational implicatures (NORMBANK, SO-CIALCHEM) respectively.

Conversational implicatures. For LLMs to properly grasp the nuances of communication, they must understand implications arising in conversations. Both LUDWIG and CIRCA include questions and indirect answers. To construct a premise, we use a template with two random names from the 2011 US Census, simulating a back-and-forth conversation. We then prompt `Gemini-Pro` with this premise and the given yes/no implied meaning of the indirect answer, asking `Gemini-Pro` to generate an implied hypothesis based on the provided meaning (See top of Figure 2).

Situational implicatures. In order to excel in applications like creative writing or mental health, LLMs must also understand implications hinging on social norms and interpersonal relationships. For NORMBANK and SOCIALCHEM, we start by using the provided behavior and context, or social situation, to create a premise that reflects the given real-world scenario. Then, we prompt `Gemini-Pro` to generate an implied entailment from the premise based on the corresponding social norm or rule-of-

thumb (See bottom of Figure 2).

Table 5 contains the premises and implied entailments derived by augmenting the examples in Table 4. Additional details on prompting procedure are provided in Appendix E.

4.2 Alternative Hypothesis Generation

After generating premises and implied entailments by augmenting implicature frames, we also want to generate the remaining hypotheses: explicit entailments, neutrals, and contradictions.

We experiment with various prompting techniques to accomplish this, analyzing results and iterating accordingly. Details on various prompting iterations are discussed in Appendix C, and our final prompt is in Appendix E.

In order for the hypotheses to be semantically similar (e.g. contradictions shouldn't have disproportionate negation) and make for a challenging NLI task, we have `Gemini-Pro` use the implied entailment as a starting point, and then replace any necessary words or phrases in order to transform the implied entailment into an alternative hypothesis that is explicit/neutral/contradictory in relation to the given premise. Figure 3 details this process.

Paraphrasing. As a final step, we have `Gemini-Pro` paraphrase all generated hypotheses to minimize artifacts in the data that could distinguish between hypothesis labels (Krishna et al., 2024; Sadasivan et al., 2024). We do not paraphrase the explicit entailments in our conversational implicature datasets, as these hypotheses are simpler and paraphrasing them might alter their true labels.

Data robustness. Past research stresses that inference in NLI datasets should depend on both the premise and hypothesis (Poliak et al., 2018). A method to test premise-dependence is fine-tuning

LUDWIG [N = 1,956]	
Premise	Diane says, "Would you like to go to a party tonight?" Sophie responds, "I am too tired."
Implied	Sophie would prefer not to attend the party this evening.
Explicit	Sophie claims to be too tired.
Neutral	The party will take place outside.
Contradiction	Sophie is excited to attend the party this evening.

CIRCA [N = 18,044]	
Premise	Teresa and Alicia are colleagues who are leaving work on a Friday at the same time. Teresa says, "Do you want to hang out later?" Alicia responds, "I could do with a stiff drink."
Implied	Alicia hopes to spend time with Teresa later.
Explicit	Alicia could do with a stiff drink.
Neutral	Alicia would benefit from a night out.
Contradiction	Alicia doesn't feel like spending time with Teresa later.

NORMBANK [N = 10,000]	
Premise	A hush fell over the elegant restaurant as a group of diners, their laughter escalating hysterically, began to fling mashed potatoes and bread rolls at each other, splattering the pristine white tablecloth and surrounding patrons with food.
Implied	The other diners were shocked by the inappropriate food fight.
Explicit	The restaurant fell silent as a group of diners started throwing food at one another.
Neutral	The diners at the restaurant had never witnessed such a big food fight.
Contradiction	The other diners were entertained by the food fight and enthusiastically joined in.

SOCIALCHEM [N = 10,000]	
Premise	With a heavy heart, Maya explained to her sister, Chloe, that she wouldn't be able to donate a kidney, bracing herself for the wave of emotions that were sure to follow.
Implied	Maya realized her choice not to donate a kidney would probably devastate Chloe.
Explicit	Maya prepared herself for Chloe's emotional reaction when she told her she couldn't be a kidney donor.
Neutral	Maya realized her decision to move to another state would probably upset Chloe.
Contradiction	Maya felt confident that Chloe would be understanding of her choice not to donate her kidney.

Table 5: Examples from INLI. We augment implicature frames from four datasets to create premises and implicitly entailed hypotheses (see Table 4 for original data) and then use `Gemini-Pro` to generate the remaining alternative hypotheses. [N] indicates the number of ⟨premise, hypothesis⟩ pairs in INLI from each dataset.

a model to label hypotheses without using the premises. Our dataset's hypothesis-only accuracy is similar to other benchmarks, suggesting minimal artifacts that dilute the difficulty of our task. Details in Appendix D.

4.3 Human Validation of Dataset

To ensure that our pipeline yields high-quality data, all six authors of this paper annotate a subset of INLI. The annotation task is similar to that of previous NLI benchmarks – annotators are only shown a premise and a hypothesis, and are instructed to label the hypothesis with either implied, explicit, neutral, or contradiction.

