ASMR: Aggregated Semantic Matching Retrieval Unleashing Commonsense Ability of LLM through Open-Ended Question Answering

Pei-Ying Lin*, **Erick Chandra***, **Jane Yung-jen Hsu**

Department of Computer Science and Information Engineering, National Taiwan University
pearlie.lin10@gmail.com, erickchandra.1@gmail.com, yjhsu@csie.ntu.edu.tw

Abstract

Commonsense reasoning refers to the ability to make inferences, draw conclusions, and understand the world based on general knowledge and commonsense. Whether Large Language Models (LLMs) have commonsense reasoning ability remains a topic of debate among researchers and experts. When confronted with multiple-choice commonsense reasoning tasks, humans typically rely on their prior knowledge and commonsense to formulate a preliminary answer in mind. Subsequently, they compare this preliminary answer to the provided choices, and select the most likely choice as the final answer. We introduce **A**ggregated **S**emantic **M**atching **R**etrieval (ASMR) as a solution for multiple-choice commonsense reasoning tasks. To mimic the process of humans solving commonsense reasoning tasks with multiple choices, we leverage the capabilities of LLMs to first generate the preliminary possible answers through open-ended question which aids in enhancing the process of retrieving relevant answers to the question from the given choices. Our experiments demonstrate the effectiveness of ASMR on popular commonsense reasoning benchmark datasets, including CSQA, SIQA, and ARC (Easy and Challenge). ASMR achieves state-of-the-art (SOTA) performance with a peak of **+15.3%** accuracy improvement over the previous SOTA on SIQA dataset.

Introduction

Large Language Models (LLMs) have demonstrated exceptional state-of-the-art (SOTA) performance in natural language processing tasks, such as machine translation (Zhu et al. 2023b), text generation (Chung, Kamar, and Amershi 2023), code generation, and information retrieval (Zhu et al. 2023c). The scaling of Large Language Models enables the emergent abilities (Wei et al. 2022b) to solve more complex and diverse tasks. Recent SOTA LLMs such as ChatGPT, Llama 2 (Touvron et al. 2023), Mistral (Jiang et al. 2023a), and GPT-3 (Brown et al. 2020) show outstanding results using few-shot in-context learning, and even with zero-shot prompting (Wei et al. 2022a). However, despite the success of LLMs across various tasks, its utility in multiple-choice question commonsense reasoning tasks is still considered as a challenging task.

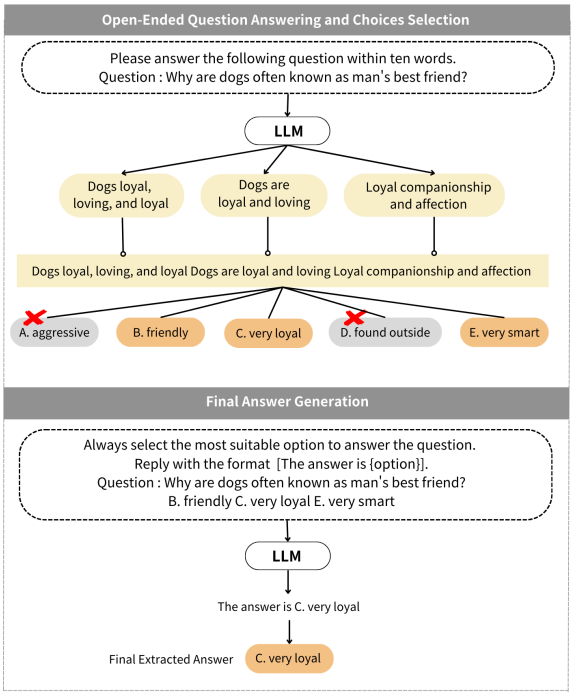

Figure 1: Our proposed ASMR framework on multiple-choice commonsense question answering task. Given a question without choices, first prompt LLM to give three possible answers. Next, aggregate these possible answers and measure the similarities between the aggregated answer and choices. Finally, use Multiple Choice Prompting (MCP) method to predict the final answer.

A multiple-choice question (MCQ) consists of two components – a stem, which defines the question or problem, and a set of answer choices. Among these choices, there is a key representing the correct answer to the question, accompanied by several distractors that present plausible yet incorrect responses. In Multiple-Choice Question-Answering (MCQA) task, the LLM is tasked to select the most suitable answer. Many studies in multiple-choice reasoning task leverage the approach of large language models with in-context learning. Existing work can be grouped into two

*These authors contributed equally.

lines. First, some work propose to present a question to an LLM, where the model independently assesses each choice and selects the option with the highest probability as its chosen answer (Brown et al. 2020; Zhao et al. 2021; Holtzman et al. 2022; Smith et al. 2022). Another line proposes multiple-choice prompting, which a question and its answer choices are all presented to an LLM, allowing model to directly compare the answer choices (Rae et al. 2022; Liévin et al. 2023; Wei et al. 2023; Wang et al. 2023). While such methods can benefit from explicitly comparing the choices, some studies (Zhao et al. 2021; Fei et al. 2023; Han et al. 2022) highlight the Large Language Models expose biases, including majority label bias, recency bias, and common token bias. Recent study (Robinson, Rytting, and Wingate 2023) observed that the second line of work outperforms the first one. Therefore, following the observation, we prompt the LLM with question and choices.

