Rule Discovery for Natural Language Inference Data Generation Using Out-of-Distribution Detection

Juyoung Han[1] **Hyunsun Hwang**[2] **Changki Lee**[2*]
[1]Department of Medical Big Data Convergence, Kangwon National University
[2]Department of Computer Science and Engineering, Kangwon National University
jo00oh82@kangwon.ac.kr hhs4322@kangwon.ac.kr leeck@kangwon.ac.kr

Abstract

Natural Language Inference (NLI) is a fundamental task in Natural Language Processing (NLP). However, adapting NLI models to new domains remains challenging due to the high cost of collecting domain-specific training data. While prior work proposed 15 sentence transformation rules to automate training data generation, these rules do not sufficiently capture the diversity of natural language. We propose a novel framework that combines Out-of-Distribution (OOD) detection and BERT-based clustering to identify premise–hypothesis pairs in the SNLI dataset that are not covered by existing rules and discover four new transformation rules from them. Using these rules with Chain-of-Thought (CoT) prompting and Large Language Models (LLMs), we generate high-quality training data and augment the SNLI dataset. Our method yields consistent performance improvements across dataset sizes, achieving +0.85%p accuracy on 2k samples and +0.15%p on 550k samples. Furthermore, a distribution-aware augmentation strategy enhances performance across all scales. Beyond manual explanations, we extend our framework to automatically-generated explanations (CoT-Ex), demonstrating that they provide a scalable alternative to human-written explanations and enable reliable rule discovery.

1 Introduction

Natural Language Inference (NLI) is a Natural Language Processing (NLP) task that involves understanding and inferring logical relationships between premise and hypothesis sentences, classifying their relationship as entailment, contradiction, or neutral. In supervised learning-based NLI research, models are trained using Premise-Hypothesis-Label (PHL) datasets. However, applying NLI models to new domains requires constructing domain-specific training data, which demands substantial time and cost. Although previous research attempted to automatically generate training data using sentence transformation rules, the existing 15 transformation rules (Varshney et al., 2022) were insufficient to comprehensively cover the diverse NLI patterns in real-world scenarios.

This paper proposes a novel approach that combines Out-of-Distribution (OOD) detection and clustering to overcome the limitations of existing rules. Our method employs a combination of Maximum Softmax Probability (MSP), Temperature Scaling (TS), and Input Preprocessing (IP) for OOD detection to identify new premise-hypothesis pairs that cannot be explained by existing rules. These identified premise-hypothesis pairs are grouped using BERT (Devlin et al., 2019) embedding-based k-means clustering (Sinaga and Yang, 2020), and as a result of manual analysis, four new transformation rules were discovered: Role Generalization (RG), Contextual Augmentation (CA), Visual Specification (VS), and Emotion Inference (EI). To further reduce reliance on manual analysis in the rule discovery stage described above, we introduced an automated rule discovery step that leverages Large Language Models (LLMs) to generate new rule candidates and validate the generated rules. Using this automated process, we successfully discovered 5 rules, three of which matched the rules identified through manual analysis. The derived rules were used to generate high-quality training data using LLMs and Chain-of-Thought (CoT) prompting (Wei et al., 2022).

The experimental results demonstrated that our method achieved consistent performance improvements regardless of the size of the training data, with improvements of 0.85%p for small-scale datasets (2k) and 0.15%p for large-scale datasets (550k), validating the effectiveness of our methodology. Furthermore, our data distribution-aware augmentation strategy showed consistent performance improvements across all dataset sizes (2k-

* Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 25983–26003
November 4-9, 2025

550k), demonstrating its effectiveness for data augmentation in NLI tasks. In summary, our contributions are as follows:

- We propose a novel (semi-)automated framework for discovering transformation rules by leveraging OOD detection and clustering techniques.
- We empirically validate the effectiveness of our proposed framework across diverse dataset scales.
- Our data distribution-aware augmentation strategy showed consistent performance improvements across all dataset sizes (2k-550k), validating its effectiveness for data augmentation in NLI tasks.

2 Related Work

Previous research related to natural language inference can be examined from two perspectives: supervised learning-based NLI and automatic generation of training data.

2.1 Supervised Learning-based NLI

In supervised learning-based natural language inference research, NLI models are trained using Premise-Hypothesis-Label (PHL) datasets (Varshney et al., 2022). With the release of large-scale datasets such as SNLI version 1.0 (Bowman et al., 2015) and the emergence of transformer-based pre-trained models (Vaswani, 2017) such as BERT, NLI performance has improved significantly. However, applying NLI models to new domains requires constructing domain-specific training data, which demands substantial time and human effort.

2.2 Automatic Generation of Training Data

To reduce data construction costs, various automatic training data generation methods have been proposed. Varshney et al. (2022) focused on using predefined sentence transformation rules with WordNet, Gensim, and ConceptNet to automatically generate hypothesis sentences from given premise sentences (Miller, 1992; Rehurek and Sojka, 2011; Speer et al., 2017). However, the existing 15 transformation rules were insufficient to comprehensively cover the diverse patterns in NLI. Cho et al. (2023) investigated using CoT and few-shot learning with LLMs, to generate data through step-by-step reasoning processes (Brown et al., 2020; Mersinias and Valvis, 2022), though ensuring the quality and diversity of generated data remains a challenge.

3 OOD Detection and New Rule Discovery for NLI Data Generation

In this research, we propose a methodology that uses OOD detection techniques to identify premise-hypothesis pairs in existing NLI training data that fall outside the patterns covered by the current 15 transformation rules. Our approach leverages OOD detection techniques based on Maximum Softmax Probability (Hendrycks and Gimpel, 2018), enhanced with Temperature Scaling (Hinton et al., 2015) and Input Preprocessing (Liang et al., 2020) to improve detection performance. The identified OOD premise-hypothesis pairs are clustered using k-means clustering, and new rules are derived from selected clusters filtered based on their cohesion scores. These derived rules are used with LLMs and CoT prompting to generate additional NLI training data, leading to demonstrated improvements in NLI model performance.

3.1 OOD Detection for Discovering New Rules in NLI Training Data Generation

This paper employs OOD detection techniques for discovering new rules for NLI training data generation through the following process:

1. We extract 15,000 premise sentences from the training set of the SNLI dataset and utilize an LLM to apply CoT prompting for each of the 15 existing rules to the extracted sentences, constructing a new Premise-Hypothesis-Label (PHL) dataset of 15,000 instances.
2. The labels in the constructed PHL dataset are modified to 'NLI Label + rule name' and the modified PHL dataset is then used for fine-tuning a pre-trained BERT-base model. The fine-tuned model takes the premise and hypothesis sentences as input and classifies the relationship between the sentence pairs into the new 15 categories (i.e., NLI label + rule name).
3. The fine-tuned model and OOD detection technique are applied to the premise-hypothesis pairs in the training set of the SNLI dataset, categorizing premise-hypothesis pairs as In-Distribution (ID) if they match existing 15 transformation rules and Out-of-Distribution (OOD) otherwise.

3.1.1 OOD Detection Using MSP

We employ Maximum Softmax Probability (MSP) as our baseline method for OOD detection. MSP leverages softmax probabilities to measure confidence scores for class predictions in deep learning classification tasks, using these scores to determine whether inputs are OOD. Specifically, MSP uses the maximum value from the model's softmax output as the OOD score, classifying inputs as ID if this score exceeds a predetermined threshold, and as OOD otherwise.

3.1.2 OOD Detection Using Temperature Scaling

A known limitation of MSP is that it tends to overestimate prediction probabilities relative to actual model accuracy. To mitigate this issue, Temperature Scaling (TS) is applied to calibrate the output probabilities. The mathematical formulation is as follows:

$$S_i(x;T) = \frac{\exp\left(f_i(x)/T\right)}{\sum_{j=1}^{N}\exp\left(f_j(x)/T\right)} \quad (1)$$

TS adjusts softmax outputs by scaling them with a temperature parameter (Guo et al., 2017), which calibrates prediction probabilities to better align with actual probabilities while preserving the model's class predictions. In this paper, we experimentally determined the optimal temperature values for OOD detection and improved detection performance by minimizing calibration error rates.

