A MISMATCHED Benchmark for Scientific Natural Language Inference

Firoz Shaik[♣] **Mobashir Sadat**[♣] **Nikita Gautam**[♦] **Doina Caragea**[♦] **Cornelia Caragea**[♣]

Computer Science
[♣]University of Illinois Chicago
[♦]Kansas State University
{fshaik8,msadat3,cornelia}@uic.edu, {ngautam,dcaragea}@ksu.edu

Abstract

Scientific Natural Language Inference (NLI) is the task of predicting the semantic relation between a pair of sentences extracted from research articles. Existing datasets for this task are derived from various computer science (CS) domains, whereas non-CS domains are completely ignored. In this paper, we introduce a novel evaluation benchmark for scientific NLI, called MISMATCHED. The new MISMATCHED benchmark covers three non-CS domains–PSYCHOLOGY, ENGINEERING, and PUBLIC HEALTH, and contains 2,700 human annotated sentence pairs. We establish strong baselines on MISMATCHED using both Pre-trained Small Language Models (SLMs) and Large Language Models (LLMs). Our best performing baseline shows a Macro F1 of only 78.17% illustrating the substantial headroom for future improvements. In addition to introducing the MISMATCHED benchmark, we show that incorporating sentence pairs having an *implicit* scientific NLI relation between them in model training improves their performance on scientific NLI. We make our dataset and code publicly available on GitHub.[1]

1 Introduction

The task of Natural Language Inference (NLI) has received significant attention, initially through several PASCAL Recognising Textual Entailment (RTE) Challenges (Dagan et al., 2006; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009) which focused on recognizing if two given sentences exhibit an entailment relationship. Subsequently, several NLI datasets (Bowman et al., 2015; Williams et al., 2018; Nie et al., 2020) have been introduced to facilitate progress on the NLI task. More recently, there has been an increasing interest in domain specific NLI tasks, including scientific NLI (Sadat and Caragea, 2022). The scientific NLI task classifies the semantic relation between a pair of sentences extracted from research articles into one of four classes—ENTAILMENT, REASONING, CONTRASTING, and NEUTRAL. This task is challenging for both Pre-trained Small Language Models (SLMs) and Large Language Models (LLMs) (Sadat and Caragea, 2024), making it suitable to serve as a challenging benchmark for evaluating the natural language understanding of state-of-the-art models. In addition, Sadat and Caragea (2024) have shown that scientific NLI can aid in improving the performance of other downstream tasks such as topic classification and citation intent classification.

To date, two datasets have been made available to facilitate research on scientific NLI—SCINLI (Sadat and Caragea, 2022), and MSCINLI (Sadat and Caragea, 2024). SCINLI is derived from papers published in the ACL anthology, related to Natural Language Processing (NLP) and computational linguistics. To introduce diversity in scientific NLI, MSCINLI is constructed using sentence pairs extracted from five different scientific domains—HARDWARE, NETWORKS, SOFTWARE & ITS ENGINEERING, SECURITY & PRIVACY, and the NEURIPS conference which is related to machine learning. The training sets of these datasets are constructed using a distant supervision method that harnesses *explicit* signals conveyed by various linking phrases. For example, if the second sentence in an adjacent sentence pair starts with "However" or "In contrast," the sentence pair is labeled as CONTRASTING. The test and development sets of both SCINLI and MSCINLI are human annotated to ensure a realistic evaluation.

Despite the diversity introduced in MSCINLI, the domains covered by the existing scientific NLI datasets are still related to only computer science (CS), while non-CS domains are completely ignored. Thus, in this paper, we propose a new evaluation benchmark for scientific NLI called MISMATCHED, which contains sentence pairs collected from 3 non-CS domains: PSYCHOLOGY, ENGI-

[1]https://github.com/fshaik8/MisMatched

Findings of the Association for Computational Linguistics: ACL 2025, pages 21524–21538

NEERING, and PUBLIC HEALTH. We constructed MISMATCHED as an out-of-domain (OOD) testbed for scientific NLI models. That is, MISMATCHED contains only development and test sets that are human annotated (of sizes 300 and 2400, respectively), without any training data. MISMATCHED is designed as an out-of-domain (OOD) benchmark for evaluating the robustness of scientific NLI models, similar to the MISMATCHED (MM) portion of MNLI (Williams et al., 2018). Like MNLI's MISMATCHED set, which uses unseen genres to test model generalization, MISMATCHED is aimed at evaluating OOD robustness when models are trained on existing scientific NLI training sets.

We establish strong baselines on MISMATCHED by fine-tuning four SLMs—BERT (Devlin et al., 2019), SCIBERT (Beltagy et al., 2019), ROBERTA (Liu et al., 2019) and XLNET (Yang et al., 2019); and by prompting four LLMs—LLAMA-2 (Touvron et al., 2023), LLAMA-3 (Grattafiori et al., 2024), MISTRAL (Jiang et al., 2023) and PHI-3 (Abdin et al., 2024) using the training sets from existing scientific NLI datasets. We find that our best performing SLM baseline with SCIBERT and best performing LLM baseline with PHI-3 show Macro F1 of only 78.17% and 57.16%, respectively, illustrating the highly challenging nature of the MISMATCHED set, and a significant amount of headroom for future improvements. In addition, given that *all* sentence pairs in the training sets of existing scientific NLI datasets are constructed using distant supervision that harnesses *explicit* relations conveyed by various linking phrases, we analyze the impact on models' performance of sentence pairs which have an *implicit* relation between them (i.e., sentence pairs which are adjacent in text and have a scientific NLI relation but the second sentence in the pair does not start with a linking phrase). We find that incorporating *implicit* relations can indeed improve the performance of scientific NLI models. Our key contributions can be summarized as follows:

- We introduce a novel MISMATCHED benchmark which is more distant from computer science domains to further enhance the diversity of scientific NLI.
- We establish strong baselines on MISMATCHED using both SLMs and LLMs and show that it presents a challenging new benchmark for out-of-domain NLI evaluation.
- We incorporate *implicit* relations in the scientific NLI model training, and show that they can improve the performance of scientific NLI models trained using only *explicit* relations.

2 Related Work

Since the introduction of the NLI task (Dagan et al., 2006), numerous datasets have been introduced that include sentence pairs from the general domain. RTE (Dagan et al., 2006) is an early dataset, which went through several iterations (Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009). The RTE dataset contains premise-hypothesis pairs, which are labeled as *entailment* or *non-entailment*. More recent datasets such as SICK (Marelli et al., 2014), SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) contain sentence pairs that are classified as *entailment*, *contradiction*, or *neutral*. The SICK dataset (Marelli et al., 2014) contains sentence pairs automatically extracted from paired image captions and video descriptions. SNLI (Bowman et al., 2015) contains premise-hypothesis pairs, where the premises are extracted from image captions, and the hypotheses are manually written by human annotators. MNLI (Williams et al., 2018) contains premise-hypothesis pairs where premises are extracted from a variety of sources such as travel guides and face-to-face conversation, while the hypotheses are manually written by human annotators, as in SNLI. ANLI (Nie et al., 2020), another NLI dataset, was constructed in an adversarial fashion with human annotators in the loop who were instructed to write sentence pairs for which the models make mistakes in their predictions.