Commonsense question answering demands Large Language Models to acquire diverse forms of commonsense knowledge and reasoning skills. While commonsense knowledge is intuitive to most people, it can be challenging to articulate explicitly. Consequently, it is also challenging for language model to learn commonsense knowledge. Do language model construct its own commonsense cognition during the pre-training?

We argue that the direct provision of all answer choices in the question prompt might limit the true potential of the LLMs to generate correct answers. Resembling what humans commonly perform, we first read the question and then have a preliminary answer in mind. The preliminary answer is then matched with the closest available choices. Inspired from this answer matching mechanism, we introduce **A**ggregated **S**emantic **M**atching **R**etrieval (ASMR), a model-agnostic choice-retrieval process to address multiple-choice commonsense reasoning task. In order to alleviate the influence of distractor choices, we first let the LLM generate a response to an open-ended question utilizing the knowledge it has acquired through pre-training. Afterwards, the question with top-k best matching choices is presented to the LLM to obtain the final answer. The entire process is illustrated in Figure 1.

Our primary **contributions** are twofold:

1. We present ASMR, a model-agnostic choice retrieval method for multiple-choice commonsense reasoning task. To the best of our knowledge, our work is the first to propose prompting the question in an open-ended fashion to unleash the commonsense ability of the LLM.
2. We empirically show that ASMR outperforms all the existing state-of-the-art methods on popular challenging datasets such as CSQA, SIQA, and ARC (Easy and Challenge). Furthermore, ablation studies are performed to show the effectiveness of our method.

Related Work

Since the birth of Transformers (Vaswani et al. 2017), advancement in the field of NLP has been growing fast. It allows the models to learn more complex tasks using larger datasets. Several work (Brown et al. 2020; Ouyang et al. 2022) show that model's number of parameters massive up-scaling could bring better performance, and furthermore with emergent abilities which are not observed in smaller models (Wei et al. 2022b). Subsequently, many Large Language Models are developed, scaled and pre-trained on very large datasets, such as Llama 2 (Touvron et al. 2023), Mistral (Jiang et al. 2023a), and GPT-3 (Brown et al. 2020). These generative models could be used in many downstream tasks, such as text classification, content generation, and information retrieval. Due to the computation cost of training, LLMs are often utilized as a pre-trained foundation model with the pre-trained weights released by the authors, and can be fine-tuned on specific tasks. Due to the emergent abilities, LLMs are able to perform in-context learning (ICL) using only few-shot examples. Moreover, recent studies observe that LLMs are also good zero-shot learners on generic tasks. This significantly changes the horizon of using LLMs, where we can no longer need any labeled samples for fine-tune training or as the ICL few-shot examples. Supporting the motivation of LLM's task transferability and zero-shot ability, our work utilizes the LLMs in a zero-shot fashion by prompting without providing any examples.

While LLMs perform well in many tasks, i.e. machine translation (Zhu et al. 2023b; Zhang, Haddow, and Birch 2023; Zhu et al. 2023a), text generation (Chung, Kamar, and Amershi 2023; Hartvigsen et al. 2022; Sahu et al. 2022), and information retrieval (Zhu et al. 2023c; Liu et al. 2023; Jiang et al. 2023b), they are still lagging in the commonsense multiple-choice question-answering (MCQA) problem (Robinson, Rytting, and Wingate 2023). There are two lines of work in the MCQA problem. First line utilizes the highest probability scored by the model to select the final answer (Brown et al. 2020; Zhao et al. 2021; Holtzman et al. 2022). Brown et al. (2020) employs two different normalization procedures to extract the score, i.e. the normalization of the sequence probability by the nth root, and the normalization of the answer probability by the unconditional probability of the answer. The other line of work prompts the LLM by presenting all the answer choices alongside the question text (Rae et al. 2022; Liévin et al. 2023; Wei et al. 2023; Wang et al. 2023). Robinson, Rytting, and Wingate (2023) show that MCQA using multiple choice prompting outperforms the first line of work.