3.1.3 OOD Detection Using Input Preprocessing

Input Preprocessing (IP) is a method that modifies the model prediction probability distribution by applying small perturbations ϵ to the input x. The mathematical formulation is:

$$\tilde{x} = x - \epsilon \text{sign}\left(-\nabla_x \log S_{\hat{y}}(x;T)\right) \quad (2)$$

This method effectively differentiates between ID and OOD data by adjusting the inputs to maximize softmax probabilities, resulting in higher probabilities for ID data and lower probabilities for OOD data. In this paper, we extend the IP technique, originally proposed for computer vision tasks, to suit the NLI task. Specifically, we applied IP to BERT word embeddings, which represent text as vectors in a high-dimensional embedding space. The input x is defined as:

$$x = WordEmbedding_{BERT}(token_{seq}) \quad (3)$$

We tuned the perturbation scale $\epsilon \in [0.01, 0.09]$, refining the search to $[0.031, 0.039]$, and selected $\epsilon = 0.033$ as optimal parameter. Similarly, we searched temperature values $T \in [10, 1000]$ and found $T = 1000$ to be the best parameter. See Appendix A for detailed results.

3.2 Derivation of New Rules through Clustering Analysis

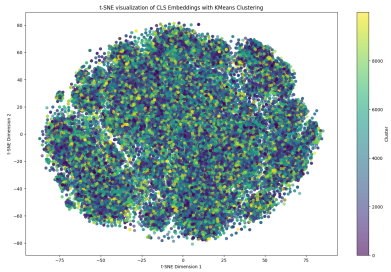
In this paper, we conducted OOD detection on 550,152 premise-hypothesis pairs from the SNLI training set, identifying 50,000 premise-hypothesis pairs that fall outside the patterns covered by the existing 15 transformation rules. For OOD detection, we employed a combined MSP+TS+IP approach and used a threshold of 0.07186 to identify 50,000 premise-hypothesis pairs as OOD. The identified OOD premise-hypothesis pairs were clustered using k-means clustering, and the cohesion of each cluster was measured based on the average Euclidean distance (Suwanda et al., 2020) between premise-hypothesis pairs within each cluster. New transformation rules were derived from high-cohesion clusters. These derived rules were then used with LLMs and CoT prompting to generate additional premise-hypothesis pairs.

3.2.1 BERT Embedding-based K-Means Clustering

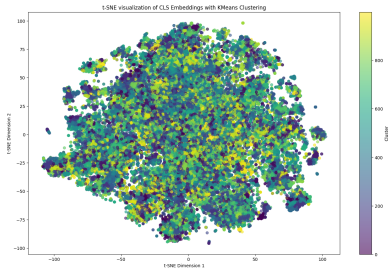
For the 50,000 premise-hypothesis pairs detected as OOD, we conducted three variants of k-means clustering using [CLS] embeddings from the fine-tuned BERT-base model as described in Section 3.1:

1. Clustering based on premise-hypothesis pairs
2. Clustering based on premise-hypothesis-explanation triples, utilizing human-annotated natural language explanations from the e-SNLI dataset (Camburu et al., 2018)
3. Clustering based exclusively on explanation sentences extracted from the premise-hypothesis-explanation triples in the e-SNLI dataset

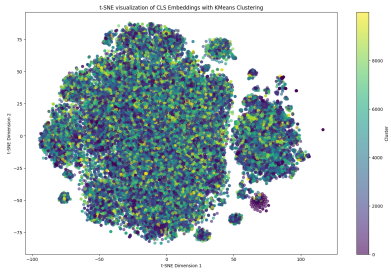
Preliminary experimental results showed that the explanation-based clustering approach achieved superior performance, and we subsequently adopted it for further experiments. Figure 1 presents t-SNE visualizations of the embedding vectors for each clustering approach: (a) premise-hypothesis pairs, (b) premise-hypothesis-explanation triplets, and



(a) Premise-Hypothesis



(b) Premise-Hypothesis-Explanation



(c) Explanation

Figure 1: t-SNE Visualizations of Sentence Embeddings

(c) explanations only. Among them, Figure 1(c) clearly illustrates that explanation-based embeddings produce the most distinct and well-defined cluster structures.

3.2.2 Cluster Cohesion Evaluation and New Rule Derivation

From the 50,000 premise-hypothesis pairs identified as OOD, we generated 10,000 clusters using k-means clustering, and new transformation rules were derived through the following step-by-step analysis:

1. To evaluate cluster cohesion, we calculated the mean Euclidean distance, referred to as the Mean Pairwise Distance (MPD), between all premise-hypothesis pairs within each cluster using sentence embedding vectors:

$$MPD = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} d(x_i, x_j)$$

where m is the number of premise-hypothesis pairs in the cluster, and $d(x_i, x_j)$ is the Euclidean distance between sentence vectors x_i and x_j. A lower MPD indicates higher cohesion, suggesting that premise-hypothesis pairs within the cluster share more similar characteristics. After sorting clusters in descending order by size, we selected the top 100 clusters with the highest cohesion.

2. We then verified the label consistency of the selected 100 clusters. If any premise-hypothesis pairs within a cluster had a label different from the others, we considered that cluster to lack a consistent transformation pattern and excluded it. Through this process, 58% of the clusters were excluded for containing inconsistent labels, leaving 42 clusters for final analysis.

3. Through manual analysis of the premise-hypothesis-explanation patterns in these 42 clusters, we derived four new rules:

   - Role Generalization (RG): Transform specific roles and occupations into general expressions
   - Contextual Augmentation (CA): Derive purpose and background information based on context
   - Visual Specification (VS): Specify visual details of subjects
   - Emotion Inference (EI): Infer emotions or states based on behaviors

   Among the four new rules, RG falls under the Entailment category, while CA, VS, and EI are classified as Neutral. Among these discovered rules, EI was derived through manual analysis of cluster 9101. This cluster comprised 20 premise-hypothesis pairs, and analysis of their premise-hypothesis-explanation patterns revealed a consistent pattern of "inferring emotions or states based on behaviors described in sentences." Detailed examples are provided in Appendix B.

3.2.3 Automated Rule Generation and Verification

To reduce manual effort in the final manual rule discovery step of Section 3.2.2, we introduce an automated rule discovery and validation procedure based on LLMs and semantic similarity evaluation. This procedure was applied to the 42 clusters identified in Section 3.2.2 and consists of the following three stages.

1. **Evaluation of Alignment with Existing Rules** For each cluster, we provide the set of premise-hypothesis-label-explanation (PHL+E) quadruples and instruct the LLM to evaluate whether the examples correspond to any of the 15 existing rules. If the proportion of unmatched examples exceeds 30% (equivalently, if the match ratio falls below 70%), the LLM is asked to generate a new rule that best explains the patterns observed in the cluster.

2. **Hypothesis Generation Using Newly Generated Rules** For clusters where new rules were generated in the previous step, we provide only the premise sentences along with the generated rule, prompting the LLM to generate new hypothesis sentences accordingly.

3. **Semantic Consistency Evaluation** To assess the semantic alignment between the original hypothesis sentences and the newly generated hypotheses from the previous step, we calculate the cosine similarity using Sentence-BERT embeddings. Clusters with an average similarity exceeding a predefined threshold (0.5) are retained as valid rule candidates.

This automated pipeline provides a scalable alternative to manual rule discovery by generating new transformation rules for clusters not covered by existing rules and validating the generated rules. We applied this procedure to 42 clusters and generated new rules for 16 clusters that exceeded the 30% unmatched-ratio threshold in the first step. To validate the generated rules, we proceeded with the second and third steps, applying a 0.5 semantic similarity threshold to these clusters, with similarity scores ranging from 0.0648 to 0.7029. As a result, 6 clusters and their corresponding rules were validated, with two clusters sharing the same rule, resulting in a total of 5 unique rules automatically generated. Notably, three of the four transformation rules derived through manual analysis (VS, CA, and EI) were also identified through this automated process. This overlap supports the validity of our approach, demonstrating that it is not only more efficient but also reliable, mitigating the limitations of manual rule discovery such as time consumption, limited scalability, and researcher subjectivity. Additional examples are provided in Appendix C.

3.3 Extension and Results with CoT-Generated Explanations

To address the limitation that explanation-based clustering is restricted to datasets with human-written explanations, we investigate the use of automatically generated explanations derived solely from premise–hypothesis pairs. Specifically, we apply CoT prompting to generate approximately 50,000 explanation sentences from the SNLI dataset and follow the same pipeline and filtering procedure as described in Sections 3.2.2 and 3.2.3. Under the human-written explanation (Human-Ex) setting, clustering the top 100 clusters yields 6 clusters with 4 new rules (CA, VS, RG, and EI), and expanding to the top 200 clusters yields 12 clusters with 5 rules (VS, CA, RG, EI, and IG). When using CoT-generated explanations (CoT-Ex), the top 100 clusters produce 13 clusters with 3 rules (VS, CA, and EI), while the top 200 clusters produce 27 clusters with 5 rules (VS, CA, EI, IG, and AFR). Across these settings, we consistently rediscover three of the four transformation rules (CA, VS, and EI). In addition, two novel rules (IG and AFR) emerge, both of which can be regarded as sub-rules of the broader CA transformation. This overlap between CoT-Ex and manually identified rules supports the validity of our approach. CoT explanations not only provide a scalable alternative to human annotations but also enable reliable rule discovery in domains without explicit explanations. More fine-grained results across different Stage 1 and Stage 3 thresholds are reported in Appendix D.

