# Connecting the Knowledge Dots:
# Retrieval-augmented Knowledge Connection for Commonsense Reasoning

**Junho Kim[1], Soyeon Bak[1], Mingyu Lee[1],**
**Minju Hong[1], Songha Kim[1], Tae-Eui Kam[1] SangKeun Lee[1,2]**
[1]Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea
[2]Department of Computer Science and Engineering, Korea University, Seoul, Republic of Korea
`{monocrat,soypark,decon9201,minjuhong,thdgk0513,kamte,yalphy}@korea.ac.kr`

## Abstract

While large language models (LLMs) have achieved remarkable performance across various natural language processing (NLP) tasks, LLMs exhibit a limited understanding of commonsense reasoning due to the necessity of implicit knowledge that is rarely expressed in text. Recently, retrieval-augmented language models (RALMs) have enhanced their commonsense reasoning ability by incorporating background knowledge from external corpora. However, previous RALMs overlook the implicit nature of commonsense knowledge, potentially leading to the retrieved documents not directly contain information needed to answer questions. In this paper, we propose **Re**trieval-augmented knowledge **Connect**ion, RECONNECT, which transforms indirectly relevant documents into a direct explanation to answer the given question. To this end, we extract relevant knowledge from various retrieved document subsets and aggregate them into a direct explanation. Experimental results show that RECONNECT outperforms state-of-the-art (SOTA) baselines, achieving improvements of +2.0% and +4.6% average accuracy on in-domain (ID) and out-of-domain (OOD) benchmarks, respectively[1].

## 1 Introduction

Commonsense reasoning is a significant challenge for modern NLP models as it requires implicit knowledge that is not directly stated in the text to ground the reasoning process (Gunning, 2018). Although LLMs considerably enhance their reasoning ability by scaling the training data and parameters (Grattafiori et al., 2024), recent studies have shown that LLMs still possess limited understanding of commonsense knowledge (Talmor et al., 2021; Ismayilzada et al., 2023). These challenges have triggered approaches to incorporate external knowledge into LLMs to complement their commonsense reasoning capabilities.

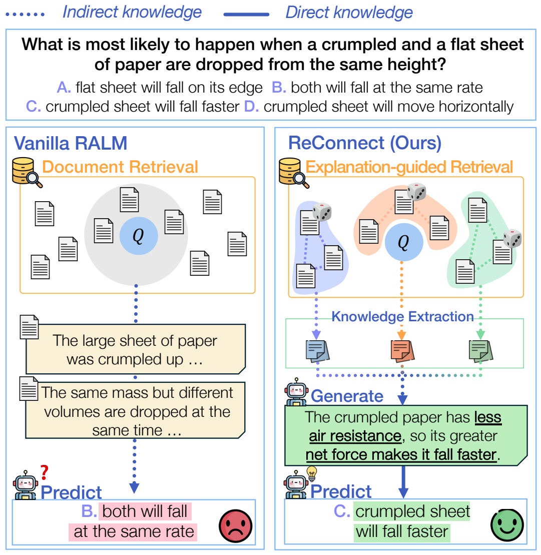
Indirect knowledge
Direct knowledge
What is most likely to happen when a crumpled and a flat sheet of paper are dropped from the same height?
A. flat sheet will fall on its edge B. both will fall at the same rate
C. crumpled sheet will fall faster D. crumpled sheet will move horizontally
Vanilla RALM
Document Retrieval
Q
The large sheet of paper was crumpled up …
The same mass but different volumes are dropped at the same time …
?
Predict
B. both will fall at the same rate
ReConnect (Ours)
Explanation-guided Retrieval
Q
Knowledge Extraction
Generate
The crumpled paper has less air resistance, so its greater net force makes it fall faster.
Predict
C. crumpled sheet will fall faster

Figure 1: Conceptual illustration of Vanilla RALM and RECONNECT (Ours). RECONNECT transforms indirect documents into a direct explanation.

To enhance the reasoning capability of LLMs, RALMs have been introduced to incorporate relevant information from external corpora into the reasoning process (Su et al., 2024; Wang et al., 2025). Recent studies employ a variety of external knowledge sources, such as textual documents (Yu et al., 2022) or exemplars of QA (Molfese et al., 2024), to supplement LLMs with the contextual grounding they often lack. These approaches have yielded notable performance gains in commonsense reasoning tasks (Yu et al., 2022; Molfese et al., 2024).

However, previous RALMs have two challenges that arise from overlooking the nature of implicit commonsense knowledge. First, the commonsense question usually does not explicitly represent the required knowledge. For example, in Figure 1, while understanding concepts like *air resistance* or *net forces* is essential to answer the given question,

---

[1] Our code and data are available at `https://github.com/JunhoKim94/ReConnect`

*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23571–23590
November 4-9, 2025 ©2025 Association for Computational Linguistics

they are not directly represented in the question. This can lead to an inevitable gap between retrieved documents and useful knowledge to answer questions. Second, because commonsense knowledge is seldom explicitly stated in the text, relying solely on finite knowledge bases may not include the direct information needed to answer questions. Such oversights can lead to indirect knowledge retrieval, which potentially undermines their generalizability.

To mitigate these challenges, we propose **Re**trieval-augmented knowledge **Connect**ion (ReConnect), a framework that bridges the gap between indirectly relevant documents and directly useful knowledge to answer commonsense questions. Instead of relying on retrieved documents, ReConnect extracts useful information and transforms it into question-specific explanations. To this end, our approach first employs *explanation-guided retrieval* to expand the original query into a detailed explanation, retrieving more contextually aligned documents. It then applies a *knowledge connection* method that selects informative subsets and integrates complementary perspectives into a coherent explanation. This two-stage process enables ReConnect to generate directly targeted knowledge that enhances reasoning beyond the simple document matching.

To validate the effectiveness of our framework, we evaluate ReConnect on diverse commonsense reasoning benchmarks. Our method consistently outperforms strong retrieval and non-retrieval baselines, achieving large-margin improvements of up to +2.0% accuracy on ID and +4.6% on OOD benchmarks. These results highlight both the effectiveness and generalizability of our approach. Our contributions include the followings:

- We identify key limitations of existing RALMs in commonsense reasoning, highlighting the gap between retrieved documents and directly useful knowledge.
- To address these limitations, we propose ReConnect, a novel framework that transforms indirectly relevant documents into question-specific explanations through explanation-guided retrieval and knowledge connection.
- We show that ReConnect achieves substantial performance gains over previous SOTA methods on both ID and OOD commonsense reasoning tasks up to +2.0% and +4.6% on average, respectively.

## 2 Related Works

### 2.1 Commonsense Knowledge Generation

Commonsense knowledge generation has emerged as a critical research area aimed at equipping LLMs with the ability to reason beyond explicitly stated facts. Early works such as Self-Talk (Shwartz et al., 2020) explicitly formulated commonsense knowledge by generating clarification questions and answers based on hand-crafted templates. GKP (Liu et al., 2022b) elicited commonsense knowledge directly from LLMs by applying a human-created few-shot prompting strategy. However, these approaches require human intervention for each dataset, limiting their generalization and scalability. To address such limitations, existing works (Liu et al., 2022a, 2023; Park et al., 2024) proposed an introspective reasoner, which trains an additional model to generate relevant knowledge for a given question automatically. Although previous works have achieved remarkable results, they require an additional fine-tuning process on training data, which can limit their reasoning ability to what LLMs encountered during training. Moreover, they incur additional costs for training a specific knowledge generation model.

### 2.2 Retrieval-augmented Language Models

RALMs have integrated external knowledge sources into LLM inference, mitigating hallucinations and enhancing reasoning ability. Recent works have primarily focused on when and what to retrieve (Jiang et al., 2023; Su et al., 2024), which dynamically decides the retrieval timing and queries in the reasoning process. Another line of work (Asai et al., 2023; Yan et al., 2024; Wang et al., 2025) introduced multiple refinement iterations or additional instruction-tuning to select relevant documents. Despite these efforts, most previous RALMs usually focused on encyclopedic knowledge for open-domain QA, which may not be suitable for commonsense QA as it requires implicit and diverse knowledge to ground the reasoning process.