The closest work to our ASMR is multiple choice prompting (MCP) (Robinson, Rytting, and Wingate 2023). MCP directly provides the question and answer choices as the prompt text to the LLM, with the selected answer choice from the generated response being the final answer. Compared to cloze-prompting (CP), which a question is passed to an LLM and the candidate answers (i.e. answer choices) are scored by the model, MCP performs better and successfully avoid the problems found in CP, such as conflation of likelihood and reliance on normalization procedures. However, MCP with the answer choices provided in the prompt text might limit the LLM to only consider the answer solely from the choices. On the contrary, LLM carries the knowledge it learns from the pre-training which could be useful when determining the answer. Therefore, different from MCP and CP, we adopt open-ended question without the an-

swer choices as the prompt text in ASMR.

Method

Problem Definition. The multiple-choice reasoning task entails assessing a given question $x_i \in X$, alongside a set of predefined answer choices $Y^{x_i} = \{y_1^{x_i}, y_2^{x_i}, ..., y_n^{x_i}\}$ to identify the most suitable or correct answer $\hat{y}^{x_i} \in Y^{x_i}$. Successful completion of this task requires not only comprehending the question's content but also applying reasoning skills to make an informed decision among the provided choices.

Guiding a Language Model (LM) through the process of executing a multiple-choice reasoning task involves inputting the question and answer choices, followed by generating the corresponding raw output. Subsequently, the answer is extracted from the raw output using the following expressions:

$\text{raw\_response} = \textsc{Generate}(x, Y)$

$y = \textsc{Extract}(\text{raw\_response})$

This process essentially instructs the language model to generate responses based on the provided question and choices and then extracts the relevant choices from the model's raw response.

Baseline Methods. As the baseline methods, recent state-of-the-art techniques such as multiple choice prompting (MCP) and self-consistency (SC) are considered.

Multiple Choice Prompting (MCP) (Robinson, Rytting, and Wingate 2023) assesses the answer choices by directly providing all choices beside the question text as the input prompt. Since LLMs suffer from majority label and recency biases (Zhao et al. 2021), demonstrating most few-shot learning is not in reality few-shot, MCP is performed using zero-shot learning. The original MCP utilizes beam search as the decoding strategy, which are the common default strategy found in the publicly available LLMs, e.g. GPT-3 (Brown et al. 2020) and Codex (Chen et al. 2021). To create a more comprehensive comparison, we test with two different decoding strategies, i.e. greedy and beam search to obtain the answer.

Self-Consistency (SC) (Wang et al. 2023) retrieves the answer by feed-forwarding the input prompt n times, and then extract the majority answer among all of the generated responses. It acts like a self-ensemble performed on a single LLM, and replaces the greedy decoding strategy used in chain-of-thought prompting. In our experiments, we employ temperature sampling (Ackley, Hinton, and Sejnowski 1985; Ficler and Goldberg 2017) as the decoding strategy for SC.

ASMR. We propose to initially prompt the model using open-ended question (i.e. prompting without answer choices) to unleash the true potential of the LLM. We utilize "`Question: {question}`" as the question prompt template where `{question}` is the question text. The obtained raw response from the LLM is extracted. SimCSE (Gao, Yao, and Chen 2022) is used to extract the embeddings of the extracted raw responses. We compare the similarity using the complement of cosine similarity, as shown in the following equation, where e_{x_1} and e_{x_2} are the embeddings to be compared, $||\cdot||_2$ is the magnitude of the embedding, and ϵ is the small value to avoid division by zero.

$$\text{cosine\_similarity}(e_{x_1}, e_{x_2}) = \frac{e_{x_1} \cdot e_{x_2}}{\max(||e_{x_1}||_2 \cdot ||e_{x_2}||_2, \epsilon)} \quad (1)$$

The top-k answer choices with the highest similarity value are selected to be included in the next prompt. The filtered choice self-guide the LLM to focus on the more potential candidate answers. We once again prompt the LLM using the template "`Question: {question}\n{choices}`", where `{choices}` contains all the filtered answer choices. The full prompt example used in each method is presented in Table 2. Eventually, the final answer is extracted from the LLM response.

We propose two variations for calculating the similarity scores as follows.

1. **Concatenation (ASMR-C)**
   The extracted raw responses from different decoding strategies are concatenated altogether as a string, delimited by space character. We extract the concatenated text embeddings, and then perform similarity score calculation with all of the answer choices. The overall architecture of ASMR-C is displayed in Figure 2.
2. **Aggregated Sum (ASMR-A)**
   The embeddings of extracted raw responses from the LLM are individually retrieved. We use each of the embeddings to calculate similarity score with the answer choices. Eventually, the similarity scores of each answer choice from different decoding strategies are aggregated by summing them all. The top-k answer choices are selected based on the aggregated similarity scores. The illustration of ASMR-A is shown in Figure 3.