Several domain-specific NLI datasets have also been introduced. For example, MEDNLI (Romanov and Shivade, 2018) was derived from medical records of patients with the premise-hypothesis pairs being annotated by experts (in the medical domain) as *entailment*, *contradiction*, or *neutral*. The NLI4CT dataset (Jullien et al., 2023) contains premise-hypothesis pairs, in the form of clinical trial reports (CTR) and statements, labeled as *entailment* or *contradiction* by human annotators. NLI4CT-P (Perturbed) (Jullien et al., 2024) is an extension of the original NLI4CT dataset (Jullien et al., 2023) and was obtained by adding a contrast set derived from perturbations to the original statements, to facilitate causal analyses.

Domain	First Sentence	Second Sentence	Class
ENGINEERING	In previous studies, GBRS has acted as a guideline to improve energy use and indoor air quality.	~~However,~~ the effectiveness of GBRS as applied to construction waste management has not been explored.	CONTRASTING
PUBLIC HEALTH	For example, sewage-associated marker genes such as Bacteroides HF183 and HPyV, and enteric viruses such as human NoV are predominantly associated with human feces or sewage.	~~Therefore,~~ these marker genes can be used as a proxy to determine the risk associated with NoV and other enteric pathogens specific to sewage.	REASONING
PSYCHOLOGY	The presence of BED in one or both parents was associated with the emotional and behavioural development in offspring.	~~Particularly,~~ the diagnosis of BED in both parents had a direct effect on infants' affective problems.	ENTAILMENT
ENGINEERING	This baffle geometry was tested for a well known seismic excitation (El Centro) and it was observed to effectively suppress free surface fluctuations and the slosh forces.	storage tank designers should ensure safe design margins and develop methodologies to overcome a wide range of possible scenarios.	NEUTRAL

Table 1: Examples of sentence pairs from MISMATCHED, extracted from different domains. The linking phrases at the beginning of the second sentence (strikethrough text in the table) are deleted after extracting the pairs and assigning the labels.

Most relevant to our work, SCINLI (Sadat and Caragea, 2022) is a scientific NLI dataset constructed from research articles published in the ACL Anthology, where sentence pairs were extracted automatically from articles based on linking phrases, and classified into one of the following four classes: *entailment*, *reasoning*, *contrasting*, and *neutral* (with manual annotations only for the test and dev sets). MSCINLI (Sadat and Caragea, 2024) is an extension of SCINLI, which was constructed from a larger variety of computer science research articles, e.g., HARDWARE, NETWORKS, etc. and contains sentence pairs labeled also with the above four classes. To further diversify the datasets and study the transferability and robustness of the models for scientific NLI in out-of-distribution settings, in this paper, we introduce test/dev sets that cover articles from three non-computer science domains, specifically PSYCHOLOGY, ENGINEERING, and PUBLIC HEALTH.

A comparison of all the datasets reviewed here is shown in Appendix A.

3 The MISMATCHED Benchmark

In this section, we describe our proposed MISMATCHED benchmark for scientific NLI. Specifically, we outline the data sources for deriving MISMATCHED, the construction process and the key statistics. Table 1 shows examples of sentence pairs from different domains and classes in our dataset.

3.1 Data Sources

Our MISMATCHED benchmark is composed of three domains—PSYCHOLOGY, ENGINEERING, and PUBLIC HEALTH. We selected these domains to extend scientific NLI beyond existing computer science focused datasets. While SCINLI (Sadat and Caragea, 2022) covers computational linguistics and MSCINLI (Sadat and Caragea, 2024) encompasses CS domains (Hardware, Networks, Software & Engineering, Security & Privacy, and NeurIPS), our new domains represent diverse non-CS scientific areas with broad real-world applications. The data sources for each of these domains are described below.

PSYCHOLOGY. Kowsari et al. (2017) constructed a dataset for topic classification of scientific papers. The dataset contains abstracts from Web of Science (WoS) papers, which belong to 7 scientific domains, including the Psychology and Engineering domains. WoS is a database that indexes global scholarly literature across sciences from various journals and academic events. We extract sentence pairs from papers in the PSYCHOLOGY domain for our MISMATCHED set.

ENGINEERING. For the ENGINEERING domain, we also utilize a subset of WoS papers from the topic classification dataset introduced by Kowsari et al. (2017). Specifically, we extract sentence pairs from "Civil Engineering", "Electrical Engineering" & "Mechanical Engineering" papers.

PUBLIC HEALTH. Three sources are used to extract sentence pairs for the PUBLIC HEALTH domain. Using twenty-five keywords related to marine water characteristics and the health risks of divers and swimmers (such as coastal water pollution and beach water contamination), we crawled about 100k abstracts from WoS. Next, the National Library of Medicine (NLM) was used to collect 200 additional abstracts for articles specific to water

Dataset	#Examples			#Words		'S' parser		Word		
	Train	Dev	Test	Prem.	Hyp.	Prem.	Hyp.	Overlap	#Domains	Agrmt.
SciNLI (ACL)	101,412	2,000	4,000	27.38	25.93	96.8%	96.7%	30.06%	1	85.8%
MSciNLI	127,320	1,000	4,000	26.84	25.85	94.4%	94.3%	30.29%	5	88.0%
MisMatched	-	300	2400	26.65	25.75	96.8%	98.2%	31.27%	3	85.7%
♢ Public Health	-	100	800	27.42	27.22	98.4%	97.8%	31.19%	1	84.3%
♢ Psychology	-	100	800	25.95	25.59	94.1%	97.7%	31.01%	1	88.3%
♢ Engineering	-	100	800	26.59	24.45	97.8%	98.8%	31.59%	1	85.6%

Table 2: Comparison of the key statistics of the MisMatched set with MSciNLI and SciNLI.

diving. Finally, 153 full-text scholarly articles and reports related to the Centers for Disease Control and Prevention (CDC) and the U.S. Environmental Protection Agency (EPA) were collected using a manual PubMed search of biomedical literature from MEDLINE, life science journals, and online books. During the initial pre-processing of the collected papers, non-English and duplicate texts were removed. Only the open-source abstracts and full texts were used to construct our dataset.

3.2 Dataset Construction

To construct our MisMatched set, we follow a procedure similar to that employed for creating the test and development sets of SciNLI and MSciNLI. In phase 1, we automatically extract and annotate sentence pairs using the distant supervision method proposed by Sadat and Caragea (2022). In phase 2, we employ human annotators to curate the final test and development sets.

Phase 1: Automatic Data Extraction and Annotation. For the Entailment, Contrasting and Reasoning classes, we automatically extract adjacent sentence pairs where the second sentence starts with a linking phrase indicative of these relations. We then remove the linking phrase from the second sentence, and assign the label based on the semantic relation indicated by the linking phrase (as shown in Table 1). For example, if the second sentence starts with "Therefore" or "As a result," we extract and annotate the sentence pair with the Reasoning label. The mapping of linking phrases to labels can be seen in Appendix B.1.

For the Neutral class, we randomly pair two non-adjacent sentences from the same paper using 3 strategies: a) BothRand: two random sentences which do not contain any linking phrases are paired; b) FirstRand: a random sentence is paired with a second sentence from the other three classes; c) SecondRand: a random sentence is paired with a first sentence from the other three classes.