In turn, RACo (Yu et al., 2022) first constructed a large-scale (20M+) commonsense corpus and proposed unified retrieval model training strategies specifically adapted for commonsense reasoning tasks. Recently, ZEBRA (Molfese et al., 2024) introduced an example-based retrieval augmentation approach, which generates knowledge tailored to each question. However, the previous works re-

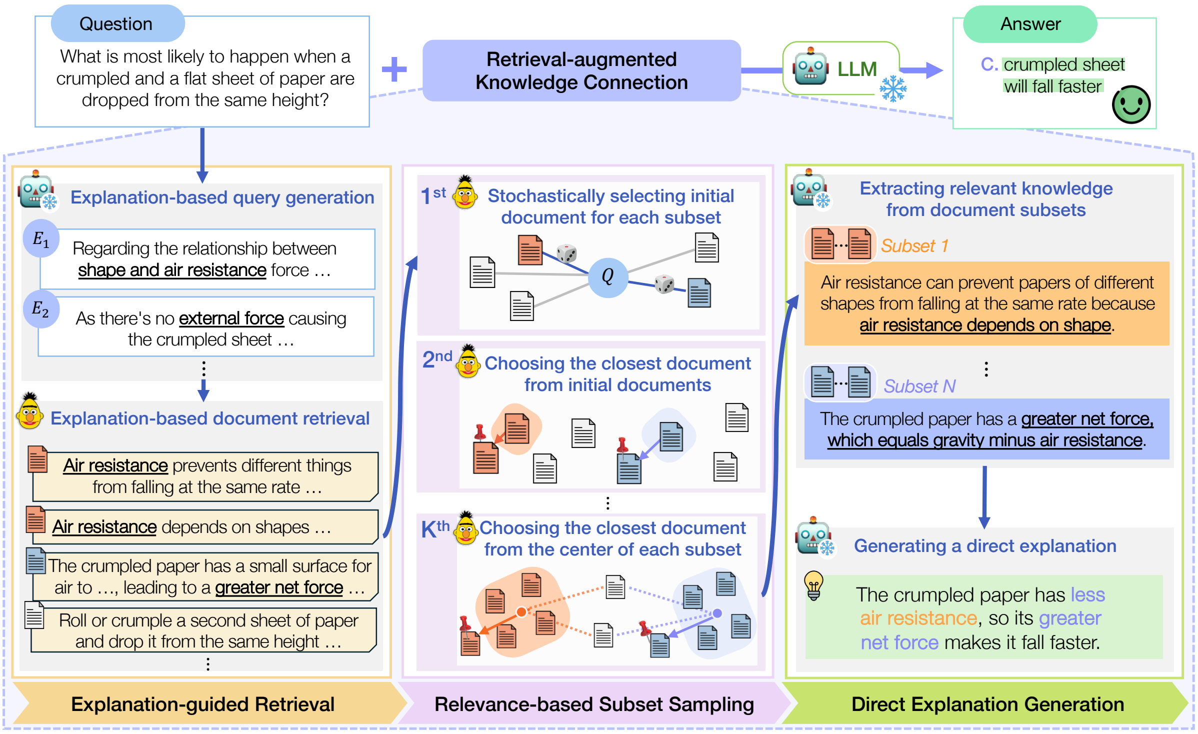
Question
What is most likely to happen when a crumpled and a flat sheet of paper are dropped from the same height?
+
Retrieval-augmented Knowledge Connection
LLM
Answer
C. crumpled sheet will fall faster
Explanation-based query generation
$E_1$
Regarding the relationship between shape and air resistance force …
$E_2$
As there's no external force causing the crumpled sheet …
Explanation-based document retrieval
Air resistance prevents different things from falling at the same rate …
Air resistance depends on shapes …
The crumpled paper has a small surface for air to …, leading to a greater net force …
Roll or crumple a second sheet of paper and drop it from the same height …
1st
Stochastically selecting initial document for each subset
Q
2nd
Choosing the closest document from initial documents
Kth
Choosing the closest document from the center of each subset
Extracting relevant knowledge from document subsets
Subset 1
Air resistance can prevent papers of different shapes from falling at the same rate because air resistance depends on shape.
Subset N
The crumpled paper has a greater net force, which equals gravity minus air resistance.
Generating a direct explanation
The crumpled paper has less air resistance, so its greater net force makes it fall faster.
Explanation-guided Retrieval
Relevance-based Subset Sampling
Direct Explanation Generation

Figure 2: Overview of ReConnect. To elicit the direct explanation from LLMs, ReConnect takes into account the queries as a form of explanation to retrieve. Based on these detailed retrieved documents, we extract knowledge from the subsets with diverse perspectives and aggregate them into a direct explanation for a given question.

main limited by the coverage of the commonsense retrieval corpus, which can lead to indirectly relevant document retrieval. While ZEBRA (Molfese et al., 2024) can mitigate such coverage issues, it requires a customized QA-formed retrieval corpus generated from LLMs, limiting their reasoning ability on the specific QA datasets. Different from the previous works, we focus on transforming diverse indirect documents into a direct explanation for each question beyond the limited coverage of the retrieval corpus.

## 3 Methodology

In this section, we elaborate ReConnect in detail. The core idea is to generate a direct explanation from the indirect documents. To achieve this, ReConnect first expands the query as explanations for gathering more detailed documents from the external knowledge base (Section 3.1). With these documents, ReConnect stochastically samples the necessary knowledge to refine it into direct explanations for answering the question (Section 3.2). Finally, augmenting these refined explanations, ReConnect predicts the answer during inference (Section 3.3). Figure 2 illustrates the overall procedure of the proposed framework.

### 3.1 Explanation-guided Retrieval

To answer commonsense reasoning questions, it is often necessary to retrieve implicit knowledge that is not directly stated in the question. Thus, we utilize model-generated explanations as queries, which include detailed reasoning processes to answer the question. We further propose a training scheme for an explanation-based retrieval model.

**Explanation-based query generation** To elicit detailed information to answer the question, we generate the explanations by prompting the LLM itself. Specifically, given the question $Q$ and choices $C = \{C_1, C_2..., C_k\}$, we generate the number of $k$ explanations $E$ by following demonstrations[2]:

**You are given a question and choices. Your task is to write comprehensive explanations that support the most likely option.**
**...**
**Question: [question]**
**Choices: [choices]**
**Explanations:**

---

[2] All detailed prompts are shown in Appendix G

Here, we replace [question] and [choices] with their corresponding QA instance in the dataset. Rather than relying solely on the question, utilizing both explanations and original questions as queries $q = \{E_1, E_2, ..., E_k, Q\}$ can provide more relevant and diverse documents in the retrieval process.

**Explanation-based retriever model** Since our objective is to find the most relevant document that explains each given question, we propose an explanation-based retriever model, which can find the document supporting the reasoning process. Specifically, we build the retriever on top of DPR (Karpukhin et al., 2020) that encodes the query $q$ and document $d$ separately and estimates the relevance by computing the similarity score between embeddings with the dot product between query and document embeddings:

$$e_q = \mathrm{R}(q), e_d = \mathrm{R}(d), \tag{1}$$

$$\mathrm{sim}(q, d) = e_q^T e_d, \tag{2}$$

where $e_q$ and $e_d$ represent embeddings of query and document, respectively. Using similarity between query and document embedding from retriever R, top-$n$ passages $D(q)$ are returned that are most similar to the input query $q$.