The following steps summarize the process of our method.

1. *Open-Ended Answer Generation*
   We prompt the model using the question x as the text. The LLM is instructed to use three different decoding strategies, i.e. greedy search (m_1), beam search (m_2), and model temperature sampling (m_3) to generate multiple open-ended answers, $Y'^m = \{y'^{m_1}, y'^{m_2}, y'^{m_3}\}$ for the given question x.
2. *Semantic Matching*
   For each generated answers $y'^m \in \{y'^{m_1}, y'^{m_2}, y'^{m_3}\}$, we calculate the semantic similarity score s using Equation 1 between the generated answer $y_i'^m$ and each answer choice $y_i \in Y$.
3. *Similarity Score Calculation*
   When using aggregated sum alternative, the similarity scores s generated from the previous step are summed up per answer choice to arrive at the final similarity scores. For concatenation alternative, this step is skipped.
4. *Relevant Answer Retrieval*
   The top-k matching answer choices with the highest similarity scores are selected.
5. *Final answer extraction*
   We prompt the LLM again using the question x with the

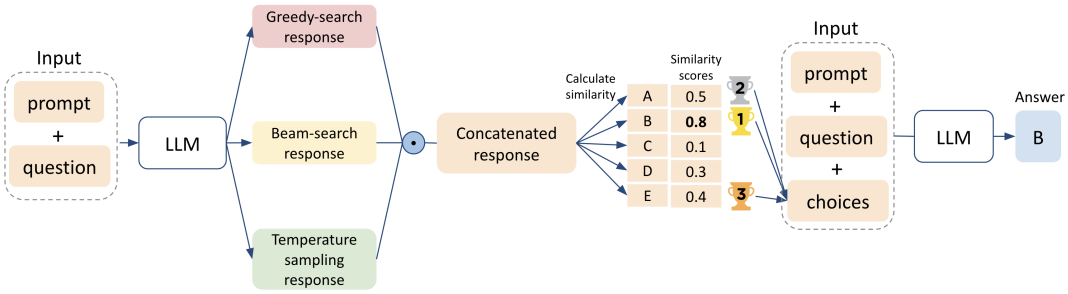

Figure 2: Overview of our approach ASMR-C. Given question x, we instruct LLM to generate three responses. Subsequently, we concatenate these responses to compute the similarity score with all of the answer choices. Afterwards, top-three similar choices are retrieved to instruct LLM to generate the final answer.

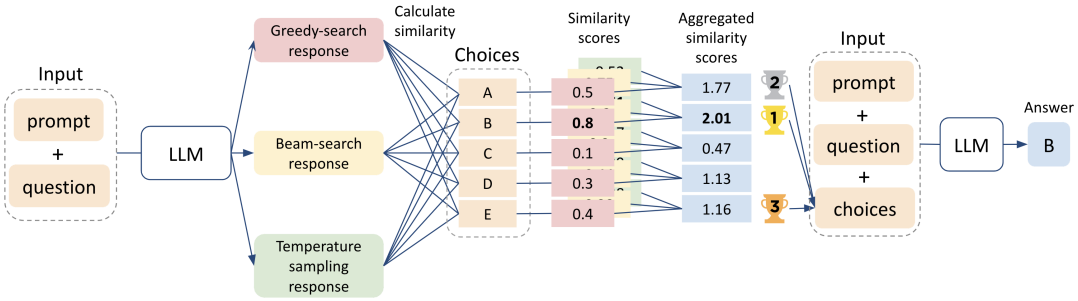

Figure 3: Overview of our approach ASMR-A. Given question x, we instruct LLM to generate three responses. Subsequently, we compute the similarity score for each response with all of the answer choice and then sum the similarity scores of each answer option to obtain the final similarity score. Then, we retrieve top three similar choices to instruct LLM to generate the final answer.

Dataset	# Answer Choices	Training	Validation
CSQA	5	9741	1221
SIQA	3	33400	1950
ARC-E	4	2250	570
ARC-C	4	1120	299

Table 1: Data statistics. CSQA, SIQA, ARC-E, ARC-C denotes CommonsenseQA, SocialIQA, ARC-Easy, ARC-Challenge respectively. "# Answer Choices" denotes the number of provided answer choices for the dataset.

top-k choices retrieved from previous step as the model input and MCP is employed to obtain the raw_output. The final answer y^* is extracted from the raw response. In this step, greedy-search decoding strategy is considered.

Experiments

Dataset. We consider three commonsense multiple-choice reasoning tasks, which are CommonsenseQA (Talmor et al. 2019), SocialIQA (Sap et al. 2019), and ARC (Clark et al. 2018). The data statistics are shown in Table 1. We evaluate our proposed method on the validation set of these tasks since they have ground-truth labels.

