Commonsense Knowledge with Negation: A Resource to Enhance Negation Understanding

Zijie Wang[1] **MohammadHossein Rezaei**[1] **Farzana Rashid**[2] **Eduardo Blanco**[1]
[1]University of Arizona [2]University of North Carolina Asheville
{zijiewang, mhrezaei, eduardoblanco}@arizona.edu frashid@unca.edu

Abstract

Negation is a common and important semantic feature in natural language, yet Large Language Models (LLMs) struggle when negation is involved in natural language understanding tasks. Commonsense knowledge, on the other hand, despite being a well-studied topic, lacks investigations involving negation. In this work, we show that commonsense knowledge with negation is challenging for models to understand. We present a novel approach to automatically augment existing commonsense knowledge corpora with negation, yielding two new corpora containing over 2M triples with *if-then* relations. In addition, pre-training LLMs on our corpora benefits negation understanding.

1 Introduction

Negation is a common and important semantic feature in natural language, appearing in approximately 25% of English sentences (Hossain et al., 2020). Despite the recent success of large language models (LLMs) across various natural language processing tasks (Achiam et al., 2023; Touvron et al., 2023), their understanding of negation remains unclear. Previous work has demonstrated that language models struggle with multiple natural language understanding tasks when negation is involved (Dobreva and Keller, 2021; Jang et al., 2022). However, these investigations have been limited to encoder-based models such as BERT (Devlin et al., 2019) and earlier LLMs such as GPT-3 (Brown et al., 2020).

Commonsense knowledge has been extensively studied, with numerous efforts focused on building commonsense knowledge bases (Speer et al., 2017; Sap et al., 2019). Commonsense reasoning has also been investigated through tasks such as question answering (Talmor et al., 2019). Despite these extensive efforts, the intersection of commonsense knowledge and negation remains underexplored.

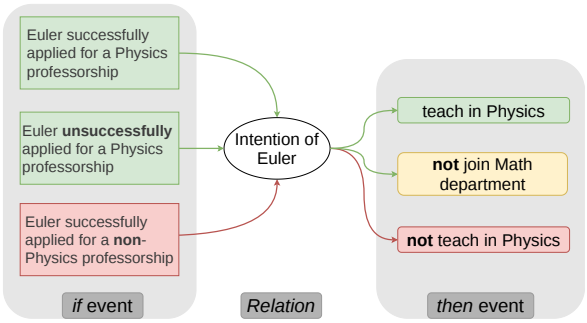

Figure 1: A commonsense knowledge triple with the *Intention* relation. Negations are added to both *if* and *then* events. Adding different negation cues results in new triples that align (same color on both sides) or conflict with (different colors) commonsense knowledge.

ATOMIC (Sap et al., 2019) is one of the largest commonsense knowledge corpora with *if-then* relations, representing commonsense knowledge as a triple <A, R, B>, where A represents the *if* event, R represents the relation, and B represents the *then* event. For example, <*the person is hungry*, *the person wants to*, *eat food*> represents the commonsense knowledge that "*if the person is hungry, then the person wants to eat food.*" To our knowledge, ANION (Jiang et al., 2021) is the only work that develops a commonsense knowledge base with negation, built on top of ATOMIC. However, their approach only negates the *if* event and generates a new *then* event via human annotation, resulting in a new triple <$\neg A$, R, B'>. This is limited because it does not consider negating the *then* event and requires significant human annotation effort.

In this work, we present a novel approach to automatically augment existing commonsense knowledge corpora with negation. Our method is motivated by the observation that negating either the *if* event, the *then* event, or both can sometimes produce new triples that still align with commonsense knowledge. As shown in the example in Figure 1, negating the *if* event *"Euler **unsuccess-***

Findings of the Association for Computational Linguistics: ACL 2026, pages 11915–11934

fully *applied for a position in Physics"* implies that *the intention of Euler* was to *"teach in Physics."* In contrast, negating the *then* event *"**not** teach in Physics"* yields a triple that conflicts with commonsense knowledge (green *if* event and red *then* event). This observation motivates us to augment both *if* and *then* events with negation, expanding existing commonsense knowledge corpora by up to three times (negating the *if* event, the *then* event, or both). In addition, our approach avoids relying on human effort to generate new events such as *"**not** join Math department"*. We further propose an automatic validation method to verify whether the new commonsense triples with negation align with commonsense knowledge. More importantly, we demonstrate that they benefit LLMs' understanding of negation in general-purpose tasks. The main contributions of this paper are:[1]

- A novel approach to develop two commonsense knowledge corpora with over 2M triples containing negation.
- An automatic method to validate commonsense triples with negation.
- Evaluation on multiple models across five benchmarks, demonstrating that our corpora improve LLMs' understanding of negation.

2 Related Work

Commonsense knowledge bases (CSKB) have been developed across multiple works with different focuses. ConceptNet (Speer et al., 2017) represents taxonomic commonsense knowledge as a graph, where each concept (i.e., word or phrase) is connected by a relation. For example, *a net* is used for *catching fish*. Sap et al. (2019) focus on relations between events rather than concepts, proposing nine *if-then* relations representing inferential knowledge. For example, if *X pays Y a compliment*, then Y wants to *return the compliment*. To our knowledge, ANION (Jiang et al., 2021) is the only work that investigates commonsense knowledge with negation. It is built by negating the *if* event from ATOMIC and manually annotating a new *then* event. For example, if *X does not pay Y a compliment*, then the effect on Y is *upset*. This approach limits the possibility of negating the *then* event and requires significant human annotation effort. Arnaout et al. (2022) develop a method to identify informative negations of commonsense concepts. The dataset complements existing CSKB as they often do not capture any negative concepts such as "gorillas are not territorial." The UNcommonsense dataset (Zhao et al., 2024) collects data involving unusual, unexpected, and unlikely situations, namely uncommonsense knowledge. The primary task asks LLMs to generate reasonable explanations given contexts with uncommon outcomes. Beyond manual efforts to create CSKB, Bosselut et al. (2019) propose COMET, a system that trains a transformer model to automatically generate the *then* event given the *if* event and relation. West et al. (2022) distill commonsense knowledge from LLMs to train a transformer for commonsense graph generation. Unlike existing works, we propose a novel approach to incorporate negation into existing CSKB. Our approach generates large-scale commonsense knowledge triples with negation while requiring no human effort.

Commonsense Knowledge for Downstream Tasks Commonsense knowledge has been shown to benefit many downstream tasks. Talmor et al. (2019) present a question answering dataset called CommonsenseQA that involves reasoning over commonsense knowledge. They develop multiple-choice questions based on the concept-relation graph from ConceptNet. Lal et al. (2022) observe that answering why-questions often requires commonsense reasoning and leverage COMET to generate relevant commonsense knowledge for this task. Guan et al. (2020) pre-train a Transformer model on external commonsense knowledge bases to improve story generation. In this work, we leverage augmented commonsense knowledge corpora with negation to improve LLMs' understanding of negation, with experimental results across three tasks demonstrating its effectiveness.

Improving Models' Negation Understanding Prior work has shown that LLMs struggle with natural language understanding tasks involving negation (Dobreva and Keller, 2021; Jang et al., 2022), motivating efforts to address this limitation. Hosseini et al. (2021) leverage unlikelihood training and synthetic data generation to improve BERT's understanding of negation. Singh et al. (2023) modify BERT's next sentence prediction task by incorporating negation cues rather than random sentences. Rezaei and Blanco (2024) show that incorporating affirmative interpretations improves performance on negation understanding benchmarks. In this work, we demonstrate that state-of-the-

[1]Code and dataset available at https://github.com/wang-zijie/commonsense_with_negation

art LLMs still lack robust negation understanding and develop two commonsense knowledge corpora augmented with negation. More importantly, pre-training on our corpora improves performance on multiple tasks requiring negation understanding.

3 Augmenting Commonsense Knowledge Triples with Negation

We propose a novel approach to automatically augment existing commonsense knowledge corpora with negation. Building on ATOMIC (Sap et al., 2019) and ANION (Jiang et al., 2021), we develop two new commonsense knowledge corpora: ¬ATOMIC and ¬ANION. We further introduce an automatic validation method that categorizes the generated commonsense triples as either *Valid* (aligned with commonsense knowledge), *Invalid* (conflicting with commonsense knowledge), or *Ambiguous* (neither). Our resulting corpora contain over 2M commonsense triples with negation.