To train our retrieval model, we first construct the training set by incorporating the explanation-based QA datasets generated from LLMs (Park et al., 2024; Molfese et al., 2024)[3]. Formally, from a dataset $D = \{Q, C, E\}$ where a question $Q$, set of choices $C$, and set of explanations $E = (E_1, E_2, ...E_n)$, we set both $Q$ and $E$ as the query $q$ and $E$ as the document $d$. We utilize noise contrastive estimation (NCE) as a training objective:

$$L_{Ret} = -log \sum_{d^+ \in D(q)} \frac{e^{\mathrm{sim}(q,d^+)}}{e^{\mathrm{sim}(q,d^+)} + e^{\mathrm{sim}(q,d^-)}} \tag{3}$$

For positive samples $D^+(q)$, we use the list of explanations $E$. Since explanations entail detailed reasoning processes to answer the questions, the retriever learns to find informative documents that have evidence to solve the question. For negative samples $D^-(q)$, we adopt the in-batch negatives method (Karpukhin et al., 2020), built using the positives of the other queries in the same batch.

## 3.2 Retrieval Knowledge Connection

While explanation-based retrieval can collect more diverse and detailed documents, the retrieved documents $D$ may still not directly explain the question due to their limited coverage. Thus, our objective is to generate a direct explanation from these indirect documents. To this end, we first sample document subsets that entail diverse perspectives and extract the relevant knowledge to solve the question.

**Relevance-based document sampling** For selecting diverse subsets of document $D_i$ from the retrieved document set $D$, we start by running the retriever model to compute an embedded representation for each data point in the pool $e_i = \mathrm{R}(d_i)$. To consider diversity, we randomly select the initial document $d_i$ from $D$. Then, we stochastically select the document by considering the following two aspects: 1) relevance to each question, $\mathrm{sim}(e_q, e_j)$, and 2) similarity among documents in each set, $\mathrm{sim}(e, e_j)$, to minimize the knowledge conflict among documents. Formally, we calculate the score $s_i$ as follows:

$$s_i = \{\mathrm{sim}(e, e_j) + \mathrm{sim}(e_q, e_j)\}_{j=1}^n \tag{4}$$

Where $e$ represents the embedding of the current document subset $D_i$ (i.e., $e = \frac{1}{N}\sum_{i \in D_i} e_i$). $e_j$ and $e_q$ represent the embedding of the question and the $j$-th document embedding, respectively. Based on the scores, we compute the sampling probability $p_i = \sigma(s_i/\tau)$ using the softmax function $\sigma$ and temperature $\tau$ to sample the most related document:

$$t \sim \mathrm{Categorical}(p_i) \tag{5}$$

$$D_i = D_i \cup \{d_t\} \tag{6}$$

where $t$ is a sampled index from a categorical random variable with a probability $p_i$. We iteratively conduct the above process to collect the subset $D_i = \{d_1, ...d_K\}$ until the maximum step $K$.

Finally, we repeat the aforementioned sampling process with a number of explanations $N$ to construct subsets of documents $D_S = \{D_1, ..., D_N\}$. By considering two aspects of scores with stochastic sampling, RECONNECT not only gathers documents that have consistent semantics but also poses more subsets that have more diverse perspectives.

**Knowledge generation** Given a question $Q$ and a list of choices $C$, and sampled document set $D_S = \{D_1, ...D_N\}$, the next step in RECONNECT

---

[3] The details are shown in Appendix A.2

is to extract the relevant knowledge from each subset that can help to answer the question $Q$. Different from the previous retrieval knowledge generation works (Molfese et al., 2024; Yu et al., 2022) that generate knowledge from a single document set, we build on top of multiple documents encompassing diverse perspectives of consistent semantics. Accordingly, our approach encourages an LLM to extract relevant knowledge from subsets of documents that represent various perspectives by generating the list of explanations $E$:

$$E = \{\text{Prompt}(Q, C, D_i) \mid D_i \in D_S\}, \quad (7)$$

**Knowledge aggregation** Finally, to generate a direct explanation for answering the question, we aggregate the generated explanations $E$ that reflect various perspectives. Formally, we generate the final explanation $E_{final}$:

$$E_{final} = \text{Prompt}(Q, C, E). \quad (8)$$

### 3.3 Knowledge-Informed Reasoning

To utilize the generated explanations, we incorporate each generated explanation $E$ into the question $Q$ and the choices $C$. Therefore, the model is asked to predict the correct answer $A$ by conditioning on the input $(Q, C, E_{final})$:

$$A = \text{argmax}_{c \in C} P(c|Q, C, E_{final}) \quad (9)$$

where $P(c|Q, C, E_{final})$ represent the probability of the choice $c$ given question, choice candidates, and explanations. Specifically, we compute the probability of the label for each choice $c$ and select the highest probability as the final prediction.

## 4 Experiments

We demonstrate the efficacy of ReConnect across diverse configurations of commonsense reasoning following these research questions:

**Q1** **(Generalizability)** Does ReConnect enable better performance than existing knowledge augmentation methods across diverse benchmarks? (§4.4)

**Q2** **(Explainability)** How does each component in ReConnect contribute to reasoning abilities of LLMs? (§4.5, §4.6)

**Q3** **(Insights)** How direct explanations by ReConnect improve the commonsense reasoning ability of LLMs? (§4.7)

### 4.1 Experiment Setup

**Language models** Consistent with the previous work (Molfese et al., 2024), we use zero-shot settings for Llama 3.1-8B Instruct model (Grattafiori et al., 2024) as our backbone. The further analysis for other LLMs is shown in Appendix E.

**Retriever corpus** We build our retriever corpus by incorporating the RACo corpus and LLM-based QA datasets, including the ZEBRA (Molfese et al., 2024) and COCONUT corpus (Park et al., 2024). Specifically, we utilize only explanations from LLM-based QA datasets to unify different types of retrieval corpora. For a fair comparison, we use the same retrieval corpus for all baselines. Detailed implementations are shown in the Appendix A.2.

**Retrieval models** We build our retriever based on E5-base-v2 (Wang et al., 2022) (109M), which is pre-trained on a large corpus of text following the previous works (Molfese et al., 2024). We fine-tune the retriever on the training set of ZEBRA-KB (Molfese et al., 2024) that contains diverse explanations generated by LLMs. We construct two explanations $e$ and $e'$ as pair-wise positive if they explain the same questions. The details are shown in Appendix A.3

### 4.2 Evaluation Benchmarks

We evaluate our methods on ID datasets, which include training sets in the retrieval corpus, and OOD datasets that do not include training sets in the retrieval corpus. For ID datasets, consistent with the previous work (Molfese et al., 2024), we evaluate on eight widely utilized commonsense QA datasets: CommonsenseQA (Talmor et al., 2019), OpenBookQA (Mihaylov et al., 2018), ARC-Easy and ARC-Challenge (Clark et al., 2018), PIQA (Bisk et al., 2020), WinoGrande (Sakaguchi et al., 2020), CommonsenseQA 2.0 (Talmor et al., 2021), and QASC (Khot et al., 2020). For OOD datasets, we evaluate on eight different commonsense reasoning datasets: Story Cloze Test (Mostafazadeh et al., 2016), HellaSwag (Zellers et al., 2019), NumerSense (Lin et al., 2020), RiddleSense (Lin et al., 2021), QuaRTz (Tafjord et al., 2019b), SciQ (Welbl et al., 2017), WSC (Levesque et al., 2012), and QuaRel (Tafjord et al., 2019a).

### 4.3 Baselines

We compare our framework with two augmentation methods using the knowledge generated from LLMs or retrieved from an external corpus.

