From Curiosity to Clarity: Exploring the Impact of Consecutive Why-Questions

Geonyeong Son [1] **Jaeyoung Lee** [1] **Misuk Kim** [1 2] *
[1]Department of Data Science, Hanyang University
[2]Department of Artificial Intelligence, Hanyang University
{geonyeongson, jaeylee, misukkim}@hanyang.ac.kr

Abstract

Humans attempt to understand the real world by asking the fundamental question "Why?" when faced with incomprehensible situations in everyday life. Such why-questions provide essential knowledge that can help in understanding these situations. In this study, we conducted an end-to-end process to verify the utility of consecutive why-questions, from constructing a large language model (LLM)-based dataset to performing quantitative evaluation and analysis. Firstly, we created a `WHY-Chain` dataset, consisting of answers generated by an LLM in response to chain-of-why-questions, including a validity check. We also incorporated objectives that effectively capture the "consecutive" characteristic of the data. Using the `WHY-Chain` dataset and two types of self-supervised objectives, we trained the pre-trained model. As a result, the refined model demonstrated improved performance on downstream tasks that require commonsense reasoning. Additionally, we conducted various ablation studies to assess the impact of different factors, confirming the scalability of the proposed approach. Lastly, we confirmed the consistency of the logical information by reasoning chain analysis of the answers generated from consecutive why-questions. Our code is available at https://github.com/GeonYeongSon/FCC.

1 Introduction

Some problems we encounter daily are difficult to understand or solve using existing frameworks or knowledge (Joo et al., 2020). To better understand and address these issues, we often ask a fundamental and universal question: "Why?" (Joo et al., 2022). This is part of a process to explore the essence of a problem (Karyawati et al., 2015; Scrivner, 2022), and it helps satisfy our curiosity about new information (Litman and Jimerson, 2004; Friston et al., 2017; Barbieri et al., 2024).

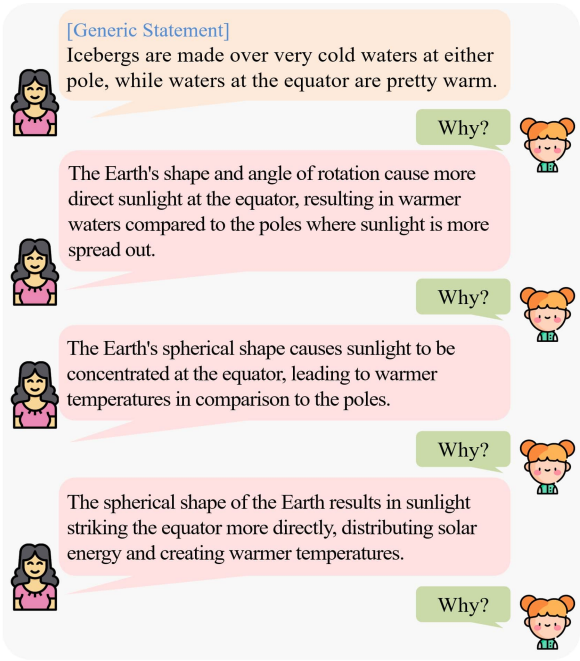

Figure 1: The example of chain-of-why-questions

Furthermore, when the simple why-question is repeated multiple times, the process of finding answers can expand our knowledge base, leading to more complex and comprehensive solutions. This process moves beyond surface-level explanations gained from a single why-question and helps solve problems by integrating external knowledge (McAuliffe et al., 2006; Gillham, 2017). For example, this situation often occurs when talking with children. As shown in Figure 1, when children are given a statement, they can gain more informative answers by asking consecutive why-questions. Previous studies have also shown that external knowledge can aid in solving problems. For instance, the `TellMeWhy` (Lal et al., 2021) task involves inferring the causes of events in narratives to provide answers (Kayesh et al., 2019; Lal et al., 2022a,b), or using external commonsense knowledge generated by models like `COMET` (Bosse-

*Corresponding Author

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 3649–3664

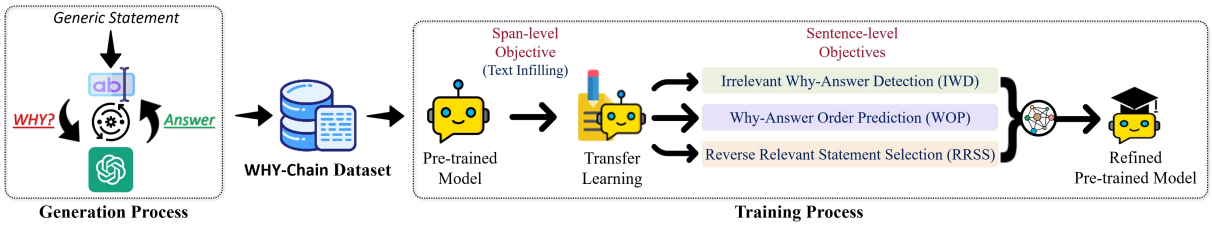

Figure 2: Overall proposed framework. We constructed the WHY-Chain dataset based on the answers generated through chain-of-why-questions for a generic statement. Furthermore, we refined the pre-trained model through transfer learning using two types of self-supervised objectives.

lut et al., 2019a) to improve answering performance (Kayesh et al., 2019; Lal et al., 2022b). However, these studies have limitations as they rely only on fragmented and surface-level knowledge.

2 Methodology

We aim to investigate the impact of using answers generated from consecutive why-questions on generic statements as an external knowledge to enhance performance on downstream tasks that require commonsense reasoning ability. To this end, the overall framework proposed in this paper is illustrated in Figure 2.

This study proposes a new framework that leverages a large language model (LLM) based on the idea that humans use consecutive why-questions to utilize broad knowledge effectively when solving problems. We aim to investigate whether extracting increasingly deeper and broader information through consecutive why-questions can enhance the performance of language models. To do this, we carried out an end-to-end process, including dataset creation, model design, quantitative evaluation, and analysis. First, using an LLM, we constructed a dataset by generating answers to chain-of-why-questions for general statements. Then, to improve understanding of the answers to these why-questions, we enhanced the language model's performance through two types of self-supervised objectives: span-level and sentence-level. Finally, we fine-tuned the pre-trained language model (e.g., T5) using consecutive why-questions for downstream tasks that require commonsense reasoning. To verify the model's generality, we conducted various ablation studies to assess the impact of our proposed objectives, model size, training order of the two objectives, number of why-questions, and the impact of the answers generated from the consecutive why-questions. We also performed a reasoning chain analysis using informativeness and correctness scores throughout the inference process.