CommonsenseQA is a 5-way multiple-choice QA task that requires commonsense knowledge for reasoning. It is based on the knowledge encoded in ConceptNet, which the questions are crowd-sourced. The validation set contains 1221 questions.

SocialIQA (Social Interaction QA) is a 3-way multiple-choice QA task, aimed at testing social and emotional intelligence, which requires commonsense for social interactions for reasoning in a variety of everyday situations. The validation set contains 1950 questions.

ARC (AI2 Reasoning Challenge) dataset is a 4-way multiple-choice question-answering task, which consists of grade-school level science questions. ARC is split into a challenge set **ARC-Challenge** and an easy set **ARC-Easy**, where ARC-Challenge contains more difficult questions that require reasoning.

Model. As Llama 2 (Touvron et al. 2023) is an advanced large language model that has been gaining a lot of attention in the technology world, and since it is available for free for research and commercial use, we employ `Llama-2-7b-chat-hf` on of our experiments. Recent LLM from Mistral (Jiang et al. 2023a), i.e. `Mistral-7B-Instruct-v0.1` is also used in our experiment to empirically show that ASMR is a model-agnostic method which can be used on any existing LLMs. For the semantic matching, we utilize the SimCSE (Gao, Yao, and Chen 2022) since its outstanding performance on the semantic textual similarity task. We run all of the experiments on an RTX 3090 GPU with an Intel Core i9-13900K

Method	Example Prompts
MCP	Question: Why are dogs often known as man's best friend? A. aggressive B. friendly C. very loyal D. found outside E. very smart Answer:
ZS-SC	Question: Why are dogs often known as man's best friend? A. aggressive B. friendly C. very loyal D. found outside E. very smart Let's think step by step. The last sentence of your reply should follow the format: [The answer is {option}].
ASMR-C and ASMR-A	**First Prompt** (in Step 1)**:** Please answer the following question within ten words. Question: Why are dogs often known as man's best friend? **Second Prompt** (in Step 5)**:** Always select the most suitable option to answer the question. Reply with the format [The answer is {option}]. Question: Why are dogs often known as man's best friend? B. friendly C. very loyal E. very smart

Table 2: Prompt examples for all methods

CPU.

Implementation Details. In order to assess the efficacy of our ASMR method, we employ the following methods as the baseline for comparison, and the design of template prompts is presented in Table 2.

MCP (Multiple Choice Prompting). We follow the MCP method described in (Robinson, Rytting, and Wingate 2023). We consider greedy-search decoding strategy for main comparison.

ZS-SC (Zero-Shot Self Consistency). Following the original Self Consistency, we sampled 10 reasoning paths using temperature sampling mechanism. The model temperature is set to 1.0, and the number of sampling is set to 10.

ASMR (Aggregated Semantic Matching Retrieval). For beam search decoding strategy, we set the number of beams to 10, and for temperature sampling, the number of sampling is 10 and the temperature equals to 1.0. For SimCSE model used in the embeddings extraction, we utilize the pre-trained weights of '`sup-simcse-bert-base-uncased`'.

Metric of Evaluation. We report the accuracy scores for all of the experiments. Accuracy score at top-k is employed for the ablation study on the number of selected answer choices.

Results

Main Results. We present our main experimental results in Table 3 and 4. "ASMR-· - topk" signifies that the top-k similar choices are retrieved as the selected choices (as mentioned in Step 5). When k equals to 1, it becomes the final answer. ASMR-C surpasses the baselines on all of the four datasets, while ASMR-A surpasses MCP on two datasets. ASMR achieves peak performance improvement of **+15.3%** using concatenation alternative with top-3 selected choices on SIQA dataset. On SIQA, ARC-E, and ARC-C datasets, ASMR-C consistently performs better than ZS-SC. Overall, ASMR-C outperforms all the other existing state-of-the-art methods. ASMR-A is a potential alternative to ASMR-C on certain dataset, for instance, CSQA.

Our competitive results demonstrate that ASMR successfully elicits the commonsense reasoning ability in LLM, aiding in the retrieval of more probable choices and further enhancing LLM's performance on MCQA tasks. Therefore, prompting the LLM using open-ended question is empiri-

	CSQA	SIQA	ARC-E	ARC-C
MCP	51.2	45.0	58.9	44.8
ZS-SC	59.2	57.2	72.5	51.5
ASMR-C - top1	**60.9**	52.8	64.9	45.2
ASMR-C - top3	58.6	**60.3**	**72.6**	**53.8**

Table 3: Accuracy (%) for ASMR-C on commonsense reasoning tasks. The best scores are boldfaced. The peak performance difference compared to the previous SOTA (i.e. MCP) is underlined.