Table 1: Comparison of the zero-shot results with retrieval methods using Llama 3.1-8B Instruct on ID benchmarks. The best and the second-best results are highlighted in **boldface** and underline, respectively.

| Method | ARC-E | ARC-C | PIQA | CSQA2 | CSQA | QASC | OBQA | WG | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| Llama 3.1-8B Instruct | 91.5 | 79.9 | 78.0 | 63.6 | 73.9 | 77.8 | 75.0 | 55.5 | 74.4 |
| *Retrieval augmentation methods* | | | | | | | | | |
| Single Retrieval | 92.3 | 80.5 | 78.5 | 65.7 | 79.5 | 75.8 | 77.0 | 55.9 | 75.6 |
| RACo (2022) | 92.0 | 79.6 | 77.8 | 67.0 | 79.2 | 76.1 | 75.4 | 56.4 | 75.4 |
| FLARE (2023) | 89.8 | 78.8 | 80.9 | 61.2 | 74.5 | 79.2 | 78.0 | 58.6 | 75.1 |
| DRAGIN (2024) | 86.8 | 76.0 | 81.0 | 60.8 | 73.2 | 79.6 | 77.8 | 60.8 | 74.5 |
| ZEBRA (2024) | 93.0 | 83.8 | **85.0** | 70.2 | 77.3 | 75.5 | 81.8 | 60.9 | 78.4 |
| *Knowledge generation methods* | | | | | | | | | |
| Self-Talk (2020) | 91.5 | 78.8 | 76.8 | 63.6 | 74.8 | 77.5 | 72.8 | 59.0 | 74.4 |
| GKP (2022b) | 91.3 | 78.5 | 76.5 | 65.8 | 74.3 | 77.4 | 70.8 | 61.0 | 74.5 |
| Rainier (2022a) | 88.7 | 75.8 | 76.7 | 57.3 | 72.5 | 74.4 | 71.2 | 58.9 | 71.9 |
| Crystal (2023) | 91.2 | 77.7 | 78.5 | 61.0 | 74.9 | 77.5 | 74.6 | 60.5 | 74.5 |
| COCONUT (2024) | 87.9 | 75.5 | 81.3 | 61.0 | 81.5 | 80.0 | 82.2 | 57.6 | 75.9 |
| ReConnect (Ours) | **94.2** | **85.0** | 84.6 | **70.4** | **82.4** | **81.3** | **83.0** | **62.4** | **80.4** |

Table 2: Comparison of the zero-shot results with retrieval methods using Llama 3.1-8B Instruct on OOD benchmarks. The best and the second-best results are highlighted in **boldface** and underline, respectively.

| Method | SCT | HellaSwag | NS | QuaRTz | RS | SciQ | WSC | QuaRel | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| Llama 3.1-8B Instruct | 92.6 | 72.1 | 36.5 | 74.7 | **63.9** | 91.9 | 68.4 | 72.8 | 71.6 |
| ZEBRA (2024) | 93.1 | 68.4 | 59.5 | 79.6 | 62.6 | 92.2 | 72.4 | 76.6 | 75.5 |
| COCONUT (2024) | 93.6 | 70.4 | 37.0 | 73.4 | 57.2 | 89.1 | 73.9 | 76.8 | 71.4 |
| ReConnect (Ours) | **94.2** | **73.0** | **79.5** | **83.1** | 62.9 | **93.6** | **77.9** | **77.2** | **80.1** |

**Retrieval augmentation** We utilize retrieval-augmentation methods as baselines, which focus on augmenting relevant knowledge from commonsense corpora. We use commonsense retrieval augmentations as our baselines, such as RACo (Yu et al., 2022) and ZEBRA (Molfese et al., 2024). We additionally compare to advanced retrieval-augmentation methods, such as FLARE (Jiang et al., 2023) and DRAGIN (Su et al., 2024), which dynamically decide queries and retrieval timing in multi-round retrieval augmentation.

**Knowledge generation** We utilize the commonsense knowledge generation methods as baselines, which involve the generated knowledge from LLMs in an unsupervised manner, such as Self-Talk (Shwartz et al., 2020), GKP (Liu et al., 2022b), and supervised training, such as Rainier (Liu et al., 2022a), Crystal (Liu et al., 2023), and COCONUT (Park et al., 2024) that require additional large training datasets and model for fine-tuning.

### 4.4 Main Results

To investigate the generalizability of our ReConnect, we evaluate the models on both ID and OOD datasets, where the training datasets are included or not in the retrieval corpus, respectively.

**ID datasets** On ID datasets, the results are presented in Table 1. We observed that ReConnect significantly outperforms the SOTA baselines, including knowledge generation and retrieval augmentation methods. Specifically, our framework achieves 2.0% better performance on average accuracy than ZEBRA, the previous SOTA. These results demonstrate that our framework successfully generates the informative knowledge to solve the given questions. Moreover, our framework consistently surpasses supervised training-based knowledge generation methods like COCONUT without fine-tuning specific generation models, which highlights the efficiency of ReConnect.

Table 3: Ablation study of ReConnect on average of ID and OOD benchmark.

| Methods | ID. | OOD. |
|---|---|---|
| ReConnect (Ours) | 80.4 | 80.1 |
| Components in ReConnect | | |
| ReConnect w/o ER | 77.2 | 76.9 |
| ReConnect w/o KC | 76.6 | 75.8 |
| ReConnect w/o ER, KC | 75.3 | 75.0 |
| Document sampling in ReConnect | | |
| ReConnect w/ Relevance (Ours) | 80.4 | 80.1 |
| ReConnect w/ Cluster | 79.2 | 79.5 |
| ReConnect w/ MMR | 79.4 | 79.1 |
| ReConnect w/ Top-K | 78.1 | 77.9 |
| ReConnect w/ Random | 78.3 | 78.8 |

**OOD datasets** To further investigate the generalizability of our method, we additionally evaluate on OOD datasets. The results are shown in Table 2. Interestingly, we observe that the knowledge generation baseline encounters difficulties in performing better than a vanilla model. These results suggest that current knowledge generation methods are not suitable for OOD datasets due to their limited coverage of the training corpus. Moreover, different from the ID dataset, ZEBRA represents marginal improvements or even performance degradation on some OOD benchmarks. Instead, ReConnect largely surpasses SOTA baselines by an average of 4.6%. This result showcases the generalizability of our framework, which can generate a direct explanation even in OOD datasets.

### 4.5 Ablation Studies

We perform ablation studies to verify the contributions of each component in ReConnect[4].

**Retrieval and generation** As shown in the upper part of Table 3, we observe that eliminating explanation-guided retrieval (ER) and knowledge connection (KC) negatively affects both ID and OOD datasets. Specifically, omitting knowledge connecting significantly hurts the performance by 3.7% and 4.3% on ID and OOD datasets, respectively. These results indicate that each component in our framework is essential to generate more high-quality explanations.

**Document sampling** To validate the design choices for document sampling, we compared with three different subset sampling schemes: 1) Cluster sampling, which samples the documents from different clusters (Wang et al., 2025); 2) MMR sampling, which re-rank the samples to prevent redundancy and provide more diverse documents, 3) Top-K sampling, which selects $K$ documents that have the highest relevance score to their queries; 4) Random sampling, which samples the documents randomly. As shown in the lower part of Table 3, adopting random sampling substantially decreases both datasets, indicating the importance of constructing document subsets. Moreover, we observed that using either top-$K$ documents or semantically diverse documents from clusters to construct subsets exhibits inferior performance. These results support our design choice to sample the document subsets with diverse perspectives, while consisting of semantically relevant documents.

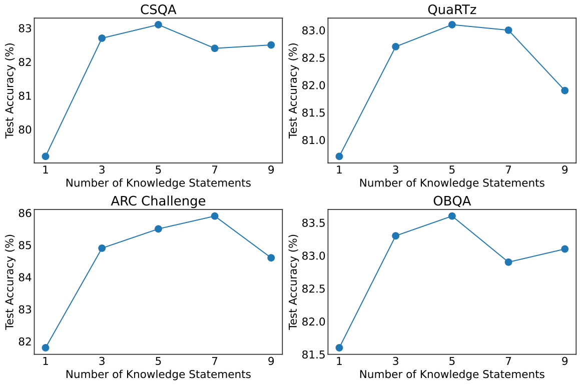

Figure 3: Performances by differentiating the number of generated knowledge per question.