Overall, our main contributions of this paper are as follows:

- To verify the positive impact of consecutive why-questions on learning, we conducted a novel end-to-end process from data construction to quantitative evaluation and analysis.
- We incorporated various sentence-level objectives suitable for the "consecutive" characteristic of the dataset and confirmed significant performance improvements in the model.
- We conducted various ablation studies for scalability and generality analysis and performed reasoning chain analysis to evaluate connections, thereby validating the effectiveness of our approach.

First, we construct a dataset called WHY-Chain through consecutive why-questions using an LLM. Then, to effectively learn the deep information in the answers to consecutive why-questions, we design a learning strategy that enhances reasoning capabilities using two types of self-supervised objectives: span-level and sentence-level. In this section, we detail each component of the framework.

2.1 WHY-Chain Dataset

2.1.1 Data Generation

We constructed a dataset based on answers generated through chain-of-why-questions starting from a generic statement. Sentences from the GenericsKB-simplewiki[1] (Bhakthavatsalam et al., 2020) dataset were used as the starting point for the initial why-question. This dataset consists of high-quality sentences commonly used in everyday life, forming a knowledge base. Using a prompt like the one shown in Figure 3, we used an LLM to

[1]https://huggingface.co/datasets/generics_kb

Pair	Premise-Hypothesis	Entailment↑	Hallucination↓		
			Entity-error	Relation-error	Unverifiability
gen.-all(answers)	gen.-c(answers)	0.896(0.159)	1.079(0.403)	1.145(0.594)	1.100(0.557)
consecutive pair	gen.-answer$_1$	0.811(0.206)	1.032(0.238)	1.072(0.459)	1.043(0.379)
	answer$_1$-answer$_2$	0.880(0.181)	1.033(0.245)	1.059(0.371)	1.020(0.275)
	answer$_2$-answer$_3$	0.897(0.173)	1.028(0.175)	1.050(0.306)	1.011(0.151)

Table 1: The average hallucination and entailment scores between generic statements and the answers to chain-of-why-questions. The entailment score ranges from 0 to 1, with higher values indicating better performance, while the hallucination score ranges from 1 to 10, with lower values being better. Parentheses indicate standard deviation.

generate answers to the why-question for a generic statement. Then, we generated further data by repeatedly asking why-questions about the previously generated answers. This process was repeated consecutively n times. In this study, we set n=3 for data generation. We used the GPT-3.5-turbo model, with user instructions provided through the prompt. The hyperparameters were set to their default values for both top-p and temperature, and the maximum output length was set to 64.

WHY-Chain Generation Prompt

Like a follow-up conversation, you are answering the "Why?" question about the previous text.

From now on, please answer with interest about this world and the complex fundamentally scientific or social phenomena that occur here in deeper and more diverse directions.

Please just tell me the reason in your answer.

Don't mention the previous words again.

It must be made into a normal sentence containing a subject and verb.

Questions###
{question} + "Why?"

Answers:

Figure 3: Instruction prompt to generate an answer to the why-question. We utilize a pre-defined instruction prompt to construct the WHY-Chain dataset, consisting of consecutive why-questions and answers, using LLM

Repeated why-questions can sometimes exceed the system's knowledge or capabilities, making it difficult to generate deep-level answers. This can cause delays in answer generation, resulting in responses that are either not in the correct format or are uninterpretable. Therefore, to ensure the reliability of the dataset, we filtered out consecutive why-questions if answers were not generated within a certain time threshold. In this study, we empirically set this threshold to 20 seconds. The statistical summary of the final WHY-Chain dataset, generated through this preprocessing process, is shown in Table 2.

Statistics	Values
Min-Mean-Max value of l_{ans}	55-127-390
Min-Mean-Max value of $N(t_{ans})$	13-20-56
Duplicated ratio of sentences	0.01%

Table 2: The statistical summary of all the answers generated during the chain-of-why-questions process. l_{ans} indicates the length of the answer, $N(t_{ans})$ indicates the number of tokens.

Through this process, we confirmed the length and token count of responses to consecutive why-questions, originated from a generic statement. Additionally, we observed that the majority of the answers were uniquely generated.

2.1.2 Data Validity Check

We conducted entailment verification and hallucination evaluation for validity checks on the WHY-Chain dataset.

Entailment Verification

For entailment verification, we used the question as the premise and the answer as the hypothesis and verified whether the premise supports the hypothesis. We used the nli-entailment-verfier-xxl[2] (Sanyal et al., 2024) model for this purpose, measuring an entailment score between 0 and 1 for each pair. The results are shown in Table 1. The gen.-all(answers) pair consists of the initial generic statement and the concatenated sentences of all the corresponding why-answers. In contrast, the consecutive pair consists of each generated answer paired with the preceding sentence used to generate it. The average entailment score across all pairs is above

[2]https://huggingface.co/soumyasanyal/nli-entailment-verifier-xxl

0.8, demonstrating that the dataset maintains a high level of logical consistency. Furthermore, Figures A1 and A2 illustrate the distribution of entailment scores, indicating that scores below 0.5 account for 4.1% in the gen.-all(answers) pair and less than 10% in the consecutive pair.

Hallucination Evaluation

Based on prompts as illustrated in Figure A3 (Lin and Chen, 2023; Li et al., 2024), we measured the hallucination score using the `GPT-3.5-turbo` model on a scale of 1 to 10. As illustrated in Table 1, we assessed three types of hallucinations in the sentences generated by the LLM. First, entity-error hallucinations arise when the generated text includes incorrect entities, such as people, dates, places, or objects, that conflict with known facts. Second, relation-error hallucinations occur when the text contains inaccurate quantitative or temporal relationships. Lastly, unverifiability hallucinations occur when the generated information cannot be validated using available sources. The results indicate that, for all pairs, the hallucination scores are close to 1, demonstrating that the dataset's pairs are factual and consistent. Additionally, as shown in Figures A4 and A5, the proportion of hallucination scores above 5.5 across all relationships in the three types of hallucinations is less than 0.2%. Furthermore, to verify the robustness of the data generation method, we generated sentences using three types of prompts with the latest LLM, `GPT-4o-mini`, and external evaluation confirmed that there were no hallucination issues in the generated data (see Table A1).

2.2 Learning Strategy for Chain-of-Why-Questions

After constructing the `WHY-Chain` dataset, we proceeded to train a pre-trained language model (e.g., `T5`) to enhance its reasoning capabilities using this data. As part of our model learning strategy, we employed two types of self-supervised objectives: a span-level text-infilling objective and sentence-level objectives. We aim to improve the pre-trained language model through transfer learning for these self-supervised objectives.

2.2.1 Span-Level Objective

Span-level text infilling is a traditional span corruption technique used in the pre-training of the `T5` model. Unlike the `BERT`(Devlin et al., 2019)-style masked language modeling objective, it replaces spans with sentinel tokens such as `<extra_id_0>` and `<extra_id_1>` for masking (see Figure 4).