Phase 2: Human Annotation. To construct the final test and development sets, we hire annotators via a crowd-sourcing platform called COGITO.[2] Note that separate annotators are hired for the three domains to ensure that the annotators have the background knowledge and expertise necessary to understand domain-specific sentences. More details on annotators (e.g., pilot batch completion, pay, etc.) are available in Appendix B.2.

We perform the human annotations in an iterative fashion. In all iterations (except last), we randomly sample a balanced (over classes) subset of sentence pairs and ask three expert annotators to assign the label based only on the context available in the two sentences in each pair. Based on the consensus of the annotators, we assign a gold label to each example. The examples for which the gold label matches with the automatically assigned label based on distant supervision are included in the MisMatched set, and the rest are discarded. For each domain, we continue the iterations until we have at least 225 examples from each of the non-Neutral classes. For the Neutral class, we notice a lower agreement rate between the gold label and the automatically assigned label in all domains. Thus, for each domain, we perform a last iteration with all sentence pairs sampled from the Neutral class to obtain (at least) 225 examples where the automatically assigned Neutral label matches with the human annotated gold label. The distribution of the automatically assigned labels is not made available to the annotators for any batch.

In total, we annotate $3,253$ sentence pairs, among which $2,791$ have an agreement between the gold label and the automatically assigned label. The annotators showed a Fleiss-κ score of 0.72 among them (see Appendix B.2 for domain-wise breakdown). The domain-wise agreement rates between the gold label and the automatically assigned label can be seen in Table 2. We report the class-wise agreement rates in Appendix B.3.

[2]https://www.cogitotech.com/

<human>: Consider the following two sentences:
Sentence1: <sentence1>
Sentence2: <sentence2>
Based only on the information available in these two sentences, which of the following options is true?
a. Sentence1 generalizes, specifies or has an equivalent meaning with Sentence2.
b. Sentence1 presents the reason, cause, or condition for the result or conclusion made Sentence2.
c. Sentence2 mentions a comparison, criticism, juxtaposition, or a limitation of something said in Sentence1.
d. Sentence1 and Sentence2 are independent.
<bot>:

Table 3: Prompt template used for LLMs. Here, <X> indicates a placeholder X, which is replaced in the actual prompt.

Data Balancing. To ensure an equal representation of the classes and the domains, we randomly downsample all classes across domains to 225 (our domain-wise target size for each class). That is, the resulting MISMATCHED set contains 2,700 examples in total, uniformly distributed over the three domains (900 in each domain).

Data Split. We split the MISMATCHED set into test and dev sets at the paper level to prevent data leakage. Specifically, we randomly split the papers in each domain ensuring that there are at least 800 and 100 examples in the test and dev sets, respectively, with both sets being balanced over classes.

3.3 Data Statistics

We report the key statistics of our MISMATCHED set in Table 2. As we can see, the per-domain test size of MISMATCHED is the same as that of MSCINLI (both 800). While the per-domain dev size of MISMATCHED is smaller compared with MSCINLI, it still contains a satisfactory number of examples to be able to perform validation and hyper-parameter tuning. We can also see that the average number of words in the sentences in MISMATCHED is similar to that of the existing datasets. In addition, for both sentences, the percentage of sentences that have an "S" root according to the Stanford PCFG Parser (3.5.2) (Klein and Manning, 2003) is over 96%. This indicates that the vast majority of sentences in our dataset are syntactically complete. We can also see that the percentage of words that overlap between the two sentences is low, similar to the existing scientific NLI datasets.

4 Baselines

Since MISMATCHED only consists of dev and test, we use the training sets of SCINLI and MSCINLI, containing $101K$ and $127K$ sentence pairs, respectively, to establish the SLM and LLM baselines. Our implementation details are in Appendix C.

4.1 Models

SLM Baselines. We fine-tune the base variants of BERT (Devlin et al., 2019), SCIBERT (Beltagy et al., 2019), ROBERTA (Liu et al., 2019) and XLNET (Yang et al., 2019) as our SLM baselines, using the training sets of SCINLI, MSCINLI, and their combination denoted as MSCINLI+.

LLM Baselines. Our selection of LLMs focused on popular, instruction-tuned models representing recent advancements suitable for prompt-based NLI and reproducible research. We experiment with several open-source LLMs, including the *Llama-2-13b-chat-hf* variant of LLAMA-2 (Touvron et al., 2023), *Llama-3.1-8B-Instruct* variant of LLAMA-3 (Grattafiori et al., 2024), *Mistral-7B-Instruct-v0.3* variant of MISTRAL (Jiang et al., 2023) and *Phi-3-medium-128k-instruct* (containing 14B parameters) variant of PHI-3 (Abdin et al., 2024). Furthermore, to benchmark against leading proprietary models, we include GPT-4O (OpenAI et al., 2024) and GEMINI-1.5-PRO (Georgiev et al., 2024). All LLMs are evaluated in both zero-shot and few-shot settings. We use the best performing prompt constructed by (Sadat and Caragea (2024)) for MSCINLI, shown in Table 3. In the zero-shot setting, no exemplars are provided to the model. In the few-shot setting, we prepend four exemplars in the prompt (one per class) to harness the LLMs' in-context learning ability.

4.2 Results & Discussion

We report the domain-wise and overall Macro F1 of the LLMs in the zero-shot setting from a single run, since we use greedy decoding and therefore, there is no randomness involved. For all other experiments (with both SLMs and LLMs), we report the average and the standard deviations of the Macro F1 scores from three separate runs. Specifically, for the few-shot LLMs, we perform 3 runs with 3 randomly sampled sets of 4 exemplars from each SCINLI, MSCINLI, and MSCINLI+ following