To define these four labels, annotators are given the definition in Section 2. They are instructed to take into account *inferability* and *implicitness* of a hypothesis when assigning a label.

Training human annotators. The annotators were first given the annotation instructions (see Appendix B) and a subset of 16 ⟨premise, hypothesis⟩ calibration examples from INLI spanning all datasets and label types. The annotators then rated the calibration examples, and afterward were encouraged to discuss their answers with the other annotators, referring to the instructions to guide the discussion. This process was meant to allow annotators to refine their understanding of the implied entailment labels before proceeding into the task.

Annotation task. For the final annotation task, we select 200 ⟨premise, hypothesis⟩ pairs from INLI (equally distributed across datasets and hypothesis types), and assign each annotator to label a different subset of 100; ultimately, we have three annotators labeling each example.

Model	Accuracy	
	Overall	**Implied Entailment**
Fine-tuned LLMs		
T5-Small	0.813	0.731
T5-Base	0.871	0.817
T5-Large	0.913	0.870
T5-XXL	0.924	0.885
8-shot Prompted LLMs		
GPT-4o	0.749	0.608
GPT-4	0.753	0.645
Claude-3-Sonnet	0.686	0.738
Mistral-Large	0.744	0.735
Gemini-Pro	0.770	0.628
Gemini-Flash	0.737	0.639

Table 6: Benchmarking LLMs on INLI. We explore fine-tuning smaller LLMs, as well as few-shot prompting larger LLMs. We see that all models perform worse on implied entailments in INLI than the rest of the dataset, indicating room for improvement in implication understanding capabilities.

Annotators have a Fleiss's kappa $\kappa = 0.711$ and a majority agreement (i.e. at least 2 out of 3 annotators agree) with the INLI labels of 0.935. Our annotator agreement is in line with that of past NLI benchmarks (ANLI: $\kappa = 0.679$ to 0.721, WANLI: $\kappa = 0.60$). Additionally, we see very high majority agreement with the INLI labels, indicating INLI examples reflect human intuition well.

5 Learning Implied Entailment

We now show the promise of INLI to imbue NLI models with understanding of implied entailment, and explore whether INLI leads to a generalizable understanding of implication. Specifically, we answer the following research questions:

- **RQ1:** Do models fine-tuned and few-shot prompted on INLI effectively reason about implied entailment?

- **RQ2:** Do models fine-tuned on INLI maintain efficacy on traditional NLI benchmarks?

- **RQ3:** Do models fine-tuned on INLI generalize across domains and datasets?

5.1 Benchmarking LLMs on INLI

How well do modern LLMs perform on INLI? We explore the capabilities of smaller LLMs that can be fine-tuned cheaply (the T5 family), as well as larger LLMs that we few-shot prompt (GPT-4,

	Training Datasets	
	Standard NLI	*3-way INLI*
SNLI	0.934	0.921
MNLI	0.916	0.914
ANLI	0.725	0.734
WANLI	0.825	0.822
3-way INLI	0.778	0.909

Table 7: Models fine-tuned on INLI retain NLI capabilities. We measure accuracy on 4 NLI benchmarks before and after finetuning T5-XXL on INLI.

Claude-3, Mistral). Results are in Table 6, and experiment details are in Appendix A.3.

Though the fine-tuned LLMs do decently well on INLI overall, the implied entailment accuracy is lower than overall accuracy for all models. Even T5-XXL, which performs the best on INLI overall, has an implied entailment accuracy of 0.885, leaving sufficient room for improvement. Note that the human agreement for implied entailment is 0.94 (see Table 11), suggesting a similar upper bound for this task.

To benchmark larger LLMs, we create an 8-shot prompt (see Appendix E) with examples spanning all hypothesis types and datasets. Though known for high performance on reasoning tasks, these larger LLMs perform worse than the fine-tuned models, and for most LLMs, the discrepancy between overall accuracy and implied entailment accuracy is even more pronounced.

Note that Gemini-Pro was the model used to build the dataset. Despite this, we see the same pattern – mediocre overall accuracy and an even lower implied entailment accuracy. These results suggest that models like GPT-4 struggle with understanding implication, but fine-tuning on INLI can significantly improve performance.

Takeaway: INLI is a challenging dataset for today's LLMs, but training on INLI improves their ability to understand implication.

5.2 INLI and Existing NLI Benchmarks

We also want to ensure that a model trained to understand implied entailment still maintains efficacy on the standard NLI task.