Negating Commonsense Knowledge Triples As discussed in Section 2, ATOMIC consists of commonsense knowledge triples with *if-then* relations connecting two events. ANION extends ATOMIC by negating the *if* event and manually annotating a new *then* event; however, it does not consider negating the *then* event. Unlike these approaches, we incorporate negation into the *if* event, the *then* event, and both. Specifically, given a commonsense triple <A, R, B> from ATOMIC, where A is the *if* event, R is the relation, and B is the *then* event, we leverage an LLM to add the logical negation cue *not* to A, B, and both, generating three new triples: <$\neg A$, R, B>, <A, R, $\neg B$>, and <$\neg A$, R, $\neg B$>. For a commonsense triple <$\neg A$, R, B'> from ANION, where B' is a new *then* event distinct from B, we negate only B' to generate <$\neg A$, R, $\neg B'$>. We do not negate $\neg A$, as doing so would yield the same *if* event as in ATOMIC. Note that we use only the training split to generate these triples.

To perform negation, we add the negation cue *not* to (1) the main verb of the event (e.g., "the person **does not** take a picture", "**not** look at the picture"), or (2) the modifier when the verb is absent (e.g., "**not** excited"). Using manually curated exemplars, we prompt Llama 3.1 70B to generate the negated events. A manual evaluation of 200 instances confirms 99% syntactic correctness.

3.1 Validating New Triples

It remains unknown whether the automatically generated triples with negation align with or conflict with commonsense knowledge. We use *Valid* and *Invalid* to denote these cases, respectively. In addition, we observe two scenarios when the triples are considered *Ambiguous*: (1) the validity is context-dependent—for example, *"If the sun is not shining, then it is daytime"* can be either *Valid* or *Invalid* because *"the sun is not shining"* could indicate heavy clouds during daytime or simply nighttime; (2) the triple lacks a clear causal connection and thus has ambiguous semantics—for example, *"If Person A does not get a gift, then it causes Person A to feel joy."*

Recent work leverages LLM-as-a-judge to evaluate LLM outputs such as synthetic datasets (Li et al., 2024). However, we demonstrate that state-of-the-art LLMs, including GPT-4o and Claude Sonnet 4, lack the ability to evaluate the validity of commonsense knowledge triples with negation. We address this limitation by training a task-specific LLM to automatically validate commonsense triples.

We train LLMs using supervised fine-tuning with three types of data: (1) *Valid* triples from existing commonsense corpora (ATOMIC and ANION), (2) *Ambiguous* triples created by randomly combining *if* events, *then* events, and relations from existing triples, similar to (Fang et al., 2022), and (3) *Invalid* triples constructed by prompting GPT-4o to generate *then* events from existing *if* events. Note that the training data are considered noisy without any manual inspection. For evaluation, we construct a benchmark comprising 200 triples per relation type from ATOMIC along with their three negated variations (7,200 triples total). We recruit two annotators to label each triple as *Valid*, *Invalid*, or *Ambiguous*, achieving an inter-annotator agreement of 0.62, indicating substantial agreement (Artstein and Poesio, 2008). Note that the training and evaluation data do not overlap, as they are sampled from the original training and test splits, respectively. Appendix A provides more details on benchmark creation and annotation.

We experiment with Llama 3.1 8B and 70B using three training data variations—differing only in whether *Valid* and *Ambiguous* triples come from ATOMIC, ANION, or both; *Invalid* triples are always synthesized via an LLM. Regardless of the training corpus source, the models are trained on

	Valid			Invalid			Ambiguous			Overall			
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Acc
Few-shot learning													
Llama 3.1 8B	0.56	0.58	0.57	0.40	0.71	0.52	0.54	0.07	0.13	0.45	0.45	0.37	0.44
Llama 3.1 70B	0.73	0.54	0.62	0.50	0.73	0.59	0.48	0.38	0.43	0.56	0.55	0.53	0.53
GPT-4o	0.71	0.48	0.51	0.54	0.65	0.59	0.51	0.54	0.51	0.53	0.54	0.52	0.54
Claude Sonnet 4	0.83	0.38	0.53	0.51	0.63	0.56	0.46	0.64	0.54	0.61	0.56	0.56	0.56
Fine-tuning													
Llama 3.1 8B	0.57	0.80	0.67	0.67	0.48	0.56	0.53	0.45	0.48	0.58	0.57	0.55	0.56
Llama 3.1 70B	**0.70**	0.76	0.73	**0.79**	0.48	0.59	0.51	0.68	0.58	0.65	0.63	0.63	0.64

Table 1: Results of validating commonsense triples with negation via (1) few-shot learning with Llama 3.1 8B, 70B, GPT-4o, and Claude Sonnet 4, and (2) fine-tuning Llama 3.1 8B and 70B. We report Precision (P), Recall (R), and F1 for each label and overall, along with overall Accuracy (Acc). Only the best fine-tuning results are reported across variants of whether *Valid* and *Ambiguous* training instances come from ATOMIC, ANION, or both; *Invalid* training data is synthesized using GPT-4o. Full results are reported in Appendix B.1.

Dataset	# Triples (%)	# w/o Neg. (%)	# with Negation (%)		
		$\langle A, R, B\rangle$	$\langle \neg A, R, B\rangle$	$\langle A, R, \neg B\rangle$	$\langle \neg A, R, \neg B\rangle$
Existing Corpora					
ATOMIC*	449k	449k	—	—	—
ANION*	142k	—	142k	—	—
Our Corpora					
¬ATOMIC	1,798k (100.0)	449k (100.0)	449k (100.0)	449k (100.0)	449k (100.0)
Valid	681k (37.9)	376k (83.7)	47k (10.5)	42k (9.2)	216k (48.0)
Invalid	463k (25.8)	8k (2.0)	128k (28.5)	286k (63.6)	41k (9.1)
Ambiguous	652k (36.3)	64k (14.3)	274k (61.0)	122k (27.2)	192k (42.9)
¬ANION	285k (100.0)	—	142k (100.0)	—	142k (100.0)
Valid	104k (36.4)	—	77k (53.9)	—	27k (18.9)
Invalid	46k (16.1)	—	8k (5.6)	—	38k (26.6)
Ambiguous	135k (47.5)	—	58k (40.5)	—	77k (54.5)
Benchmark (¬ATOMIC)	7,200 (100.0)	1,800 (100.0)	1,800 (100.0)	1,800 (100.0)	1,800 (100.0)
Valid	2,329 (32.3)	1,287 (71.5)	113 (6.3)	118 (6.5)	811 (45.1)
Invalid	2,150 (29.9)	41 (2.3)	823 (45.7)	1,184 (65.8)	102 (5.6)
Ambiguous	2,721 (37.8)	472 (26.2)	864 (48.0)	498 (27.7)	887 (49.3)

Table 2: Statistics of commonsense triples in existing corpora, our corpora, and the benchmark. ATOMIC contains no triples with negation, and ANION negates only the *if* event. * denotes subsets of the datasets: training splits from each corpus, excluding underspecified triples from ATOMIC and using only the logical negation split from ANION.

5,400 training triples (200 triples per relation per label) using QLoRA (Dettmers et al., 2023) with 4-bit quantization. Both training and evaluation data are verbalized from commonsense triples to natural language if-then statements. Further experimental details including the verbalization mapping can be found in Appendix B.

Table 1 reports the validation results on our benchmark. Claude Sonnet 4 outperforms the other LLMs, though it achieves only 0.56 F1 score overall. Fine-tuned Llama 3.1 70B model outperforms all proprietary models (F1 score: 0.63 vs. 0.56; Accuracy: 0.64 vs. 0.56). Llama 3.1 8B shows lower performance, as expected for a smaller model. We consider precision for *Valid* and *Invalid* triples the more critical metric, as only triples with these two labels are used for training (Section 4). Lower recall is tolerable since it only reduces training set size. Our best LLM judge achieves 0.70 and 0.79 precision for *Valid* and *Invalid* triples, respectively. Moreover, empirical experiments (Section 5) demonstrate that the commonsense corpora synthesized using our LLM judge effectively improve LLMs' negation understanding.

3.2 Analysis of Commonsense Knowledge Corpora with Negation

We develop two new corpora with negation, ¬ATOMIC and ¬ANION, which are validated by our LLM judge (Table 1, the last row) with three labels, *Valid*, *Invalid*, and *Ambiguous*. Table 2 presents statistics for the existing corpora (ATOMIC

and ANION), our corpora, and our benchmark.

We use only a subset of the existing corpora: the training split, excluding underspecified triples (e.g., PersonX sees ___ in the water) from ATOMIC and using only the logical negation split from ANION. Our approach augmenting commonsense triples with negation is effective: 37.9% of the commonsense triples from ¬ATOMIC (36.4% from ¬ANION) are identified as *Valid*, aligning with commonsense knowledge, while fewer triples are identified as *Invalid* (¬ATOMIC: 25.8%; ¬ANION: 16.1%). More importantly, they are augmented with negation. As we demonstrate in Section 5, both *Valid* and *Invalid* triples improve LLMs' ability to understand negation.