MODEL	PSYCHOLOGY	ENGINEERING	PUBLIC HEALTH	OVERALL
BERT$_{SciNLI}$	68.59 ± 2.8	69.26 ± 2.3	66.57 ± 2.6	68.16 ± 2.5
BERT$_{MSciNLI}$	68.00 ± 1.4	69.23 ± 2.1	66.34 ± 1.2	67.89 ± 1.2
BERT$_{MSciNLI+}$	71.16 ± 0.9	73.52 ± 0.1	69.47 ± 1.3	71.41 ± 0.6
SCIBERT$_{SciNLI}$	76.24 ± 1.5	74.36 ± 1.4	78.14 ± 2.0	76.27 ± 1.6
SCIBERT$_{MSciNLI}$	76.98 ± 1.2	76.56 ± 0.8	77.97 ± 0.8	77.66 ± 0.8
SCIBERT$_{MSciNLI+}$	$\mathbf{79.18 \pm 0.4}$	76.50 ± 0.8	$\mathbf{78.79 \pm 0.3}$	$\mathbf{78.17 \pm 0.2}$
ROBERTA$_{SciNLI}$	75.76 ± 0.1	75.12 ± 0.7	75.34 ± 1.5	75.43 ± 0.5
ROBERTA$_{MSciNLI}$	75.05 ± 1.2	76.07 ± 0.8	74.89 ± 1.1	75.38 ± 1.0
ROBERTA$_{MSciNLI+}$	77.91 ± 0.3	$\mathbf{77.63 \pm 0.3}$	$\mathbf{78.79 \pm 1.0}$	78.11 ± 0.3
XLNET$_{SciNLI}$	73.61 ± 0.8	72.61 ± 0.7	73.23 ± 2.0	73.19 ± 1.0
XLNET$_{MSciNLI}$	73.24 ± 2.2	74.31 ± 1.0	73.19 ± 0.4	73.60 ± 1.2
XLNET$_{MSciNLI+}$	76.40 ± 1.0	75.44 ± 2.1	75.54 ± 0.9	76.49 ± 1.3
PHI-3$_{zs}$	55.38 ± 0.00	53.15 ± 0.00	49.31 ± 0.00	52.95 ± 0.00
PHI-3$_{fs-SciNLI}$	57.98 ± 1.31	55.46 ± 1.02	53.53 ± 0.77	55.84 ± 0.98
PHI-3$_{fs-MSciNLI}$	58.64 ± 1.11	$\mathbf{56.76 \pm 0.57}$	$\mathbf{55.68 \pm 0.25}$	$\mathbf{57.16 \pm 0.59}$
PHI-3$_{fs-MSciNLI+}$	$\mathbf{59.02 \pm 0.34}$	55.53 ± 0.80	55.54 ± 0.93	56.87 ± 0.25
LLAMA-2$_{zs}$	26.37 ± 0.00	32.71 ± 0.00	27.25 ± 0.00	28.98 ± 0.00
LLAMA-2$_{fs-SciNLI}$	43.92 ± 0.93	49.11 ± 1.54	45.09 ± 2.84	46.24 ± 1.71
LLAMA-2$_{fs-MSciNLI}$	44.83 ± 2.75	50.26 ± 1.63	45.45 ± 1.33	47.09 ± 1.88
LLAMA-2$_{fs-MSciNLI+}$	43.54 ± 2.00	49.05 ± 1.56	44.03 ± 2.04	45.79 ± 1.84
LLAMA-3$_{zs}$	33.67 ± 0.00	37.00 ± 0.00	30.87 ± 0.00	33.95 ± 0.00
LLAMA-3$_{fs-SciNLI}$	51.18 ± 1.11	46.88 ± 0.48	45.68 ± 2.26	48.19 ± 1.03
LLAMA-3$_{fs-MSciNLI}$	52.66 ± 1.15	47.54 ± 0.39	45.85 ± 0.72	48.94 ± 0.10
LLAMA-3$_{fs-MSciNLI+}$	53.92 ± 1.01	50.18 ± 1.10	48.13 ± 0.62	51.02 ± 0.51
MISTRAL$_{zs}$	31.14 ± 0.00	34.70 ± 0.00	25.85 ± 0.00	30.63 ± 0.00
MISTRAL$_{fs-SciNLI}$	44.26 ± 2.69	44.58 ± 2.59	39.79 ± 3.87	43.02 ± 3.03
MISTRAL$_{fs-MSciNLI}$	47.04 ± 1.82	47.12 ± 2.82	43.68 ± 2.68	46.09 ± 2.40
MISTRAL$_{fs-MSciNLI+}$	44.73 ± 0.37	45.46 ± 0.71	41.75 ± 2.14	44.09 ± 0.87
GPT-4O$_{zs}$	52.42 ± 0.00	50.12 ± 0.00	47.26 ± 0.00	50.26 ± 0.00
GPT-4O$_{fs-SciNLI}$	63.33 ± 1.52	61.34 ± 0.46	61.62 ± 0.50	62.29 ± 0.51
GPT-4O$_{fs-MSciNLI}$	62.65 ± 2.31	57.94 ± 1.84	58.61 ± 0.72	59.96 ± 1.62
GPT-4O$_{fs-MSciNLI+}$	63.62 ± 1.57	61.06 ± 0.83	$\mathbf{62.96} \pm 1.22$	62.73 ± 0.98
GEMINI-1.5-PRO$_{zs}$	54.28 ± 0.00	58.49 ± 0.00	51.59 ± 0.00	55.55 ± 0.00
GEMINI-1.5-PRO$_{fs-SciNLI}$	63.50 ± 1.92	61.94 ± 1.41	62.69 ± 1.03	62.78 ± 1.41
GEMINI-1.5-PRO$_{fs-MSciNLI}$	63.09 ± 0.86	61.74 ± 0.91	62.53 ± 0.59	62.51 ± 0.54
GEMINI-1.5-PRO$_{fs-MSciNLI+}$	$\mathbf{63.68} \pm 1.70$	$\mathbf{62.57} \pm 2.00$	62.51 ± 1.20	$\mathbf{62.95} \pm 1.50$

Table 4: Macro F1 scores (%) of the SLM and LLM baselines on different domains. Here, the subscript with the SLMs denotes the dataset used for fine-tuning the model. A subscript of zs with LLMs indicates zero-shot setting, and $fs - X$ indicates few-shot setting with four exemplars (one per class) from dataset X. Best scores within SLM, Open-Source LLM, and Proprietary LLM baselines per domain and overall are shown in **bold**.

the procedure detailed in Appendix D. For SLM, we perform 3 runs with 3 different random seeds.

The results can be seen in Table 4. Our findings are described below.

Fine-tuning SLMs on combined training sets yields better MISMATCHED performance As we can see from the results, the SLMs fine-tuned on SCINLI and MSCINLI generally show a similar performances on the MISMATCHED set. The performance shows consistent improvements across domains when the SLMs are fine-tuned on MSCINLI+, which is the combination of the training sets of SCINLI and MSCINLI. Therefore, fine-tuning the models using a training set with larger size and higher diversity enhances its robustness in an OOD setting. However, given that the best performing model with SCIBERT shows a Macro F1 of only 78.17%, there is a substantial headroom for future improvements.

Domain-specific pre-training is more useful for MISMATCHED than 'better' pre-training methods The results show that SCIBERT, ROBERTA and XLNET outperform BERT by a substantial margin in all domains. Note that the only difference between BERT and SCIBERT is that BERT was pre-trained using generic text from Wikipedia and BookCorpus, whereas SCIBERT was pre-trained using scientific text. Thus, the domain-specific pre-training of SCIBERT aids in achieving a better performance than BERT on MISMATCHED. Both ROBERTA and XLNET were pre-trained using general domain text similar to BERT. However, stronger (better) pre-training methods were employed in pre-training these two models and we observe their performance improvements for MISMATCHED over BERT. We can also see that XLNET shows a substantially lower Macro F1 than SCIBERT and ROBERTA. While the best performance results shown by ROBERTA and SCIBERT (both when fine-tuned on MSCINLI+), are almost

identical, SCIBERT outperforms ROBERTA in several cases (e.g., when they are fine-tuned on SCINLI or MSCINLI separately). Thus, domain-specific pre-training (on scientific documents) results in a better performance for MISMATCHED than 'better' (more robust) pre-training methods.