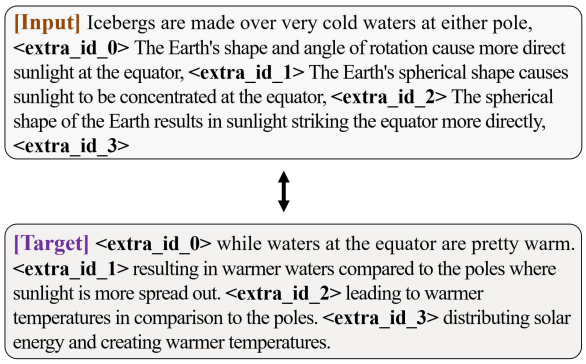

Figure 4: Illustration of the span-level text infilling objective. It involves masking parts of the input text and instructing them to be predicted. In contrast to the `WHY-Chain`, when training the model with `GenericsKB-simplewiki`, there is a key difference in those responses generated through consecutive why-questions (which contain more complex information) are not incorporated as input text.

In this way, spans consisting of multiple tokens in the input text are masked with sentinel tokens, and in the target text, the unmasked parts of the input text are masked. The span-level text infilling technique is essential for training to accurately and effectively capture the answers, which are progressively expanded through consecutive why-questions. We first train the model using a traditional span corruption technique before training sentence-level objectives that reflect the "consecutive" characteristic.

2.2.2 Sentence-Level Objectives

The `WHY-Chain` dataset has a chain structure, so we propose three sentence-level objectives that can reflect the "consecutive" characteristic. We performed transfer learning on the model trained with span-level text infilling using these sentence-level objectives. This section provides a detailed explanation of the three sentence-level objectives.

Irrelevant Why-Answer Detection (IWD)

Irrelevant why-answer detection (IWD) is the task of identifying answers that are not generated through chain-of-why-questions starting from a generic statement. As shown in Figure 5, two of the three candidates are answers to the first and second why-questions for the generic statement, while the remaining one is the answer to the first why-question of a different generic statement. This

process enables the model to filter out information that is difficult to derive from the original generic statement, allowing it to select relevant information at the sentence level.

[Input] Find a candidate that is not related to the statement:
Statement: Icebergs are made over very cold waters at either pole, while waters at the equator are pretty warm

Candidates:
A. The Earth's shape and angle of rotation cause more direct sunlight at the equator, resulting in warmer waters compared to the poles where sunlight is more spread out.
B. The Earth's spherical shape causes sunlight to be concentrated at the equator, leading to warmer temperatures in comparison to the poles.
C. Parasites evolve to maximize their reproductive success by adapting to their host's biology, leading to a co-evolutionary relationship.

[Target] **[C]**

Figure 5: IWD objective example. The task is to select the sentence least related to the statement from the candidates.

Why-Answer Order Prediction (WOP)

Why-answer order prediction (WOP) is a task for predicting the sequence of a series of answers generated through chain-of-why-questions from the initial generic statement. The candidates in Figure 6 were created by randomly shuffling three consecutive why-answers from the generic statement. This helps the model develop the ability to understand and infer causal or sequential relationships between questions and answers. Additionally, it enhances the model's ability to maintain consistency in long conversations and to understand the overall context.

[Input] Correctly align the sentences following the statement:
Statement: Icebergs are made over very cold waters at either pole, while waters at the equator are pretty warm.

[1] The Earth's spherical shape causes sunlight to be concentrated at the equator, leading to warmer temperatures in comparison to the poles. **[2]** The spherical shape of the Earth results in sunlight striking the equator more directly, distributing solar energy and creating warmer temperatures. **[3]** The Earth's spherical shape causes sunlight to be concentrated at the equator, leading to warmer temperatures in comparison to the poles.

[Target] **[3] [1] [2]**

Figure 6: WOP objective example. It involves leading to correctly reordering randomly shuffled sentences.

Reverse Relevant Statement Selection (RRSS)

Reverse Relevant Statement Selection (RRSS) is a task that contrasts with the IWD objective, where, given all the answers generated through the chain-of-why-questions, the goal is to select the corresponding generic statement. As shown in Figure 7, the candidates consist of the correct generic statement and a different one. This objective, which involves reverse training by predicting the cause from the result, can improve the model's ability to better understand the correlation between causes and effects. It also enhances the model's generalization ability and its capacity to comprehend the multifaceted meaning of text.

[Input] Choose a candidate that encompasses the why-answers:
Why-answers: The Earth's shape and angle of rotation cause more direct sunlight at the equator, resulting in warmer waters compared to the poles where sunlight is more spread out. The Earth's spherical shape causes sunlight to be concentrated at the equator, leading to warmer temperatures in comparison to the poles. The spherical shape of the Earth results in sunlight striking the equator more directly, distributing solar energy and creating warmer temperatures.

Candidates:
A. Icebergs are made over very cold waters at either pole, while waters at the equator are pretty warm.
B. Buffalo is on the eastern side of Lake Erie near the Niagara River.

[Target] **[A]**

Figure 7: RRSS objective example. The task is to identify the generic statement corresponding to the why-answers.

2.3 Learning Through Combined Objectives

By utilizing the span-level text infilling objective from Section 2.2.1 and the three objectives from Section 2.2.2, the model can learn from detailed span-level to comprehensive sentence-level. The final loss used for training with sentence-level objectives is as follows:

$$\mathcal{L}_{total} = \frac{1}{3}(\mathcal{L}_{IWD} + \mathcal{L}_{WOP} + \mathcal{L}_{RRSS}). \quad (1)$$

This final objective function not only reflects the characteristics of the span-level objective but also enables the model to be trained in a way that incorporates the specific characteristics of the data through three novel sentence-level objectives.

3 Experimental Setup

3.1 Data

The `WHY-Chain` dataset consists of [generic statement, question, why-answer] pairs and contains 37,455 instances. The generic statement represents the initial generic sentence, the question denotes the answer to the preceding why-question, and the why-answer denotes the answer generated through the current why-question. Additionally, we employed five commonsense reasoning datasets—`COPA`, `CODAH`, `OBQA`, `CSQA`, and `PIQA`—along with the `WHY-Chain` dataset to observe the impact of commonsense on answering. Detailed descriptions of each dataset are provided in Appendix C.