4 Commonsense Knowledge with Negation for Downstream Tasks

We have presented a novel method to automatically augment existing commonsense knowledge corpora with negation. Our two corpora, ¬ATOMIC and ¬ANION, contain over 2M commonsense knowledge triples augmented with negation, validated by an LLM judge. Beyond this contribution, we demonstrate that our corpora improve LLMs' ability to understand negation. Specifically, we evaluate on five benchmarks across three tasks: question answering (QA), natural language inference (NLI), and information retrieval (IR).

4.1 Benchmarks Evaluating LLMs' Negation Understanding

CondaQA (Ravichander et al., 2022) is a contrastive QA dataset requiring understanding of negation cues in passages to answer questions. Each question is paired with a passage containing the answer, with answers being either *Yes*, *No*, *Don't know*, a span in the question, or a span in the context. Following Ravichander et al. (2022), we evaluate CondaQA using two metrics: *accuracy* and *group consistency*. The term *group* refers to the original passage and either all three or one of the edited passages. *Group consistency* measures the percentage of questions answered correctly for all passages in a group, which is arguably more important than accuracy, as robustness against negation requires correctly answering questions with both original and negated passages.

NLI with Negation is introduced by Hossain et al. (2020). It contains three NLI benchmarks developed from existing benchmarks: RTE (Dagan et al., 2005), SNLI (Bowman et al., 2015), and MNLI (Williams et al., 2018). The authors show that existing NLI benchmarks contain few negation cues and develop new benchmarks by negating the main verbs in premises and hypotheses to create three new pairs from each original pair. The new pairs are manually annotated to obtain labels.

NevIR (Weller et al., 2024) addresses the weakness of neural information retrieval (IR) models in understanding negation. The dataset is constructed using contrastive query-document pairs from CondaQA, where each pair of queries and documents is nearly identical except for a crucial negation. An IR model is expected to rank documents based on queries by correctly understanding negation. Pairwise accuracy serves as the evaluation metric: the model must correctly rank documents for both queries, flipping the ranking when given the negated query. Although the dataset's primary purpose is to evaluate IR model performance, it is essentially a binary classification task, as only two documents are provided for each query. Still, it is considered a challenging benchmark as models are required to understand negation from long context (documents).

4.2 Training LLMs with Commonsense Knowledge with Negation

Although over half of our generated triples are identified as *Invalid* (Table 2), meaning they conflict with commonsense knowledge, they remain useful for teaching LLMs to understand negation. Specifically, we design a training objective that enables models to learn from both *Valid* and *Invalid* triples, and how negating the *if* or *then* event affects triple validity. We construct training corpora by selecting contrastive triples from ¬ATOMIC following the patterns below. We select triples if:

- the original triple <*A*, *R*, *B*> is *Valid*, <¬*A*, *R*, *B*> is *Invalid*, and <*A*, *R*, ¬*B*> is *Valid*;
- the original triple <*A*, *R*, *B*> is *Valid*, <¬*A*, *R*, *B*> is *Valid*, and <*A*, *R*, ¬*B*> is *Invalid*;
- the original triple <*A*, *R*, *B*> is *Invalid*, <¬*A*, *R*, *B*> is *Valid*, and <*A*, *R*, ¬*B*> is *Invalid*; or
- the original triple <*A*, *R*, *B*> is *Invalid*, <¬*A*, *R*, *B*> is *Invalid*, and <*A*, *R*, ¬*B*> is *Valid*.

For ¬ANION, we select triples where the original and negated triples have different labels. This approach balances the distribution of *Valid* and *Invalid* triples and, more importantly, highlights two patterns for models to compare: (1) negating either the *if* or *then* event keeps the new triple's validity

	# Params.	Accuracy (Δ%)	Group Consistency			
			All	Par.	Sco.	Aff.
Fully supervised						
UnifiedQA-v2-large (Ravichander et al., 2022)	770M	66.7	30.2	64.0	43.7	46.5
RoBERTa-large	355M	64.9	29.6	61.9	41.4	45.8
Pre-trained with						
ATOMIC + ANION		67.0 (+3.2)	32.9	65.6	46.3	48.1
Best of (¬ATOMIC, ¬ANION)		**68.5** (+5.5)	**34.3**	**66.4**	**47.6**	**50.1**
In-context learning						
Zero-shot						
OpenAI o1	—	65.3	24.9	67.4	43.8	38.6
Few-shot						
InstructGPT + COT (Ravichander et al., 2022)	—	66.3	27.3	64.2	45.1	44.9
GPT-4o	—	72.9	34.4	78.7	52.6	47.9
Llama 3.1 70B	70B	77.5	43.3	83.7	61.8	54.9
Llama 3.1 8B	8B	68.7	31.5	69.0	48.1	44.4
Pre-trained with						
ATOMIC + ANION		67.6 (-1.6)	27.5	69.5	44.3	41.4
Best of (¬ATOMIC,¬ANION)		**71.5** (+4.1)*	**33.5**	**72.9**	**48.7**	**46.4**
Qwen2 7B	7B	65.1	24.9	65.9	39.7	39.2
Pre-trained with						
ATOMIC + ANION		64.1 (-1.5)	22.8	65.2	38.4	37.7
Best of (¬ATOMIC, ¬ANION)		69.7 (+7.1)*	32.7	71.2	46.3	45.5

Table 3: Results evaluating CondaQA with two settings: fully supervised (top) and in-context learning (bottom). The best result for each model is in bold. Delta (Δ) indicates the percent change in accuracy compared to the off-the-shelf model. An asterisk * indicates a statistically significant improvement (McNemar's test (McNemar, 1947), $p < 0.05$) over both the off-the-shelf model and the model trained with existing corpora.

(either *Valid* or *Invalid*), and (2) negating either the *if* or *then* event flips the new triple's validity. We do not include <¬*A*, *R*, ¬*B*> as double negations result in more complex semantics. The final training set consists of 89k triples from ¬ATOMIC and 76k triples from ¬ANION. Finally, the model is trained to predict whether a triple is *Valid* or *Invalid*, using the validation results (Section 3.1) as ground truth.

As a baseline, we also construct a training dataset without augmented negation. This is done by sampling the commonsense triples from ATOMIC and ANION as *Valid* instances, while reusing the *Invalid* triples synthesized by an LLM (Section 3.1). Although ANION contains triples with negated *if* events, they differ significantly from our corpora as they do not preserve the original *then* event. Note that we always sample the same number of triples from ATOMIC and ANION as ¬ATOMIC and ¬ANION.

5 Experiments

We evaluate our corpora on negation understanding using an encoder-based model (RoBERTa-large) and two LLMs (Llama 3.1 8B and Qwen2 7B) across three tasks: (1) CondaQA, a QA task (Section 5.1); (2) NLI with Negation, an NLI task (Section 5.2); and (3) NevIR, an IR task (Section 5.3).

All three models are first pre-trained on our commonsense corpora, ¬ATOMIC, ¬ANION, or both. As a baseline, models are pre-trained on existing corpora (ATOMIC and ANION). We then evaluate on downstream tasks in two settings: (1) fully-supervised fine-tuning on the downstream task's training split for encoder-based models, and (2) zero-shot and few-shot in-context learning for LLMs. Following standard practices, we use a zero-shot prompt with the OpenAI o1 model and a few-shot prompt with GPT-4o and open-source LLMs. Pre-trained LLMs are trained and evaluated locally using QLoRA (Dettmers et al., 2023) due to computational resource limitations. Appendix C reports experimental details including the prompts and hyperparameters.

5.1 Evaluating with CondaQA

Table 3 reports results on CondaQA using two metrics: *accuracy* and *group consistency* (Ravichander et al., 2022). For pre-trained models, we report only the best results among three corpus configurations. Appendix D.1 provides the complete results.