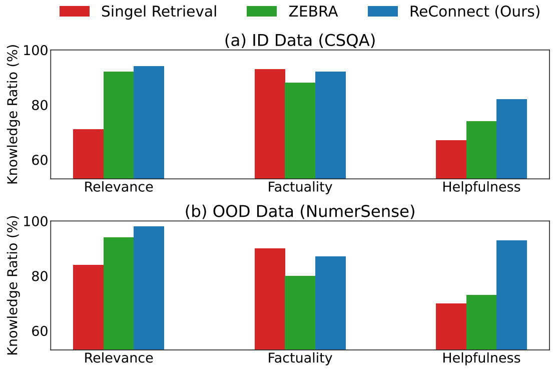

Figure 4: Comparison of the quality of generated knowledge with respect to three aspects.

### 4.6 Impact of Knowledge Connection

**Quantity of generated knowledge** We first analyze the impact of the number of generated knowl-

---

[4] Further ablation studies are shown in Appendix B

Table 4: Case studies for demonstrating how ReConnect works. This example highlights ReConnect's ability to transform fragmented indirect knowledge into coherent and direct knowledge aligned with the given question. Gray highlights indicate retrieved documents; darker shades reflect more supporting documents.

| **NumerSense** | |
|---|---|
| Question | A kid can count up to _ on one hand typically.<br>A. no, B. zero, C. one, D. two, E. three, F. four, G. five, H. six, I. seven, J. eight, K. nine, L. ten |
| Answer | G. five |
| Documents | (1) *A child can count.*<br>(2) *kids can count on their fingers.*<br>(3) *thumb is used to oppose of four fingers.* |
| ReConnect (Ours) | Typically, *a child can count* up to a certain *number using their fingers*, which is often related to the number of fingers they have on one hand. Considering the external knowledge that *a thumb is used to oppose four fingers*, *a child can use their thumb to count up* *to four fingers* on the same hand. This suggests that a child can count up to five using their thumb and four fingers on one hand. |

edge $K$ per question on four different datasets[5]. The results are shown in Figure 3. We observe that performance increases as the number of generated explanations grows. These results demonstrate that our ReConnect successfully generates diverse rationales that are helpful for commonsense reasoning. However, performance is usually saturated over five augmentations and begins to decline, which may be due to the generation of noise knowledge (Liu et al., 2022b; Lee et al., 2024).

**Quality of generated knowledge** To investigate the quality of our generated explanations, we compare our knowledge connection methods to augment the retrieved document (i.e., single retrieval) and example-based generated knowledge (i.e., ZEBRA (Molfese et al., 2024)). Following the standard metrics in the knowledge generation field (Liu et al., 2022b; Molfese et al., 2024), we evaluate the three metrics with human evaluation[6], consisting of 1) relevance, whether the knowledge is relevant to the question, 2) factuality, whether the knowledge is factually correct, and 3) helpfulness, whether the knowledge helps to answer the question directly.

The results are shown in Figure 4. Our observation reveals that ReConnect achieves the highest relevance and helpfulness while maintaining the factuality score close to the retrieved documents. These results indicate that ReConnect not only reduces the semantic gaps between documents and questions but also mitigates hallucination problems. Moreover, while the example-based generation represents better quality than retrieved documents in the ID dataset, its quality significantly dropped in the OOD dataset. However, ReConnect consistently maintains the high quality of knowledge even on OOD datasets. This demonstrates the generalizability of our methods, which can generate relevant knowledge beyond the retrieval corpus.

### 4.7 Case Studies

We illustrate how our framework bridges fragmented, indirect knowledge from multiple documents into unified, direct answers to a given question. Individually, these documents provide only implicit clues that are insufficient on their own. By integrating this partial knowledge, our approach synthesizes coherent explanations that lead to correct predictions. The results are shown in Table 4. For example, in the question "a kid can count up to _ on one hand typically," the retrieved documents mention that "kids count on their fingers" and that "the thumb is used to oppose four fingers." While these statements do not directly state a number, our method connects them to infer the answer "five". These results demonstrate our method's ability to consolidate fragmented and indirect information into coherent, direct knowledge, thereby improving relevance for answering the given question[7].

## 5 Conclusion

We have proposed ReConnect, a novel retrieval augmentation framework to enhance the commonsense reasoning ability of LLMs. To this end, we have introduced explanation-guided retrieval and knowledge connecting methods to bridge the knowledge gap between indirectly relevant documents and useful knowledge to answer. Through

---

[5] Analysis for inference costs are in Appendix D

[6] Detailed settings are represented in Appendix A.5

[7] Further case study results are shown in Appendix F

extensive experiments, we have demonstrated that RECONNECT has significantly outperformed the previous SOTA methods across various commonsense benchmarks. Moreover, we have verified that RECONNECT successfully generated more relevant and helpful knowledge than other baselines.

## 6 Limitations

While we have demonstrated that RECONNECT effectively improves the commonsense reasoning ability of LLMs by refining the indirect documents into direct explanations, there are some limitations that present promising avenues for future research.

**Exploration on Other Types of Reasoning** While we successfully improve the commonsense reasoning abilities of LLMs, it is an open question whether our framework can be applied to other types of complex reasoning strategies, such as logical reasoning, arithmetic reasoning, and symbolic reasoning. Nevertheless, based on recent evidence that the knowledge gap problem between questions and retrieved documents can also happen in complex reasoning (Ma et al., 2023; Wang et al., 2023), we believe that RECONNECT is expected to demonstrate performance improvements on other types of reasoning tasks as well.

**Exploration on Other Forms of QA** We have verified the effectiveness of our framework in the form of multi-choice questions consistent with previous commonsense reasoning studies (Molfese et al., 2024; Li et al., 2024). Therefore, the applicability of our framework to another form of QA, such as open-ended QA, remains under-explored in this work. We leave the exploration of retrieval augmentation for various forms of QA as a future research direction in this field.

**Inference Costs for Knowledge Generation** Our framework requires additional computational costs to generate indirect documents into direct explanations using LLMs. Besides the inference costs, our study mainly focuses on enhancing the complex reasoning ability of LLMs, which remains challenging for current AI systems. This goal aligns with current research in commonsense or complex reasoning, which utilizes more computational power to achieve stronger intelligence, such as using the huge size of LLMs (e.g., GPT-3 175B) as knowledge sources (Liu et al., 2022b), employing multiple agents (Wang et al., 2024), or implementing multiple reasoning processes (Yao et al., 2023b,a). Nonetheless, we believe that investigating the efficiency of the reasoning process for LLMs is a significant research direction and should be addressed as future work.

## Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2025-00517221 and No.RS-2024-00415812) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2024-00439328, Karma: Towards Knowledge Augmentation for Complex Reasoning (SW Starlab), No.RS-2024-00457882, AI Research Hub Project, and No.RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University)).

## References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *Association of the Advanced of Artificial Intelligence (AAAI)*, pages 7432–7439. AAAI Press.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*.

Tri Dao. 2024. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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

David Gunning. 2018. Machine common sense concept paper. *arXiv preprint arXiv:1810.07528*.

Mete Ismayilzada, Debjit Paul, Syrielle Montariol, Mor Geva, and Antoine Bosselut. 2023. CRoW: Benchmarking commonsense reasoning in real-world tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9785–9821. Association for Computational Linguistics.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proc. the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7969–7992. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proc. the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *Association of the Advanced of Artificial Intelligence (AAAI)*, pages 8082–8090. AAAI Press.

Hojae Lee, Junho Kim, and SangKeun Lee. 2024. Mentor-kd: Making small language models better multi-step reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17643–17658.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Association of the Advanced of Artificial Intelligence (AAAI)*.