Method	Dataset	Accuracy (official dev)				
		COPA	CODAH	OBQA	CSQA	PIQA
T5-base(STI)	GenericsKB-simplewiki	69.00	58.45	57.00	61.34	68.55
T5-base(STI)	WHY-Chain(proposed)	69.80	60.41	59.80	61.75	69.63
T5-base(STI, total-loss)	WHY-Chain(proposed)	**73.40**	**62.23**	**61.80**	**63.14**	**71.27**
Flan-T5-base(STI)	GenericsKB-simplewiki	81.60	68.71	64.00	72.24	74.21
Flan-T5-base(STI)	WHY-Chain(proposed)	81.80	69.24	64.40	73.05	74.48
Flan-T5-base(STI, total-loss)	WHY-Chain(proposed)	**82.40**	**69.42**	**65.20**	**73.14**	**74.81**
UnifiedQA-v2-base(STI)	GenericsKB-simplewiki	71.20	59.35	59.20	58.15	70.40
UnifiedQA-v2-base(STI)	WHY-Chain(proposed)	71.80	59.53	59.40	59.13	70.43
UnifiedQA-v2-base(STI, total-loss)	WHY-Chain(proposed)	**72.60**	**60.61**	**60.20**	**59.71**	**70.48**

Table 3: The performance of various refined pre-trained language models depends on using the dataset and objective function. STI indicates the span-level text infilling objective, and total-loss indicates to the three sentence-level objectives.

3.2 Pre-Trained Language Model

In our experiments, we aim to enhance the popular pre-trained model, T5 (Raffel et al., 2020), by applying answers to the chain-of-why-questions. To evaluate the generalization and versatility of our proposed approach, we utilized five models as pre-trained language models: T5-small, T5-base, T5-large, UnifiedQA-v2-base (Khashabi et al., 2022), and Flan-T5-base (Chung et al., 2022). A detailed explanation of the models is provided in Appendix B.1.

3.3 Training

We refined the pre-trained language model using the WHY-Chain dataset. To train the model, we utilized two types of self-supervised objectives to proceed with transfer learning. Our models were implemented using pytorch (Paszke et al., 2019) and huggingface's pytorch transformers (Wolf et al., 2020). Transfer learning and fine-tuning details are included in the Appendix B.2.

4 Experimental Results

4.1 Commonsense Reasoning Results

We experimented with the refined pre-trained model on the commonsense reasoning task using the WHY-Chain dataset, and the results are presented in Table 3. As shown in Table 3, the model pre-trained on the WHY-Chain dataset and trained solely with the span-level text infilling objective demonstrated improved commonsense reasoning performance compared to the model pre-trained exclusively on the GenericsKB dataset, which consists of generic statements. This emphasizes the contribution of the WHY-Chain dataset to enhancing commonsense reasoning capabilities. Furthermore, the highest performance was achieved when the model was trained with three sentence-level objectives (total-loss) that leverage the "consecutive" characteristic of the WHY-Chain dataset. This demonstrates that training objectives effectively capturing the characteristic of the WHY-Chain dataset lead to performance improvements. Moreover, consistent performance enhancements were observed across all three models: T5-base, UnifiedQA-v2-base, and Flan-T5-base. This demonstrates that the proposed approach is both generalizable and versatile across various models.

Method	Accuracy (official dev)			
	COPA	OBQA	CSQA	PIQA
KnowBERT	69.40	58.50	53.88	66.61
ERNIE-base	68.90	58.90	54.06	66.47
COMET	69.10	51.20	45.32	60.73
Ours(T5-base)	73.40	61.80	63.14	71.27
Ours(Flan-T5-base)	82.40	65.20	73.14	74.81
Ours(UnifiedQA-v2-base)	72.60	60.20	59.71	70.48

Table 4: Performance comparison with popular knowledge-enhanced pre-trained models. Our models are based on the WHY-Chain dataset and incorporate STI and total-loss.

Next, we conducted comparative experiments with three popular knowledge-enhanced pre-trained models that utilize

Method	Dataset	Accuracy (official dev)				
		COPA	CODAH	OBQA	CSQA	PIQA
T5-small(STI)	GenericsKB-simplewiki	49.80	41.17	50.80	45.21	49.51
T5-small(STI)	WHY-Chain	49.80	43.88	51.00	45.74	49.56
T5-small(STI, total-loss)	WHY-Chain	**50.60**	**46.94**	**51.20**	**45.95**	**50.65**
T5-large(STI)	GenericsKB-simplewiki	75.80	73.38	65.30	69.93	73.04
T5-large(STI)	WHY-Chain	77.20	73.56	65.40	69.93	73.91
T5-large(STI, total-loss)	WHY-Chain	**77.80**	**65.60**	**67.80**	**70.59**	**75.22**

Table 5: Performance of commonsense reasoning across model sizes. These results were obtained by training on the WHY-Chain dataset with two types (span, sentence) of self-supervised objectives.

external knowledge: KnowBERT (Peters et al., 2019), ERNIE-base (Zhang et al., 2019), and COMET (Bosselut et al., 2019b). As shown in Table 4, the performance of our proposed model is superior. This demonstrates that the diverse and complex information obtained through consecutive why-questions has significantly improved the model's reasoning ability.

4.2 Ablation Study

Impact of Model Size

We examined the impact of model size, T5-small and T5-large, on the proposed approach, and the results are shown in Table 5. We observed performance improvements across all datasets, regardless of model size. This demonstrates the size-invariant effectiveness of the proposed approach, highlighting its general applicability.

Comparative Experiments with Retrieved Data

We extracted three statements similar to the generic statement from the GenericsKB dataset to construct a comparison dataset, Retrieval, and conducted comparative experiments. The process of constructing the Retrieval data is detailed in Appendix D. The results are shown in Table 6.

Dataset (Objectives)	Accuracy (official dev)				
	COPA	CODAH	OBQA	CSQA	PIQA
Retrieval (STI)	69.20	60.20	57.40	61.02	68.99
Retrieval (proposed)	69.80	60.25	58.60	61.82	69.15
WHY-Chain (STI)	69.80	60.41	59.80	61.75	69.63
WHY-Chain (proposed)	**73.40**	**62.23**	**61.80**	**63.14**	**71.27**

Table 6: Performance comparison of the Retrieval and WHY-Chain datasets, both starting from the same simplewiki statements but differing in subsequent sentences.

The experimental results demonstrate that the WHY-Chain dataset has a significantly positive impact compared to the dataset composed of similar knowledge to the generic statement. Additionally, it was found that the proposed objectives are particularly effective for training on the WHY-Chain dataset.

Impact of Why-Question Count

We conducted a performance comparison based on the number of consecutive why-questions. As shown in Table 7, the best performance across all commonsense reasoning datasets was achieved when N=3. This suggests that with N=2, the model does not fully capture the characteristics and relationships in the data. In contrast, with N=4, the increased complexity of the proposed objective tasks may introduce challenges that limit the model's learning.