RoBERTa-large benefits from pre-training on both existing and our commonsense corpora, with our corpora yielding higher performance. Pre-training on our corpora also outperforms

	# Params.	NLI with Negation			NevIR
		RTE-Neg	SNLI-Neg	MNLI-Neg	
Fully supervised					
BERTNOT (Hosseini et al., 2021)	110M	74.5	46.0	60.9	—
RoBERTa-large-NSP (Rezaei and Blanco, 2025)	355M	87.2	56.5	69.9	—
stsb-roberta-large (Weller et al., 2024)	355M	—	—	—	24.9
MonoT5 3B (Nogueira et al., 2020)	3B	—	—	—	50.6
RoBERTa-large	355M	84.7	56.0	**69.9**	24.5
Pre-trained with					
ATOMIC + ANION		86.2	56.5	69.4	29.1
Best of (¬ATOMIC, ¬ANION)		**88.1***	**58.3**	69.7	**34.3***
In-context learning					
Zero-shot					
OpenAI o1	—	87.5	75.9	74.6	59.7
Few-shot					
GPT-4o	—	86.9	74.8	75.0	61.7
Llama 3.1 70B	70B	78.9	69.1	65.9	58.8
Llama 3.1 8B	8B	60.0	54.4	47.0	30.6
Pre-trained with					
ATOMIC + ANION		65.3	57.5	51.1	37.9
Best of (¬ATOMIC, ¬ANION)		81.3*	68.2*	63.9*	**42.2***
Qwen2 7B	7B	71.7	60.7	59.5	29.8
Pre-trained with					
ATOMIC + ANION		78.3	66.5	63.0	33.8
Best of (¬ATOMIC, ¬ANION)		**82.3***	**72.4***	**67.5***	39.8*

Table 4: Results evaluating three NLI benchmarks with negation cues (accuracy) and NevIR (pairwise accuracy) in two settings: (1) fully supervised (top) and (2) in-context learning (bottom). The best results for each model are in bold. An asterisk * indicates a statistically significant improvement (McNemar's test (McNemar, 1947), $p < 0.05$) over both the off-the-shelf model and the model trained with existing corpora.

UnifiedQA-v2-large (Ravichander et al., 2022), despite the latter being a larger model (770M vs. 355M). Surprisingly, proprietary models (OpenAI o1 and GPT-4o) yield worse results than the less powerful open-source Llama 3.1 70B (65.3 vs. 72.9 vs. 77.5). We hypothesize that OpenAI o1 with a few-shot prompt may achieve higher performance, though at significantly greater token cost. The two LLMs pre-trained on our corpora consistently outperform the base model, with Llama 3.1 8B outperforming Qwen2 7B. Pre-trained Llama 3.1 8B even achieves competitive performance with GPT-4o (71.5 vs. 72.9). More importantly, most improvements from our corpora are statistically significant over both the off-the-shelf baseline and models trained with existing corpora (indicated with * in Table 3). In contrast, pre-training on existing corpora yields worse results than the off-the-shelf baseline. Note that the baseline with existing corpora uses a single configuration: *Valid* triples from ATOMIC and ANION, and synthesized *Invalid* triples.

Importantly, pre-training on our corpora does not degrade performance with non-negation instances. The Affirmative Edit setting in CondaQA (column Aff.) requires the model to reason over passages with negation removed, and we observe no drop in performance, indicating that pre-trained models remain capable of handling affirmative text. Appendix D.2 provides extra results with CommonsenseQA (Talmor et al., 2019), a standard commonsense benchmark, further confirming this finding.

5.2 Evaluating with NLI Benchmarks

We further evaluate on three NLI benchmarks with negation: RTE-Neg, SNLI-Neg, and MNLI-Neg. We include additional fully-supervised baselines: BERTNOT (Hosseini et al., 2021), RoBERTa-large-NSP (Rezaei and Blanco, 2025), stsb-roberta-large (Weller et al., 2024), and MonoT5 3B (Nogueira et al., 2020). Table 4 (NLI with Negation) reports results using *accuracy*. Pre-training RoBERTa-large on our corpora yields the best results among all models and outperforms all baselines. However, only one result yields statistically significant improvement over the off-the-shelf baseline. We hypothesize this is due to potential overfitting—models are further fine-tuned on each benchmark's training split when the tasks are as simple as classification. Moreover, the off-the-shelf

	CondaQA (Accuracy)	NLI with Negation			NevIR
		RTE-Neg	SNLI-Neg	MNLI-Neg	
Llama 3.1 8B	68.7	60.0	54.4	47.0	30.6
Pre-trained with					
¬ATOMIC	70.6	81.2	**68.5**	63.8	38.1
<¬*A*, *R*, *B*>	69.3	76.8	64.0	53.5	34.3
<*A*, *R*, ¬*B*>	67.6	74.0	65.5	54.6	35.6
¬ANION	**71.3**	**81.3**	68.2	**63.9**	**42.2**
<¬*A*, *R*, ¬*B*>	68.7	70.2	62.9	51.6	35.3
Qwen2 7B	65.1	71.7	60.7	59.5	29.8
Pre-trained with					
¬ATOMIC	**69.7**	**82.3**	**72.4**	**67.5**	**39.8**
<¬*A*, *R*, *B*>	65.8	74.1	65.5	65.1	35.8
<*A*, *R*, ¬*B*>	64.3	75.7	68.3	62.4	34.4
¬ANION	69.2	78.9	66.1	62.5	36.3
<¬*A*, *R*, ¬*B*>	65.4	71.0	54.3	52.5	30.1

Table 5: Ablation results evaluating CondaQA, NLI, and NevIR, with pre-trained LLMs on different types of negations: *if* event, *then* event, or both.

models already achieve strong results.

In-context learning with LLMs shows more expected trends. The OpenAI o1 model achieves the highest results overall, followed by GPT-4o. LLMs pre-trained on our corpora demonstrate statistically significant improvements over both the off-the-shelf baseline and models pre-trained on existing corpora across all three benchmarks. Notably, Qwen2 7B even outperforms the larger Llama 3.1 70B model. Unlike CondaQA, pre-training on existing corpora also yields benefits, despite being significantly smaller than our corpora. Note that the existing corpora include negated triples from ANION, which benefit models' negation understanding on the simpler NLI task.

5.3 Evaluating with NevIR

Following Weller et al. (2024), we perform fully-supervised fine-tuning with RoBERTa-large on STS-B (Cer et al., 2017) instead of NevIR's training split. Table 4 (NevIR) reports the results using pairwise accuracy—the model needs to correctly rank documents for both queries (with and without negation). Although NevIR is a retrieval task, it only requires ranking between two documents for each query and essentially becomes binary classification. In fact, RoBERTa-large pre-trained on our corpora outperforms a customized IR model baseline (stsb-roberta-large (Weller et al., 2024), 34.3 vs. 24.9). Again, pre-training on our corpora yields statistically significant improvements over off-the-shelf models and models trained with existing corpora. The results are consistent across all three models. All our models significantly underperform OpenAI o1, GPT-4o, and even MonoT5 3B (Nogueira et al., 2020). We hypothesize that pre-training a retrieval model with our corpora as a starting point would yield greater benefits.

5.4 Ablation Studies

Pre-training with Individual Negation Types We further investigate whether training with individual negation types contributes differently to performance improvements. For ¬ATOMIC, we compare the complete corpus against training with triples that only negate the *if* event (<¬*A*, *R*, *B*>) or the *then* event (<*A*, *R*, ¬*B*>). For ¬ANION, we compare against training with triples that negate both events (<¬*A*, *R*, ¬*B*>).

Table 5 reports the results with two LLMs. Note that results with the complete corpus are not directly comparable to Table 3 and 4, which report only the best configuration. Training with individual negation types consistently underperforms the complete corpus across all benchmarks, demonstrating that the patterns within our corpora are critical (Section 4.2). Notably, negating the *if* event alone yields higher results than the off-the-shelf model, suggesting that certain negation types still provide partial benefits for negation understanding.

For ¬ANION, training with triples negating both events yields substantially worse performance than the complete corpus. We hypothesize that double negation alone is insufficient, as models benefit from first learning simpler single-negation patterns before generalizing to more complex negations.

	# Triples	CondaQA (Accuracy)	NLI with Negation			NevIR
			RTE-Neg	SNLI-Neg	MNLI-Neg	
Llama 3.1 8B	n/a	68.7	60.0	54.4	47.0	30.6
Pre-trained with ¬ATOMIC + ¬ANION	1K + 1K	67.7	70.8	57.1	53.5	31.8
	10K + 10K	69.3	81.0	67.7	62.9	34.5
	89K + 76K	71.5	81.3	67.9	63.7	38.2
Qwen2 7B	n/a	65.1	71.7	60.7	59.5	29.8
Pre-trained with ¬ATOMIC + ¬ANION	1K + 1K	65.2	76.0	65.2	63.1	34.9
	10K + 10K	66.6	81.7	72.2	67.4	36.0
	89K + 76K	66.6	82.1	72.2	67.5	38.5

Table 6: Ablation on training dataset size. We train Llama 3.1 8B and Qwen2 7B with two subsets of ¬ATOMIC + ¬ANION: (1) 2K triples (1K from each dataset) and (2) 20K triples (10K each). Training with the 2K subset yields significantly worse performance than the complete dataset, while training with the 20K subset shows substantial improvement and tends to saturate on simpler tasks such as NLI.