4 Experiments

4.1 OOD Detection for discovering new rules

To evaluate OOD detection performance for discovering new rules, we extracted 500 PHL triples from the SNLI test set. These triples were manually examined to determine if they were covered by the existing 15 transformation rules, labeling them as ID if included and OOD if not. The performance of OOD detection was evaluated using the following three metrics:

- FPR at 95% TPR (95FPR): Measures the False Positive Rate (FPR) when True Positive Rate (TPR) is 95%. This indicates the rate at which OOD samples are misclassified as ID when the model identifies ID samples with 95% accuracy. Lower values indicate better performance.

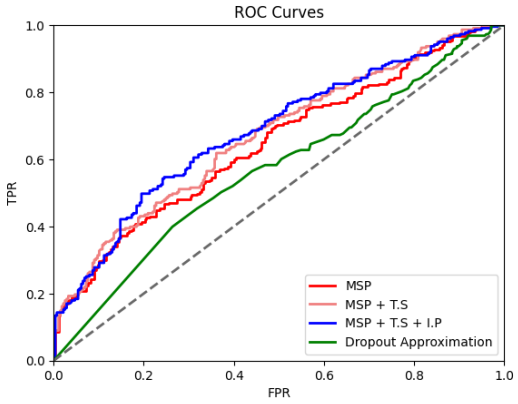


Figure 2: The ROC curves of the baseline (red), MSP and our method (blue), MSP+TS+IP models

95FPR(↓)		93FPR(↓)		AUROC(↑)	
MSP	MSP+TS+IP	MSP	MSP+TS+IP	MSP	MSP+TS+IP
93.1	84.1	91.0	80.1	66.5	68.8

Table 1: OOD Detection Performance of Baseline and Proposed Models

- FPR at 93% TPR (93FPR): Measures FPR at 93% TPR threshold. As with 95FPR, lower values indicate superior OOD detection performance.
- AUROC(%): Calculates the area under the TPR-FPR curve across all classification thresholds (Davis and Goadrich, 2006), comprehensively evaluating the model's OOD detection capability. Higher AUROC values indicate better discrimination between ID and OOD samples (Humblot-Renaux et al., 2023).

Figure 2 and Table 1 compare the OOD detection performance between the baseline (MSP) and our proposed model (MSP+TS+IP). AUROC analysis shows that our proposed model consistently outperformed the baseline. Specifically, as shown in Table 1, our model improved AUROC from 66.5% to 68.8%, with FPR decreasing by 9 percentage points at 95% TPR and 10.9 percentage points at 93% TPR. These results demonstrate the effectiveness of combining temperature scaling and input preprocessing for enhancing OOD detection in the NLI task. Appendix A details threshold selection, data partitioning, and the evaluation of Monte Carlo Dropout (Gal and Ghahramani, 2016) for OOD detection. Table 2 presents examples of OOD detection using the proposed model (MSP+TS+IP). The model accurately classified premise-hypothesis pairs matching the existing 15 transformation rules, such as Paraphrasing (PA), as ID, while classifying premise-hypothesis pairs that deviate from existing rule patterns as OOD.

Examples	Label	Type
P: The girls walk down the street H: Girls set down in the street	PA	ID
P: A young man in a heavy brown winter coat stands in front of a blue railing with his arms spread H: The young man is at his grandmothers house	–	OOD

Table 2: Examples of OOD Detection by the Proposed Model

4.2 Automatic Rule-based NLI Data Generation with Chain-of-Thought Prompting

Using the 15 existing transformation rules and the 4 newly derived transformation rules introduced in Section 3.2.2, we extracted premise sentences from the SNLI training set and applied all 19 rules to generate new PHL triples. To implement the transformation rules, we employed the GPT-4o-mini model with 3-shot CoT prompting, which produced between 4 and 2,307 PHL triples per rule depending on the experimental setting. As shown in Table 3, the generated premise–hypothesis pairs effectively capture the characteristics of each rule and exhibit logical relationships consistent with their corresponding NLI labels. To evaluate the quality of the generated data, we produced 200 PHL triples for each rule and randomly sampled 40 triples per rule for manual assessment, measuring accuracy based on whether the premise–hypothesis relationships adhered to their intended rule patterns. The evaluation results presented in Table 4 show that the four newly derived rules (RG, CA, VS, EI) achieved high accuracy (97.5%–100%), comparable to the existing rules, with an overall average accuracy of 97.63% across all rules. Detailed CoT prompt examples for both the existing 15 rules and the newly derived four rules are provided in Appendix E, demonstrating that our proposed method can reliably generate high-quality NLI training data.

4.3 NLI Performance Analysis

To evaluate the effectiveness of new rules, we generated PHL data by extracting premise sentences from the SNLI training set and applying our transformation rules using GPT-4o-mini with CoT prompting. We generated 1,000 PHL triples

Rule	CoT Generation Result
RG	P: Concert goers watch as a guitarist performs on stage H: Concert goers watch as a musician performs on stage L: Entailment
CA	P: A man with a big red bowl is walking toward a brown donkey H: A man with a big red bowl is walking toward a brown donkey to feed it L: Neutral
VS	P: A guy is reading a newspaper H: A guy is reading a crumpled newspaper L: Neutral
EI	P: Three smiling children are running indoors H: Three smiling children are running indoors because they are excited L: Neutral

Table 3: Generated PHL Examples Using the Newly Derived Rules

Category		Accuracy(%)
Entailment	HS	100
	PS	100
	CT	65
	PA	100
	ES	100
	***RG**	97.5
Contradiction	CW-adj	100
	CW-noun	100
	CV	100
	NS	100
	SOS	97.5
	IrH	100
	NI	100
Neutral	AM	100
	Con	97.5
	SSNCV	100
	***CA**	97.5
	***VS**	100
	***EI**	100
Average Generation Accuracy		97.63

Table 4: Generation Accuracy of Premise-Hypothesis-Label Data Using CoT Prompting. Asterisk(*) denotes the four newly derived rules.

Dataset Augmentation Method	BERT-base
	Avg (Std.Dev)
Original(550,152)	89.85 (± 0.362)
Original + RG(1,000)	90.02 (± 0.054)
Original + CA(1,000)	90.07 (± 0.077)
Original + VS(1,000)	89.97 (± 0.253)
Original + EI(1,000)	89.93 (± 0.254)

Table 5: The Performance Improvement Effect of Using Each New Rule as a Data Augmentation Method

Dataset	Size	BERT-base	
		Best	Average (Std.Dev)
SNLI	550,152	90.35	89.85 (± 0.362)
SNLI + 15 rules	554,652 (+4,500)	90.51	89.95 (± 0.359)
SNLI + 19 rules (Ours)	555,852 (+5,700)	**90.59**	**90.00** (± 0.399)

Table 6: NLI Performance Comparison with Integrated Rules.

per rule (for both 15 existing and 4 new rules) and augmented the SNLI training set (550,152 examples) with these generated triples. The BERT-base model was trained using negative log-likelihood (NLL) loss and the Adam optimizer, with a batch size of 32 and a learning rate of 3e-5. Hyperparameters optimized on the SNLI validation set. NLI performance was evaluated using accuracy. For each of five different random seeds, the model was trained for 25 epochs and the highest accuracy on the SNLI test set (10,000 examples) was recorded. The final result was computed as the average of these five highest accuracies.

Table 5 presents the results of using each new rule as a data augmentation method. After adding 1,000 generated examples per rule to the SNLI training set, the performance of each rule was evaluated individually. The experimental results showed that Contextual Augmentation (CA) achieved the highest performance improvement (+0.22%p), with all rules contributing to performance enhancement.

Table 6 presents the performance when integrating the 15 existing rules and 4 new rules as a data augmentation method. Compared to the baseline SNLI dataset, we evaluated the performance when applying existing 15 rules (300 examples per rule,

total 4,500) and all 19 rules including the new 4 rules (300 examples per rule, total 5,700) as a data augmentation method. When all 19 rules were applied, we achieved the best performance of 90.59% and an average performance of 90.00%(±0.39), demonstrating the effectiveness of the proposed rules.