To do this, we first fine-tune a T5-XXL model on the four NLI benchmarks listed in Table 2, and assess performance. These preliminary results are in the first column of Table 7. We then augment INLI to fit the standard 3-label NLI task, where

	Training Subset	Accuracy	Generalization Subset	Accuracy
Intra-Domain Generalization	NORMBANK	0.917	SOCIALCHEM	0.795
	SOCIALCHEM	0.913	NORMBANK	0.850
Cross-Domain Generalization	LUDWIG, CIRCA	0.965	SOCIALCHEM, NORMBANK	0.695
	SOCIALCHEM, NORMBANK	0.913	LUDWIG, CIRCA	0.796
Cross-Dataset Generalization	NORMBANK, LUDWIG, CIRCA	0.945	SOCIALCHEM	0.804
	SOCIALCHEM, LUDWIG, CIRCA	0.942	NORMBANK	0.851

Table 8: INLI supports generalization across datasets and domains. We run various generalization experiments and show how INLI can improve LLM ability to understand implication in new environments.

explicit and implicit entailments are collapsed into the "entailed" label. Lastly, we fine-tune the model on augmented INLI, using early stopping to prevent forgetting (Barone et al., 2017). Results can be found in the second column of Table 7, and experiment details are in Appendix A.4.

Encouragingly, we see that performance on INLI increases, while performance on the NLI benchmarks roughly stays the same. Performance on ANLI also increases slightly, suggesting that understanding implication is useful for the challenging examples in ANLI.

Takeaway: Results indicate LLMs trained to infer implication will retain previous reasoning capabilities on other tasks.

5.3 Generalization Capabilities of INLI

A key goal of INLI is helping LLMs understand implication. So, it is crucial that models trained on INLI have the ability to generalize to unseen domains and datasets.

In order to assess whether INLI is useful for generalizable understanding of implied entailment, we perform a series of generalization experiments to assess how LLMs that learn implication in one environment can apply it to a different environment. We specifically test the following three generalization capabilities:

1. **Intra-domain:** Generalization from one dataset to another in the same domain. We train on NORMBANK and test generalization to SOCIALCHEM, and vice versa.[3]

2. **Cross-domain:** Generalization from one domain to another. We train on the situational subset of INLI and test on the conversational subset, and vice versa.

3. **Cross-dataset:** Generalization to an unseen dataset. We train on three datasets and test on the fourth, using NORMBANK and SOCIALCHEM as our held-out datasets.

Results are in Table 8 and experiment details are in Appendix A.5. We see impressive generalization results across all experiments – for instance, a model that has never seen NORMBANK but is fine-tuned on the other three datasets in INLI has better accuracy on NORMBANK than GPT-4 or Claude-3 prompted using examples from all four datasets.

Takeaway: Models trained on INLI can generalize to understand implication in new domains and environments.

6 Related Work

Structured Implicatures: Most past work that focuses on implicatures requires a fixed input structure – indirect QA (George and Mamidi, 2020; Louis et al., 2020; Takayama et al., 2021; Damgaard et al., 2021), scalar implicatures focusing on logical consistency (Zheng et al., 2021; Jeretic et al., 2020), pairwise entity selection (Hosseini et al., 2023), etc. However, their fixed input and output structure limits utility; INLI is unstructured and thus more useful for generalizable implicature understanding.

Implicature Frames: A parallel line of work analyzes how the implicit meaning of text changes based on context – social norms (Sap et al., 2019; Forbes et al., 2020), cultural norms (Ziems et al., 2023; Rai et al., 2024), offensive speech (Zhou et al., 2023), etc. We pull from this rich body of work for our situational implicature frames.

Implicature Understanding: Previous work has attempted to measure implicature understanding in LLMs – via chain-of-thought (Kim et al., 2023), explanation generation (Yue et al., 2024a),

---

[3]We choose NORMBANK and SOCIALCHEM as CIRCA and LUDWIG are quite similar in content/structure, so our situational datasets more accurately assess generalization.

comparison to humans (Qiu et al., 2023), and intent classification (Zhang et al., 2024). Findings vary from "LLMs fail to understand implicatures" (Yue et al., 2024b; Qiu et al., 2023) to "LLMs show promise." (Kim et al., 1998; Zhang et al., 2024). This discrepancy further motivates the need for an implication-focused benchmark like INLI.

Commonsense NLI: Past commonsense NLI research (Rudinger et al., 2020; Zhang et al., 2017) integrates various types of world knowledge and analyzes downstream likelihood of entailment. These works, along with certain NLI datasets (Zellers et al., 2019; Bhagavatula et al., 2020; Bisk et al., 2019) emphasize understanding physical and temporal constructs for success but do not address the distinction between explicit and implied entailment. Ghosal et al. (2021) touches on implicitness but is limited to causal phrases derived from predefined relations, making it inapplicable to free-form reasoning. INLI focuses on explicit vs. implied information in free-form premises, a boundary not captured by Commonsense NLI. This distinction is highlighted by performance discrepancies (e.g. GPT-4 achieves 95% on HellaSwag (Zellers et al., 2019), but 75% on INLI).

7 Conclusion

In this paper, we formalized the implied entailment task, providing an evaluation scheme for implied entailments. We created the Implied NLI (INLI) dataset, a stepping stone to help LLMs reason about and infer implication in text. We demonstrated the utility of INLI in working towards models that understand the subtleties of human language.