N	Accuracy (official dev)				
	COPA	CODAH	OBQA	CSQA	PIQA
2	69.80	60.61	58.00	61.75	69.86
3	**73.40**	**62.23**	**61.80**	**63.14**	**71.27**
4	68.60	59.71	60.20	61.59	69.48

Table 7: Performance comparison based on the number of why-questions. N is the number of consecutive 'why-questions.

Impact of Each Sentence-Level Objective

We evaluated the impact of the three sentence-level objectives on the performance of downstream tasks, as detailed in Table 8. This evaluation demonstrates that each sentence-level objective enhances performance, indicating that the additional learning strategies are well-suited to the "consecutive" characteristic of the tasks.

Objectives	Accuracy (official dev)				
	COPA	CODAH	OBQA	CSQA	PIQA
STI, total-loss	**73.40**	**62.23**	**61.80**	**63.14**	**71.27**
STI, w/o IWD	72.80	60.97	60.40	61.83	69.75
STI, w/o WOP	70.00	60.43	60.20	62.24	70.13
STI, w/o RRSS	71.80	60.97	61.20	62.57	70.89

Table 8: Model performance based on sentence-level objectives. The pre-trained model utilized T5-base and was trained on the WHY-Chain dataset.

Impact of Objective Order

We compared the performance based on the training order of two types of self-supervised objectives. Table 9 shows better performance when transfer learning is applied from the span-level objective to the sentence-level objectives. This is likely because the span-level text infilling objective follows the traditional pre-training approach, while the three sentence-level objectives are structurally more similar to downstream tasks.

Objective Order	Accuracy (official dev)				
	COPA	CODAH	OBQA	CSQA	PIQA
sentence->span	72.00	60.07	60.32	62.42	70.46
span->sentence	**73.40**	**62.23**	**61.80**	**63.14**	**71.27**

Table 9: Performance comparison based on the training order of two types of self-supervised objectives using T5-base.

5 Reasoning Chain Analysis

We performed a reasoning chain analysis on the WHY-Chain dataset, examining the relationship between the generic statement and answers generated through chain-of-why-questions using informativeness and correctness scores. We believe that chain-of-why-questions and multi-step reasoning are similar in deeply exploring information for better understanding and resolving situations (Prasad et al., 2023). Therefore, we analyze these two scores to evaluate the connection between the generic statement and the answer to the current why-question.

5.1 Informativeness Score

The informativeness score indicates how well the generated why-question answers help derive the generic statement. This score is calculated for the current answer using the overall flow from the generic statement to the current why-question answer, as shown in Equation 2.

$$S_{\text{i}} = \frac{1}{N}\sum_{k=1}^{N}\left(p_{input}(k) - p_{ref}(k)\right), \qquad (2)$$

where $p_z = \log\sigma(f(z)),\ for\ z = \{input, ref\}$, *input* represents the concatenation of previous answers, the current answer, the term "*Therefore*," and the target "*generic statement*", and *ref* is obtained by removing the current answer from *input* and concatenating the remaining components. f uses GPT-2-xl[3] (Radford et al., 2019) model. k is the index for each token in the label sequence, and N is the number of tokenized labels in the target sentence.

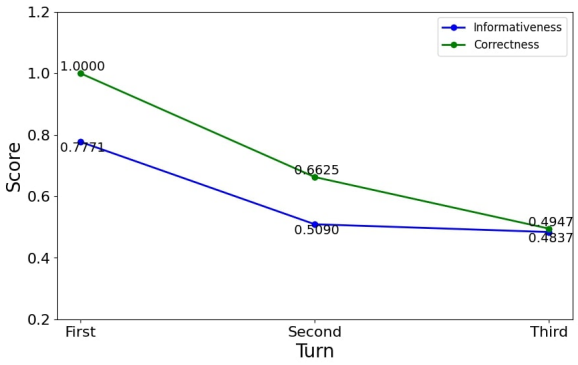

Figure 8: Average correctness and informativeness scores. The graph represents a decrease as the turn progresses.

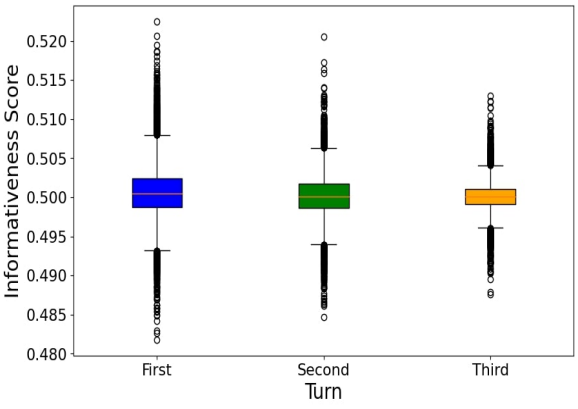

Figure 9: Distribution of Informativeness scores. Turn indicates the ordinal number of the why-question. The average score is approximately 0.5, indicating that the answers generated through consecutive why-questions are indirectly but coherently connected within the overall flow of the generic statement.

[3]https://huggingface.co/openai-community/gpt2-xl

As illustrated in Figure 8, each turn exhibits a sufficiently large positive value, indicating that the current answer has enough informativeness to derive the generic statement. Additionally, as shown in Figure 9, the generated answers are indirectly but coherently connected within the overall flow of the generic statement.

5.2 Correctness Score

The correctness score evaluates the level of contradiction in the answer to the current why-question. By comparing the current answer to the preceding why-question answers and the target generic statement, we can simultaneously assess both intra-connection and inter-connection. For this purpose, we used the exponential moving average (EMA) score, a widely used technical analysis tool for identifying tendencies. We were motivated by (Cai et al., 2021; Morales-Brotons et al., 2024), where the EMA score captures temporal dynamics, reduces fluctuations in data and weights, and enhances stability, ultimately leading to performance improvements. Based on this, we adopted the EMA score to ensure a stable evaluation while tracking the logical consistency between responses and reflecting evolving patterns in conversations with temporal flow. The correctness score based on the EMA score, as shown in Equation 3.

$$S_{\mathrm{c}}^{t} = w_t(1-\max_{r\in R}\{p(a_t,r)\})+(1-w_t)S_{\mathrm{c}}^{t-1}, \quad (3)$$

where S_{c}^{t} represents the correctness score of the current turn t, and $w_t = 2/\{len(A_{1:t-1})+2\}$ denotes the weight of the current turn where $A_{1:t-1}$ is the set of all past turn answers. Adding 2 to the denominator in w_t ensures that the weight remains below 1 even when there are no past turns. a_t is the answer of the current turn, and R is a set where each element of $A_{1:t-1}$ is concatenated with a generic statement. Finally, p represents the probability distribution calcuated by the `DeBERTa-v3-large-mnli-fever-anli-ling-wanli`[4] (Laurer et al., 2024) model. As shown in Equation 3, using w_t allows us to give greater importance to recent data while still incorporating information from all past periods. The results of the correctness score are shown in Figure 8, where the score tends to decrease over time. This suggests that no contradictory sentences were generated during the answer generation process, indicating low volatility and high stability in the data.