Pre-training with Different Data Sizes Our full training dataset consists of 89k triples from ¬ATOMIC and 76k triples from ¬ANION (Section 4.2). We conduct two ablations on the effect of training data size on downstream task performance: (1) training with a 2,000-triple subset (randomly sampled and balanced with 1,000 triples each from ¬ATOMIC and ¬ANION), and (2) training with a 20,000-triple subset (10,000 each). Table 6 reports the results compared to off-the-shelf models and models trained with the complete dataset. Similarly, Table 3 and Table 4 only report the best result among ¬ATOMIC, ¬ANION, or both. Appendix D reports comparable results with [¬ATOMIC + ¬ANION].

Training with only 2K triples already improves over the off-the-shelf models on NLI tasks, though the gains are modest. Scaling to 20K triples yields substantial improvements, with NLI performance approaching that of the full dataset (e.g., RTE improves from 70.8 to 81.0 for Llama 3.1 8B, compared to 81.3 with the full training set). However, NevIR continues to benefit from additional training data beyond 20K (34.5 → 38.2 for Llama 3.1 8B, 36.0 → 38.5 for Qwen2 7B), suggesting that more complex negation reasoning tasks benefit from larger training sets.

Pre-training with Randomly Labelled Data We further validate the importance of our LLM judge by training models on randomly labelled data. Training with randomly labelled data substantially degrades performance compared to LLM-validated data, particularly on NevIR where performance drops well below the off-the-shelf baseline. Due to space limitations, we report the full results and analyses in Appendix D.3; Appendix E further provides an error analysis categorizing improvements by negation type (Hossain et al., 2020) and case studies illustrating specific reasoning patterns improved by our approach.

6 Conclusion

Negation and commonsense knowledge are both common and important in human language. Previous work shows that models struggle when negation appears in natural language understanding tasks. However, few works have investigated commonsense knowledge with negation.

We present an approach to automatically augment existing commonsense knowledge corpora with negation, contributing over 2M commonsense knowledge triples with negation. We show that pre-training models with our corpora is beneficial in understanding negation. This holds true across three models and five benchmarks. We further conduct ablation studies and analyses that provide additional evidence and insights into the performance improvements.

Limitations

We work on two existing commonsense knowledge corpora with limited focus—they only contain *if-then* relations. It would be more comprehensive to investigate commonsense knowledge in different forms. In addition, we only consider the logical negation cue *not*. Future work should consider various negation cues, including semantic negation.

We only experiment with two relatively small LLMs (a 7B and an 8B model) to study the benefit of our corpora for improving negation under-

standing. Due to limited computational resources, we choose not to conduct experiments with larger LLMs (e.g., 70B). Moreover, the LLMs are trained and evaluated with 4-bit quantization due to the same resource limitation.

NevIR was developed to evaluate information retrieval tasks involving negation. We experiment with NevIR as a classification task using either classifier models or general-purpose LLMs. Future work should consider models specifically designed for information retrieval.

Ethics Statement

Data Sources and Collection We collect the datasets (ATOMIC, ANION, CondaQA, NLI with Negation, and NevIR) via links provided by the authors.

Data from ATOMIC (Sap et al., 2019) is used under the Creative Commons Attribution 4.0 International License; CondaQA (Ravichander et al., 2022) under Apache-2.0 License; NLI with Negation (Hossain et al., 2020) and NevIR (Weller et al., 2024) under MIT License. ANION (Jiang et al., 2021) does not specify its license.

Acknowledgments

We thank the reviewers for their insightful comments.

The OpenAI Researcher Access Program provided credits to conduct this research.

This material is based upon work supported by the National Science Foundation under Grant No. 2310334. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Hiba Arnaout, Simon Razniewski, Gerhard Weikum, and Jeff Z. Pan. 2022. Uncommonsense: Informative negative knowledge about everyday concepts. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 37–46, New York, NY, USA. Association for Computing Machinery.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, page 177–190, Berlin, Heidelberg. Springer-Verlag.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

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

Radina Dobreva and Frank Keller. 2021. Investigating negation in pre-trained vision-and-language models. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 350–362, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianqing Fang, Quyet V. Do, Hongming Zhang, Yangqiu Song, Ginny Y. Wong, and Simon See. 2022. PseudoReasoner: Leveraging pseudo labels for commonsense knowledge base population. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3379–3394, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312, Online. Association for Computational Linguistics.

Myeongjun Jang, Frank Mtumbuka, and Thomas Lukasiewicz. 2022. Beyond distributional hypothesis: Let language models learn meaning-text correspondence. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2030–2042, Seattle, United States. Association for Computational Linguistics.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. 2021. "I'm not mad": Commonsense implications of negation and contradiction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4380–4397, Online. Association for Computational Linguistics.

Yash Kumar Lal, Niket Tandon, Tanvi Aggarwal, Horace Liu, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2022. Using commonsense knowledge to answer why-questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Abhilasha Ravichander, Matt Gardner, and Ana Marasovic. 2022. CONDAQA: A contrastive reading comprehension dataset for reasoning about negation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8729–8755, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

MohammadHossein Rezaei and Eduardo Blanco. 2024. Paraphrasing in affirmative terms improves negation understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 602–615, Bangkok, Thailand. Association for Computational Linguistics.

MohammadHossein Rezaei and Eduardo Blanco. 2025. Making language models robust against negation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8123–8142, Albuquerque, New Mexico. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Rituraj Singh, Rahul Kumar, and Vivek Sridhar. 2023. NLMs: Augmenting negation in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13104–13116, Singapore. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Orion Weller, Dawn Lawrie, and Benjamin Van Durme. 2024. NevIR: Negation in neural information retrieval. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2274–2287, St. Julian's, Malta. Association for Computational Linguistics.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

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.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Wenting Zhao, Justin Chiu, Jena Hwang, Faeze Brahman, Jack Hessel, Sanjiban Choudhury, Yejin Choi, Xiang Li, and Alane Suhr. 2024. UNcommonsense reasoning: Abductive reasoning about uncommon situations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8487–8505, Mexico City, Mexico. Association for Computational Linguistics.

A Details for Creating and Annotating the Benchmark

We create the benchmark by sampling 200 triples per relation (9 relations in total) from ATOMIC's test split. The benchmark includes 7,200 triples augmented with three types of negation per original triple (negating the *if* event, *then* event, or both). We ask two annotators, one graduate student and one with a PhD degree, to validate each triple using three labels: *Valid*, *Invalid*, and *Ambiguous*. Table 7 shows the instructions we provide to annotators.

B Details for Validating Commonsense Knowledge Triples

We train a task-specific LLM to validate augmented commonsense triples. The model is trained with the same amount of data regardless of the corpus source. Specifically, we sample 1,800 triples per label from the training split of each source corpus (in total 5,400). For *Valid* and *Ambiguous* triples, we sample 100 triples per relation per corpus if the sources are ATOMIC and ANION; otherwise, we sample 200 triples per relation from the single corpus.

For *Invalid* triples, we use GPT-4o to generate *then* events given sampled *if* events from ATOMIC and ANION to construct *Invalid* triples. Table 8 provides the prompt.

We train the task-specific LLM judge based on Llama 3.1 Instruct 8B and 70B using QLoRA (Dettmers et al., 2023) with 4-bit quantization, which are hosted locally on two H100 GPUs with a total of 160 GB memory. Llama 3.1 8B is trained for 5 epochs with a learning rate of 5e-6 with 1 hour of training time, while Llama 3.1 70B is trained for 1 epoch with a learning rate of 2e-5 with 16 hours of training time.

Table 9 lists the mapping we use to convert commonsense knowledge triples to natural language *if-then* statements. The models are trained and evaluated with the *if-then* statements instead of the triples.

Given a triple <A, R, B> and its three negated variations <$\neg A$, R, B>, <A, R, $\neg B$>, and <$\neg A$, R, $\neg B$>, the annotation task is to determine if each triple aligns with real-world commonsense knowledge. We use label *Valid* to represent a triple that always aligns with real-world commonsense knowledge; label *Invalid* to represent a triple that always conflicts with real-world commonsense knowledge; and label *Ambiguous* to represent a triple for two cases: (1) the interpretation of the triple is ambiguous, meaning it can either align or conflict with real-world commonsense knowledge, or (2) the *if* event does not relate to the *then* event by the relation.

Below are some examples.
Valid: If PersonX takes a picture, then PersonX wants to look at the picture.
Invalid: If PersonX takes a picture, then PersonX wants to not look at the picture.
Ambiguous: If PersonX takes a picture, then PersonX wants to take a nap.
Ambiguous: If PersonX opens the window, then PersonX wants to breathe.

Table 7: Annotation instructions for the benchmark.

You are an expert in commonsense reasoning and knowledge generation. Your task is to generate a then event complementing the given if event so that the if-then statement conflicts with commonsense knowledge.
These invalid statements should be clearly wrong or illogical based on everyday commonsense knowledge.

Given the following incomplete if-then statement:
Statement: If event, then relation ...
Generate the then event within a phrase.