	CSQA	SIQA	ARC-E	ARC-C
MCP	51.2	45.0	58.9	44.8
ZS-SC	59.2	57.2	**72.5**	**51.5**
ASMR-A - top1	**61.3**	53.6	67.4	47.5
ASMR-A - top3	58.9	**59.7**	70.7	49.5

Table 4: Accuracy (%) for ASMR-A on commonsense reasoning tasks. The best scores are boldfaced.

cally proven to be effective and crucial step to enable full potential of LLM's commonsense reasoning ability.

Response Analysis. We further analyze the response outputs from the LLM for each dataset and method, as depicted in Table 9 in Appendix. One of the success yet challenging examples by ASMR-C but a failure for MCP on CSQA dataset has the question prompt being: "*Why are dogs often known as man's best friend?*". ASMR-C successfully answers with the correct keywords, i.e. "*loyal*" for most if not all of the decoding strategies, however, MCP fails and instead chooses the incorrect choice "*friendly*". Three of the answers provided, i.e. *friendly*, *very loyal*, and *very smart* seem to be plausible and highly probable answers. Nonetheless, the well-known and unique characteristics of dog is being very loyal to human, which means it requires commonsense knowledge to answer the question correctly. From this example, it is shown that LLMs have the commonsense ability unlocked when performing open-ended question, whereas it tends to generate incorrect answer when restricted by certain pre-defined choices.

Model-Agnostic Property. To ensure that ASMR produces consistent performance across other LLMs, we conducted experiments using Mistral on two representative datasets: CSQA and ARC-E. As shown in Table 5, for CSQA task, ASMR outperforms MCP by 13.6%. For ARC-E task, ASMR outperforms MCP by 4.1%. Hence, it is evident that ASMR can be practically applied on different large language models.

Ablation Study. We conduct ablation study to further demonstrate the effectiveness of each component of our method.

Effect of Different Decoding Strategies. ASMR-C performance is investigated using different decoding strategies on step 1. We consider greedy search, beam search, temperature sampling, and the aggregation of the three strategies. For beam search, we set the number of beams to 10, and for temperature sampling, we set number of sampling to 10

	CSQA	ARC-E
MCP	51.5	74.7
ASMR-C - top3	64.8	**78.8**
ASMR-A - top3	**65.1**	78.1

Table 5: Accuracy(%) for ASMR-C on commonsense reasoning tasks using Mistral as the LLM. The best scores are boldfaced.

Strategy	CSQA	SIQA	ARC-E	ARC-C
Greedy search (G)	57.5	53.4	63.5	45.2
Beam search (B)	58.6	52.0	64.7	43.8
Tmp. sampling (S)	58.1	52.1	63.0	44.8
Aggregated (GBS)	**61.3**	**53.6**	**67.4**	**47.5**

Table 6: Top-1 Accuracy (%) on four commonsense reasoning tasks using ASMR-C. The best scores are boldfaced.

with the temperature equals to 1.0. The number of sampling in temperature sampling of the aggregated strategies is set to 1. We measure the effect by calculating the top-1 accuracy, which is the proportion of instances that the selected answer matches the answer key. The results are presented in Table 6. The results show that by aggregating the responses from three decoding strategies, it outperforms those three individual strategies. Compared to temperature scaling, the computation overhead of the aggregation is very small, i.e. 3 forward passes in contrast to k passes in the individual temperature sampling, while the observed performance increase is up to 2.7%.

Effect of the Number of Selected Answer Choices. We compare the number of selected answer choices using ASMR-C by using three strategies. The first is to retrieve the most likely (top-1) answer choice, which has the highest similarity score, as the final answer. For this method, the top-k accuracy equals the final accuracy. The second is to obtain the choices whose similarity score is higher than the average similarity score of all choices. The third is to retrieve top-3 choices measured by similarity scores.

We calculate the top-k accuracy as the proportion of instances which the selected choices contain the correct answer, and calculate the final accuracy after step 5 using ASMR. The results are presented in Table 7. Our finding indicates that the higher the top-k accuracy, the higher the final accuracy, except for CommonsenseQA. Overall, retrieving the top-3 filtered answer choices leads to the highest accuracy, consequently enhancing ASMR's performance in multiple-choice commonsense reasoning tasks.

Effect of ASMR with Self-Consistency. We compare the performance of step 5 in ASMR using greedy-search decoding strategy versus Self-Consistency (temperature sampling). We utilize ASMR-C to retrieve top-3 likely choices and instruct LLM to generate the final answer. This experiment is performed to test the efficacy and orthogonal compatibility of ASMR with other existing work. The results show that ASMR with ZS-SC generally outperforms ZS-SC alone, as depicted in Table 8.