Jiachun Li, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Daojian Zeng, Kang Liu, and Jun Zhao. 2024. Focus on your question! interpreting and mitigating toxic cot problems in commonsense reasoning. In *Proc. the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9206–9230. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pretrained language models. In *Proc. the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868. Association for Computational Linguistics.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL*, pages 1504–1515. Association for Computational Linguistics.

Jiacheng Liu, Skyler Hallinan, Ximing Lu, Pengfei He, Sean Welleck, Hannaneh Hajishirzi, and Yejin Choi. 2022a. Rainier: Reinforced knowledge introspector for commonsense question answering. In *Proc. the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8938–8958. Association for Computational Linguistics.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022b. Generated knowledge prompting for commonsense reasoning. In *Proc. the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3154–3169. Association for Computational Linguistics.

Jiacheng Liu, Ramakanth Pasunuru, Hannaneh Hajishirzi, Yejin Choi, and Asli Celikyilmaz. 2023. Crystal: Introspective reasoners reinforced with self-feedback. In *Proc. the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11557–11572. Association for Computational Linguistics.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the variance of the adaptive learning rate and beyond. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proc. the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5303–5315. Association for Computational Linguistics.

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 *Proc. the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2381–2391. Association for Computational Linguistics.

Francesco Maria Molfese, Simone Conia, Riccardo Orlando, and Roberto Navigli. 2024. ZEBRA: Zero-shot example-based retrieval augmentation for commonsense question answering. In *Proc. the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 22429–22444. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849. Association for Computational Linguistics.

Jun-Hyung Park, Mingyu Lee, Junho Kim, and SangKeun Lee. 2024. Coconut: Contextualized commonsense unified transformers for graph-based commonsense augmentation of language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5815–5830. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Association of the Advanced of Artificial Intelligence (AAAI)*, pages 8732–8740. AAAI Press.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proc. the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629. Association for Computational Linguistics.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proc. the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 12991–13013. Association for Computational Linguistics.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. 2019a. Quarel: A dataset and models for answering questions about qualitative relationships. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7063–7071.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019b. Quartz: An open-domain dataset of qualitative relationship questions. In *Proc. the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5940–5945. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proc. the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4149–4158. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. Commonsenseqa 2.0: Exposing the limits of AI through gamification. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proc. the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9414–9423. Association for Computational Linguistics.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proc. the Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Zilong Wang, Zifeng Wang, Long Le, Huaixiu Steven Zheng, Swaroop Mishra, Vincent Perot, Yuwei Zhang, Anush Mattapalli, Ankur Taly, Jingbo Shang, et al. 2025. Speculative rag: Enhancing retrieval augmented generation through drafting. In *Proc. the International Conference on Learning Representations (ICLR)*.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*. Association for Computational Linguistics.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *Proc. the International Conference on Learning Representations (ICLR)*. OpenReview.net.

Wenhao Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang, and Meng Jiang. 2022. Retrieval augmentation for commonsense reasoning: A unified approach. In *Proc. the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4364–4377. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proc. the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4791–4800. Association for Computational Linguistics.

# Appendix

## A Implementation Details

We provide detailed implementations, including hyperparameters, dataset statistics, retrieval model training, hardware details, and human evaluations.

### A.1 Detailed Hyper-parameters

For experiments on Llama 3.1 Instruct, we conduct a hyperparameter search on number of retrieved documents $K$ of $\{1, 3, 5, 7, 9\}$, the number of knowledge $N$ of $\{1, 3, 5, 7, 9\}$, and $\tau$ of $\{1.0, 1.5, 2.0\}$. For our main results, we utilize $K = 5$ and $N = 3$. For baselines, we utilize the best performance settings represented in their original paper. Specifically, we utilize the number of knowledge $N = 10$ for COCONUT (Park et al., 2024), Crystal (Liu et al., 2023), and Rainier (Liu et al., 2022a). For Self-Talk (Shwartz et al., 2020) and GKP (Liu et al., 2022b), we set the maximum number of knowledge $N = 20$. For retrieval augmentation baselines, we utilize the number of retrieved documents $K = 5$ for including ZEBRA (Molfese et al., 2024), and $K = 3$ for DRAGIN (Liu et al., 2023), and FLARE (Liu et al., 2022a). We report the average results for all methods across the three random seeds.

### A.2 Retrieval Corpus

Table 5: The number of documents in different corpora.

| | Number of documents |
|---|---|
| RACo | 20,251,133 |
| COCONUT | 3,124,702 |
| ZEBRA-KB | 190,273 |
| ReConnect (Ours) | 23,566,108 |

To construct a broad range of retrieval corpora, we incorporate the three different types of knowledge bases. Specifically, we utilize a document-based corpus, RACo (Yu et al., 2022), and a question-explanation-based corpus, COCONUT (Park et al., 2024) and ZEBRA-KB (Molfese et al., 2024), which are generated by LLMs. For question-explanation-based corpora, we utilize only explanations as documents. The detailed statistics are shown in the Table 5. To prevent the knowledge leakage from the retrieval corpus, we exclude the commonsense benchmark datasets (CBD) from the RACo corpus for ID benchmark evaluation settings [8]. It is noteworthy that we didn't include any development or test sets to prevent the shortcut arising from the direct knowledge linking between questions and correct answers. Moreover, we exclude the OOD benchmark (e.g., HellaSwag) to construct our retrieval corpus.

For a fair comparison, we utilize the same retrieval corpus for all RALM baselines except ZEBRA. ZEBRA can be utilized with their specific corpus, which consists of questions, choices, and explanations. In addition, we observed that utilizing the incorporated other retrieval corpus significantly negatively affects their performance. Thus, we used ZEBRA-KB as the retrieval corpus for the ZEBRA baseline.

### A.3 Retrieval Model Training

We utilize ZEBRA-KB (Molfese et al., 2024) as our training corpus for retrieval models. Specifically, we utilize the questions and explanations for constructing training sets. We randomly split the training and test sets as 90% and 10%, respectively. Following the previous work (Molfese et al., 2024), we select 64 positive samples for each explanation and use up to 200 negative samples per batch. We set the maximum sequence lengths as 256 for questions and explanations, and train our models for 25K steps using RAdam (Liu et al., 2020). For learning rate and weight decay, we use 1e-5 and 1e-2, respectively.

### A.4 Hardware Details

We implement all baselines and our ReConnect using four RTX A6000 GPUs for experiments. For all experiments, we implement with PyTorch, building on FlashAttention2 (Dao, 2024) and bfloat16 mixed-precision, to ensure efficient inference.

### A.5 Human Evaluation

Following the previous study on knowledge generation (Park et al., 2024; Molfese et al., 2024), we asked three NLP experts to annotate the quality of the knowledge generated. We randomly sample 100 examples from each dataset. Each annotator selects the "True" or "False" along three metrics, consisting of relevance, actuality, and helpfulness. The annotation process is conducted in a blind manner that each annotator is not aware of the labels assigned by the other annotators, following the previous work (Molfese et al., 2024). We report the results by utilizing majority voting.

---

[8] We utilize the 21,622,807 documents for ID benchmarks

Table 6: Ablation study of ReConnect for retrieval models on the average of ID and OOD benchmarks.

| Methods | ID. | OOD. |
|---|---|---|
| ReConnect (Ours) | 80.4 | 80.1 |
| ReConnect w/ RACo | 79.1 | 77.5 |
| ReConnect w/ DPR | 79.3 | 78.1 |
| ReConnect w/ ZEBRA | 78.7 | 77.8 |

Table 7: Ablation study of knowledge connection methods in ReConnect for the average of ID and OOD benchmarks.

| Methods | ID. | OOD. |
|---|---|---|
| ReConnect (Ours) | 80.4 | 80.1 |
| ReConnect w/o KE | 76.8 | 77.6 |
| ReConnect w/o KA | 78.1 | 77.4 |
| ReConnect w/o KE, KA | 76.6 | 75.8 |

## B Further Ablation Studies

To further explore the effectiveness of ReConnect and our design choice, we further provide the analysis for the retrieval model and the knowledge aggregation method.