4.4 Distribution-aware Data Augmentation

To understand the actual distribution of each transformation rule in the SNLI dataset, we modified the labels of the PHL triples generated using 15 and 19 rules, respectively, to their corresponding rule names. Then, we trained a BERT-base model and used it to classify 10,000 samples from the SNLI validation set, analyzing the distribution of each rule. Figure 3 and 4 visually demonstrate the distributions of both rule sets. IH, ES, PA, and COUNT consistently showed major proportions in both rule sets. Based on these findings, we conducted experiments with two data augmentation approaches:

- Uniform Distribution (w/o Distribution): Generate an equal number of examples per rule
- Distribution-aware: Generate examples reflecting the actual rule distribution in SNLI dev-set

To evaluate the performance of the two data augmentation methods, we conducted experiments by augmenting 4,500 examples across various training set sizes (2k, 10k, 50k, 550k) under the same conditions. In the 'Uniform distribution' approach, we generated 300 examples per rule for the 15-rule set and approximately 237 per rule for the 19-rule set. In the 'Distribution-aware' approach, we generated between 4 and 2,307 examples based on the actual distribution, as detailed in Appendix F.

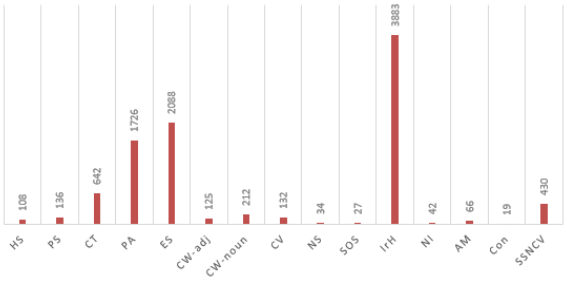


Figure 3: Distribution Analysis of SNLI Validation Set for Basic 15 Rules

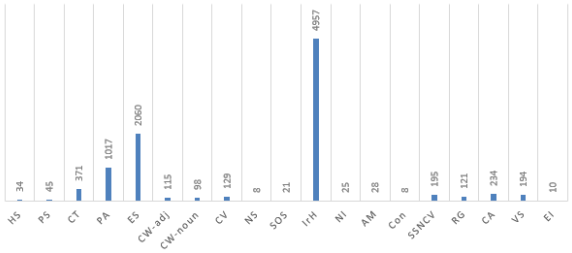


Figure 4: Distribution Analysis of SNLI Validation Set for Extended 19 Rules

		BERT-base Average (Std.Dev)			
Datasets	Types	550k	50k	10k	2k
Original	-	89.85 (± 0.362)	86.06 (± 0.240)	82.29 (± 0.347)	76.31 (± 0.513)
Original + 15rules (+4,500)	Uniform	89.95 (± 0.359)	86.14 (± 0.255)	82.49 (± 0.392)	76.71 (± 0.806)
	Distribution -aware	89.99 (± 0.158)	86.19 (± 0.222)	82.62 (± 0.178)	76.43 (± 0.883)
Original + 19rules (+4,500) ∗Ours	Uniform	90.00 (± 0.405)	86.17 (± 0.224)	82.64 (± 0.383)	76.94 (± 0.835)
	Distribution -aware	90.00 (± 0.134)	**86.24** (± 0.175)	**82.72** (± 0.258)	**77.16** (± 0.392)

Table 7: Performance Comparison across Dataset Sizes and Augmentation Strategies. '∗Ours' (19 rules with Distribution-aware) achieves the highest performance across all experimental settings.

Table 7 compares performance improvements across different dataset sizes and augmentation strategies. Results show that distribution-aware augmentation performed better across all dataset sizes, achieving maximum improvement (+0.85%p) with small datasets (2k). Additionally, applying all 19 rules, including the newly discovered rules, consistently outperformed the 15-rule approach.

Figure 5 shows overall performance changes by dataset size. For detailed performance graphs for each individual dataset size, see Appendix G. Data augmentation effects were more pronounced with smaller datasets, with distribution-aware data augmentation showing superior performance across all sizes. These experimental results demonstrate the effectiveness of augmentation strategies that reflect actual data distribution in NLI tasks, suggesting the importance of considering real training data distribution patterns in data augmentation.

Figure 5: Performance Comparison by Data Size and Augmentation Strategy. Our proposed method (Original + 19 rules with Distribution-aware, shown in mint) consistently achieves the best performance across all data sizes.

5 Conclusion

In this paper, we proposed a novel framework for discovering sentence transformation rules to generate natural language inference (NLI) training data. Our method combines Out-of-Distribution (OOD) detection and clustering to identify premise–hypothesis pairs in the SNLI dataset that are not covered by existing rules. From these, we discovered four new transformation rules (RG, CA, VS, EI) in addition to the existing 15, and automatically generated high-quality training examples using large language models (LLMs) with Chain-of-Thought (CoT) prompting.

Beyond human-written explanations, we further extend our approach to CoT-generated explanations, demonstrating that CoT explanations provide a scalable alternative to human annotations and enable reliable rule discovery in domains where explicit explanations are unavailable.

Experimental results confirm that augmenting the training set with rule-based examples improves model performance, with distribution-aware augmentation strategies proving particularly effective in low-resource settings (2k). Overall, our framework offers an efficient and scalable solution for adapting NLI models to new domains with limited labeled data and shows potential for broader applicability to other NLP tasks, such as discovering new error types in Grammatical Error Correction.

Limitation

While our method automates the rule discovery process using LLMs to evaluate clusters, generate rule candidates, and validate them via semantic similarity, it still depends on several design choices such as similarity thresholds and filtering criteria. We conducted additional experiments under multiple threshold settings to mitigate this issue, but fully removing such dependencies remains an open challenge.

Although we generated NLI training data based on 15 existing rules and 4 newly discovered rules, a comprehensive validation of whether the generated data matches the quality and diversity of human-annotated examples is still lacking. In particular, potential biases introduced by LLM-generated data remain underexplored.

We further extended our framework to automatically generated explanations (CoT-Ex), showing their potential as a scalable alternative to human-written explanations. However, the consistency and reliability of CoT explanations across domains, datasets, and languages remain to be rigorously validated.

Finally, our augmentation strategy has been evaluated only on the SNLI dataset, and further studies are needed to establish its generalizability to other domains, datasets, and tasks.

Acknowledgments

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2022-II220369, (Part 4) Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge).

References

Noureddine Bouhmala. 2016. How good is the euclidean distance metric for the clustering problem. In *2016 5th IIAI international congress on advanced applied informatics (IIAI-AAI)*, pages 312–315. IEEE.

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.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Heejin Cho, Changki Lee, and Kyoungman Bae. 2023. Generating premise-hypothesis-label triplet using chain-of-thought and program-aided language models. In *Proceedings of the 35th Annual Conference on Human and Cognitive Language Technology*, pages 352–357. Korean Institute of Information Scientists and Engineers (KIISE).

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

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.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Dan Hendrycks and Kevin Gimpel. 2018. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Preprint*, arXiv:1610.02136.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Galadrielle Humblot-Renaux, Sergio Escalera, and Thomas B Moeslund. 2023. Beyond auroc & co. for evaluating out-of-distribution detection performance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3880–3889.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. 2023. Detecting out-of-distribution data through in-distribution class prior. In *International Conference on Machine Learning*, pages 15067–15088. PMLR.

Shruti Kapil and Meenu Chawla. 2016. Performance evaluation of k-means clustering algorithm with various distance metrics. In *2016 IEEE 1st international conference on power electronics, intelligent control and energy systems (ICPEICES)*, pages 1–4. IEEE.

Shiyu Liang, Yixuan Li, and R. Srikant. 2020. Enhancing the reliability of out-of-distribution image detection in neural networks. *Preprint*, arXiv:1706.02690.

Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. 2023. What makes chain-of-thought prompting effective? a counterfactual study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1448–1535, Singapore. Association for Computational Linguistics.

Michail Mersinias and Panagiotis Valvis. 2022. Mitigating dataset artifacts in natural language inference through automatic contextual data augmentation and learning optimization. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 427–435, Marseille, France. European Language Resources Association.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Kristina P Sinaga and Miin-Shen Yang. 2020. Unsupervised k-means clustering algorithm. *IEEE access*, 8:80716–80727.

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, page 4444–4451.

R Suwanda, Zulfahmi Syahputra, and Elvi M Zamzami. 2020. Analysis of euclidean distance and manhattan distance in the k-means algorithm for variations number of centroid k. *Journal of Physics: Conference Series*, 1566(1):012058.

Neeraj Varshney, Pratyay Banerjee, Tejas Gokhale, and Chitta Baral. 2022. Unsupervised natural language inference using PHL triplet generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2003–2016, Dublin, Ireland. Association for Computational Linguistics.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

A Additional OOD Detection Details

A.1 Threshold Selection for OOD Detection

To ensure transparency in our OOD detection experiments, we provide a detailed explanation of the threshold selection procedure.