	CSQA		SIQA		ARC-E		ARC-C	
	Top-k Acc	Acc	Top-k Acc	Acc	Top-k Acc	Acc	Top-k Acc	Acc
top 1	60.9	**60.9**	52.7	52.7	64.9	64.9	45.2	45.2
>avg	83.9	58.1	67.7	56.2	80.4	70.8	66.6	48.2
top 3	89.6	58.6	N/A	**60.3**	91.1	**72.6**	81.3	**53.8**

Table 7: Top-k accuracy (%) of the selected answer choices and the final answer accuracy (%) using ASMR-C on four commonsense reasoning tasks. The best scores are boldfaced.

	ZS-SC	ASMR-C - top3 + ZS-SC
CSQA	59.2	**62.1**
SIQA	57.2	57.1
ARC-E	72.5	**73.7**
ARC-C	51.5	**52.8**

Table 8: Accuracy (%) comparison on four commonsense reasoning tasks. The best scores are boldfaced. The results show that ASMR with ZS-SC in general outperforms ZS-SC alone.

Conclusions

In this work, we introduced ASMR, a model-agnostic choice-retrieval process for multiple-choice questions on commonsense reasoning tasks, where the Large Language Model (LLM) first generate a set of preliminary possible answers on open-ended questions. Based on these preliminary answers, top-k likely choices measured by the similarity scores are selected. We prompt the LLM with the question once again with the filtered choices. Eventually, the final answer choice is extracted.

ASMR using concatenation method achieves state-of-the-art performance on challenging datasets such as CSQA, SIQA, ARC-Easy, and ARC-Challenge. The peak performance is observed on SIQA with +15.3% accuracy improvement over previous SOTA method. The main results, along with the ablation studies, demonstrates the efficacy of ASMR, where using open-ended question to prompt the LLM is crucial to unlock its underpinning commonsense reasoning ability.

Future directions include experimenting using various prompt templates for the question prompts and evaluating those prompts on different datasets and models. Further investigation on the commonsense reasoning ability could be performed, especially on the future open-source LLMs that clearly mention or release their pre-training data sources.

Appendix

A.1 ASMR Case Study for Each Dataset

Dataset	Question and Choice	MCP	ASMR-C
CSQA	Why are dogs often known as man's best friend? A. aggressive B. friendly C. very loyal D. found outside E. very smart *Correct answer*: C	B	Greedy: Dogs loyal, loving, and loyal Beam search: Dogs are loyal and loving Temp. sampling: Loyal companionship and affection *Extracted final answer:* C
SIQA	Alex called my parents to inform them that I was safe at their house. How would you describe Alex? A. a close friend of mine B. relieved C. happy *Correct answer*: A	B	Greedy: Caring friend Beam search: Concerned friend Temp. sampling: Concerned friend *Extracted final answer:* A
ARC-E	Which type of graph would best display the changes in temperature over a 24 hour period? A. line graph B. pictograph C. circle (pie) graph D. stem-and-leaf graph *Correct answer*: A	D	Greedy: A line graph would best display temperature changes Beam search: Line graph Temp. sampling: A line graph *Extracted final answer:* A
ARC-C	About how long does it take for the Moon to complete one revolution around Earth? A. 7 days B. 30 days C. 90 days D. 365 days *Correct answer*: B	C	Greedy: Moon completes revolution in 27.3 days Beam search: Moon takes 27.3 days to orbit Earth Temp. sampling: The Moon takes 27.3 days to orbit Earth *Extracted final answer:* B

Table 9: Example of the response output from the LLM (using Llama 2) for each method and dataset

References

Ackley, D. H.; Hinton, G. E.; and Sejnowski, T. J. 1985. A learning algorithm for Boltzmann machines. *Cognitive science*, 9(1): 147–169.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. d. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Chung, J. J. Y.; Kamar, E.; and Amershi, S. 2023. Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions. *arXiv preprint arXiv:2306.04140*.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv:1803.05457.

Fei, Y.; Hou, Y.; Chen, Z.; and Bosselut, A. 2023. Mitigating Label Biases for In-context Learning. arXiv:2305.19148.

Ficler, J.; and Goldberg, Y. 2017. Controlling linguistic style aspects in neural language generation. *arXiv preprint arXiv:1707.02633*.

Gao, T.; Yao, X.; and Chen, D. 2022. SimCSE: Simple Contrastive Learning of Sentence Embeddings. arXiv:2104.08821.

Han, Z.; Hao, Y.; Dong, L.; Sun, Y.; and Wei, F. 2022. Prototypical Calibration for Few-shot Learning of Language Models. arXiv:2205.10183.

Hartvigsen, T.; Gabriel, S.; Palangi, H.; Sap, M.; Ray, D.; and Kamar, E. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. arXiv:2203.09509.