We hope INLI will provide a tool to benchmark LLMs and improve their ability to understand implication. We encourage future researchers to explore how INLI can help LLMs read between the lines and better support real-world communication.

Limitations

There are various limitations with the data we augmented to build INLI. Since our objective was to build a dataset that mirrored human communication, we focused on the situational and conversational domains. However, this may lead to a lack of domain diversity poor generalization to formal prose or highly specific text (e.g. medical, legal).

Since we chose to augment existing data for more reliable implied entailments, any flaws with the original datasets (typos, incorrect labels, etc.)

may be propagated up in INLI. Though we had annotators vet a subset of INLI, the entirety of our dataset has not been human-vetted; thus, some examples may be erroneous.

Additionally, the pipeline we used to generate premises, implied entailments, and alternative hypotheses may result in little diversity amongst examples, as generating multiple statements using a single prompt tends to produce similar outputs. Though we attempt to tackle this issue with paraphrasing, our dataset does not wholly capture real-world diversity of language.

At a higher level, implications are subjective, and personal/cultural context plays a large role in how humans understand implied language. Given a premise in INLI, the implied entailment may follow for some readers, but not all. For instance, people from different cultural backgrounds have a different understanding of norms and behaviors, and might disagree with some examples in INLI.

Ethics Statement

The datasets built on in this work, specifically NORMBANK and SOCIALCHEM, focus on social norms/situations and therefore may contain sensitive content and potentially offensive themes. However, we feel is important for LLMs to reason about implications in a diverse variety of settings and situations, so we chose to not filter out these examples.

We also recognize that data produced by large pre-trained language models risks including hallucinations, misinformation, toxicity, and other social harms. As with any generated dataset, INLI is susceptible to these issues.

Additionally, INLI contains names in the vast majority of its premises, which are randomly sampled from the most popular names in the 2011 U.S. Census (Bureau, 2011). Any premises in INLI are entirely fictitious and bear no resemblance to specific people or events in the real world.

References

Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich. 2017. Regularization techniques for fine-tuning in neural machine translation. *CoRR*, abs/1707.09920.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

U.S. Census Bureau. 2011. 2011 census. U.S. Department of Commerce.

Cathrine Damgaard, Paulina Toborek, Trine Eriksen, and Barbara Plank. 2021. "I'll be there for you": The one with understanding indirect answers. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 1–11, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. Rarr: Researching and revising what language models say, using language models.

Elizabeth Jasmi George and Radhika Mamidi. 2020. Conversational implicatures in english dialogue: Annotated dataset. *Procedia Computer Science*, 171:2316–2323.

Deepanway Ghosal, Pengfei Hong, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. Cider: Commonsense inference for dialogue explanation and reasoning.

Paul Grice. 1991. *Studies in the Way of Words*. Harvard University Press.

Siri Gullestad and Bjørn Killingmo. 2019. *The theory and practice of psychoanalytic therapy: Listening for the subtext*. Routledge.

Shreya Havaldar, Matthew Pressimone, Eric Wong, and Lyle Ungar. 2023. Comparing styles across languages.

Mohammad Javad Hosseini, Yang Gao, Tim Baumgärtner, Alex Fabrikant, and Reinald Kim Amplayo. 2024a. Scalable and domain-general abstractive proposition segmentation.

Mohammad Javad Hosseini, Andrey Petrov, Alex Fabrikant, and Annie Louis. 2024b. A synthetic data approach for domain generalization of nli models. *arXiv preprint arXiv:2402.12368*.

Mohammad Javad Hosseini, Filip Radlinski, Silvia Pareti, and Annie Louis. 2023. Resolving indirect referring expressions for entity selection.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models imppressive? learning implicature and presupposition. *arXiv preprint arXiv:2004.03066*.

Donghoon Kim, Yigang Pan, and Heung Soo Park. 1998. High-versus low-context culture: A comparison of chinese, korean, and american cultures. *Psychology & Marketing*, 15(6):507–521.

Zae Myung Kim, David E. Taylor, and Dongyeop Kang. 2023. "is the pope catholic?" applying chain-of-thought reasoning to understanding conversational implicatures.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2024. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36.

Andrew K. Lampinen, Ishita Dasgupta, Stephanie C. Y. Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L. McClelland, Jane X. Wang, and Felix Hill. 2022. Can language models learn from explanations in context?

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation. In *Conference on Empirical Methods in Natural Language Processing*.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. "i'd rather just go to bed": Understanding indirect answers. *arXiv preprint arXiv:2010.03450*.

Andrei Moldovan. 2019. Can entailments be implicatures? *Philosophical Insights into Pragmatics*, pages 43–62.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference.

Zhuang Qiu, Xufeng Duan, and Zhenguang Cai. 2023. Does ChatGPT resemble humans in processing implicatures? In *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop*, pages 25–34, Nancy, France. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Sunny Rai, Khushang Jilesh Zaveri, Shreya Havaldar, Soumna Nema, Lyle Ungar, and Sharath Chandra Guntuku. 2024. Social norms in cinema: A cross-cultural analysis of shame, pride and prejudice.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2024. Can ai-generated text be reliably detected?