Fine-tuned Small Language Models outperform Prompt-based Large Language Models We can observe from Table 4 that the SLMs perform much better than even the leading prompt-based LLMs (such as GEMINI-1.5-PRO and GPT-4O) on all three domains. The average performance gap is approximately 15% between the highest performing SLM (SCIBERT) and the top-performing LLM. GEMINI-1.5-PRO and GPT-4O outperform open-source models in all few-shot settings, with GEMINI-1.5-PRO achieving the strongest overall performance in zero-shot settings, surpassing both GPT-4O and all open-source models. Among open-source LLMs (Table 4), PHI-3 demonstrates the best performance, outperforming LLAMA-2, LLAMA-3, and MISTRAL in both zero-shot and few-shot settings, indicating strong complex reasoning capabilities. Notably, PHI-3 few-shot with MSCINLI exemplars shows the best performance among open-source models. While GPT-4O's zero-shot capability was below PHI-3, it still outperformed other open-source baselines. The superiority of proprietary models is particularly evident in few-shot settings, where both GEMINI-1.5-PRO and GPT-4O show similar high performance and significantly outperform all open-source models, suggesting superior in-context learning ability for scientific NLI tasks. We show results with fine-tuned Llama-2 in Appendix E.

4.3 Analysis

In-Domain vs. Out-of-Domain Given that we establish our baselines using the training sets of SCINLI and MSCINLI, i.e., sentence pairs from CS domains, the baseline performance results reported on MISMATCHED in Table 4 are in the out-of-domain (OOD) setting. We now compare the OOD performances with the in-domain (ID) performance of the respective models. The ID performance is calculated by evaluating the model on the test set of the dataset it is trained on. We choose both SCIBERT and ROBERTA because of their strong performance on MISMATCHED. The results can be seen in Table 5.

First, we can observe that RoBERTa which is trained in a robust way shows a consistent drop in performance from ID to OOD especially when the model is fine-tuned individually on SciNLI or MSciNLI showing about 2-3% performance drop. These results demonstrate that the OOD is more challenging for RoBERTa. Second, training RoBERTa on increased diversity data (i.e., MSciNLI+) lowers the gap in performance between ID and OOD. These results show the impact of data diversity on model training and generalization to OOD data. Third, we can observe that while the SciBERT model (which is trained on scientific documents) performs worse than RoBERTa on ID data, its scientific knowledge that is learned from large training sets of research papers during its pre-training is beneficial for the scientific OOD data and in fact, it helps the SciBERT model to achieve similar performance with that of RoBERTa. These results show that scientific knowledge that is learned during the pre-training of SciBERT is retained and leveraged in the OOD setting and is more beneficial in OOD than training the model in a more robust way but on general (not specifically scientific) data. Similar to RoBERTa, we observe that the results with SciBERT in OOD when trained with increased diversity data (i.e., MSciNLI+) show that this diversity is beneficial on the OOD data.

Model	ID	OOD (MISMATCHED)
SCIBERT$_{SciNLI}$	77.11	76.27
SCIBERT$_{MSciNLI}$	76.66	77.66
SCIBERT$_{MSciNLI+}$	77.38	78.17
ROBERTA$_{SciNLI}$	78.24	75.43
ROBERTA$_{MSciNLI}$	77.02	75.38
ROBERTA$_{MSciNLI+}$	78.77	78.11

Table 5: Macro F1 (%) shown by SCIBERT and ROBERTA in ID and OOD (MISMATCHED) settings.

#Shot	PSY	ENGG	PH	OVERALL
4-SHOT	58.64	56.76	55.68	57.16
8-SHOT	58.80	57.15	56.82	57.71
12-SHOT	59.63	58.29	56.69	58.32
16-SHOT	59.56	57.97	56.50	58.14

Table 6: 4-shot, 8-shot, 12-shot and 16-shot Macro F1 (%) by PHI-3. Here, PSY: PSYCHOLOGY, ENGG: ENGINEERING, and PH: PUBLIC HEALTH.

Few-shot Scaling Experiments While proprietary models (GEMINI-1.5-PRO and GPT-4O) demonstrated superior performance, we selected PHI-3 for few-shot scaling experiments as the best performing open-source model. This choice enables comprehensive analysis of in-context learning mechanisms with full reproducibility and extensive experimentation without API constraints. We used

Model \ Dataset	Sentence Input	PSYCHOLOGY	ENGINEERING	PUBLIC HEALTH	Macro Ave.
RoBERTa$_{MSciNLI+}$	BOTH SENTENCES	77.91	77.63	78.79	78.11
RoBERTa$_{MSciNLI+}$	ONLY 2ND SENTENCE	53.17	58.59	52.05	54.64
SciBERT$_{MSciNLI+}$	BOTH SENTENCES	79.18	76.50	78.79	78.17
SciBERT$_{MSciNLI+}$	ONLY 2ND SENTENCE	56.68	58.12	54.54	56.50

Table 7: Comparison of Macro F1 scores (%) for RoBERTa and SciBERT on MISMATCHED domains when using both premise and hypothesis sentences versus only the hypothesis sentence as input.

Model	Contrasting	Reasoning	Entailment	Neutral	Macro Ave.
SciBERT					
Psychology	81.60 ± 0.9	74.15 ± 1.1	79.97 ± 1.5	80.99 ± 0.2	79.18 ± 0.4
Engineering	80.98 ± 0.4	76.50 ± 1.1	75.09 ± 1.4	73.43 ± 1.0	76.50 ± 0.8
Public Health	80.25 ± 0.4	74.55 ± 0.3	80.09 ± 1.0	80.31 ± 0.9	78.79 ± 0.3
MisMatched	80.94 ± 0.5	75.09 ± 0.6	78.44 ± 0.3	78.22 ± 0.6	78.17 ± 0.2
Phi-3					
Psychology	70.28 ± 1.37	40.60 ± 2.95	62.44 ± 0.62	61.27 ± 0.94	58.65 ± 1.10
Engineering	71.35 ± 1.30	47.10 ± 2.90	51.20 ± 2.07	57.42 ± 0.31	56.77 ± 0.57
Public Health	67.53 ± 0.73	44.08 ± 4.09	52.51 ± 2.27	58.62 ± 2.11	55.68 ± 0.25
MisMatched	69.67 ± 0.40	44.07 ± 3.27	55.92 ± 0.89	59.00 ± 0.97	57.16 ± 0.59

Table 8: Class-wise F1 (%) and their macro averages (%) of our best performing SLM and LLM baselines on each domain in MisMatched and their combination.

exemplars from MSciNLI given its superior performance on MisMatched among all settings (as discussed in 4.2). Our experiments investigate the impact of increasing few-shots from 4 to 8, 12, and 16 on PHI-3's performance, with results shown in Table 6. Results show that 12-shots achieve slightly better performance than 4-shots and 8-shots, while performance drops at 16-shots.

Hypothesis-only Baseline Experiment To verify whether our dataset contains spurious correlations or not, i.e., any stylistic artifacts that are present only in the hypotheses and are indicative of the label (without the need for the premise), we compare hypothesis-only models against full premise-hypothesis models using RoBERTa and SciBERT fine-tuned on MSciNLI+ in Table 7. We chose MSciNLI+ as the training set because fine-tuned models (RoBERTa and SciBERT) achieved their highest performance on MisMatchedwhen trained on MSciNLI+ compared to SciNLI or MSciNLI alone (Table 4). Results show significant performance degradation when using only the hypothesis compared to the full premise-hypothesis input, demonstrating that premise-hypothesis understanding is critical for model performance. Thus, our dataset does not exhibit hypothesis-only artifacts.