Table 8: Prompt to generate if-then statements that conflict with commonsense knowledge. They are considered *Invalid* triples for the validation task and downstream task.

Relation	Verbalization
oEffect	the effect of {object} is
oReact	the reaction of {object} is
oWant	{object} want
xAttr	the attribute of PersonX is
xEffect	the effect of PersonX is
xIntent	the intention of PersonX is
xNeed	PersonX needs
xReact	the reaction of PersonX is
xWant	PersonX wants

Table 9: The mapping from commonsense knowledge triples with nine *if-then* relations to natural language statements. {object} indicates the object in the *if* event.

B.1 Additional Results on Validating Commonsense Knowledge Triples

Table 12 reports additional results for validating commonsense triples with negation, including F1 for each relation and overall Accuracy (Acc). Among individual relations, xIntent and xAttr consistently yield higher F1 across models, while oEffect and oReact are more challenging. Fine-tuning with ATOMIC generally outperforms ANION as the source of *Valid* and *Ambiguous* training data.

You are a helpful assistant. In this task, you are expected to write answers to questions involving reasoning about negation.
The answer to the question should be "yes", "no", "don't know", or a phrase in the passage. Questions can only have one correct answer.
Only output [YES], [NO], [DON'T KNOW] or a short phrase in the passage.

{4 exemplars sampled from the few-shot learning split of CondaQA}

Table 10: 4-shot prompts to evaluate LLMs with CondaQA. We randomly sample 4 exemplars from the few-shot learning split provided by Ravichander et al. (2022).

	Learning Rate	Batch Size	Epochs
RTE-neg	2e-5	8	15
SNLI-neg	2e-5	32	3
MNLI-neg	2e-5	64	4

Table 11: Hyperparameters for fully-supervised fine-tuning with three NLI benchmarks: RTE-neg, SNLI-neg, and MNLI-neg.

C Experimental Details for Downstream Tasks

We train three models with either existing corpora or our commonsense knowledge corpora. RoBERTa-large (Liu et al., 2019) is further fine-tuned with the training split of the specific downstream task, using a batch size of 128, a learning rate of 1e-6, and early stopping with patience of 3 epochs and a maximum of 5 epochs.

For the two LLMs (Llama 3.1 8B (Grattafiori et al., 2024) and Qwen2 7B (Yang et al., 2024)), we adopt a standard instruction fine-tuning paradigm, where the training input is formatted as: [instruction, verbalized commonsense triple, output label], and the model is trained using the causal language modeling objective. Specifically, they are trained for 2 epochs using a batch size of 16; Llama 3.1 8B uses a learning rate of 5e-6 and Qwen2 7B uses 2e-5. We also use QLoRA (Dettmers et al., 2023) with 4-bit quantization for both models. Each model takes approximately 4 hours to train on one H100 GPU with 80 GB memory. They are further evaluated using few-shot prompting. GPT-4o is called via OpenAI's API and Claude Sonnet 4 is called

	oEffect	oReact	oWant	xAttr	xEffect	xIntent	xNeed	xReact	xWant	All Relations	
										F1	Acc
Few-shot learning											
Llama 3.1 8B	0.34	0.38	0.34	0.39	0.35	0.50	0.31	0.34	0.36	0.37	0.44
Llama 3.1 70B	0.48	0.45	0.53	0.56	0.56	0.56	0.42	0.52	0.57	0.53	0.53
GPT-4o	0.52	0.46	0.54	0.54	0.44	0.57	0.51	0.54	0.57	0.52	0.54
Claude Sonnet 4	0.40	0.37	0.57	0.66	0.59	0.61	0.56	0.54	0.63	0.56	0.56
Fine-tuning											
Llama 3.1 8B with											
ATOMIC	0.51	0.57	0.49	0.63	0.53	0.62	0.50	0.53	0.52	0.55	0.56
ANION	0.52	0.56	0.50	0.59	0.51	0.59	0.45	0.56	0.50	0.53	0.55
ATOMIC + ANION	0.38	0.44	0.46	0.57	0.47	0.56	0.48	0.51	0.47	0.48	0.51
Llama 3.1 70B with											
ATOMIC	0.59	0.60	0.61	0.72	0.62	0.69	0.59	0.62	0.63	0.63	0.64
ANION	0.54	0.54	0.56	0.68	0.51	0.60	0.43	0.50	0.55	0.55	0.56
ATOMIC + ANION	0.56	0.58	0.61	0.71	0.60	0.66	0.57	0.62	0.63	0.62	0.62

Table 12: Complete results of validating commonsense triples with negation. We report F1 for each relation and overall, along with overall Accuracy (Acc). We further report the complete fine-tuning results across variants that differ in whether *Valid* and *Ambiguous* instances come from ATOMIC, ANION, or both. The results complement Table 1 in the main paper.

via the AWS Bedrock API.

C.1 Experimental Details for CondaQA

For fully-supervised evaluation, we train LLMs with CondaQA's training split using a batch size of 8 and a learning rate of 1e-5. We use early stopping with a patience of 3 epochs and a maximum of 5 epochs.

Table 10 reports the 4-shot prompts for evaluating LLMs with CondaQA. We reuse the prompts with minimal edits and sample 4 exemplars from the few-shot learning split provided by Ravichander et al. (2022).

C.2 Experimental Details for NLI Benchmarks

Table 11 reports the hyperparameters for fully-supervised fine-tuning LLMs with NLI datasets.

Table 13 reports the 4-shot prompts used to evaluate LLMs with RTE-neg dataset, and Table 14 reports the 4-shot prompts for evaluating with MNLI-neg and SNLI-neg datasets. All exemplars are sampled from their training split.

C.3 Experimental Details for NevIR

As NevIR does not provide any training data, following Weller et al. (2024), we perform fully-supervised fine-tuning with LLMs on STS-B (Cer et al., 2017) using a learning rate of 2e-5 and a batch size of 32 for 4 epochs.

Table 15 reports the 4-shot prompts for evaluating with NevIR.

D Additional Results on Downstream Tasks

D.1 Complete Results for Downstream Tasks

Table 3 and Table 4 report only the best results among three training corpora configurations using either ¬ATOMIC, ¬ANION, or both. Table 16 and Table 17 provide complete results for all three configurations.

D.2 Evaluation on CommonsenseQA

We evaluate off-the-shelf and pre-trained models on CommonsenseQA (Talmor et al., 2019), a standard commonsense reasoning benchmark without negation, to verify that pre-training on our negated corpora does not degrade general commonsense reasoning ability. Table 18 reports the results. Both models maintain comparable performance after pre-training, with Llama 3.1 8B slightly improving (71.8 → 72.7) and Qwen2 7B showing a minor decrease (80.7 → 79.5). These results demonstrate that pre-training on negated corpora does not lead to catastrophic forgetting of general commonsense knowledge.

D.3 Ablation on Randomly Labelled Data

Although our LLM judge achieves relatively high precision on validating commonsense triples with negation, the resulting data are still noisy. To quantify the impact of data quality, we train models on a noisier dataset—with randomly assigned labels—where the validation accuracy for *Valid* and *Invalid*

You are a helpful assistant. You are given a pair of sentences: a premise and a hypothesis. Your task is to determine the relationship between the two sentences.
[entailment]: The premise guarantees the truth of the hypothesis.
[not_entailment]: The premise does not guarantee the truth of the hypothesis.
Format your answer by only outputting [entailment] or [not_entailment].

Premise: Edward VIII became King in January of 1936 and abdicated in December.
Hypothesis: King Edward VIII abdicated in December 1936.
Response: [entailment]

Premise: Oil prices fall back as Yukos oil threat lifted.
Hypothesis: Oil prices rise.
Response: [not_entailment]

Premise: World Bank programs have been heavily criticized for many years for resulting in poverty.
Hypothesis: The World Bank is criticized for its activities.
Response: [entailment]

Premise: The cost of the consumer of the United States fell in June.
Hypothesis: U.S. consumer spending dived in June.
Response: [not_entailment]

Table 13: 4-shot prompts for evaluating LLMs with RTE-neg dataset. The exemplars are chosen from the training split.

You are a helpful assistant. You are given a pair of sentences: a premise and a hypothesis. Your task is to determine the relationship between the two sentences.
[entailment]: The hypothesis is definitely true given the premise.
[contradiction]: The hypothesis is definitely false given the premise.
[neutral]: It is not possible to determine whether the hypothesis is true or false just from the premise.
Format your answer by only outputting [entailment], [contradiction], or [neutral].