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

D. Sperber and D. Wilson. 1986. *Relevance: Communication and Cognition*. Language and thought series. Harvard University Press.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Junya Takayama, Tomoyuki Kajiwara, and Yuki Arase. 2021. DIRECT: Direct and indirect responses in conversational text corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1980–1989, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models.

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.

Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. 2024a. Do large language models understand conversational implicature – a case study with a chinese sitcom.

Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. 2024b. Do large language models understand conversational implicature – a case study with a chinese sitcom.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence?

Fengzhu Zeng and Wei Gao. 2024. Justilm: Few-shot justification generation for explainable fact-checking of real-world claims. *Transactions of the Association for Computational Linguistics*, 12:334–354.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal Common-sense Inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Yuwei Zhang, Siffi Singh, Sailik Sengupta, Igor Shalyminov, Hang Su, Hwanjun Song, and Saab Mansour. 2024. Can your model tell a negation from an implicature? unravelling challenges with intent encoders.

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. Grice: A grammar-based dataset for recovering implicature and conversational reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. Cobra frames: Contextual reasoning about effects and harms of offensive statements. *arXiv preprint arXiv:2306.01985*.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. Normbank: A knowledge bank of situational social norms. *arXiv preprint arXiv:2305.17008*.

A Experiment Details

In this section we provide details on the exact experimental setup used in our experiments in order to aid reproducibility of this work.

A.1 NLI Benchmarks Contain Few Implicatures

To fine-tune the implicitness detection model, we use implied entailments and explicit entailments from the training set of INLI. We fine-tune a T5-XXL model using 3 different learning rates: $1e-6, 5e-6, 1e-5$, as they work well for similar previous experiments (Hosseini et al., 2024b). We fine-tune for 50,000 steps using a batch size of 32 and ultimately select the checkpoint with the highest validation accuracy upon experiment completion for the results reported in Table 2.

We then apply this model to the test sets of SNLI, MNLI, ANLI, and WANLI, and report the % of test set classified as implicit.

Manual validation of classifier. Two authors manually annotate a subset of 80 examples to ensure the model is generalizing correctly. We randomly select 10 labeled implied entailments and 10 labeled explicit entailments from each of our four existing benchmarks to ensure a balanced subset. Given a ⟨premise, hypothesis⟩ pair, annotators are instructed to label the hypothesis as either implicitly entailed or explicitly entailed. We then calculate classifier accuracy by averaging the accuracy with respect to Annotator 1's labels and Annotator 2's label. Accuracies and inter-annotator agreements are given in Table 9.

A.2 NLI Models Rarely Generalize to Implicatures

We fine-tune a different T5-XXL model for every benchmark in Table 3, using the entirety of the training set for each benchmark. We use a learning rate of $5e-6$ and fine-tune for 50,000 steps using a batch size of 32 and select the checkpoint with the highest validation accuracy. We then run the four fine-tuned models on the test set of INLI and report accuracy on the explicit entailment and implied entailment subsets.

A.3 Benchmarking LLMs on INLI

To fine-tune the famliy of T5 models, we use a learning rate of $5e-5, 1e-4, 5e-4, 1e-3$ for XXL, Large, Base, Small respectively. We fine-tune for 50,000 steps using a batch size of 32 for

Dataset	Classifier Accuracy	Agreement (Cohen's Kappa)
SNLI	0.925	0.875
MNLI	0.889	0.780
ANLI	0.917	0.756
WANLI	0.917	0.667
Overall	0.920	0.768

Table 9: Human validation for implicitness detection classifier. We report for what % of labels the annotators agree with the classifier, and the inter-annotator agreement for each dataset.

XXL and 64 for the other models. For each model, we use the checkpoint with the highest validation accuracy. We report results on the full test set of INLI and the implied entailment subset.

To query the larger LLMs, we use an 8-shot prompt (See Appendix E) and use the open-source API corresponding to each model. For all LLMs, temperature is set to 1. Results in Table 6 reflect models queried during August 1st-15th, 2024.

A.4 INLI and Existing NLI Benchmarks

We fine-tune a T5-XXL model on the combination of SNLI, MNLI, ANLI, and WANLI training sets. We use 3 different learning rates: $1e-6, 5e-6, 1e-5$. We fine-tune for 50,000 steps using a batch size of 32 and select the checkpoint with the highest validation accuracy upon experiment completion for the results shown in Table 7. We report results on the test sets of all 4 NLI benchmarks.

Next, we collapse INLI into a 3-label dataset by mapping our implied entailment and explicit entailment labels to "entailed." Lastly, we fine-tune the same model on INLI for 5,000 steps, implementing early stopping as a strategy to prevent forgetting, a phenomenon observed during transfer learning (Barone et al., 2017). We report results once again on the test sets of all 4 NLI benchmarks.