Class-wise Performance We report the per-class F1 scores of our best performing SLM baseline, SciBERT (fine-tuned using MSciNLI+) and best performing LLM baseline, PHI-3 (in the few-shot setting with MSciNLI exemplars) in Table 8. We can see that generally, both types of models show lower F1 scores for the Reasoning class compared with the other classes. Therefore, recognizing a Reasoning relation between sentences is more challenging than recognizing other scientific NLI relations. We provide an in-depth analysis of the "reasoning" relation in Appendix F.

5 Harnessing Implicit Relations

Existing training sets for scientific NLI datasets (i.e., SciNLI and MSciNLI) only include sentence pairs where the relation between them is made *explicit* with linking phrases. We posit that, if two sentences are adjacent, there can potentially be a scientific NLI relation between them despite the second sentence not starting with a linking phrase. We define the relation between these sentence pairs as an *implicit* relation. Here, we propose to incorporate adjacent sentences with *implicit* relations in model training and analyze their impact on models' performance. We detail below the data sources from which we extract implicit sentence pairs, how we annotate them, and how we use them in model training.

Data The implicit sentence pairs are sourced from the research papers from SciNLI, MSciNLI and MisMatched separately. For each dataset, we extract the adjacent sentence pairs in which none of the sentences contain any linking phrases as the examples potentially containing an *implicit* Entailment/Contrasting/Reasoning relation. For the Neutral class, we randomly pair

Model \ Dataset	SciNLI	MisMatched
SciBERT$_{MS+}$	79.04	78.17
SciBERT$_{MS+\ +\ Impl}$	**79.44**	**79.66**
Phi-3$_{fs-Expl\{SciNLI\}}$	59.67 ± 1.92	55.84 ± 0.98
Phi-3$_{fs-Impl\{SciNLI\}}$	$\mathbf{61.41} \pm 1.11$	$\mathbf{56.56} \pm 1.47$
Phi-3$_{fs-Expl\{MSciNLI\}}$	59.88 ± 1.15	57.16 ± 0.59
Phi-3$_{fs-Impl\{MSciNLI\}}$	$\mathbf{60.58} \pm 0.43$	$\mathbf{57.50} \pm 0.06$
Phi-3$_{fs-Expl\{MisMatched\}}$	58.57 ± 1.29	56.96 ± 0.68
Phi-3$_{fs-Impl\{MisMatched\}}$	$\mathbf{61.03 \pm 0.40}$	$\mathbf{58.26 \pm 0.25}$

Table 9: Performance comparison between models utilizing *implicit* relations with models only using *explicit* examples. Here, MS+: MSciNLI+, Expl: *explicit* and Impl: *implicit*.

two non-adjacent sentences selected from the other three classes. For SciNLI and MSciNLI, we extracted the implicit pairs from papers that are part of the training set, whereas for MisMatched, we extracted the implicit pairs from papers that are not utilized to construct its test and development sets. We extracted $\approx 210K$ and $\approx 120K$ implicit sentence pairs for SciNLI/MSciNLI and MisMatched respectively, with the number of implicit relations being about twice as large as explicit relations. We provide examples of implicit sentence pairs extracted from different domains in our MisMatched dataset in Appendix G.

Implicit Relation Annotation Next, we identify the implicit scientific NLI relation among the extracted sentence pairs in three steps: a) assign pseudo-labels to the extracted sentence pairs based on the predictions made by the SciBERT model fine-tuned on MSciNLI+; b) filter the examples based on a confidence (i.e., the probability for the predicted pseudo-label by the model) threshold of 0.6; and c) filter the examples where a Contrasting/Entailment/Reasoning label is predicted for a non-adjacent sentence pair or a Neutral label for an adjacent sentence pair.

Incorporating Implicit Relations We incorporate *implicit* relations in model training by experimenting with SciBERT and Phi-3 and evaluating their performance on the test sets of SciNLI and MisMatched. For SciBERT, we first fine-tune an *out-of-the-box* model using the selected *implicit* examples from the same domain as the test set (i.e., when the test set is MisMatched, the implicit examples are from papers from the MisMatched domains). We then continue fine-tuning the model using the *explicit* examples from MSciNLI+. For Phi-3, we randomly sample four examples (one from each class) from the selected *implicit* set, and use them as the exemplars in the few-shot setting.

Results Table 9 shows a comparison between the models that use only *explicit* examples with their counterparts that incorporate *implicit* examples. As we can see, the Macro F1 of SciBERT improves by 1.5% for MisMatched when implicit relations are incorporated into model training. In addition, the performance of Phi-3 also shows improvement in Macro F1 when *implicit* examples are used as the few-shot exemplars compared to *explicit* examples from SciNLI, MSciNLI, MisMatched used as exemplars. Given that all sentence pairs from SciNLI, MSciNLI and MSciNLI+ are out-of-domain for MisMatched, incorporating in-domain *implicit* relations into models' training helps improve its performance. Interestingly, when Phi-3 with few-shot exemplars from MisMatched is evaluated on SciNLI, we can see an improvement of 2.46 (from 58.57 to 61.03) which demonstrates the benefits of using implicit relations that make the model more robust and capable to generalize better. Thus, given the improvements for both datasets, we can conclude that sentence pairs with *implicit* relations can be a valuable resource for exposing scientific NLI models to more diverse data that can further improve the performance.

6 Conclusion & Future Directions

In this paper, we introduce a MisMatched testbed for scientific NLI, derived from non-CS domains unlike the existing datasets. We establish strong baselines on the MisMatched set with both SLMs and LLMs using the training sets from SciNLI and MSciNLI. Our results show that the best performing baseline achieves a Macro F1 of only 78.17%, illustrating the substantial room for future improvements. Furthermore, we show that sentence pairs containing an *implicit* scientific NLI relation can aid in improving the performance of two scientific NLI benchmarks. In our future work, we will develop domain adaptation methods for scientific NLI to improve the performance on the MisMatched set.

Acknowledgments

We thank US-NSF for support from grant IIS-2107518 and UIC Discovery Partners Institute which supported the research and the computation in this study. Research reported in this publication was also partially supported by the CNAP Center of Biomedical Research Excellence of the NIH under grant No. P20GM113109.

Limitations

Our MISMATCHED benchmark indeed enhances the diversity in scientific NLI to non-CS domains. However, there are numerous scientific domains and disciplines (e.g., Physics, Chemistry, etc.) that are not covered by our dataset. Therefore, a future research direction is to study scientific NLI to other non-CS domains that can serve as a more robust and generalized benchmark.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. *TAC*, 7(8):1.

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.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit Optimizers via Block-wise Quantization. *arXiv preprint arXiv:2110.02861*. Published as a conference paper at ICLR 2022.

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.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.

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*.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, volume 7, pages 785–794.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Maël Jullien, Marco Valentino, and André Freitas. 2024. Semeval-2024 task 2: Safe biomedical natural language inference for clinical trials. *arXiv preprint arXiv:2404.04963*.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and André Freitas. 2023. Nli4ct: Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan. Association for Computational Linguistics.

Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S Gerber, and Laura E Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 364–371. IEEE.

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*.

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

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

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*.

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

Mobashir Sadat and Cornelia Caragea. 2022. SciNLI: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7399–7409, Dublin, Ireland. Association for Computational Linguistics.