We first randomly sampled 500 PHL (Premise–Hypothesis–Label) triplets from the SNLI test set and manually annotated whether each hypothesis conformed to one of the 15 existing transformation rules. This subset served as a gold set for evaluating the performance of various OOD detection methods, as described in Section 4.1

We evaluated three configurations—MSP, MSP+TS, and MSP+TS+IP—on this gold set and compared their AUROC scores (Table 1). Among these, MSP+TS+IP achieved the highest AUROC (0.6880) and was selected as the final method for OOD detection. To determine the decision threshold, we set the value to 0.07155 to ensure 95% precision on the 500-sample gold set. However, given that the SNLI training set contains over 550,000 examples, this threshold was deemed unsuitable for global application, as it could overfit to the small manually labeled subset.

To better align with the full training distribution and enhance generalizability, we reapplied the MSP+TS+IP method to the entire SNLI training set. We then determined a new threshold based on the bottom 10% of softmax probabilities, approximately 50,000 examples, which yielded a threshold value of 0.07186. This final threshold was used to filter OOD samples in subsequent experiments.

This adaptive thresholding strategy mitigates the risk of overfitting on a small evaluation set and ensures robustness across the full dataset.

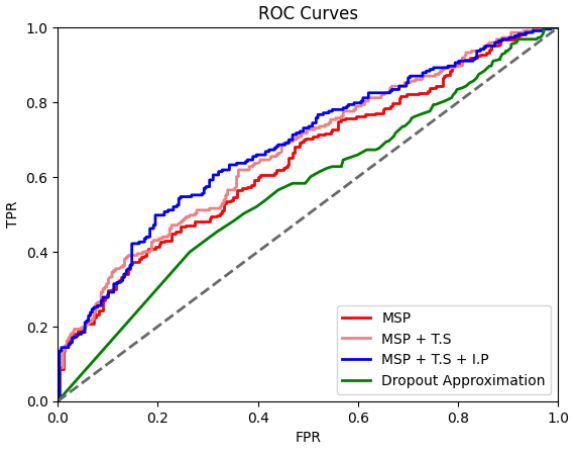


Figure 6: Visual comparison of ROC curves for Monte Carlo Dropout and MSP-Based OOD Detection.

A.2 Evaluation of Uncertainty-based Method: Monte Carlo Dropout

To evaluate the effectiveness of alternative uncertainty-based methods for OOD detection, we conducted additional experiments using Monte Carlo Dropout. Following the approach proposed by Gal and Ghahramani (2016), we applied dropout at inference time to estimate predictive uncertainty.

We tested dropout rates ranging from 0.1 to 0.5, enabling dropout during inference and averaging predictions over 100 stochastic forward passes. While Monte Carlo Dropout is a widely used method for uncertainty estimation, it consistently underperformed in our setting. The best AUROC score (0.6154) was achieved with a dropout rate of 0.1, which was still notably lower than that of our final method, MSP + TS + IP (AUROC 0.6880).

These results suggest that, although Monte Carlo Dropout is a valid uncertainty estimation technique, the combined approach of MSP + TS + IP yields better separation between in-distribution and out-of-distribution samples in the NLI setting. Accordingly, we excluded dropout-based methods from our final pipeline. The full AUROC scores and corresponding ROC curves are presented in Figure 6 and Table 8.

A.3 Hyperparameter Tuning for Temperature and Epsilon

To ensure reproducibility and clarify the rationale for hyperparameter selection, we conducted extensive tuning experiments for temperature scaling (T) and input perturbation scale (ϵ) used in MSP-based OOD detection.

Experiments	AUROC
MSP	0.6651
MSP + T.S	0.6745
MSP + T.S + I.P	**0.6880**
Dropout	0.6154

Table 8: AUROC scores for different OOD detection methods.

Temperature (T): We tested values ranging from 10 to 1000 and evaluated performance using AUROC on the 500-sample gold set described in Section 4.1. The best result was observed at $T = 1000$, achieving an AUROC of 0.6745. See Table 9 for detailed results.

T Value	AUROC
MSP (No T)	0.6651
10	0.6736
100	0.6741
500	0.6743
600	0.6742
700	0.6743
800	0.6742
900	0.6742
1000	**0.6745**

Table 9: AUROC for different temperature (T) values in temperature scaling.

Epsilon (ϵ): We first explored a coarse range $[0.01, 0.09]$, as summarized in Table 10, and then conducted a finer search within $[0.031, 0.039]$, shown in Table 11. The highest AUROC (0.6880) was achieved at $\epsilon = 0.033$, which was adopted as the final perturbation strength.

B Examples of Emotion Inference (EI) Rule

The Emotion Inference (EI) rule, which "infers emotions or states from actions described in sentences," was derived from analyzing premise-hypothesis sentence pairs in cluster 9101, as shown in Table 12. Among the 20 sentence pairs in the cluster, the following 5 examples effectively demonstrate the characteristics of this rule. For example, in the first example, we identified a consistent pattern where an underlying emotional state is inferred from the behavioral description "The little boy is at the side of the river throwing rocks"

ϵ Value	AUROC
0.01	0.6783
0.02	0.6864
0.03	**0.6870**
0.04	0.6833
0.05	0.6801
0.06	0.6772
0.07	0.6754
0.08	0.6726
0.09	0.6684

Table 10: AUROC for coarse-grained search over perturbation scale ϵ.

ϵ Value	AUROC
0.031	0.6872
0.032	0.6875
0.033	**0.6880**
0.034	0.6876
0.035	0.6868
0.036	0.6844
0.037	0.6835
0.038	0.6834
0.039	0.6834

Table 11: AUROC for fine-grained search over perturbation scale ϵ.

to derive "A boy is bored outdoors"

C Prompt Design and Evaluation of Automated Rule Discovery

C.1 Prompt Design for Automated Rule Discovery

In this section, we present the prompt structure used to automate the discovery of new sentence transformation rules. The prompt was designed to evaluate whether each cluster is explainable by any of the existing 15 transformation rules and to induce a new rule if not.

C.1.1 Existing 15 Transformation Rules

We provided the LLM with a list of 15 transformation rules categorized by NLI labels. Each rule was defined with a short description and illustrative example.

Entailment

1. **Hypernym Substitution**: Replace nouns with their hypernyms (e.g., *dog* $\rightarrow$ *animal*)

Premise	Hypothesis	Explanation	Label
The little boy is at the side of the river throwing rocks	A boy is bored outdoors	Sentence 1 is missing boy is bored	neutral
A young bride purses her lips	She is angry	Sentence 1 is missing person is angry	neutral
A barefoot boy is crying	The boy is hurt	Sentence 1 is missing boy is hurt	neutral
A screaming man playing hand-ball makes a throw by jumping into the air	A man is very athletic	Sentence 1 is missing man is very athletic	neutral
A young Indian boy is lean into a wall, wearing a red and white striped shirt, and covering his eyes	boy is crying	Sentence 1 is missing boy is crying	neutral

Table 12: The Emotion Inference (EI) rule, which "infers emotions or states from actions described in sentences".

2. **Pronoun Substitution**: Replace nouns with pronouns (e.g., *two men* → *they*)
3. **Counting**: Count nouns sharing a common hypernym (e.g., *a bike and a car* → *two automobiles*)
4. **Paraphrasing**: Rephrase the sentence using synonymous expressions (e.g., *bench* → *seat*)
5. **Extracting Snippets**: Retain only the core semantic content (e.g., *a person with red shirt* → *a person*)

Contradiction

6. **Contradictory Words-adj**: Replace adjectives with antonyms (e.g., *big* → *small*)
7. **Contradictory Words-noun**: Replace nouns with different nouns (e.g., *piano* → *violin*)
8. **Contradictory Verb**: Replace verbs with antonyms (e.g., *walk* → *drive*)
9. **Number Substitution**: Replace numerical expressions (e.g., *two* → *seven*)
10. **Subject Object Swap**: Swap the subject and object positions (e.g., *clock, pillow* → *pillow, clock*)
11. **Irrelevant Hypothesis**: Sample a completely unrelated sentence (no CoT applied)
12. **Negation Introduction**: Apply negation to the sentence (e.g., *covered* → *did not cover*)

Neutral

13. **Adding Modifiers**: Add modifiers to nouns (e.g., *bird* → *small bird*)
14. **ConceptNet**: Add spatial or relational information (e.g., *eating the grass* → *eating the grass in the yard*)
15. **Same Subject but Non-Contradictory Verb**: Replace verbs with synonymous alternatives and add arbitrary nouns (e.g., *sleeping* → *laying + chair*)

C.1.2 Prompt Template for Cluster Evaluation

Each cluster was evaluated using the following prompt structure:

> *The following is a list of 15 transformation rules used for natural language inference (NLI) data generation. For each sentence pair in the given cluster, check whether the transformation from premise to hypothesis aligns with one of the 15 rules. If at least 70% of the pairs match a single rule (equivalently, if fewer than 30% are unmatched), name the rule. otherwise (unmatched ratio $\geq$ 30%), define a new rule that best explains the transformation pattern observed in this cluster. Please output only the rule name if matched, or a newly proposed rule name and its description if unmatched.*

C.1.3 Cluster Input Example

The following is a representative example used in the evaluation prompt. Each line contains a premise, hypothesis, label and explanation.