[4]https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli

6 Conclusion

We constructed a `WHY-Chain` dataset by generating answers to chain-of-why-questions. Next, we conducted a validity check on the dataset through entailment verification and hallucination evaluation. To refine the pre-trained language model, we applied two types of self-supervised objectives that capture the characteristics of the data. The experimental results indicated that the refined model achieved improved performance in commonsense reasoning tasks, and various ablation studies confirmed the scalability of the approach. Furthermore, using informativeness and correctness scores, reasoning chain analysis validated the positive connections within the data. Our findings suggest that consecutive why-questions effectively enhance the model's understanding by extracting deeper knowledge from the statements. In future research, we plan to expand the dataset by incorporating various question techniques that humans use to acquire knowledge in the real world beyond just Why-questions. Furthermore, we aim to design multimodal datasets that include text and other elements, such as images. Based on the enhanced dataset, we seek to apply it broadly to various tasks. For instance, we plan to extend its application to fact-checking tasks, which focus on precisely identifying the essence of claims and providing reliable verification results. Additionally, we intend to explore Retrieval-Augmented Generation (RAG) tasks, which effectively analyze the implicit information within queries to understand the core issue better and generate high-quality responses tailored to users' inquiries.

Limitations

Pre-training LLMs requires increasingly significant time and computational resources as the volume of generated data expands. In this paper, GPT-3.5-turbo was utilized to generate answers to why-questions. However, the computational cost may vary depending on the model used. For instance, employing a higher-performance model such as GPT-4o (Achiam et al., 2023) may incur additional API or GPU computation costs.

Ethics Statement

It has been confirmed that the WHY-Chain dataset generated through the large language model does not contain any socially controversial or ethically incorrect data.

Acknowledgements

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2020-II201373, Artificial Intelligence Graduate School Program(Hanyang University)) and National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS2024-0040780)

References

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

Paolo Barbieri, Pietro Sarasso, Fabio Lodico, Alice Aliverti, Kou Murayama, Katiuscia Sacco, and Irene Ronga. 2024. The aesthetic valve: how aesthetic appreciation may switch emotional states from anxiety to curiosity. *Philosophical Transactions of the Royal Society B*, 379(1895):20220413.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. Genericskb: A knowledge base of generic statements. *arXiv preprint arXiv:2005.00660*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

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

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

Zhaowei Cai, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Zhuowen Tu, and Stefano Soatto. 2021. Exponential moving average normalization for self-supervised and semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 194–203.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

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.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Karl J Friston, Marco Lin, Christopher D Frith, Giovanni Pezzulo, J Allan Hobson, and Sasha Ondobaka. 2017. Active inference, curiosity and insight. *Neural computation*, 29(10):2633–2683.

Andrew Gillham. 2017. Beyond 'crude pragmatism'in sports coaching: Insights from cs peirce, william james, and john dewey: A commentary. *International Journal of Sports Science & Coaching*, 12(1):53–55.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Sehrang Joo, Sami Yousif, and Frank Keil. 2020. Implicit questions shape information preferences. In *CogSci*.

Sehrang Joo, Sami R Yousif, and Frank C Keil. 2022. Understanding "why:" how implicit questions shape explanation preferences. *Cognitive Science*, 46(2):e13091.

AAIN Eka Karyawati, Edi Winarko, Azhari Azhari, and Agus Harjoko. 2015. Ontology-based why-question analysis using lexico-syntactic patterns. *International Journal of Electrical and Computer Engineering*, 5(2):318.

Humayun Kayesh, Md. Saiful Islam, and Junhu Wang. 2019. Event causality detection in tweets by context word extension and neural networks. In *2019 20th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT)*, pages 352–357.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. Unifiedqa-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2021. TellMeWhy: A dataset for answering why-questions in narratives. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 596–610, Online. Association for Computational Linguistics.

Yash Kumar Lal, Horace Liu, Niket Tandon, Nathanael Chambers, Ray Mooney, and Niranjan Balasubramanian. 2022a. Analyzing the contribution of commonsense knowledge sources for why-question answering. In *ACL 2022 Workshop on Commonsense Representation and Reasoning*.

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

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. *Preprint*, arXiv:2401.03205.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Jordan A Litman and Tiffany L Jimerson. 2004. The measurement of curiosity as a feeling of deprivation. *Journal of personality assessment*, 82(2):147–157.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Eilish McAuliffe, Cindy Van Vaerenbergh, et al. 2006. Guiding change in the irish health system.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. 2024. Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. ReCEval: Evaluating reasoning chains via correctness and informativeness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10066–10086, Singapore. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Soumya Sanyal, Tianyi Xiao, Jiacheng Liu, Wenya Wang, and Xiang Ren. 2024. Are machines better at complex reasoning? unveiling human-machine inference gaps in entailment verification. *Preprint*, arXiv:2402.03686.

Coltan Scrivner. 2022. Curiosity: A behavioral biology perspective.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2018. Conceptnet 5.5: An open multilingual graph of general knowledge. *Preprint*, arXiv:1612.03975.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. *Preprint*, arXiv:2309.03409.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

A WHY-Chain Dataset

A.1 Entailment Verification Visualization

Figures A1 and A2 show the distribution of entailment scores, indicating that the proportion of scores below 0.5 is 4.1% for the gen.-all(answers) pair and less than 10% for the consecutive pair.

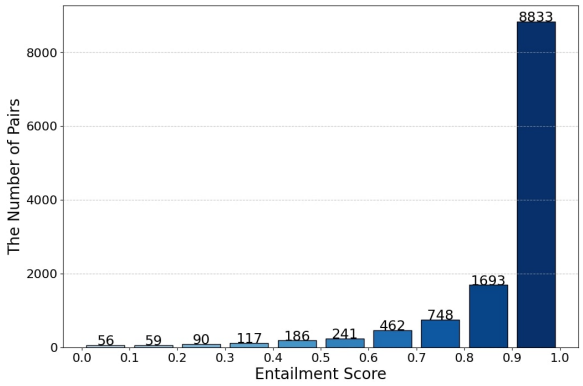

Figure A1: Distribution of entailment scores for generic statement and all answer pairs. The proportion of entailment scores below 0.5 is 4.1%.

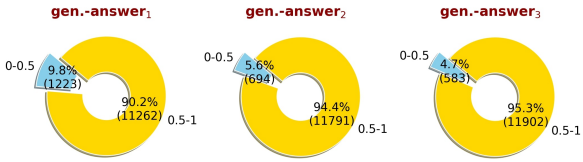

Figure A2: Distribution of entailment scores for each consecutive pair. All three types of relationships maintain a low proportion of entailment scores below 0.5, with the proportion staying under 10%.