Premise: One of our number will carry out your instructions minutely.
Hypothesis: A member of my team will execute your orders with immense precision.
Response: [entailment]

Premise: Fun for adults and children.
Hypothesis: Fun for only children.
Response: [contradiction]

Premise: He turned and smiled at Vrenna.
Hypothesis: He smiled at Vrenna who was walking slowly behind him with her mother.
Response: [neutral]

Premise: The famous tenements (or lands) began to be built.
Hypothesis: The land remained deserted.
Response: [contradiction]

Table 14: 4-shot prompts for evaluating LLMs with SNLI-neg and MNLI-neg datasets. The exemplars are chosen from the training split.

triples is approximately 0.50.

Table 19 reports the results. Training with randomly labelled data substantially degrades performance compared to LLM-validated data across all tasks. Most notably, NevIR drops well below the off-the-shelf baseline for both models (18.9 vs. 30.6 for Llama 3.1 8B; 12.0 vs. 29.8 for Qwen2 7B), indicating that noisy labels can be actively harmful for complex negation reasoning. For NLI tasks, randomly labelled data still yields some improvement over the baseline (e.g., RTE: 70.8 vs. 60.0 for Llama 3.1 8B), suggesting that the model learns partial negation patterns from the triple structure itself. However, the gains are far smaller than with validated data (e.g., RTE: 70.8 vs. 81.3). These results validate the importance of our LLM judge: even imperfect labels substantially outperform random ones.

You are a helpful assistant. You are given a query and two documents. Your task is to choose the document that has the answer for the query.
Output [Doc1] if the first document has the answer for the query, or [Doc2] if the second document has the answer.

Query: Which mayor did more vetoing than anticipated?
Doc1: In his first year as mayor, Medill received very little legislative resistance from the Chicago City Council. While he vetoed what was an unprecedented eleven City Council ordinances that year, most narrowly were involved with specific financial practices considered wasteful and none of the vetoes were overridden. He used his new powers to appoint the members of the newly constituted Chicago Board of Education and the commissioners of its constituted public library. His appointments were approved unanimously by the City Council.
Doc2: In his first year as mayor, Medill received very little legislative resistance from the Chicago City Council. While some expected an unprecedented number of vetoes, in actuality he only vetoed eleven City Council ordinances that year, and most of those were narrowly involved with specific financial practices he considered wasteful and none of the vetoes were overridden. He used his new powers to appoint the members of the newly constituted Chicago Board of Education and the commissioners of its constituted public library. His appointments were approved unanimously by the City Council.
Response: [Doc1]

Query: Which mayor did less vetoing than anticipated?
Doc1: In his first year as mayor, Medill received very little legislative resistance from the Chicago City Council. While he vetoed what was an unprecedented eleven City Council ordinances that year, most narrowly were involved with specific financial practices considered wasteful and none of the vetoes were overridden. He used his new powers to appoint the members of the newly constituted Chicago Board of Education and the commissioners of its constituted public library. His appointments were approved unanimously by the City Council.
Doc2: In his first year as mayor, Medill received very little legislative resistance from the Chicago City Council. While some expected an unprecedented number of vetoes, in actuality he only vetoed eleven City Council ordinances that year, and most of those were narrowly involved with specific financial practices he considered wasteful and none of the vetoes were overridden. He used his new powers to appoint the members of the newly constituted Chicago Board of Education and the commissioners of its constituted public library. His appointments were approved unanimously by the City Council.
Response: [Doc2]

Query: Which Swiss cantons do not have official churches?
Doc1: Switzerland has no official state religion, though most of the cantons (except Geneva and Neuchâtel) recognise official churches, which are either the Roman Catholic Church or the Swiss Reformed Church. These churches, and in some cantons also the Old Catholic Church and Jewish congregations, are financed by official taxation of adherents.
Doc2: Switzerland has no official state religion, though most of the cantons (except Neuchâtel) recognise official churches, which are either the Roman Catholic Church or the Swiss Reformed Church. These churches, and in some cantons also the Old Catholic Church and Jewish congregations, are financed by official taxation of adherents.
Response: [Doc1]

Query: Which Swiss canton does not have official churches?
Doc1: Switzerland has no official state religion, though most of the cantons (except Geneva and Neuchâtel) recognise official churches, which are either the Roman Catholic Church or the Swiss Reformed Church. These churches, and in some cantons also the Old Catholic Church and Jewish congregations, are financed by official taxation of adherents.
Doc2: Switzerland has no official state religion, though most of the cantons (except Neuchâtel) recognise official churches, which are either the Roman Catholic Church or the Swiss Reformed Church. These churches, and in some cantons also the Old Catholic Church and Jewish congregations, are financed by official taxation of adherents.
Response: [Doc2]

Table 15: 4-shot prompts to evaluate LLMs with NevIR.

E Error Analysis

We conduct a detailed error analysis comparing the off-the-shelf Llama 3.1 8B with the model pre-trained on ¬ATOMIC + ¬ANION across all downstream tasks. Our analysis categorizes the types of negation reasoning errors that are corrected by training with our augmented corpora.

E.1 Error Types in NLI with Negation

We examine all 940 examples across RTE-Neg, SNLI-Neg, and MNLI-Neg where the pre-trained model answers correctly but the off-the-shelf model does not. Following the negation taxonomy of Hossain et al. (2020), we categorize these improvements by negation type in Table 20.

Verbal Negation in Premise Only (34.4%) In these cases, the premise contains an explicit verbal negation cue (e.g., *not*, *never*, *no*) but the hypothesis does not. The off-the-shelf model often treats the negated premise as if the negation were absent. For example, given the premise *"The prosecutor did not tell the court that the incident had caused 'distress' to one of the children."* and the hypothesis *"The prosecutor told the court that 'distress' in one of the children is associated with the incident."*, the off-the-shelf model predicts *entailment*, ignoring

	# Params.	Accuracy (Δ%)	Group Consistency			
			All	Par.	Sco.	Aff.
Fully supervised						
UnifiedQA-v2-large (Ravichander et al., 2022)	770M	66.7	30.2	64.0	43.7	46.5
RoBERTa-large	355M	64.9	29.6	61.9	41.4	45.8
Further pre-trained with						
Original commonsense triples from						
ATOMIC + ANION		67.0 (+3.2)	32.9	65.6	46.3	48.1
Negated commonsense triples from						
¬ATOMIC		67.7 (+4.3)	32.8	66.1	46.3	48.6
¬ANION		67.9 (+4.6)*	33.3	66.0	45.8	50.1
¬ATOMIC + ¬ANION		**68.5** (+5.5)*	**34.3**	**66.4**	**47.6**	**50.1**
In-context learning						
Zero-shot						
OpenAI o1	—	65.3	24.9	67.4	43.8	38.6
Few-shot						
InstructGPT + COT (Ravichander et al., 2022)	—	66.3	27.3	64.2	45.1	44.9
GPT-4o	—	72.9	34.4	78.7	52.6	47.9
Llama 3.1 70B	70B	77.5	43.3	83.7	61.8	54.9
Llama 3.1 8B	8B	68.7	31.5	69.0	48.1	44.4
Further pre-trained with						
Original commonsense triples from						
ATOMIC + ANION		67.6 (-1.6)	27.5	69.5	44.3	41.4
Negated commonsense triples from						
¬ATOMIC		70.6 (+2.8)	32.8	71.8	48.6	45.0
¬ANION		71.3 (+3.8)*	33.4	72.5	**49.6**	45.7
¬ATOMIC + ¬ANION		**71.5** (+4.1)*	**33.5**	**72.9**	48.7	**46.4**
Qwen2 7B	7B	65.1	24.9	65.9	39.7	39.2
Further pre-trained with						
Original commonsense triples from						
ATOMIC + ANION		64.1 (-1.5)	22.8	65.2	38.4	37.7
Negated commonsense triples from						
¬ATOMIC		**69.7** (+7.1)*	**32.7**	**71.2**	**46.3**	**45.5**
¬ANION		69.2 (+6.2)*	29.9	68.8	45.4	44.1
¬ATOMIC + ¬ANION		66.6 (+2.3)	28.7	68.8	43.7	40.8

Table 16: Complete results evaluating CondaQA with two settings: fully supervised (top) and in-context learning (bottom). We experiment with three training configurations on our corpora. The best result for each model is in bold. Delta (Δ) indicates the percent change in accuracy compared to the off-the-shelf model. An asterisk * indicates a statistically significant improvement (McNemar's test (McNemar, 1947), $p < 0.05$) over both the off-the-shelf model and the model trained with existing corpora.

"did not" in the premise. The pre-trained model correctly predicts *not_entailment*.

Verbal Negation in Hypothesis Only (23.4%) These examples contain verbal negation in the hypothesis but not in the premise. The off-the-shelf model struggles to determine whether a non-negated premise supports or contradicts a negated hypothesis.