**Retrieval models** In the RALM paradigm, the choice of the retriever plays an important role in retrieving relevant passages from a corpus. Thus, we compare our retrieval model to three different baselines, including 1) RACo, indicating the DPR models trained commonsense datasets following (Yu et al., 2022), 2) DPR, indicating the DPR backbone models (Wang et al., 2022) without fine-tuning on commonsense retrieval corpus, and 3) ZEBRA, which trains based on similarity between commonsense questions in ZEBRA-KB (Molfese et al., 2024). The results are shown in Table 6.

We observed that our proposed retrieval model surpasses previous baselines on both ID and OOD benchmarks. These results demonstrate that explanation-based training methods can further improve commonsense retrieval ability by learning the relations between the detailed reasoning process and questions. Moreover, we observed that our ReConnect surpasses the ZEBRA retriever model, which learns the relevance of commonsense questions. These results support the hypothesis that relying solely on questions for retrieval can cause the knowledge gap between the question and the direct knowledge needed to solve the question.

Table 8: The number of retrieved documents from different corpora

| Methods | RACo | LLM-generated |
|---|---|---|
| w/o Fine-tuning | 89.5% | 10.5% |
| ReConnect | 90.3% | 9.7% |

Table 9: Quantitative analysis for inference costs

| Methods | #Tokens | #LLM calls |
|---|---|---|
| Single RAG | 1 | 1 |
| ZEBRA | 104.5 | 2 |
| ReConnect (N=1) | 343.1 | 4 |
| ReConnect (N=2) | 469.4 | 5 |
| ReConnect (N=3) | 562.5 | 6 |
| ReConnect (N=5) | 821.1 | 8 |
| ReConnect (N=7) | 1036.4 | 10 |

**Components in knowledge connection** For the proposed knowledge connection method, we extract knowledge from diverse document subsets and aggregate them into a single explanation. We conduct further ablation studies for knowledge extraction (KE) and knowledge aggregation (KA). Specifically, "w/o KE" aggregates the knowledge without extracting knowledge from diverse subsets, and "w/o KA" directly utilizes the generated $N$ knowledge for answering the question.

The results are shown in Table 7. We observed that both components contribute to the performance improvements. These results highlight that our knowledge extraction and aggregation successfully extract and refine the knowledge from indirectly relevant documents into a question-specific explanation. Moreover, omitting KE significantly hurts the performance on both ID and OOD datasets. This demonstrate that extracting the knowledge from informative subsets can further improve the performance rather than generating the explanation from documents directly.

## C Style Matching Analysis

Since our retrieval models are fine-tuned on LLM-generated explanations, this may cause the style-matching problem, which can lead to overly selecting LLM-generated documents. Thus, we additionally compare the number of retrieved documents between LLM-generated (i.e., ZEBRA and COCONUT) and human-written text (i.e., RACo) on w/o fine-tuned and our fine-tuned retrieval models.

Table 10: Comparison of the zero-shot results with retrieval methods using Mistral-7B Instruct on OOD benchmarks. We report the accuracy on the development set. The best and the second-best results are highlighted in **boldface** and underline, respectively.

| Method | ARC-C | OBQA | CSQA | PIQA | QASC | NS | Quartz | SCT | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| Mistral-7B Instruct | 73.1 | 73.0 | 68.5 | 76.4 | 67.3 | 37.5 | 72.9 | 90.9 | 70.0 |
| ZEBRA (2024) | 77.5 | 77.0 | 72.9 | 81.3 | 69.2 | 51.0 | 74.4 | 92.1 | 74.4 |
| COCONUT (2024) | 73.2 | 76.2 | 77.4 | 76.0 | 69.9 | 36.0 | 72.1 | **93.2** | 71.8 |
| ReConnect (Ours) | **78.7** | **80.4** | **77.7** | **82.1** | **71.7** | **67.0** | **76.6** | 92.5 | **78.3** |

Table 11: Comparison of the zero-shot results with retrieval methods using Llama 3.2-3B Instruct on various commonsense benchmarks. We report the accuracy on the development set. The best and the second-best results are highlighted in **boldface** and underline, respectively.

| Method | ARC-C | OBQA | CSQA | PIQA | QASC | NS | Quartz | SCT | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| Llama 3.2-3B Instruct | 72.7 | 69.9 | 67.7 | 71.6 | 69.5 | 26.0 | 68.7 | 84.9 | 66.4 |
| ZEBRA (2024) | 76.3 | 75.5 | 72.4 | 77.6 | 72.3 | 38.5 | 75.2 | 89.0 | 72.1 |
| COCONUT (2024) | 70.4 | 75.7 | 76.0 | 75.1 | 77.0 | 25.5 | 66.4 | 89.3 | 69.4 |
| ReConnect (Ours) | **78.6** | **76.2** | **76.3** | **79.0** | **78.0** | **63.0** | **75.5** | **90.1** | **77.1** |

The results are shown in Table 8. Given the different scale of the human-written and AI-generated corpus (20M vs. 3.3M), both retrieval models retrieve diverse documents. Moreover, we observed that our model retrieves a similar number of documents compared to the non-fine-tuned retrieval models. This result demonstrates that our fine-tuning has less effect on style-matching problems. We speculate that this is because the backbone of the retrieval model (Wang et al., 2022) has already learned various types of text styles by pre-training on an enormous human-written corpus, such as Reddit or Common Crawl.

## D Inference Costs Analysis

We additionally provide the detailed quantitative analysis for computational costs in Figure 3. Since the costs of our method can be controlled by the number of N (i.e., the number of generated knowledge), we compare the baselines and our method by varying the number of N. We report the average number of tokens and LLM calls per a single instance in four commonsense reasoning benchmarks, respectively. The results are shown in Table 9. While our methods require more computational costs, we observed that utilizing more computational costs tends to improve the performance of our ReConnect as shown in Figure 3.

## E Results on Various LLMs

To verify the generalizability of our proposed models, we conduct additional experiments on various types and sizes of LLMs, including Mistral v0.3-7B Instruct[9] and Llama 3.2-3B Instruct[10]. Specifically, we evaluate our methods on eight different commonsense benchmarks, including ID and OOD datasets. The results are shown in Tables 10 and 11. We observed that our ReConnect consistently outperforms the SOTA baselines by a large margin. These results demonstrate the generalizability of our proposed methods.

## F Further Case Studies

We provide additional case studies of how our ReConnect successfully transforms the indirect documents into a direct explanation. Specifically, we conduct three different case studies for explanation-based queries, comparison to the SOTA baseline, and indirectly relevant document transformation. The results are shown in Tables 12, 13, and 14.

---

[9] `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`

[10] `https://huggingface.co/meta-llama/Llama-3.2-3B`

## G Knowledge Generation Prompts

In RECONNECT, we utilize LLMs to query expansion, knowledge extraction, knowledge aggregation, and question answering. Following the previous work (Molfese et al., 2024), we construct the prompt with the system, assistant, and user message for all prompts. We report the detailed prompts for each generation in Tables 15, 16, 17, and 18.