A.2 The Average of Hallucination Scores

We externally evaluated hallucinations using sentences generated through three distinct prompting strategies. The first approach employed the `G-Eval`(Liu et al., 2023) method, leveraging its evaluation steps for the prompting strategy. The second approach adopted the `Let's think step by step`(Yang et al., 2024) prompting strategy. The third approach utilized the `Take a deep breath and work on this problem step-by-step` (Kojima et al., 2022) prompting strategy.

A.3 Hallucination Evaluation Visualization

For hallucination evaluation using LLM, we developed instruction prompts based on finely categorized types of hallucinations, following the methodologies described in (Lin and Chen, 2023; Li et al., 2024). Additionally, as shown in Figures A4 and A5, the proportion of hallucination scores above 5 is below 0.4% for all pairs, indicating very low hallucination rates.

Hallucination Instruction Prompt

You are an excellent evaluator who grades.

Given a question and answer pair, decide how likely it is that the answer is a hallucination.

For each question and answer pair, rate the response from 1 to 10 for each hallucination.

Here, 1 means it is very unlikely to be a hallucination, and 10 means it is very likely to be a hallucination.

Each criterion is defined as follows:

Entity-error Hallucination: This type of hallucination refers to the situations where the generated text of LLMs contains erroneous entities, such as person, date, location, and object, that contradict with the world knowledge.

Relation-error Hallucination: This type of hallucination refers to instances where the generated text of LLMs contains wrong relations between entities such as quantitative and chronological relation.

Unverifiability Hallucination: The information generated by LLMs cannot be verified by available information sources.

Example of the output format:

```
{{
  "Entity-error Hallucination": value,
  "Relation-error Hallucination": value,
  "Unverifiability Hallucination: value
}}
```

Content###
Question: {question} + Answer: {answer}

Figure A3: An instruction prompt for generating hallucination scores between question and answer pairs using the LLM.

B Training

B.1 Pre-trained Language Models

We experiment with `T5-small`, which consists of 60 million parameters, 6 layers, a 512 hidden state size, 2048 feed-forward hidden state size, and 8 attention heads; `T5-base`, consisting of 220 million parameters, 12 layers, a 768 hidden state size, 3072 feed-forward hidden states, and 12 attention heads; and `T5-large`, consisting 770 million parameters, 24 layers, a 1024 hidden state size, 4096 feed-forward hidden states, and 16 attention heads. `Flan-T5` is a model fine-tuned through instruction-based training, possessing robust generalization capabilities to handle a wide range of tasks. In contrast, `UnifiedQA` is a widely used model for question-answering tasks, trained by integrating multiple QA formats. This model consistently performs strongly across various question types, demonstrating robust generalization capabilities and providing reliable results across diverse contexts.

Strategy	Pair	Premise-Hypothesis	Hallucination↓		
			Entity-error	Relation-error	Unverifiability
G-Eval	gen.-all(answers)	gen.-c(answers)	1.119(0.616)	1.553(0.932)	1.964(1.034)
	consecutive pair	gen.-answer$_1$	1.072(0.450)	1.365(0.722)	1.579(0.863)
		answer$_1$-answer$_2$	1.047(0.350)	1.419(0.728)	1.492(0.773)
		answer$_2$-answer$_3$	1.049(0.333)	1.457(0.741)	1.509(0.791)
Let's think step by step	gen.-all(answers)	gen.-c(answers)	1.143(0.649)	1.607(0.919)	2.341(1.075)
	consecutive pair	gen.-answer$_1$	1.215(0.668)	1.613(0.895)	2.174(1.119)
		answer$_1$-answer$_2$	1.131(0.479)	1.649(0.824)	2.149(0.989)
		answer$_2$-answer$_3$	1.125(0.477)	1.68(0.83)	2.18(0.993)
Take a deep breath and work on this problem step-by-step	gen.-all(answers)	gen.-c(answers)	1.122(0.627)	1.602(0.864)	2.428(1.046)
	consecutive pair	gen.-answer$_1$	1.168(0.595)	1.551(0.847)	2.191(1.059)
		answer$_1$-answer$_2$	1.086(0.361)	1.573(0.783)	2.138(0.945)
		answer$_2$-answer$_3$	1.094(0.395)	1.624(0.799)	2.192(0.962)

Table A1: The hallucination score ranges from 1 to 10, with higher scores indicating a greater degree of hallucination. As a result, for all four pairs, none of the hallucination evaluation criteria exceeded a score of 2.5, indicating very low levels of hallucination. This demonstrates that no significant hallucination issues were detected after verifying the data quality through various prompting methods.

B.2 Training Details

At first, for span-level text infilling objective, we employed the Adafactor(Shazeer and Stern, 2018) optimizer with epoch 1, batch size 8, and learning rate 2e-5. Subsequently, we performed transfer learning using three sentence-level objectives. We used the Adafactor optimizer with epoch 1, batch size 8, and learning rate 2e-5. Utilizing the refined pre-trained language model, we fine-tuned the model for downstream tasks that require commonsense reasoning. For this fine-tuning process, we used the AdamW (Loshchilov and Hutter, 2017) optimizer with maximum sequence length 256, adam epsilon 1e-8, epoch 20, and batch size 8. The hyperparameter settings for each commonsense reasoning dataset are presented in Table A2. All experiments were conducted identically on NVIDIA GeForce RTX 3090 4ea.

C Dataset Properties

We used the following five datasets to evaluate the model's performance on downstream tasks that require commonsense reasoning ability.

COPA (Roemmele et al., 2011) The Choice Of Plausible Alternatives dataset is used to assess commonsense causal reasoning. Each question is formulated with a premise and two alternatives of cause or effect, structured as multiple-choice QA, where the alternative that more plausibly shares a causal relationship with the premise is to be selected.

CODAH (Chen et al., 2019) The COmmonsense Dataset Adversarially-authored by Humans is an adversarially curated dataset for commonsense testing, consisting of multiple-choice QA that includes sentence completion questions describing scenarios.

CSQA (Talmor et al., 2019) CommonsenseQA is a dataset created by extracting concepts with semantic relationships from ConceptNet (Speer et al., 2018), a knowledge graph designed to represent commonsense relationships.

OBQA (Mihaylov et al., 2018) The OpenBook Question Answering dataset models open-book exams, comprising questions about scientific facts in a 4-way multiple-choice QA format.

PIQA (Bisk et al., 2020) The Physical Interaction Question Answering dataset evaluates physical knowledge, focusing on the affordances and interactions provided by everyday objects, presented as a question with two possible solutions.