Table 13 (Cluster 9101) shows an example cluster used in the evaluation.

C.1.4 Output Interpretation Example

If the proportion of unmatched examples in a cluster is 30% or higher, the model is instructed to

Premise	Hypothesis	Label	Explanation
The little boy is at the side of the river throwing rocks	A boy is bored outdoors	neutral	Sentence 1 is missing boy is bored
A man is sleeping inside on a bench with his hat over his eyes	A man fell asleep on a bench because he was drunk	neutral	Sentence 1 is missing he was drunk
A young bride purses her lips	She is angry	neutral	Sentence 1 is missing person is angry

Table 13: Example premise-hypothesis-label-explanation triples from Cluster 9101.

generate a new rule. For example (Cluster 9101):

> **Proposed Rule (Inferred Contextual Attribute)**: Construct the hypothesis by inferring contextual attributes of a person that are not explicitly mentioned in the premise (e.g., identity, emotional state, physical condition), based on commonsense and situational cues. (e.g., *The boy is throwing rocks* → *The boy is bored*)

C.2 Evaluation Results of Automated Rule Discovery

C.2.1 Cluster Filtering and Validation Summary

We applied our automated rule discovery procedure to 42 clusters. In the **After Matching** stage, 16 clusters were flagged as requiring novel rules due to an unmatched ratio above 30% with the existing transformation rules. In the **After Similarity Eval.** stage, new hypotheses were generated based on LLM-proposed rules and evaluated using Sentence-BERT. Among the 16 clusters, 6 passed the semantic consistency threshold (0.5) and were retained as valid rule candidates. Notably, 4 of these 6 clusters (890, 1527, 2007, 9101) overlapped with rules (VS, CA, EI) that were independently derived via manual analysis. Details of each stage are summarized in Table 14.

C.2.2 Semantic Comparison of Hypotheses

To evaluate the semantic consistency of hypotheses generated through the LLM-based rule application process, we compared them to the original hypotheses. The original hypotheses were written by human annotators, while the generated ones were produced by prompting the LLM with only the premise sentences and the automatically derived rule.

Table 15 presents a side-by-side comparison of the original and generated hypotheses. Cluster 9101 is shown here as a representative example, using the inferred rule *Inferred_Contextual_Attribute*. Sentence-level semantic similarity was computed using Sentence-BERT, and the average cosine similarity across these pairs was used as the validation criterion in Stage 3 of the automated rule discovery pipeline.

C.2.3 Automatically Discovered Rules and Alignment with Manual Rules

Table 16 presents the six transformation rules retained after Stage 3 of the automated rule discovery pipeline. These rules were automatically generated by prompting the LLM with clusters that did not align with any of the 15 existing transformation rules. Each rule was subsequently validated for semantic consistency. The table reports the cluster ID, rule name, rule description, and whether the rule aligns with a manually defined rule. Four rules—*Generalization Refinement*, *Situation Inference*, *Action Addition*, and *Inferred Contextual Attribute*—show alignment with existing manual rules (VS, CA, EI), indicating a high degree of consistency between automated and human-authored rule definitions.

D Detailed Threshold-based Results

In this appendix, we present the detailed results of our threshold-based experiments for automated rule discovery. We varied similarity thresholds at Stage 1 (0.3, 0.5, 0.7) and Stage 3 (0.3, 0.35, 0.4, 0.5) under two different settings: Human-Ex (human-written explanations) and CoT-Ex (Chain-of-Thought generated explanations).

D.1 Human-Ex (e-SNLI Dataset)

For the Human-Ex setting, we initially selected the top-100 clusters based on cohesion scores, where 42 clusters passed the label consistency filter. When extending to the top-200 clusters, an additional 24 clusters met the same criterion, although with lower semantic coherence. Among these, one cluster (IG) corresponded to a novel transformation rule not included in the original four, while the remaining clusters aligned with one of the 15

Stage	Remaining	Selected Clusters	Notes
Initial	42	–	–
After Matching	16	233, 384, 890, 1527, 1739, 2007, 2456, 3072, 3201, 3334, 4581, 5084, 5881, 6020, 7546, 9101	Unmatched $\geq 30\%$
After Similarity Eval.	6	890, 1527, 2007, 6020, 7546, 9101	Similarity ≥ 0.5
Manual Overlap	4	890 (VS), 1527 & 2007 (CA), 9101 (EI)	Matches manual rules

Table 14: Cluster selection and validation results at each stage of the automated rule discovery pipeline.

Original Hypothesis	Generated Hypothesis
A boy is bored outdoors	The boy is curious about nature
The man is asian	The man is highly focused and skilled
A man fell asleep on a bench because he was drunk	The man is exhausted from work
The man has made a lot of money	The man is a street performer earning money
The man is fat	The man is a construction worker
boy is crying	The boy is shy and nervous
The man is homeless	The man is struggling financially
A boy confirms he finds rock climbing easy	The boy is confident in his climbing skills
the children are white	The children are energetic and playful
The child she is holding is not hers	The child feels safe in her arms
The woman is homeless	The woman is living in poverty
She is angry	She is feeling anxious about her wedding
The boy is hurt	The boy is emotionally distressed
The boy only skateboards at night	The boy enjoys the freedom of night rides
the horse is leaping to see his girlfriend	The horse is well-trained and competitive
A man is very athletic	The man is physically very fit
THe man is fishing	The man is preparing to fish
The man is good at guitar	The man is an experienced street musician
the man is a spy	The man is a motorcycle enthusiast
The driver is bored	The driver is tired after a long shift

Table 15: Comparison of Original and LLM-generated hypotheses for Cluster 9101 under the inferred rule *Inferred_Contextual_Attribute*, with Sentence-BERT similarity scores from Stage 3.

existing rules or the previously discovered four new rules (RG, CA, VS, EI). Across multiple threshold settings at Stage 1 and Stage 3, we consistently identified 3–5 new transformation rules, indicating that the automated rule discovery process is relatively robust to threshold variations on the e-SNLI dataset. The detailed results across different thresholds are shown in Table 17.

D.2 CoT-Ex (SNLI Dataset)

For the CoT-Ex setting, we generated approximately 50,000 explanations using Chain-of-Thought prompting. We applied the same clustering and filtering pipeline as in Section 3.2.3 and analyzed the top-100 and top-200 clusters under multiple threshold settings. The results show consistent rediscovery of three previously identified rules (CA, VS, EI), as well as the emergence of two novel rules (IG, AFR) in the top-200 clusters. These findings indicate that CoT-generated explanations serve as a scalable and reliable alternative to human annotations for rule discovery. The detailed results across different thresholds are presented in Table 18.

E Examples of CoT Prompt-based Sentence Transformation Rule Application

E.1 Data Generation Process

We generated premise-hypothesis-label data using the GPT-4o-mini model with rule-specific CoT prompts (Madaan et al., 2023) for 1,000 premise sentences randomly sampled from the SNLI training set. The Stanford Natural Language Inference

Cluster ID	Rule	Rule Description	Matches Manual Rule
890	Generalization Refinement	Explicitly add visual attributes such as color, size, or condition to objects mentioned in the premise when generating the hypothesis.	VS
1527	Situation Inference	Generate a hypothesis by inferring unstated intentions, purposes, or situational context from the premise.	CA
2007	Action Addition	Generate a hypothesis by adding an implied purpose-driven action to the scenario in the premise.	CA
6020	Implicit Attribute Addition	Generate a hypothesis by inferring a person's social identity from contextual clues and commonsense knowledge, focusing on who the person is, not why they act.	–
7546	Implicit Intention Inference	Generate a hypothesis by inferring the agent's internal motivations solely from their actions, without assuming any unstated background context.	–
9101	Inferred Contextual Attribute	Generate a hypothesis by inferring the subject's emotional or physical state based on actions described in the premise.	EI

Table 16: Final six automatically generated transformation rules that passed semantic consistency validation. Three of these rules (VS, CA, EI) align with manually defined rules.

(SNLI) dataset (version 1.0) (Bowman et al., 2015) is publicly available for research purposes under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) License. This license permits sharing and adaptation of the dataset with proper attribution and distribution under the same terms.