Holtzman, A.; West, P.; Shwartz, V.; Choi, Y.; and Zettlemoyer, L. 2022. Surface Form Competition: Why the Highest Probability Answer Isn't Always Right. arXiv:2104.08315.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023a. Mistral 7B. arXiv:2310.06825.

Jiang, Z.; Xu, F. F.; Gao, L.; Sun, Z.; Liu, Q.; Dwivedi-Yu, J.; Yang, Y.; Callan, J.; and Neubig, G. 2023b. Active Retrieval Augmented Generation. arXiv:2305.06983.

Liu, J.; Jin, J.; Wang, Z.; Cheng, J.; Dou, Z.; and Wen, J.-R. 2023. RETA-LLM: A Retrieval-Augmented Large Language Model Toolkit. arXiv:2306.05212.

Liévin, V.; Hother, C. E.; Motzfeldt, A. G.; and Winther, O. 2023. Can large language models reason about medical questions? arXiv:2207.08143.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Rae, J. W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; Young, S.; Rutherford, E.; Hennigan, T.; Menick, J.; Cassirer, A.; Powell, R.; van den Driessche, G.; Hendricks, L. A.; Rauh, M.; Huang, P.-S.; Glaese, A.; Welbl, J.; Dathathri, S.; Huang, S.; Uesato, J.; Mellor, J.; Higgins, I.; Creswell, A.; McAleese, N.; Wu, A.; Elsen, E.; Jayakumar, S.; Buchatskaya, E.; Budden, D.; Sutherland, E.; Simonyan, K.; Paganini, M.; Sifre, L.; Martens, L.; Li, X. L.; Kuncoro, A.; Nematzadeh, A.; Gribovskaya, E.; Donato, D.; Lazaridou, A.; Mensch, A.; Lespiau, J.-B.; Tsimpoukelli, M.; Grigorev, N.; Fritz, D.; Sottiaux, T.; Pajarskas, M.; Pohlen, T.; Gong, Z.; Toyama, D.; de Masson d'Autume, C.; Li, Y.; Terzi, T.; Mikulik, V.; Babuschkin, I.; Clark, A.; de Las Casas, D.; Guy, A.; Jones, C.; Bradbury, J.; Johnson, M.; Hechtman, B.; Weidinger, L.; Gabriel, I.; Isaac, W.; Lockhart, E.; Osindero, S.; Rimell, L.; Dyer, C.; Vinyals, O.; Ayoub, K.; Stanway, J.; Bennett, L.; Hassabis, D.; Kavukcuoglu, K.; and Irving, G. 2022. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. arXiv:2112.11446.

Robinson, J.; Rytting, C. M.; and Wingate, D. 2023. Leveraging Large Language Models for Multiple Choice Question Answering. arXiv:2210.12353.

Sahu, G.; Rodriguez, P.; Laradji, I. H.; Atighehchian, P.; Vazquez, D.; and Bahdanau, D. 2022. Data Augmentation for Intent Classification with Off-the-shelf Large Language Models. arXiv:2204.01959.

Sap, M.; Rashkin, H.; Chen, D.; LeBras, R.; and Choi, Y. 2019. SocialIQA: Commonsense Reasoning about Social Interactions. arXiv:1904.09728.

Smith, S.; Patwary, M.; Norick, B.; LeGresley, P.; Rajbhandari, S.; Casper, J.; Liu, Z.; Prabhumoye, S.; Zerveas, G.; Korthikanti, V.; Zhang, E.; Child, R.; Aminabadi, R. Y.; Bernauer, J.; Song, X.; Shoeybi, M.; He, Y.; Houston, M.; Tiwary, S.; and Catanzaro, B. 2022. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. arXiv:2201.11990.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. arXiv:1811.00937.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.;

Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171.

Wei, J.; Bosma, M.; Zhao, V.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022a. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022b. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. Survey Certification.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Zhang, B.; Haddow, B.; and Birch, A. 2023. Prompting Large Language Model for Machine Translation: A Case Study. arXiv:2301.07069.

Zhao, T. Z.; Wallace, E.; Feng, S.; Klein, D.; and Singh, S. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. arXiv:2102.09690.

Zhu, W.; Liu, H.; Dong, Q.; Xu, J.; Huang, S.; Kong, L.; Chen, J.; and Li, L. 2023a. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. arXiv:2304.04675.

Zhu, W.; Liu, H.; Dong, Q.; Xu, J.; Kong, L.; Chen, J.; Li, L.; and Huang, S. 2023b. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

Zhu, Y.; Yuan, H.; Wang, S.; Liu, J.; Liu, W.; Deng, C.; Dou, Z.; and Wen, J.-R. 2023c. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.