A.5 Generalization Capabilities of INLI

We fine-tune six T5-XXL models as follows: Using the subset of INLI's training set defined in the first column of Table 8, we fine-tune the model with a learning rate of $5e-6$ and a batch size of 32. We then report accuracy on the subset of INLI's test set defined in the second column of Table 8.

Dataset	Premise	Implied Entailment	Cognitive Step
LUDWIG	Diane says, "Would you like to go to a party tonight?" Sophie responds, "I am too tired."	Sophie would prefer not to attend the party this evening.	*Logical Reasoning:* If someone is tired, then they would not attend a social event.
CIRCA	Teresa and Alicia are colleagues who are leaving work on a Friday at the same time. Teresa says, "Do you want to hang out later?" Alicia responds, "I could do with a stiff drink."	Alicia hopes to spend time with Teresa later.	*Conversational Pragmatics:* Alicia is suggesting she and Teresa grab a drink together.
NORMBANK	A hush fell over the elegant restaurant as a group of diners, their laughter escalating hysterically, began to fling mashed potatoes and bread rolls at each other, splattering the white tablecloth and surrounding patrons with food.	The other diners were shocked by the inappropriate food fight.	*World Knowledge:* A hush falls over a room when something shocking is happening.
SOCIALCHEM	With a heavy heart, Maya explained to her sister, Chloe, that she wouldn't be able to donate a kidney, bracing herself for the wave of emotions that were sure to follow.	Maya realized her choice not to donate a kidney would probably devastate Chloe.	*World Knowledge:* When someone asks a family member for an organ donation, they hope the answer is yes.

Table 10: We specifically select datasets with both conversational and situational implications for a diverse set of cognitive deductions that need to be made. The implications in Circa and Ludwig primarily require conversational pragmatics and logical reasoning to deduce, while the implications in SocialChem and NormBank rely on world knowledge. We show a categorization of cognitive deductions for the examples in Table 5.
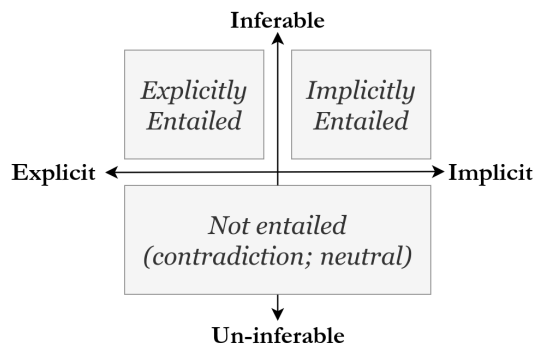

Figure 4: Defining implied entailment. We propose implied entailments should be evaluated via two axes – inferability and implicitness.

B Annotation Study

The six annotators chosen for this task are all experts in NLI, and therefore have an established understanding of entailment, neutrality, and contradiction. Annotation instruction mainly focuses on distinguishing implied entailments from explicit entailments.

Annotators are shown Figure 4 and given the following evaluation scheme to evaluate premises and hypotheses in INLI:

Instructions. Implied entailments should be defined along two axes: inferability and implicitness. An implied entailment must be *inferable* (i.e. able to be reasonably inferred from the premise) and *implicit* (i.e. requiring a cognitive step based on

Label	% Majority Agreement
Explicit entailment	0.96
Implied entailment	0.94
Neutral	0.94
Contradiction	0.90
Overall	0.935

Table 11: Class-wise annotation results. For each class in INLI, we report the % of majority annotations that matched the given label.

implicit shared context to infer). The inferability axis measures to what degree a hypothesis can be supported by a premise. For instance, in Figure 4, hypotheses that can be inferred, either explicitly or implicitly, are entailed. The implicitness axis measures to what extent a hypothesis requires cognitive steps to infer. Hypotheses that are not directly expressed through lexical syntax or semantics are implicit.

The overall majority agreements (i.e. at least 2 out of 3 annotators agree with the given label) for each class are given in Table 11.

C Developing our Prompting Framework

In this section, we will cover the variety of prompting frameworks we experimented with before settling on the framework detailed in Section 4. We will discuss the intuition behind developing each iteration of the framework and what worked/did

Dataset	Accuracy			
	Entailed (Implicit / Explicit)	Neutral	Contradiction	Overall
SNLI (Bowman et al., 2015)	0.706	0.697	0.724	0.709
MNLI (Williams et al., 2018)	0.570	0.661	0.702	0.642
ANLI (Nie et al., 2020)	0.532	0.645	0.401	0.526
WANLI (Liu et al., 2022)	0.501	0.595	0.301	0.516
INLI (ours)	0.684 / 0.727	0.496	0.475	0.595

Table 12: Hypothesis-only baseline. We fine-tune a `T5-XXL` model to label the hypothesis, while entirely excluding the premise. We find the hypothesis-only baseline for our dataset is in line with existing benchmarks, indicating that inference on our dataset does depend on both premise and hypothesis.

not work in regards to the final generated data for each iteration.