Table 12: Case study for explanation-based query expansion.

| | |
|---|---|
| **(a) PIQA** | |
| Question | To make a hood for a cloak<br>A. Sew two pieces of fabric together., B. Drape fabric over your head. |
| Answer | A |
| Queries (Ours) | (1) To make a hood for a cloak, you would need to create a separate piece of fabric that can be attached to the rest of the cloak This would involve cutting out a hood shape from the fabric, and then sewing it to the top of the cloak ...<br>(2) Draping fabric over your head is a structure for a functional hood, and it would not be a practical ... |
| Single Retrieval | (1) There was a woman wearing a cloak with a hood. ✗<br>(2) Cloaks are made of materials. ✗<br>(3) cloak is related to hood. ✗ |
| ReConnect (Ours) | (1) A wash cloth is small enough to be sewn onto a bath towel to make a hood ✓<br>(2) A bath mat is too big to be sewn onto a bath towel to make a hood ✓<br>(3) How do you use your hood on a hoodie? place it over your head. ✓ |
| **(b) QuaRTz** | |
| Question | Reid puts the same amount of gas in a 20ml container and a 10ml container. The container with the densest gas will be the _ container.<br>A. 20ml, B. 10ml |
| Answer | B |
| Queries (Ours) | (1) Becasue the 10ml container has a smaller volume ...<br>(2) Density is mass per unit volume, so a smaller volume means a higher density ... |
| Single Retrieval | (1) Gas: have mass, do not have definite volume, have no definite shape but take the shape of their container, and are the least dense. ✓<br>(2) What is the density of an object that has a mass of 20 grams and a volume of 5 mL ✗<br>(3) (1) gas to liquid (2) gas to solid (3) solid to liquid (4) liquid to gas At constant temperature ... ✗ |
| ReConnect (Ours) | (1) Density is equal to the amount of mass per unit of volume ✓<br>(2) The lower the density, the less mass per unit volume ✓<br>(3) Gas: have mass, do not have definite volume, have no definite shape but take the shape of their container, and are the least dense. ✓ |

Table 13: Case study for comparison to SOTA baseline.

| | |
|---|---|
| **(a) PIQA** | |
| Question | How do I open a water bottle if I can't do it with my hands<br>A. Wrap a small cloth or napkin around it then shake it, B. Wrap a small cloth or napkin around it then twist |
| Answer | B |
| ZEBRA | Explanation 1: Wrapping a cloth around the bottle and shaking it can create enough force to break the seal and open the bottle.<br>Explanation 2: Wrapping a cloth around the bottle and twisting it is unlikely to generate enough force to open the bottle, as the twisting motion is not strong enough.<br>Prediction: A ✗ |
| RECONNECT (Ours) | Explanation: To open a water bottle without using your hands, you can create enough force to dislodge the cap by applying friction. This can be achieved by wrapping a small cloth or napkin around the bottle, which increases the surface area in contact with the cap. * When you shake the bottle (Option A), the friction generated between the cloth and the cap can cause the cap to loosen and eventually come off. Alternatively, twisting the bottle while it's wrapped in a cloth (Option B) can also create enough friction to dislodge the cap. Both methods are plausible solutions, but twisting the bottle while wrapped in a cloth may be more effective in creating the necessary force to open the bottle.<br>Prediction: B ✓ |
| **(b) QuaRTz** | |
| Question | Jim, the manager of the new nuclear energy plant, realized that the energy cost of running the plant was rising. This means that the rate that the resource will be used will<br>A. decrease, B. increase |
| Answer | B |
| ZEBRA | Explanation 1: If the energy cost of running the plant is rising, it becomes more expensive to use the resource.<br>Explanation 2: As the cost increases, people are likely to use the resource less, so the rate of use will decrease. The rising cost will not cause people to use the resource more, so the rate of use will not increase.<br>Prediction: A ✗ |
| RECONNECT (Ours) | Explanation: External knowledge references suggest higher energy cost, which implies increased usage. Higher energy cost will cause the price of energy to go up, likely increasing demand. Energy prices will probably go up, indicating increased usage of the resource. However, the external knowledge reference "As the demand for energy has increased, so has the demand for new fuel" (implying increased usage) contradicts the idea that higher energy cost would lead to decreased usage. Therefore, the most likely explanation is that higher energy cost will cause the price of energy to go up, likely increasing demand. This is supported by the external knowledge reference that higher energy cost implies increased usage.<br>Prediction: B ✓ |

Table 14: Case study for demonstrating how ReConnect works.

| **(a) PIQA** | |
|---|---|
| Question | Prevent drying paint.<br>A. Place paint brushes in newspaper until ready for use again., B. Place paint brushes in ziploc until ready for use again. |
| Answer | B |
| Documents | (1) Keeping the brush bristles in an airtight container will prevent the paint from drying out and becoming unusable.<br>(2) Newspaper absorbs excess paint moisture. |
| ReConnect (Ours) | Explanation: To prevent paint from drying out, it is essential to maintain a moist environment. Newspaper may not provide an airtight environment, which is necessary to keep the paint moist. In fact, newspaper can absorb excess paint moisture, which may not be ideal for preventing paint from drying out. On the other hand, a ziploc bag provides an airtight environment, which is necessary to keep the paint moist and prevent it from drying out. This is supported by the external knowledge reference, which states that keeping the brush bristles in an airtight container will prevent the paint from drying out and becoming unusable. Therefore, the most likely option to prevent drying paint is to place paint brushes in a ziploc bag until ready for use again.<br>Prediction: B ✓ |

| **(b) QuaRTz** | |
|---|---|
| Question | A piece of wire that is short will have _ resistance than one that is very long.<br>A. more, B. less |
| Answer | B |
| Documents | (1) Resistance in wire depends on how thick and how long it is, and what it's made of.<br>(2) Similarly, a long wire allows less of the current being applied to how than a short wire does. |
| ReConnect (Ours) | Explanation: Based on the given explanations, it can be inferred that the resistance of a wire is directly related to its length. The longer the wire, the more resistance it will have. Conversely, a short wire will have less resistance due to its shorter length. Therefore, a piece of wire that is short will have less resistance than one that is very long.<br>Prediction: B ✓ |

Table 15: Prompt for the query expansion step.

**System**: You are given a question and $N$ options. Based on the reasoning skills, your task is to write comprehensive explanations that support the most likely option.
Note that:
* There is always one option that is correct and more likely than the others.
* The explanations must support only the most likely option and refute all the others.
* The explanations must be simple and concise.
Do you understand the task?

**Assistant**: Yes, I understand. Please provide the question and the possible choices

**User**:
Question:
{question}
Choices:
{choices}

**Assistant:** Explanations:

Table 16: Prompt for the knowledge extraction step.

**System**: Below are the external knowledge references:
{sampled documents}
You are given a question, $N$ options, and external knowledge references. Based on the given external references, your task is to extract the most relevant information that supports the most likely option.
Note that:
* It is crucial to critically evaluate the information provided in the external knowledge, recognizing that some of it may be irrelevant to the question.
* Do not simply replicate the given knowledge, but you should offer a refined and comprehensive explanation.
* The explanations must be simple and concise.
Do you understand the task?

**Assistant**: Yes, I understand. Please provide the question and the possible choices

**User**:
Question:
{question}
Choices:
{choices}

**Assistant:** Explanations:

Table 17: Prompt for the knowledge aggregation step.

**System**: Below is a list of explanations:
{list of explanations}
You are given a question, $N$ options, and a list of explanations. Based on the given explanations, your task is to synthesize these explanations into a single, high-quality explanation that supports the most likely option.
Note that:
* It is crucial to critically evaluate the information provided in the external knowledge, recognizing that some of it may be incorrect.
* Do not simply replicate the given knowledge, but you should offer a refined and accurate explanation.
* There is always one option that is correct.
Do you understand the task?

**Assistant**: Yes, I understand. Please provide the question and the possible choices

**User**:
Question:
{question}
Choices:
{choices}

**Assistant:** Aggregated Explanation:

Table 18: Prompt for the question answering step.

| |
|---|
| **System**: You are a helpful assistant for question answering. You are given a question, $N$ options, and a list of explanations. Based on the given explanations, your task is to choose the label corresponding to the best answer for the question.<br>Do you understand the task? |
| **Assistant**: Yes, I understand. Please provide the question and the possible choices |
| **User**:<br>Question:<br>{question}<br>Choices:<br>{choices}<br>Explanations:<br>{aggregated explanations} |
| **Assistant:** Answer: |