E.2 Illustrative Examples for Existing Transformation Rules

Table 19 shows the rule definitions and their associated illustrative examples for the existing 15 transformation rules. These examples are simplified Q/A demonstrations of how each rule is applied, and can serve as few-shot instances when constructing CoT prompts for data generation.

E.3 Illustrative Examples for New Transformation Rules

Table 20 shows the rule definitions and their associated illustrative examples for the 4 newly discovered transformation rules. As with the existing rules, these examples serve as concise demonstrations of rule application. They also provide canonical cases that can be incorporated into CoT prompting to guide large language models in generating high-quality NLI data.

F Data Augmentation Distribution by Rules

We augmented 4,500 samples for each dataset size (2k, 10k, 50k, 550k) using two approaches: uniform and distribution-aware. The uniform distribution allocated samples equally across rules, while the distribution-aware followed the rule frequencies observed in the SNLI validation set. These approaches were applied to both 15-rule and 19-rule sets. The following sections detail the augmentation distribution for each approach:

F.1 Uniform Distribution-based Data Augmentation for 15 Rules

Table 21 shows the data augmentation distribution where 300 samples were uniformly allocated to each of the 15 rules, resulting in a total of 4,500 augmented samples.

F.2 Distribution-aware Data Augmentation for 15 Rules

Table 22 shows the data augmentation following the distribution observed in the SNLI validation set. Samples were allocated proportionally to each rule's frequency, ranging from 4 to 2,307 samples per rule.

F.3 Uniform Distribution-based Data Augmentation for 19 Rules

Table 23 shows the uniform distribution approach for 19 rules, where 4,500 samples were evenly dis-

Top-k Clusters	#Filtered Clusters	Stage 1 (Threshold)	Stage 3 (Threshold)	#Clusters	#Clusters with New Rules	New Rules
100	42	0.3	0.3	10	6 (VS(2), CA(2), RG(1), EI(1))	4 (VS, CA, RG, EI)
			0.35	9	6 (VS(2), CA(2), RG(1), EI(1))	4 (VS, CA, RG, EI)
			0.4	4	3 (VS(1), CA(1), RG(1))	3 (VS, CA, RG)
			0.5	0	0	0
		0.5	0.3	11	6 (VS(2), CA(2), RG(1), EI(1))	4 (VS, CA, RG, EI)
			0.35	9	5 (VS(2), CA(2), RG(1))	3 (VS, CA, RG)
			0.4	6	3 (VS(2), RG(1))	2 (VS, RG)
			0.5	0	0	0
		0.7	0.3	9	4 (VS(2), CA(1), EI(1))	4 (VS, CA, RG, EI)
			0.35	5	3 (VS(2), CA(1))	2 (VS, CA)
			0.4	4	2 (VS(1), CA(1))	2 (VS, CA)
			0.5	0	0	0
200	66	0.3	0.3	27	12 (VS(2), CA(3), RG(4), EI(2), IG(1))	5 (VS, CA, RG, EI, IG)
			0.35	23	10 (VS(2), CA(3), RG(3), EI(1), IG(1))	5 (VS, CA, RG, EI, IG)
			0.4	12	6 (VS(1), CA(1), RG(3), IG(1))	4 (VS, CA, RG, IG)
			0.5	2	0	0
		0.5	0.3	27	12 (VS(2), CA(3), RG(4), EI(2), IG(1))	5 (VS, CA, RG, EI, IG)
			0.35	22	9 (VS(2), CA(3), RG(3), IG(1))	4 (VS, CA, RG, IG)
			0.4	11	3 (VS(2), RG(1))	2 (VS, RG)
			0.5	2	0	0
		0.7	0.3	23	9 (VS(2), CA(2), RG(2), EI(2), IG(1))	5 (VS, CA, RG, EI, IG)
			0.35	16	7 (VS(2), CA(2), RG(2), IG(1))	4 (VS, CA, RG, IG)
			0.4	8	2 (VS(1), CA(1))	2 (VS, CA)
			0.5	2	0	0

Table 17: Detailed results for Human-Ex (human-written explanations) under varying Stage 1 and Stage 3 thresholds.

Top-k Clusters	#Filtered Clusters	Stage 1 (Threshold)	Stage 3 (Threshold)	#Clusters	#Clusters with New Rules	New Rules
100	65	0.3	0.3	25	13 (VS(7), CA(5), EI(1))	3 (VS, CA, EI)
			0.35	17	8 (VS(4), CA(4))	2 (VS, CA)
			0.4	9	4 (VS(2), CA(2))	2 (VS, CA)
			0.5	7	1 (CA(1))	1 (CA)
		0.5	0.3	23	12 (VS(7), CA(5))	2 (VS, CA)
			0.35	17	8 (VS(4), CA(4))	2 (VS, CA)
			0.4	10	4 (VS(2), CA(2))	2 (VS, CA)
			0.5	7	1 (CA(1))	1 (CA)
		0.7	0.3	22	11 (VS(7), CA(4))	2 (VS, CA)
			0.35	15	8 (VS(4), CA(4))	2 (VS, CA)
			0.4	9	3 (VS(2), CA(1))	2 (VS, CA)
			0.5	6	0	0
200	107	0.3	0.3	47	27 (VS(8), CA(10), EI(2), IG(1), AFR(1))	5 (VS, CA, EI, IG, AFR)
			0.35	31	16 (VS(5), CA(9), IG(1), AFR(1))	4 (VS, CA, IG, AFR)
			0.4	16	10 (VS(3), CA(7))	2 (VS, CA)
			0.5	7	1 (CA(1))	1 (CA)
		0.5	0.3	43	21 (VS(8), CA(10), EI(1), IG(1), AFR(1))	5 (VS, CA, EI, IG, AFR)
			0.35	28	15 (VS(5), CA(9), IG(1))	3 (VS, CA, IG)
			0.4	15	8 (VS(2), CA(6))	2 (VS, CA)
			0.5	7	1 (CA(1))	1 (CA)
		0.7	0.3	40	20 (VS(8), CA(9), EI(1), IG(1), AFR(1))	5 (VS, CA, EI, IG, AFR)
			0.35	23	14 (VS(5), CA(9))	2 (VS, CA)
			0.4	14	8 (VS(2), CA(5))	2 (VS, CA)
			0.5	6	0	0

Table 18: Detailed results for CoT-Ex (CoT-generated explanations) under varying Stage 1 and Stage 3 thresholds.

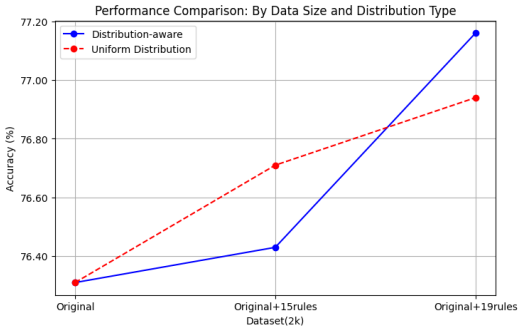


Figure 7: Data Augmentation Effects on 2k Dataset. The lines represent uniform distribution sampling (red) and our proposed distribution-aware sampling (blue)

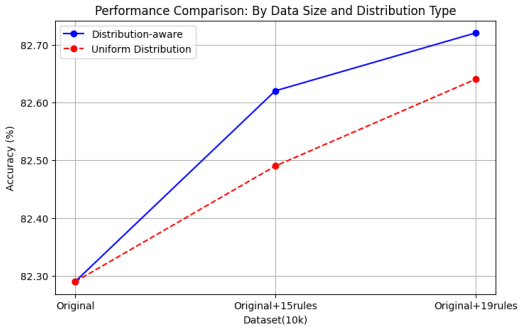


Figure 8: Data Augmentation Effects on 10k Dataset. The lines represent uniform distribution sampling (red) and our proposed distribution-aware sampling (blue)

tributed, resulting in approximately 237 samples per rule.

F.4 Distribution-aware Data Augmentation for 19 Rules

Table 24 shows the distribution-aware approach for 19 rules. The number of samples was determined by the distribution ratios observed in the SNLI validation set, with differential allocation based on each rule's prevalence.

G Detailed Performance by Dataset Size

Figures 7–10 present detailed comparisons of data augmentation performance for each dataset size: 2k, 10k, 50k, and 550k, respectively. These figures expand on the overall trend shown in Figure 5, highlighting how the effects of distribution-aware augmentation vary across different data scales. In particular, the performance improvements are most pronounced in the smallest dataset (2k), while the relative gains diminish as the dataset size increases. These results further support our claim that distribution-aware augmentation is especially beneficial in low-resource settings.