Mobashir Sadat and Cornelia Caragea. 2024. MsciNLI: A diverse benchmark for scientific natural language inference. In *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

A Datasets for NLI

Table 10 shows a comparison of relevant datasets in terms of sources from which data was collected, domains covered, classes, in-domain (ID) and out-of-domain (OOD) training, real or synthetic (generated) hypothesis and dataset size (as number of sentence pairs).

B Details on Data Annotation

B.1 Linking Phrases Used in Distant Supervision

The linking phrases and their classes used in the distant supervision method for automatically extracting and annotating sentence pairs in MISMATCHED can be seen in Table 11.

B.2 Details about Annotators and Inter-Annotator Agreement

We hire separate annotators for each of the three domains in our dataset via a cloud-sourcing platform called COGITO[3]. For each domain, we complete 3 pilot batches containing 52 sentence pairs (balanced over classes). After each pilot batch, we provide feedback to the annotators on their work and ask them for their acknowledgement of our feedback before starting the next batch. The annotators are paid at a rate of $0.6/sample.

The inter-annotator agreement varied across domains, as shown in Table 12. PSYCHOLOGY showed the highest agreement (FLEISS-K = 0.78), followed by ENGINEERING (0.70) and PUBLIC HEALTH (0.65). The variation in agreement rates likely reflects the differing complexity and ambiguity levels inherent to scientific texts across these domains.

B.3 Class-wise Agreement Rates

The total number of sentence pairs annotated for each class and the agreement rate between the gold label and automatically assigned label are shown in Table 13. As we can see, for the CONTRASTING, REASONING and ENTAILMENT classes, there is a very high agreement between the human annotated gold label and the automatically annotated label based on distant supervision. This indicates that the annotators possesses a solid understanding of the scientific NLI task. In contrast, the agreement rate for the NEUTRAL class is low (only 68.3%) compared to the $> 93\%$ agreement rates for the other classes. This is because, unlike SCINLI and MSCINLI (where sentence pairs are extracted from full text of the papers), most sentence pairs in MISMATCHED are extracted from abstracts of the papers. Given the small number of sentences in paper abstracts, even non-adjacent sentences remain related in many cases resulting in a low agreement for the NEUTRAL class.

C Implementation Details

SLM Baselines We utilize the huggingface[4] implementations for our SLM baselines in the experiments. For these models, we concatenate the sentence in each pair with a `[SEP]` token between them and append a `[CLS]`. We then project the representation for the `[CLS]` token with a weight matrix $\mathbf{W} \in \mathbb{R}^{d \times 4}$. This projection is then sent as the input to a softmax activation to get the predicted probability distribution over the four classes.

Each model is fine-tuned for five epochs on different training sets (SCINLI, MSCINLI, MSCINLI+). Early stopping with a patience of 2 epochs is employed while fine-tuning the SLMs. We use the Macro F1 score on the development set of MISMATCHED as the early stopping criteria. For all SLM baselines, we use a learning rate of $2e-5$ and a mini-batch size of 64. We fine-tune the models using the Adam (Kingma and Ba, 2014) optimizer, and the cross-entropy loss.

LLM Baselines For open-source LLMs (LLAMA-2, LLAMA-3, MISTRAL and PHI-3), we utilize the Hugging Face library, employing a greedy decoding strategy with no random sampling and a maximum generated token limit of 40. Proprietary models GPT-4O and GEMINI-1.5-PRO were evaluated via their respective official APIs, specifying the model identifiers as "gpt-4o" and "models/gemini-1.5-pro" respectively. For GPT-4O, deterministic output was ensured by setting temperature=0.0. For GEMINI-1.5-PRO, default API generation settings were used without specifying temperature or other generation parameters. Our evaluation scripts for both proprietary models incorporated retry logic (up to 3 attempts upon API failure).

Fine-Tuned LLM Baseline For the fine-tuned LLAMA-2 experiments (results presented in Appendix E), we employed Parameter-Efficient Fine-Tuning (PEFT) using Low-Rank Adaptation

[3] https://www.cogitotech.com/

[4] https://huggingface.co/

Dataset	Source/Domains	Classes	ID	OOD	Hypothesis	≈ Size
RTE (Wang et al., 2018)	Wikipedia and news sources	2 | *entailment*, *non-entailment*	✓	✗	Synthetic	2,500
SICK (Marelli et al., 2014)	Image captions and video descriptions	3 | *entailment*, *contradiction* *neutral*	✓	✗	Synthetic	10,000
SNLI (Bowman et al., 2015)	Image captions	3 | *entailment*, *contradiction* *neutral*	✓	✗	Synthetic	570,000
MULTINLI (Williams et al., 2018)	Nine sources from second OANC release (Face-to-face, government, letter, etc.) & Fiction (mystery, humor, western, etc.)	3 | *entailment*, *contradiction* *neutral*	✓	✓	Synthetic	433,000
ANLI (Nie et al., 2020)	Wikipedia, news, fiction, spoken text, etc.	3 | *entailment*, *contradiction* *neutral*	✓	✗	Synthetic	170,000
MEDNLI (Romanov and Shivade, 2018)	MIMIC-III, clinical notes (Past Medical History)	3 | *entailment* *contradiction* *neutral*	✓	✗	Real	14,000
NLI4CT dataset (Jullien et al., 2023)	Breast cancer clinical trial reports (U.S. National Library of Medicine)	2 | *entailment* *contradiction*	✓	✗	Synthetic	2,400
NLI4CT-P (Jullien et al., 2024)	Breast cancer clinical trial reports (U.S. National Library of Medicine)	2 | *entailment* *contradiction*	✓	✗	Synthetic	8,600
SCINLI (Sadat and Caragea, 2022)	Research articles from ACL Anthology	4 | *entailment* *reasoning* *contrasting* *neutral*	✓	✗	Real	101,000
MSCINLI (Sadat and Caragea, 2024)	Computer science research articles, HARDWARE, NETWORKS, SOFTWARE & ITS ENGINEERING, etc.	4 | *entailment* *reasoning* *contrasting* *neutral*	✓	✗	Real	127,000
MISMATCHED (ours)	Research articles from PUBLIC HEALTH PSYCHOLOGY and ENGINEERING	4 | *entailment* *reasoning* *contrasting* *neutral*	✗	✓	Real	2,700

Table 10: Comparison of relevant NLI datasets. The *Source/Domains* column indicates the sources of data collection and/or the domains covered by the dataset. The *Classes* column indicates the number of classes, followed by specific classes in the dataset. The *ID* and *OOD* columns indicate if the dataset is *in-domain* (i.e., contains both training and test data for some domains) and/or *out-of-domain* (i.e., contains only test data for some domains. *Hypothesis* refers to the fact that the hypothesis is *Real* (extracted directly from existing text) or *Synthetic* (written or re-written by human annotators). Finally, the last column, ≈ *Size* refers to the approximate numbers of pairs in the dataset (note that some datasets may have a smaller number of premises).

Class	Linking Phrases
CONTRASTING	'However', 'On the other hand', 'In contrast', 'On the contrary'
REASONING	'Therefore', 'Thus', 'Consequently', 'As a result', 'As a consequence', 'From here, we can infer'
ENTAILMENT	'Specifically', 'Precisely', 'In particular', 'Particularly', 'That is', 'In other words'

Table 11: Linking phrases used to extract sentence pairs and their corresponding classes.