All framework iterations below generated hypotheses for LUDWIG, CIRCA, NORMBANK, & SOCIALCHEM101 separately. We also used the same 20 few-shot examples (5 per dataset) across all iterations.

Few-shot examples. In order to determine high-quality hypotheses based on the premises in the few-shot examples, the authors consulted in-house entailment experts. Together, the authors and their colleagues brainstormed what would be good implied entailments, explicit entailments, neutrals, and contradictions based on the premises, and settled on the final examples shown in Appendix E.

C.1 Iteration 1: Separate generation of each hypothesis

The first attempt of our data generation framework was straightforward and simple. In order to ensure each hypothesis was indeed valid with respect to the premise, we thought it best to create a different prompt for each type of hypothesis (i.e. explicit entailment, neutral, and contradiction). An example in our few-shot prompt contained the example premise, followed by the example hypothesis.

What worked. The generated hypotheses mimicked the few-shot examples in length, complexity, and sentence structure.

What didn't work. Many generated hypotheses were not valid with respect to the premise (e.g. a generated "contradiction" was actually neutral based on premise).

C.2 Iteration 2: Adding justifications

In order to improve the validity of the generated hypotheses, we decided to have the model generate

justifications, or explanations of why the hypothesis is indeed contradictory, explicit, etc. Having LLMs provide explanations of their generations (Lampinen et al., 2022) has been shown to increase accuracy and relevance of generations, especially in the domains of inference and claim extraction (Wang and Shu, 2023; Zeng and Gao, 2024). An example in our few-shot prompt contained the example premise, followed by the example hypothesis, and a 1-2 sentence justification of why the hypothesis was validly labeled with respect to the premise.

What worked. The hypotheses were indeed valid with respect to the premise. Incorporating justification significantly increased the correctness of the generated hypotheses.

What didn't work. Amongst the four hypotheses corresponding to the premise, there was a lot of inconsistency. We noticed neutral hypotheses were much shorter and contradictions nearly always had negation.

C.3 Iteration 3: All-at-once generation

We wanted to limit any artifacts in the data that a model might learn that separate hypotheses into classes, irrespective of the premise. To fix the inconsistencies resulting from Iteration 2, we had the model generate all hypotheses at once. An example in our few-shot prompt contained the example premise, followed by the four corresponding hypotheses, and a 1-2 sentence justification for each hypothesis explaining why the hypothesis was validly labeled with respect to the premise.

What worked. The inconsistencies in sentence length and sentence structure were resolved.

What didn't work. We noticed other inconsistencies like word choice (e.g. NORMBANK impli-

catures tended to have words like "inappropriate" or "disrespectful") and theme (e.g. neutral hypotheses tended to be more outlandish).

C.4 Iteration 5: All-at once generation via modification

Our last and final iteration aimed to fix the inconsistencies in word choice and theme of the generated hypotheses. To do this, we instructed `Gemini-Pro` to use the implied entailment as a starting point and modify key words/phrases to transform the implied hypothesis into a contradiction, a neutral, and an explicature.

What worked. The hypotheses were much more homogeneous. Given we wanted to create a challenging benchmark, the homogeneity made it harder for the model to disentangle hypotheses without using the premise to reason.

D Data Robustness

Past work from (Poliak et al., 2018) highlights a desired condition for all NLI datasets: *inference should depend on both premise and hypothesis.* One way to measure whether inference is premise-dependent is to fine-tune a model to label the hypotheses, while entirely excluding the premises.

Table 12 contains the hypothesis-only baseline for our dataset, along with those of other NLI benchmarks. We see our overall hypothesis-only accuracy is in line with other benchmarks, indicating our data does not have many artifacts that may make the task less challenging.

We do see a higher accuracy in identifying explicit entailments; this is expected, as all of our situational implied entailments focus on violating social norms or reasoning about social rules-of-thumb. However, such social constructs are all implied by nature; therefore, none of our situational explicatures focus on social constructs, while the other hypotheses do.

E Prompts

Here, we provide the prompt templates used for all of generation tasks described in this paper. For the INLI generation pipeline, prompts vary for each dataset, we provide the full list of examples used here.[4]

---

[4] https://github.com/google-deepmind/inli/tree/main/Resources/Prompts

E.1 Implicature Augmentation

Prompt for Stage 1 of our pipeline to build INLI: Implicature augmentation. This prompt is used to augment implicature frames from NORMBANK and SOCIALCHEM into premises and implicatures. We use a different set of 5 examples for each dataset.

```
You are an expert in English reading
comprehension.  You also understand social
situations well.  Your task is, given a setting,
behavior, and details about a socially taboo
situation, construct a passage that depicts
the situation being interpreted as taboo in
a real-life setting.  Then, extract a logical
implicature about why the behavior described in
the passage is seen as taboo. An "implicature" is
defined as a statement that can be inferred from
the passage, but is not explicitly stated in the
passage. There are many possible ways to do this
task. Here are a few examples, with explanations.

[-- begin example --]
<Original Data>
<Passage>
<Implicature>
<Justification>
[-- end example --]
```