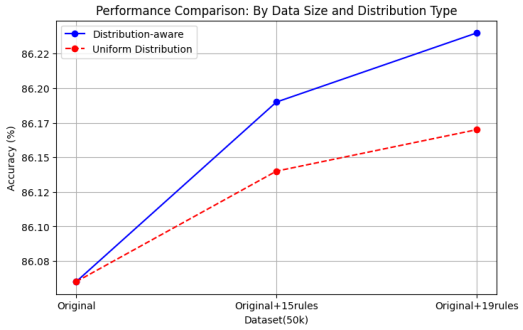


Figure 9: Data Augmentation Effects on 50k Dataset. The lines represent uniform distribution sampling (red) and our proposed distribution-aware sampling (blue)

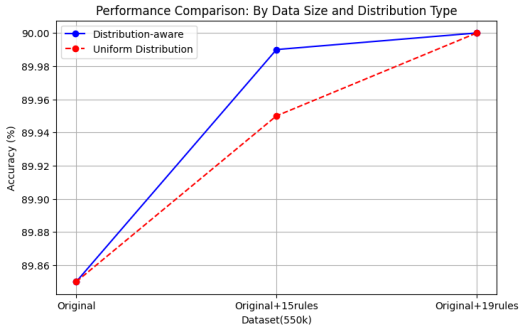


Figure 10: Data Augmentation Effects on 550k Dataset. The lines represent uniform distribution sampling (red) and our proposed distribution-aware sampling (blue)

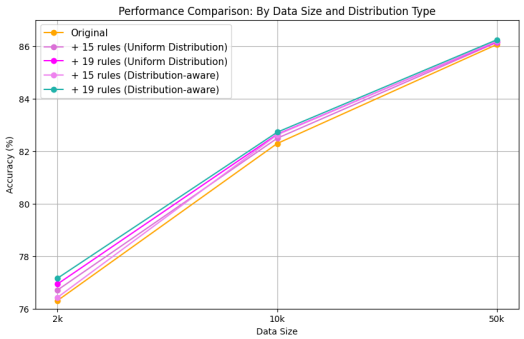


Figure 11: Reproduced from the main paper (Figure 5): Overall performance comparison across dataset sizes (2k, 10k, 50k, 550k).

Label	Rule-Name	Explanation	Illustrative Example
Entailment	Hypernym Substitution (HS)	Generate a hypothesis by replacing nouns with their hypernyms.	Q: In a sentence {a black dog is sleeping}, replace 'dog' with its hypernym 'animal'. A: {a black animal is sleeping}.
	Pronoun Substitution (PS)	Generate a hypothesis by replacing nouns with pronouns.	Q: In a sentence {Two men are sitting on a blue truck}, replace 'Two men' with a pronoun. A: {They are sitting on a blue truck}.
	Counting (CT)	Generate a hypothesis by expressing the number of nouns that share a common hypernym.	Q: In a sentence {A motorbike and a car are parked}, replace with the hypernym and count. A: {two automobiles are parked}.
	Paraphrasing (PA)	Generate a hypothesis by paraphrasing the sentence using synonyms.	Q: In a sentence {A brown purse is sitting on a green bench}, replace with synonyms. A: {A brown bag is perched atop a green seat}.
	Extracting Snippets (ES)	Generate a hypothesis by retaining only the core meaning of the sentence.	Q: In a sentence {A person with a red shirt is running near the garden}, remove modifiers. A: {A person is running near the garden}.
Contradiction	Contradictory Words-adj (CW-adj)	Generate a hypothesis by replacing adjectives with their antonyms.	Q: In a sentence {He lives in a big house}, replace 'big' with its antonym. A: {He lives in a small house}.
	Contradictory Words-noun (CW-noun)	Generate a hypothesis by replacing nouns with contradictory nouns.	Q: In a sentence {She is playing the piano}, replace 'piano' with another noun. A: {She is playing the violin}.
	Contradictory Verb (CV)	Generate a hypothesis by replacing verbs with their antonyms.	Q: In a sentence {A girl is walking}, replace 'walking' with an antonym. A: {A girl is driving}.
	Number Substitution (NS)	Generate a hypothesis by replacing numbers with different numbers.	Q: In a sentence {two cars are parked on the sidewalk}, replace 'two' with another number. A: {seven cars are parked on the sidewalk}.
	Subject Object Swap (SOS)	Generate a hypothesis by swapping the subject and object.	Q: In a sentence {A clock is standing on a pillar}, swap subject and object. A: {A pillar is standing on a clock}.
	Irrelevant Hypothesis (IrH)	Generate a hypothesis by sampling an unrelated sentence. (CoT is not applied in this case)	Q: In a sentence {A sign for an ancient monument is on the roadside}, generate an unrelated hypothesis. A: {A man goes to strike a tennis ball}.
	Negation Introduction (NI)	Generate a hypothesis by introducing negation.	Q: In a sentence {Empty fog covered the streets at night}, negate the verb. A: {Empty fog did not cover the streets at night}.
Neutral	Adding Modifiers (AM)	Generate a hypothesis by adding modifiers to nouns	Q: In a sentence {This is a bird sitting on a twig}, add a modifier. A: {This is a small bird sitting on a twig}.
	ConceptNet (Con)	Generate a hypothesis by adding spatial or relational information.	Q: In a sentence {Three horses are eating grass}, add a location. A: {Three horses are eating grass in the yard}.
	Same Subject but Non-Contradictory Verb (SSNCV)	Generate a hypothesis by replacing verbs with synonyms and adding arbitrary nouns.	Q: In a sentence {A child is sleeping in a bed}, replace the verb and add a noun. A: {A child is laying in a bed with a chair nearby}.

Table 19: Definitions and illustrative examples for the 15 existing transformation rules. Each rule is described with its definition and a corresponding example.

Label	Rule-Name	Explanation	Illustrative Example
Entailment	Role Generalization (RG)	Generate a hypothesis by replacing specific roles with general categories.	Q: In a sentence {A baseball player is diving to catch a ball}, generalize 'baseball player'. A: {An athlete is diving to catch a ball}.
Neutral	Contextual Augmentation (CA)	Generate a hypothesis by adding implicit purposes or background.	Q: In a sentence {A man is playing the saxophone on the street}, add a plausible purpose. A: {A man is playing the saxophone on the street to collect donations}.
	Visual Specification (VS)	Generate a hypothesis by adding visual characteristics.	Q: In a sentence {A man is wearing a straw hat}, add a plausible visual trait. A: {A man is wearing a dirty straw hat}.
	Emotion Inference (EI)	Generate a hypothesis by inferring emotions or states from actions.	Q: In a sentence {A boy is throwing rocks by the river}, infer an emotional state. A: {A boy is throwing rocks by the river because he is bored}.

Table 20: Definitions and illustrative examples for the 4 newly proposed transformation rules. Each rule is described with its definition and a corresponding example.

	Rule-Name	# Samples
Entailment	HS	300
	PS	300
	CT	300
	PA	300
	ES	300
Contradiction	CW-adj	300
	CW-noun	300
	CV	300
	NS	300
	SOS	300
	IrH	300
	NI	300
Neutral	AM	300
	Con	300
	SSNCV	300

Table 21: Uniform Distribution-based Data Augmentation for 15 Rules
(Total: 4,500 samples, 300 samples per rule)

	Rule-Name	# Samples
Entailment	HS	50
	PS	63
	CT	299
	PA	803
	ES	972
Contradiction	CW-adj	58
	CW-noun	99
	CV	61
	NS	16
	SOS	13
	IrH	1806
	NI	20
Neutral	AM	31
	Con	9
	SSNCV	200

Table 22: Distribution-aware Data Augmentation for 15 Rules
(Total: 4,500 samples, allocated based on SNLI validation set distribution)

	Rule-Name	# Samples
Entailment	HS	237
	PS	237
	CT	237
	PA	237
	ES	237
	RG	237
Contradiction	CW-adj	237
	CW-noun	237
	CV	237
	NS	236
	SOS	237
	IrH	237
	NI	237
Neutral	AM	237
	Con	236
	SSNCV	237
	CA	237
	VS	237
	EI	236

Table 23: Uniform Distribution-based Data Augmentation for 19 Rules
(Total: 4,500 samples, 236 or 237 samples per rule)

	Rule-Name	# Samples
Entailment	HS	16
	PS	21
	CT	173
	PA	473
	ES	958
	RG	56
Contradiction	CW-adj	54
	CW-noun	45
	CV	60
	NS	4
	SOS	10
	IrH	2307
	NI	12
Neutral	AM	13
	Con	4
	SSNCV	91
	CA	109
	VS	90
	EI	4

Table 24: Distribution-aware Data Augmentation for 19 Rules
(Total: 4,500 samples, allocated based on SNLI validation set distribution)