D Construction for Retrieval Data

First, from the GenericsKB dataset composed of over 3.5M entries (Bhakthavatsalam et al., 2020), we filtered out sentences that were identical to the generic statements in the GenericsKB-simplewiki dataset. Next, we employed the sentence-transformers (Reimers and Gurevych, 2019) and FAISS[5] (Douze et al., 2024;

[5]https://github.com/facebookresearch/faiss

	COPA	CODAH	OBQA	CSQA	PIQA
Learning rate	[1e-4, 2e-4, 3e-4, 1e-5, 3e-5]	[1e-4, 2e-4, 3e-4, 5e-4]	[1e-4, 2e-4, 3e-4, 3e-5]	[1e-4, 2e-4, 3e-4, 3e-5]	[1e-4, 2e-4, 3e-4, 3e-5]

Table A2: Fine-tuning hyperparameters

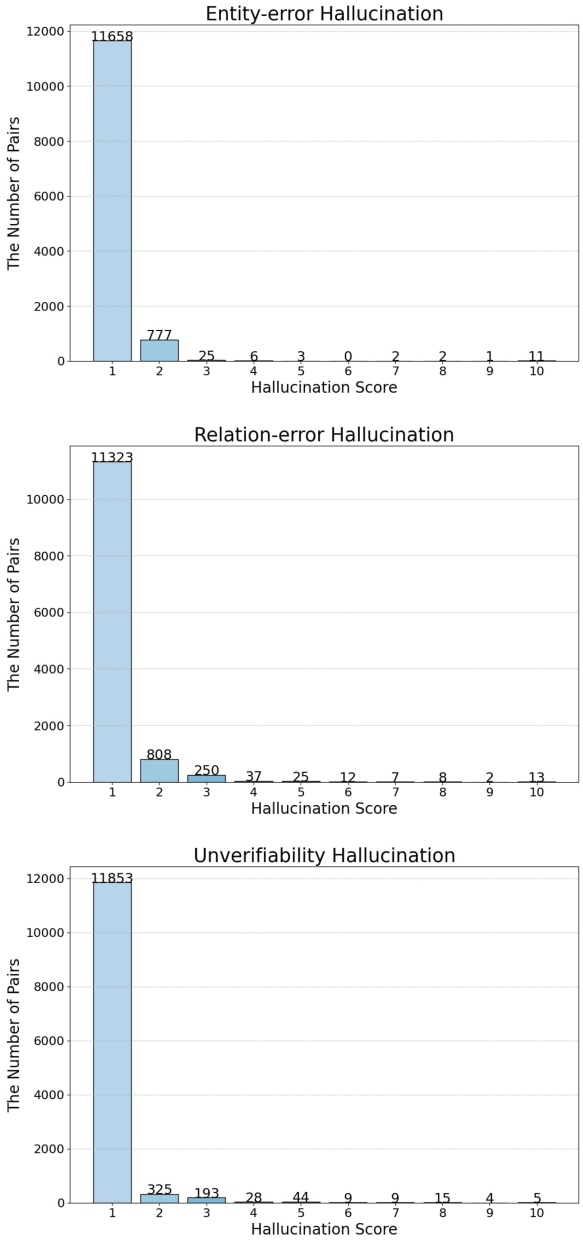

Figure A4: Distribution of hallucination scores for generic statement and all answer pairs. The proportion of all three types of hallucination scores above 5.5 is below 0.4%.

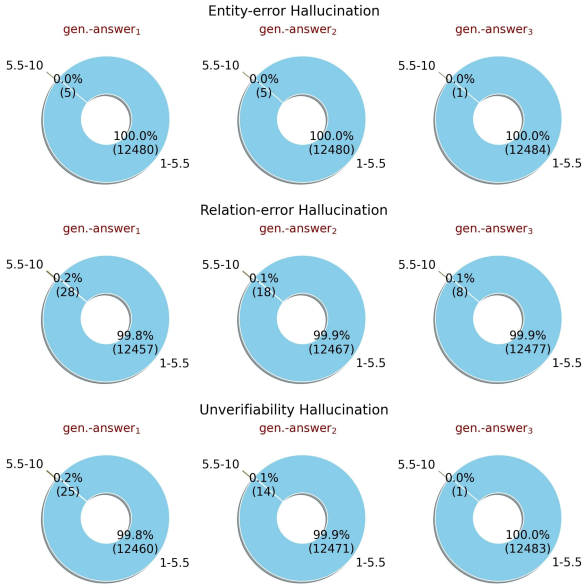

Figure A5: Distribution of hallucination scores for each consecutive pair. The proportion of scores above 5.5 for all relationships in the three types of hallucinations is below 0.2%.

Johnson et al., 2019) libraries to retrieve generic statements from the filtered GenericsKB dataset that were similar to the generic statements in the GenericsKB-simplewiki dataset. Utilizing this constructed dataset, in the same manner as the WHY-Chain, we performed transfer learning on the model and evaluated its performance on tasks requiring commonsense reasoning ability.

E Dataset Example

Figures A6 and A7 are examples from the WHY-Chain dataset.

Generic Statement: Human ears are good at hearing low pitches, but bad at hearing high pitches.		
No	**Question**	**Why-answer**
1	Human ears are good at hearing low pitches, but bad at hearing high pitches.	The structure of the human ear, specifically the inner ear, is optimized for detecting lower frequency sounds due to the size and stiffness of the basilar membrane.
2	The structure of the human ear, specifically the inner ear, is optimized for detecting lower frequency sounds due to the size and stiffness of the basilar membrane.	To efficiently detect lower frequency sounds in the environment, the human ear has evolved to have an inner ear structure that optimizes the detection of these frequencies by utilizing.
3	To efficiently detect lower frequency sounds in the environment, the human ear has evolved to have an inner ear structure that optimizes the detection of these frequencies by utilizing.	Selective pressure favored humans with the ability to detect lower frequency sounds for survival and communication purposes.

Figure A6: An example of WHY-Chain dataset

Generic Statement: Creation myths are usually part of religions and mythologies.		
No	**Question**	**Why-answer**
1	Creation myths are usually part of religions and mythologies.	Creation myths are often present in religions and mythologies to provide explanations for the origins of the universe and human existence.
2	Creation myths are often present in religions and mythologies to provide explanations for the origins of the universe and human existence.	Humans have a natural curiosity and desire to understand the world around them, leading to the creation of myths to make sense of complex phenomena.
3	Humans have a natural curiosity and desire to understand the world around them, leading to the creation of myths to make sense of complex phenomena.	Humans seek patterns and explanations to reduce uncertainty and fear, fostering a sense of control over their environment.

Figure A7: An example of WHY-Chain dataset