Negation Interaction (41.3%) The largest category involves examples where both the premise and hypothesis contain negation, requiring the model to reason about the interaction between multiple negation cues. For instance, given *"This growing number of titles does not leave publishing houses with less time and attention to edit and market books."* and *"Publishing houses cannot give less attention to editing books."*, the off-the-shelf model predicts *entailment*. However, the correct label is *neutral*, as the negated premise no longer supports the negated hypothesis. The pre-trained model correctly identifies this relationship.

Affixal Negation and Other (0.9%) A small fraction of improvements involve affixal negation (e.g., *un-*, *dis-*) or other negation types. This low proportion reflects the dominance of verbal negation cues in NLI benchmarks (Hossain et al., 2020).

E.2 Analysis by Negation Cue Type in CondaQA

CondaQA annotates each example with a negation cue word. Following the negation taxonomy of Hossain et al. (2020), we categorize these cues into five linguistic types and report accuracy for

	# Params.	NLI with Negation			NevIR
		RTE-Neg	SNLI-Neg	MNLI-Neg	
Fully supervised					
BERTNOT (Hosseini et al., 2021)	110M	74.5	46.0	60.9	—
RoBERTa-large-NSP (Rezaei and Blanco, 2025)	355M	87.2	56.5	69.9	—
stsb-roberta-large (Weller et al., 2024)	355M	—	—	—	24.9
MonoT5 3B (Nogueira et al., 2020)	3B	—	—	—	50.6
RoBERTa-large	355M	84.7	56.0	69.9	24.5
Further pre-trained with					
Original commonsense triples from					
ATOMIC + ANION		86.2	56.5	69.4	29.1
Negated commonsense triples from					
¬ATOMIC		81.5	56.6	**70.9**	29.8
¬ANION		85.2	57.1	69.3	30.5
¬ATOMIC + ¬ANION		**88.1***	**58.3**	69.7	**34.3***
Zero-shot					
OpenAI o1	—	87.5	75.9	74.6	59.7
Few-shot					
GPT-4o	—	86.9	74.8	75.0	61.7
Llama 3.1 70B	70B	78.9	69.1	65.9	58.8
Llama 3.1 8B	8B	60.0	54.4	47.0	30.6
Further pre-trained with					
Original commonsense triples from					
ATOMIC + ANION		65.3	57.5	51.1	37.9
Negated commonsense triples from					
¬ATOMIC		81.2*	**68.5***	63.8*	38.1
¬ANION		**81.3***	68.2*	**63.9***	**42.2***
¬ATOMIC + ¬ANION		81.3*	67.9*	63.7*	38.2
Qwen2 7B	7B	71.7	60.7	59.5	29.8
Further pre-trained with					
Original commonsense triples from					
ATOMIC + ANION		78.3	66.5	63.0	33.8
Negated commonsense triples from					
¬ATOMIC		**82.3***	**72.4***	**67.5***	**39.8***
¬ANION		78.9	66.1	62.5	36.3
¬ATOMIC + ¬ANION		82.1*	72.2*	67.5*	38.5*

Table 17: Complete results evaluating three NLI benchmarks with negation cues (accuracy) and NevIR benchmark (pairwise accuracy) in two settings: (1) fully supervised (top) and (2) in-context learning (bottom). We experiment with three training configurations on our corpora. The best result for each model is in bold. An asterisk * indicates a statistically significant improvement over both the off-the-shelf model and the model trained with existing corpora.

	CommonsenseQA
Llama 3.1 8B	71.8
Pre-trained with	
¬ATOMIC + ¬ANION	72.7
Qwen2 7B	80.7
Pre-trained with	
¬ATOMIC + ¬ANION	79.5

Table 18: Results on CommonsenseQA (accuracy), a non-negated commonsense reasoning benchmark. Pre-training on our negated corpora does not degrade performance.

the off-the-shelf and pre-trained Llama 3.1 8B in Table 21.

Pre-training with our corpora yields the largest improvement for verbal negation cues (+6.5 points), which aligns with our training data where negation is introduced by adding *not* to events. Affixal negation also shows substantial gains (+5.9 points), indicating effective transfer from explicit to morphological negation. Diminisher cues (e.g., *rarely*, *barely*, *few*) improve by +2.3 points, suggesting that learning explicit negation patterns helps models better handle attenuated assertions. Implicit negation cues show a modest improvement (+1.8 points), indicating that implicit negation such as *lack* or *prevent* partially benefits from explicit negation training but may require additional training signals. The Other category (n=24) is too small for reliable estimates.

E.3 Case Studies

We present representative case studies illustrating specific reasoning patterns improved by our ap-

	CondaQA (Accuracy)	NLI with Negation			NevIR
		RTE-Neg	SNLI-Neg	MNLI-Neg	
Llama 3.1 8B	68.7	60.0	54.4	47.0	30.6
Pre-trained with					
Randomly labelled ¬ATOMIC + ¬ANION	67.1	70.8	57.5	50.9	18.9
¬ATOMIC + ¬ANION	71.5	81.3	67.9	63.7	38.2
Qwen2 7B	65.1	71.7	60.7	59.5	29.8
Pre-trained with					
Randomly labelled ¬ATOMIC + ¬ANION	65.2	76.1	62.0	61.9	12.0
¬ATOMIC + ¬ANION	66.6	82.1	72.2	67.5	38.5

Table 19: Ablation on randomly labelled training data. We train Llama 3.1 8B and Qwen2 7B with randomly labelled ¬ATOMIC + ¬ANION. While the randomly labelled dataset demonstrates marginal benefits on CondaQA and NLI tasks, training with the validated dataset significantly outperforms. In addition, randomly labelled data is detrimental to NevIR performance.

Error Type	Count	Percentage
Verbal negation in premise only	323	34.4%
Verbal negation in hypothesis only	220	23.4%
Negation interaction (both P & H)	388	41.3%
Affixal negation	2	0.2%
Other	7	0.7%

Table 20: Distribution of error types corrected by pre-training with our commonsense corpora, across 940 improved NLI examples (Llama 3.1 8B). Categories follow the negation taxonomy of Hossain et al. (2020).

proach.

Case 1: Negation Scope in NLI

> **Premise:** *"A barefoot young girl in a pink gown is **not** asleep on a hard wood floor cuddling her baby doll."*
> **Hypothesis:** *"A girl is playing with her doll outside."*
> **Gold:** neutral
> **Off-the-shelf:** contradiction
> **Pre-trained:** neutral

The off-the-shelf model interprets *"not asleep"* as contradicting *"playing outside"*, failing to recognize that negating *asleep* does not specify what the girl is doing or where she is. The pre-trained model correctly identifies that the negated premise leaves the hypothesis unresolved.

Case 2: Negation and Entailment Direction

> **Premise:** *"Organic fertilizers like vermi compost are **not** used for increasing the quality, fertility and mineral content of the soil."*
> **Hypothesis:** *"Organic fertilizers are used as soil enhancers."*
> **Gold:** not_entailment
> **Off-the-shelf:** entailment
> **Pre-trained:** not_entailment

The off-the-shelf model associates *organic fertilizers* with *soil enhancers* based on world knowledge, completely ignoring the negation in the premise. Pre-training on commonsense triples with negation teaches the model that a negated *if* event does not entail the original *then* event.

Case 3: Affixal Negation in CondaQA

> **Cue:** *unmyelinated*
> **Question:** *"Does the wording of the passage suggest that axons have myelin sheaths while neurons typically do not?"*
> **Gold:** No
> **Off-the-shelf:** Yes
> **Pre-trained:** No

The prefix *un-* in *unmyelinated* reverses the meaning, and the pre-trained model correctly identifies that the passage does not support the hypothesis.

Case 4: Negation in Information Retrieval (NevIR)

> **Q1:** *"Whose ship is attacked by **unfamiliar** enemies?"*
> **Q2:** *"Whose ship is attacked by **familiar** enemies?"*
> **Expected:** Q1→Doc1, Q2→Doc2

NevIR requires the model to distinguish between a query and its negated counterpart. The model must identify which document contains information matching the specific polarity of each query—a task that directly requires understanding whether an event is negated.

Category	Examples	Count	Off-the-shelf	Pre-trained
Verbal	*not, never, no, none*	1,902	67.6	74.1
Affixal	*un-, in-, dis-, -less*	3,553	62.1	68.0
Implicit	*lack, without, prevent*	1,502	67.9	69.7
Diminisher	*rarely, barely, few*	259	66.0	68.3
Other		24	79.2	75.0

Table 21: CondaQA accuracy by negation cue category (Llama 3.1 8B). Categories follow the negation taxonomy of Hossain et al. (2020): *Verbal*: explicit negation words; *Affixal*: morphological negation (prefixes/suffixes); *Implicit*: words conveying negation without explicit markers; *Diminisher*: words that reduce the degree of an assertion.