Domain	PSY	ENG	PH
FLEISS-K	0.78	0.70	0.65

Table 12: Inter-annotator agreement (FLEISS-K) by domain. Here, PSY: PSYCHOLOGY, ENGG: ENGINEERING, and PH: PUBLIC HEALTH.

(LoRA). The model was fine-tuned specifically on the SCINLI training dataset. Key hyperparameters were configured as follows: LORA rank (r) was set to 16 with alpha of 32, and LORA dropout was set to 0.05. The model underwent training for 3 epochs with a learning rate of $2e-3$. We used a per-device batch size of 32 with 4 gradient accumulation steps, resulting in an effective batch size of 128. Training employed the adamw_bnb_8bit (Dettmers et al., 2021) optimizer with mixed precision (fp16) training. The fine-tuned model was then evaluated on both SCINLI and MISMATCHED test sets to assess cross-domain performance, with detailed results provided in Table 15 of Appendix E.

Computational Cost. We fine-tune each SLM baseline using a single NVIDIA RTX A5000 GPU. It takes $\approx$ 2 hours to fine-tune each SLM on SCINLI and MSCINLI, and $\approx$ 4 hours to fine-tune them on MSCINLI+. For our LLM baselines (LLAMA-2, LLAMA-3, MISTRAL and PHI-3), we utilize one NVIDIA A100-SXM4-80GB GPU. The inference time for all LLMs for MISMATCHED is ≈ 0.25 hours in the zero-shot setting, and ≈ 3.5 hours in the few-shot (4-shot) setting. The few-shot experiments for SCINLI require ≈ 4 hours to complete.

D Few-shot Exemplar Selection

To ensure robust and reliable few-shot performance evaluation, we employed a systematic approach for exemplar selection and ordering across all experiments.

Class	#Annotated	Agreement
CONTRASTING	744	93.5%
REASONING	744	93.4%
ENTAILMENT	744	96.2%
NEUTRAL	1021	68.3%
Overall	3253	85.7%

Table 13: Number of sentence pairs annotated manually for each class and their agreement rate between the gold labels and automatically assigned labels.

EXEMPLAR SELECTION AND ORDERING: For each k-shot experiment, we conducted 3 independent runs to obtain reliable results. In each run, we randomly selected k exemplars (one from each class for balanced representation). The same set of k exemplars was used consistently throughout that entire run for all test examples. The order of exemplars in the prompt was kept identical across all test instances within each run. Final results reported in our tables represent the mean performance and standard deviation computed across these 3 independent runs.

MSCINLI+ EXEMPLAR HANDLING: Given that MSCINLI+ combines SCINLI and MSCINLI datasets, we implemented specific procedures to ensure exemplars truly represent this combined nature. For each independent run on MSCINLI+, we: (1) randomly selected initial candidate exemplars separately from SCINLI and MSCINLI datasets, (2) formed 4-shot prompt combinations from these candidates with the strict requirement that each combination must include at least one exemplar from both original datasets (SCINLI and MSCINLI), and (3) selected three such combinations for our three independent runs. This approach guaranteed that MSCINLI+ exemplars always reflected the diverse nature of the combined dataset rather than being dominated by examples from a single source dataset.

E Results with Fine-Tuned Llama-2

We show results of fine-tuned Llama-2 on SciNLI using LoRA. The Macro F1 of this fine-tuned LLM can be seen in Table 15.

As we can see, while the performance improves substantially over the prompt based version of the model, there are still differences across the datasets. The in-domain Macro F1 of this model on SciNLI is 83.83%, which drops to 82.87% for MisMatched. These results further illustrate the unique linguistic characteristics of the two datasets.

Domain	First Sentence	Second Sentence	Class
ENGINEERING	Tools to predict its vibratory and acoustic performance at the design stage need to be developed.	an improved finite element model has been developed to analyse the vibration behaviour of a Permanent Magnet Synchronous Machine of a lift installation using the finite element software ABAQUS.	REASONING
PSYCHOLOGY	This literature review provides information for identifying children who have been abused and neglected but exposes the need for a comprehensive screening instrument or protocol that will capture all forms of child abuse and neglect.	screening needs to be succinct, user-friendly, and amenable for use with children at every point of care in the healthcare system.	CONTRASTING

Table 14: Examples of implicit sentence pairs from MISMATCHED, extracted from different domains. Unlike explicit relations marked by linking phrases (as shown in Table 1), these pairs contain implicit discourse relations without explicit connective markers.

SciNLI	MisMatched
83.83%	82.87%

Table 15: Results of Llama-2 fine-tuned on SciNLI.

F Analysis of the "Reasoning" Relation

We provide here an in-depth analysis of the "reasoning" relation which is more challenging than the other relations in our MISMATCHED dataset. Specifically, we show a confusion matrix between the true labels and the predicted labels by SciBERT (our best performing baseline) on the MISMATCHED test set in Table 16.

True \ Predicted	C	R	E	N
C	532	23	30	15
R	60	428	79	33
E	55	32	485	28
N	71	62	39	428

Table 16: Confusion matrix of SciBERT on MisMatched. C: Contrasting; R: Reasoning; E: Entailment; N: Neutral.

As we can see, the "reasoning" relation is often mistaken with "entailment" by the model. In addition, a fair number of "reasoning" relations are also mistaken as "contrasting" by the model. This results in a lower Macro F1 for the "reasoning" class compared to the other classes.

G *Implicit* Relations

Novelty of Implicit Relations. The "implicit" relations as defined here can help open new directions of research, e.g., to improve discourse coherence analysis by suggesting linking phrases between contiguous sentences for better reading comprehension and natural language understanding.

Examples of Implicit sentence pairs from MISMATCHED Table 14 illustrates representative examples from ENGINEERING and PSYCHOLOGY domains, where REASONING and CONTRASTING relations must be inferred without explicit connective markers (i.e., without explicit linking phrases between the two sentences).

Further Details on Experimental Setup for *Implicit* Relations In our experiments with *implicit* relations in Section 5, for MISMATCHED, SCINLI and MSCINLI, we utilize the SCIBERT model fine-tuned as our baseline to predict the labels of the extracted sentence pairs which potentially contain *implicit* relations. However, for predicting the label for the sentence pairs extracted for SCINLI, we fine-tune a separate SCIBERT model using MSCINLI+ for training and the development set from SCINLI for early stopping. All other implementation details (e.g., learning rate, batch size) are the same as for the SLM baselines.

After selecting the implicit relations based on the models' (fine-tuned on MSCINLI+) predictions, we fine-tune an out-of-the-box SCIBERT model on these selected examples using the same hyperparameters as for the SLM baselines.

The last checkpoint of the model fine-tuned on sentence pairs with *implicit* relations is further fine-tuned on MSCINLI+. Specifically, we initialize the language model layers of SCIBERT from the model fine-tuned in the previous step. However, the weight matrix $\mathbf{W} \in \mathbb{R}^{d \times 4}$ (which projects the `[CLS]` representation to get the probability distribution over the classes) is reinitialized randomly. Furthermore, we use a lower learning rate of $2e-6$ for fine-tuning the models in this step.