E.2 Alternative Hypothesis Generation

Prompt for Stage 2 of our pipeline to build INLI: Alternative hypothesis generation. This prompt is used to generate explicit entailments, neutrals, and contradictions, given the passages and implied entailments generated in Stage 1. We use a different set of 5 examples for each dataset.

```
You are an expert in English reading
comprehension.  Given a passage depicting a
conversation/social situation, there are four
possible types of statements that can be extracted
from the passage.

These four statements are: an implied statement,
an explicit statement, a neutral statement, and a
contradictory statement. An "implied statement"
is defined as a statement that can be inferred
from the passage, but is not explicitly stated in
the passage. An "explicit statement" is defined as
a statement that is entailed in the passage, and
can be proven true using the passage as support.
A "neutral statement" is defined as a statement
that is neither entailed nor contradicted by the
```

passage. A neutral statement can be true given the passage, but can also be false given the passage. A "contradictory statement" is defined as a statement that can be directly refuted and proven false using the passage as support.

Given a passage and an implied statement, your task is to generate explicit, neutral, and contradictory statements, using the implied statement as a starting point. You may add, remove, and replace any necessary phrases from the implied statement, so that the generated statements are explicit, neutral, and contradictory with respect to the passage.

Here are few example passages, the four statements corresponding to the passage, and justifications explaining each of the statements:

[-- begin example --]
<Passage>
<Implied statement>
<Explicit statement + justification>
<Neutral statement + justification>
<Contradictory statement + justification>
[-- end example --]

E.3 Benchmarking LLMs on INLI

Prompt to benchmark GPT-4, GPT-4o, Claude-3-Sonnet, and Mistral-Large on INLI. We select 8 random examples from the training subset of INLI, ensuring that these examples include all statement types and datasets.

Given a premise and a hypothesis, your task is to label whether the hypothesis is a valid inference from the premise. Specifically, you will need to assign one of four labels to the hypothesis:
- Implicature: The hypothesis is a valid inference from the passage, but it is NOT explicitly stated in the passage.
- Explicature: The hypothesis is a valid inference from the passage, and it is explicitly stated in the passage.
- Contradiction: The hypothesis is NOT a valid inference from the passage, and is contradicted by the passage.
- Neutral: The hypothesis can neither be supported nor contradicted by the passage.

Here are some examples:

[-- begin examples --]
Premise: Ashley felt a pang of guilt as she accepted a date with Jake, knowing her friend Lisa had briefly dated him a few months prior.
Hypothesis: Ashley is aware that dating Jake might make Lisa, her friend, upset.
Label: Implicature

Premise: Diana wants to know what activities Paula likes to do during weekends. Diana says, "Do you like local theater?" Paula responds, "Only musicals."
Hypothesis: Paula likes local theater, but only musicals.
Label: Explicature

Premise: As the professor dismissed the class, a classmate approached Daniel, but Daniel, eager to get to his next class, rushed past with a quick "excuse me" and a fleeting wave.
Hypothesis: Daniel thinks that socializing during class is inappropriate.
Label: Neutral

Premise: Bryon has just travelled from a different city to meet Rodney. Rodney says, "Do you want some time on your own to explore?" Bryon responds, "I'm happy for company."
Hypothesis: Bryon desires some solitary time for exploration.
Label: Contradiction

Premise: Corrine says, "You're gonna be seeing a lot of me. You're sure you don't mind?" Russel responds, "It's good what you're doing."
Hypothesis: Russel is fine with seeing a lot of Corrine.
Label: Implicature

Premise: A bead of sweat trickled down Edgar's temple as he finished mopping the floor by the ellipticals. He paused, catching his breath just as Mr. Robinson, the gym's quiet janitor, awkwardly swung his leg over the machine and began slowly pedaling.
Hypothesis: Edgar observed Mr. Robinson, the gym custodian, clumsily mount the exercise bike and start pedaling at a leisurely pace.
Label: Explicature

32289

Premise: Frieda says, "Have you had a chance to wear your new shirt yet?" Genevieve responds, "I have been trying to exchange it for a larger size."
Hypothesis: Genevieve likes to wear shirts that are pre-owned.
Label: Neutral

Premise: A group of teenagers giggled as they snapped photos of a family struggling to handle their overexcited toddler, their laughter echoing through the primate house. Dr. Lewis, the zoo's veterinarian, frowned, shaking his head as he walked past.
Hypothesis: Dr. Lewis smiled as the teenagers photographed unsuspecting strangers at the zoo.
Label: Contradiction
[— end examples —]

[your task]
Given a premise and a hypothesis, your task is to label the hypothesis as one of the four labels: Implicature, Explicature, Contradiction, or Neutral. Your response should be only one word, the name of the label.
Premise:{}
Hypothesis:{}
Label: