

# ARR: Question Answering with Large Language Models via Analyzing, Retrieving, and Reasoning

Yuwei Yin

University of British Columbia  
yuweiyin@cs.ubc.ca

Giuseppe Carenini

University of British Columbia  
carenini@cs.ubc.ca

## Abstract

Large language models (LLMs) have demonstrated impressive capabilities on complex evaluation benchmarks, many of which are formulated as question-answering (QA) tasks. Enhancing the performance of LLMs in QA contexts is becoming increasingly vital for advancing their development and applicability. This paper introduces **ARR**, an intuitive, effective, and general QA solving method that explicitly incorporates three key steps: **analyzing** the intent of the question, **retrieving** relevant information, and **reasoning** step by step. Notably, this paper is the first to introduce *intent analysis* in QA, which plays a vital role in ARR. Comprehensive evaluations across 10 diverse QA tasks demonstrate that ARR consistently outperforms the baseline methods. Ablation and case studies further validate the positive contributions of each ARR component. Furthermore, experiments involving variations in prompt design indicate that ARR maintains its effectiveness regardless of the specific prompt formulation. Additionally, extensive evaluations across various model sizes, LLM series, and generation settings solidify the effectiveness, robustness, and generalizability of ARR.<sup>1</sup>

## 1 Introduction

Large language models (LLMs) (Zhao et al., 2023; Min et al., 2023; Minaee et al., 2024) have been a transformative technique in Natural Language Processing (NLP) owing to their excellent text generation and conversation abilities (Hurst et al., 2024; Anthropic, 2024; Team et al., 2024a). Challenging benchmarks for language model evaluation have significantly driven LLM advancements (Chang et al., 2024), with most designed as multiple-choice question-answering (MCQA) tasks (Robinson and Wingate, 2023) requiring answer selection from given options (Clark et al., 2018; Liu et al., 2020; Hendrycks et al., 2021). Recent LLM benchmarks

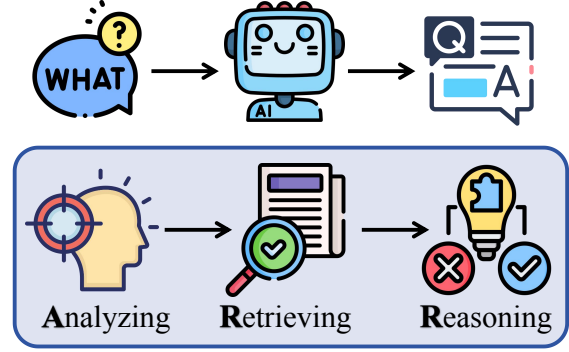


Figure 1: **ARR motivation.** To answer a question, we often need to **analyze** the question’s intent, **retrieve** relevant information, and **reason** step by step.

demand extensive commonsense, world knowledge, and complex reasoning (Srivastava et al., 2023; Suzgun et al., 2023; Wang et al., 2024b), posing significant challenges for LLMs. Optimizing LLM performance in QA tasks is increasingly crucial for their continued development.

As illustrated in Figure 1, answering complex questions typically involves three key steps: (1) analyzing the question’s intent (Adams, 1986; Mele, 1989; Mele and Moser, 1994) to obtain a thorough context understanding, a clear problem-solving target, and a purposeful planning guide, (2) retrieving relevant information from context, external sources, or memory for supportive reference (Jones and Steinhardt, 2022; Shi et al., 2023), and (3) systematically applying inductive and deductive reasoning (Clark, 1969; Johnson-Laird, 1999; Heit, 2000; Hayes and Heit, 2018). Therefore, we hypothesize that an effective solution should direct LLMs to complete these key steps. To verify this hypothesis, we propose a refined QA framework, **ARR**, which explicitly incorporates these three elements: **Analyzing**, **Retrieving**, and **Reasoning**.

As a general framework, ARR can be implemented by simply prompting LLMs to follow the three steps or more elaborately by collecting such

<sup>1</sup>Source code: <https://github.com/YuweiYin/ARR>

three-step data for LLM training. In this work, we investigate ARR as a test-time prompting method because this is the most natural and direct approach to assess its effectiveness when applied to pre-trained foundation models like LLMs. Specifically, ARR adopts the following *answer trigger sentence* at the beginning of LLM’s output: “*Let’s analyze the intent of the question, find relevant information, and answer the question with step-by-step reasoning.*” This explicit and structured approach is expected to enhance the performance across diverse QA tasks and various models.

To evaluate our ARR method, we test the performance (accuracy) of open-weights LLMs (Dubey et al., 2024) on 10 diverse QA datasets, covering reading comprehension (Clark et al., 2019; Liu et al., 2020), commonsense reasoning (Talmor et al., 2019; Sap et al., 2019), world knowledge (Welbl et al., 2017; Mihaylov et al., 2018; Clark et al., 2018), and multitask understanding (Suzgun et al., 2023; Hendrycks et al., 2021; Wang et al., 2024b). Compared to the Direct Answer (DA) method without a specific trigger sentence and zero-shot Chain-of-Thought (CoT) method (Kojima et al., 2022) with a generic prompt (“*Let’s think step by step.*”), **ARR consistently improves QA performance across all datasets**, demonstrating its effectiveness and superiority. Additionally, ablation studies show that each component of ARR (Analyzing, Retrieving, and Reasoning) outperforms the baselines, confirming their individual positive contributions. Notably, **Intent Analysis**—first introduced by ARR—yields the largest performance gain on average, highlighting the critical role of intent analysis in question answering. Moreover, experiments on five distinct prompt variants—each representing a paraphrased version of the original ARR prompt—demonstrate that ARR consistently remains effective irrespective of the particular prompt design.

Furthermore, we conduct extensive experiments across various settings to assess the generalizability of our method. ARR consistently outperforms alternatives across different model sizes, LLM series (architectures), generation temperatures, and few-shot scenarios. These comprehensive experiments and analyses further solidify its effectiveness, robustness, and adaptability. Beyond quantitative results, we provide case studies (Appendix C) to reveal problems in the baselines such as intent misunderstanding, context misuse, and faulty reasoning. The key contributions of this work are as follows:

1. We propose ARR, an intuitive, effective, and general QA framework of three key components: intent analysis, information retrieval, and logical reasoning.
2. Comprehensive experiments across diverse QA tasks demonstrate that ARR consistently outperforms baseline methods. Ablation and case studies further validate the positive contributions of each component.
3. Additional extensive experiments on various settings solidify the effectiveness and generalizability of ARR across different model sizes, LLM series, and generation configurations.

## 2 Related Work

### 2.1 LLM Prompting

Recent large language models (LLMs) (Dubey et al., 2024; Lambert et al., 2024; Liu et al., 2024) are pre-trained on large-scale text corpora curated from the Internet (Soldaini et al., 2024; Penedo et al., 2024; Weber et al., 2024). Their advanced text understanding and generation capabilities (Hurst et al., 2024; Anthropic, 2024; Team et al., 2024a) have significantly revolutionized the field of natural language processing (NLP). Consequently, the NLP paradigm is shifting toward a framework comprising pre-training, post-training, and prompting (Liu et al., 2023), with post-training focusing on aligning models with human preferences (Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023) rather than fine-tuning for specific downstream tasks (Devlin et al., 2019). After the training stages, LLMs can generate satisfactory responses to natural language instructions and questions, highlighting the growing importance of prompt design (White et al., 2023; Giray, 2023; Sahoo et al., 2024). In this work, we implement ARR in a prompting manner to empower LLMs.

### 2.2 LLM Reasoning

Recent LLM research increasingly emphasizes reasoning abilities (Qiao et al., 2023; Sun et al., 2023). Chain-of-Thought (CoT) (Kojima et al., 2022; Wei et al., 2022) is a prompting strategy that enhances problem-solving by guiding LLMs to generate intermediate reasoning steps. Building on CoT, various reasoning techniques have emerged (Zhou et al., 2023a,b; Wang et al., 2023a; Yasunaga et al., 2024; Wang and Zhou, 2024). Some studies explore optimal reasoning paths through self-

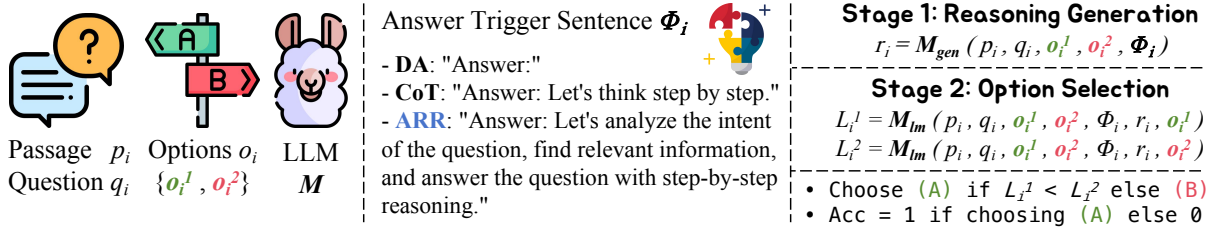


Figure 2: **Question answering with LLMs.** We first obtain rationale  $r_i$  by reasoning generation and then select the optimal option via evaluating the language modeling losses of different context-option combinations.

consistency (Wang et al., 2023c; Chen et al., 2023) or tree-like searches (Yao et al., 2023), while others investigate self-refinement (Madaan et al., 2023), self-correction (Huang et al., 2024; Tyen et al., 2024; Chen et al., 2024), self-verification (Cobbe et al., 2021; Li et al., 2023b; Lightman et al., 2024), and self-evolution (Guan et al., 2025; Lee et al., 2025) mechanisms. Beyond prompting and generation-based approaches, post-training methods (Chu et al., 2025), particularly those leveraging reinforcement learning (RL) (Sutton and Barto, 2018), have been developed to enhance reasoning capabilities (Shao et al., 2024; Wang et al., 2024a; Setlur et al., 2024; Xu et al., 2025). As a structured reasoning-enhancing method, ARR effectively complements existing research by guiding LLMs through three essential steps: intent analysis, information retrieval, and logical reasoning.

### 2.3 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) enhances output quality by retrieving relevant information from pre-processed knowledge sources (Gao et al., 2023). The retrieving component of our ARR method is inspired by the traditional “external RAG” approach (Lewis et al., 2020), which retrieves relevant information from the explicit context or outer sources, and realizes instead a form of “internal RAG,” which utilizes language models as implicit knowledge bases (Petroni et al., 2019; Jiang et al., 2020) and extracts references from memory (training data) (Carlini et al., 2021; Shi et al., 2024). This retrieval mechanism is essential for enhancing LLM performance in question answering, as irrelevant information can significantly degrade accuracy (Jones and Steinhardt, 2022; Shi et al., 2023; Yoran et al., 2024).

## 3 Question Answering with LLMs

This section presents a formally defined multiple-choice question-answering workflow using large

language models. Our pipeline combines ideas from the two-step prompting introduced by Kojima et al. (2022) and the multiple-choice selection method proposed by Robinson and Wingate (2023).

### 3.1 Question Answering Data

In this work, we consider multiple-choice question-answering (MCQA) tasks with one correct answer, where the model is asked to answer the question by selecting an option from a list of choices. Formally, let  $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$  be an MCQA dataset, where  $\mathcal{X} = \{X_1, X_2, \dots, X_n\}$  is the input information,  $\mathcal{Y} = \{y_1, y_2, \dots, y_n\}$  is the corresponding correct-choice label ( $y_i \in \mathbb{R}$ ), and  $n$  is the number of instances in  $\mathcal{D}$ .

In closed-book QA tasks,  $X_i = \{q_i, o_i\}$ , where  $q_i$  is the  $i$ -th question, and  $o_i = \{o_i^j\}_{j=1}^m$  is the option list with  $m$  choices. In open-book QA tasks,  $X_i = \{p_i, q_i, o_i\}$ , where  $p_i$  is the  $i$ -th passage provided by the task. Then, we obtain the input prompt  $x_i$  for LLMs as follows:

$$x_i = \begin{cases} \mathbf{P}(p_i, q_i, o_i), & \text{Open-book QA} \\ \mathbf{P}(q_i, o_i), & \text{Closed-book QA} \end{cases} \quad (1)$$

where  $\mathbf{P}(\cdot)$  denotes the prompt function which concatenates the string objects in  $X_i$  using line breaks as the delimiter ( $\Delta = “\n”$ ). Thus,  $\mathbf{P}(p_i, q_i, o_i)$  is:

$$p_i \Delta q_i \Delta o_i^1 \Delta o_i^2 \Delta \dots \Delta o_i^m \Delta \phi$$

The answer trigger sentence  $\phi$  is the only difference between the proposed ARR method and baseline methods in each experiment. Figure 2 presents each  $\phi$  used in Direct Answer (DA), zero-shot CoT, and our ARR methods. For simplicity,  $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$  is used in the rest of the paper for both open- and closed-book QA.

### 3.2 Multiple-Choice Question Answering

**Stage 1: Reasoning Generation (RG).** Let  $\tilde{x}_i$  be the tokenized representation of text  $x_i$ . The

decoder-only Transformer-based (Vaswani et al., 2017; Radford et al., 2018) LLM  $\mathcal{M}$  takes  $\tilde{x}_i$  as input and generate a new token after each timestep. The model freely generates the text response by

$$r_i = \mathcal{M}(\tilde{x}_i), \quad (2)$$

where  $r_i$  may contain the analysis, reasoning, and answer. Then, we combine the original text input  $x_i$ , the generated response  $r_i$ , and each choice  $o_i^j$  in the option list  $o_i$  as follows:

$$z_i^j = \mathbf{P}(x_i, r_i, o_i^j). \quad (3)$$

**Stage 2: Option Selection (OS).** Let  $\tilde{z}_i^j = [t_i^{j:1}, t_i^{j:2}, \dots, t_i^{j:L}] \in \mathbb{R}^L$  be the tokenized  $z_i^j$ , where  $L$  is the number of effective tokens that are not used for word masking or sequence padding. To select an option, we feed the model  $\mathcal{M}$  and obtain the cross-entropy loss (Shannon, 1948, 1951; Jurafsky and Martin, 2025) of each  $\tilde{z}_i^j$  as follows:

$$\mathcal{L}_i^j = - \sum_k \log \Pr(t_i^{j:k} | t_i^{j:<k}; \Theta), \quad (4)$$

where  $\Theta$  is the parameters of  $\mathcal{M}$ ,  $t_i^{j:k}$  is the  $k$ -th token, and  $t_i^{j:<k}$  denotes all the previous tokens before  $t_i^{j:k}$ . Hence, for each option  $o_i^j$  in  $o_i = \{o_i^j\}_{j=1}^m$ , we have a corresponding cross-entropy loss  $\mathcal{L}_i^j$ . Then, the option with the lowest loss value is selected, i.e.,

$$\hat{y}_i = \arg \min_{j \in \{1, 2, \dots, m\}} \{\mathcal{L}_i^j\}_{j=1}^m. \quad (5)$$

Thus, the overall accuracy is calculated by

$$\alpha = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(y_i = \hat{y}_i), \quad (6)$$

where  $\alpha \in [0, 1]$  and the indicator function  $\mathbb{I}(\cdot)$  returns 1 if  $y_i = \hat{y}_i$  or 0 otherwise.

## 4 Experimental Setup

This section introduces the experimental setup, including datasets, models, and evaluation settings.<sup>2</sup>

### 4.1 Datasets

As mentioned in § 3.1, we consider 10 multiple-choice QA tasks with questions  $q_i$  and options  $o_i$ . Reading comprehension tasks (Chen, 2018) explicitly provide passages  $p_i$  to base on. The model  $\mathcal{M}$

QA Dataset	Split	Size	# Tok.	# Class
BoolQ	Valid	3,270	145	2
LogiQA	Test	651	192	4
CSQA	Valid	1,221	43	5
SIQA	Valid	1,954	51	3
SciQ	Test	1,000	132	4
OBQA	Test	500	55	4
ARC	Test	3,548	59	4
BBH	Test	5,281	112	2–18
MMLU	Test	13,842	108	4
MMLU-Pro	Test	12,032	186	10

Table 1: **QA dataset statistics.** “# Class” is the number of options  $m$ , “Size” is the total number of data items for evaluation, and “# Tok.” is the average number of tokens per instance (zero-shot prompt), tokenized by the LLaMA (Dubey et al., 2024) tokenizer.

is asked to answer the question by choosing one from the option list. We consider a wide range of QA benchmarks to evaluate the capabilities of  $\mathcal{M}$  in different aspects, including reading comprehension, commonsense reasoning, world knowledge, and multitask understanding. The dataset statistics are presented in Table 1.

#### 4.1.1 Reading Comprehension

**BoolQ.** BoolQ (Clark et al., 2019) is a question answering dataset for yes/no questions. It evaluates the performance of  $\mathcal{M}$  on reading comprehension.

**LogiQA.** LogiQA (Liu et al., 2020) is a reading comprehension dataset that requires  $\mathcal{M}$  to have logical reasoning for question-answering.

#### 4.1.2 Commonsense Reasoning

**CSQA.** Commonsense QA (Talmor et al., 2019) examines  $\mathcal{M}$  on commonsense question-answering problems constructed using information from ConceptNet (Speer et al., 2017).

**SIQA.** SocialIQA (Sap et al., 2019) is a large-scale QA benchmark for commonsense reasoning about social situations, which probes emotional and social intelligence in everyday situations.

#### 4.1.3 World Knowledge

**SciQ.** SciQ (Welbl et al., 2017) provides scientific supports for  $\mathcal{M}$  to answer the multiple-choice science questions.

**OBQA.** OpenBookQA (Mihaylov et al., 2018) asks  $\mathcal{M}$  to answer the question based on the given elementary level science facts and broad commonsense knowledge.

<sup>2</sup>Please refer to Appendix A for more experiment details.



Method	Reading		Commonsense		World Knowledge			Multitask Understanding			Avg.
	BoolQ	LogiQA	CSQA	SIQA	SciQ	OBQA	ARC	BBH	MMLU	MMLU-Pro	
w/o RG	77.86	35.64	50.37	47.49	91.20	69.80	64.61	50.26	45.54	29.60	56.24
DA	84.16	35.79	72.97	69.55	85.90	72.20	82.59	52.19	60.68	38.75	65.48
CoT	84.65	38.10	73.71	68.12	93.70	78.20	84.31	58.40	62.08	40.10	68.14
<b>ARR</b>	<b>86.33</b>	<b>39.02</b>	<b>74.94</b>	<b>70.98</b>	<b>94.40</b>	<b>80.00</b>	<b>84.84</b>	<b>59.01</b>	<b>63.51</b>	<b>42.72</b>	<b>69.58</b>

Table 2: **Main experiments.** The zero-shot performance (Accuracy %) of the LLaMA3-8B-Chat model on various QA benchmarks using different answer trigger sentences  $\phi$ . (1) w/o RG: directly selecting an option without Reasoning Generation; (2) DA (Direct Answer):  $\phi = \text{“Answer:”}$ ; (3) CoT (Kojima et al., 2022):  $\phi = \text{“Answer: Let’s think step by step.”}$ ; (4) ARR: our method that elicits intent analysis, information retrieval, and logical reasoning.

**ARC.** AI2 Reasoning Challenge (Clark et al., 2018) contains grade-school science questions and is divided into a Challenge and an Easy set.

#### 4.1.4 Multitask Understanding

**BBH.** BIG-Bench Hard (Suzgun et al., 2023) is a suite challenging tasks filtered from BIG-Bench (Srivastava et al., 2023). Solving these problems often requires multi-step reasoning.

**MMLU.** MMLU (Hendrycks et al., 2021) comprehensively measures the multitask accuracy of  $\mathcal{M}$  on 57 tasks including elementary mathematics, history, computer science, and more.

**MMLU-Pro.** MMLU-Pro (Wang et al., 2024b) extends the mostly knowledge-driven MMLU benchmark by integrating more challenging, reasoning-focused questions and expanding the choice set from four to ten options.

## 4.2 Models

Our experiments adopt open-weights, decoder-only, and Transformer-based (Vaswani et al., 2017) LLMs. We mainly employ LLaMA3-8B-Chat (Dubey et al., 2024), an instruction-following LLM with 8 billion model parameters, and use the model implementation provided by Hugging Face Transformers (Wolf et al., 2020). In generalizability experiments, we also explore LLaMA3-Chat models of different sizes in § 6.1 and 7B-Chat models of different LLM series in § 6.2, i.e., Qwen2.5 (Yang et al., 2024), Gemma (Team et al., 2024b,c), and Mistral (Jiang et al., 2023).

## 4.3 Evaluation

To evaluate the QA performance of LLMs, we apply a two-step process including **reasoning generation** and **option selection**, as mentioned in § 3.2. First, we let the model freely generate text responses that may include their analysis, reasoning,

and answer choice. Then, we concatenate the input and output in the first stage with each choice from the given option list, pass each concatenation to the model, and select the option with the lowest cross-entropy loss. The loss corresponds to the perplexity of language models: A lower loss means a lower perplexity and a higher confidence. Length normalization is not applied because the options are mostly in the A/B/C/D, Yes/No, or True/False format. As the datasets in our experiments are all multiple-choice QA tasks, we adopt accuracy as the evaluation metric, which is calculated by Eq. 6.

## 5 Main Experiments

### 5.1 QA Performance

The main experiments test the zero-shot QA performance of LLaMA3-8B-Chat (Dubey et al., 2024) on diverse QA datasets. The only difference between Direct Answer (DA), zero-shot CoT (Kojima et al., 2022), and ARR is the answer trigger sentence  $\phi$  shown in Figure 2. The results in Table 2 demonstrate that our ARR method boosts the DA method by a large margin, with an improvement of +4.1 points on average. In addition, ARR consistently outperforms zero-shot CoT prompting across all QA datasets, highlighting its universal superiority in various task types including reading comprehension, commonsense reasoning, world knowledge, and multitask understanding<sup>3</sup>. Moreover, the “w/o RG” method, which directly selects options without relying on rationales ( $r_i$  in Eq. 2), performs significantly worse, emphasizing the benefits of our two-stage QA approach.

### 5.2 Ablation Study

To better understand the performance gains shown in Table 2, we conduct an ablation study to explore the efficacy of each component of the ARR

<sup>3</sup>Please refer to Appendix C for detailed case studies.

	A	R	R	Answer Trigger Sentence $\phi$
①	✓	✓	✓	Answer: Let's analyze the intent of the question, find relevant information, and answer the question with step-by-step reasoning.
②	✓			Answer: Let's analyze the intent of the question, and answer the question.
③		✓		Answer: Let's find relevant information, and answer the question.
④			✓	Answer: Let's answer the question with step-by-step reasoning.
⑤				Answer:

Table 3: **Ablation study prompts.** The answer trigger sentences  $\phi$  used in different ARR ablation study settings.

	Ablation			Reading		Commonsense		World Knowledge			Multitask Understanding			Avg.
	A	R	R	BoolQ	LogiQA	CSQA	SIQA	SciQ	OBQA	ARC	BBH	MMLU	MMLU-Pro	
①	✓	✓	✓	<b>86.33</b>	<b>39.02</b>	74.94	<b>70.98</b>	94.40	80.00	84.84	<b>59.01</b>	63.51	42.72	69.58
②	✓			86.09	38.40	<b>75.76</b>	70.78	94.30	<b>86.80</b>	<b>85.83</b>	57.08	63.66	42.54	<b>70.12</b>
③		✓		85.35	37.79	75.59	68.01	92.80	81.20	85.33	58.27	<b>63.73</b>	<b>43.08</b>	69.12
④			✓	85.87	38.86	74.53	68.01	<b>94.50</b>	82.60	85.03	58.96	61.77	41.11	69.12
⑤				84.16	35.79	72.97	69.55	85.90	72.20	82.59	52.19	60.68	38.75	65.48

Table 4: **Ablation study results.** The accuracy scores (%) of the LLaMA3-8B-Chat model on diverse QA datasets using different answer trigger sentences  $\phi$  ( Analyzing , Retrieving , and Reasoning ).

method, i.e., analyzing, retrieving, and reasoning. Specifically, we test the model’s QA performance using the five different answer trigger sentences  $\phi$  in Table 3. Table 4 reports the accuracy scores of LLaMA3-8B-Chat under different ablation cases, where ① is the full version of ARR and ⑤ is equivalent to the “DA” method in Table 2. In ②, ③, and ④,  $\phi$  only contains one single component, i.e., analyzing, retrieving, and reasoning, respectively.

We observe that all the single-component ARR settings (②, ③, and ④) outperform the DA method (⑤) by a large margin, which verifies that each ARR component contributes positively. Furthermore, the complete ARR method (①) has a higher accuracy score than the Retrieving-only (③) and Reasoning-only (④) methods, meaning the intent analysis benefits the other two “R” parts. Notably, the Intent Analysis component (②) brings the greatest improvement gain, suggesting the significance of analyzing the intent of the question.

Observing Table 4, each of ARR-full (①), Analyzing-only (②), Retrieving-only (③), and Reasoning-only (④) settings outperforms CoT and DA in Table 2. Moreover, Analyzing-only (70.12%) and Retrieving-only (69.12%) settings (i.e., excluding the “Reasoning” component) beat CoT (68.14%), showing the fundamental difference between our ARR method and existing prompting methods. The experimental results shed light on future exploration on incorporating ARR steps, especially intent analysis, into problem solving.

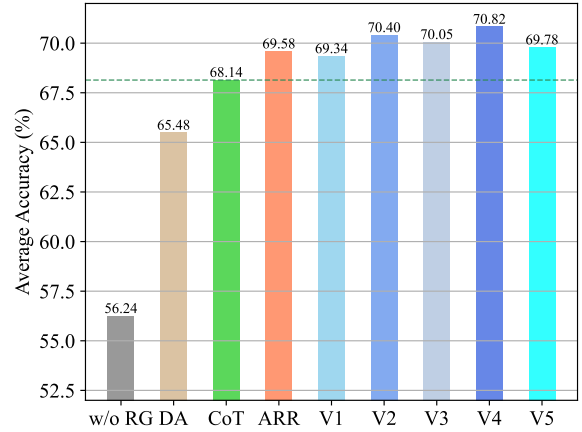


Figure 3: **Experiments on prompt variants.** The average performance (Accuracy %) of the LLaMA3-8B-Chat model on 10 QA datasets using different ARR prompt variants (“V1”–“V5”).

### 5.3 Prompt Variants

To demonstrate that ARR works effectively irrespective of specific prompt design, we conduct experiments on different ARR prompt variants. The original ARR prompt (as in Figure 2) is paraphrased into five different versions<sup>4</sup> by GPT-4o (Hurst et al., 2024). Figure 3 shows the average performance of each method over 10 QA datasets, and the full results are presented in Table 12. As illustrated, the proposed ARR method, regardless of its prompt implementation, remains a consistent advantage over the baselines, substantiating that ARR is a general framework for question answering.

<sup>4</sup>Please refer to Appendix B for details on variants.

Size	Method	BBH	MMLU	MMLU-Pro	Avg.	$\Delta$
1B	DA	35.88	<b>43.27</b>	21.62	33.59	0
	CoT	36.30	41.10	22.74	33.38	-0.21
	<b>ARR</b>	<b>39.02</b>	42.70	<b>23.49</b>	<b>35.07</b>	+1.48
3B	DA	45.65	48.26	30.88	41.60	0
	CoT	46.89	46.80	30.03	41.24	-0.36
	<b>ARR</b>	<b>51.97</b>	<b>52.82</b>	<b>33.39</b>	<b>46.06</b>	+4.46
8B	DA	52.19	60.68	38.75	50.54	0
	CoT	58.40	62.08	40.10	53.53	+2.99
	<b>ARR</b>	<b>59.01</b>	<b>63.51</b>	<b>42.72</b>	<b>55.08</b>	+4.54

Table 5: **Model size experiments.** The zero-shot performance (Accuracy %) of LLaMA3-Chat models of different sizes on multitask QA datasets.

## 6 Generalizability

The main experiments in § 5 have validated the effectiveness of our ARR method quantitatively and qualitatively. To verify the generalizability of ARR, we conduct additional extensive experiments under different configurations on three challenging, reasoning-intense, and multitask benchmarks introduced in § 4.1.4: BBH, MMLU, and MMLU-Pro.

### 6.1 Model Sizes

We evaluate the LLaMA3-Chat models of different sizes, i.e., 1B, 3B, and 8B (default) parameters, on multitask QA tasks. As the accuracy scores (%) shown in Table 5, our ARR method brings solid performance gains over the DA method and consistently outperforms zero-shot CoT. For the 1B model, ARR slightly underperforms the DA method on MMLU, likely due to the weaker instruction-following ability in smaller models. Still, our ARR method achieves overall performance improvements over the DA in the 1B setting. Observing the improvements over the DA method, larger models benefit more from ARR. The results conform to the scaling laws of language models (Kaplan et al., 2020), demonstrating the potential of ARR when applied to larger models.

### 6.2 LLM Series

To verify the effectiveness of our ARR method on open models other than LLaMA3 (Dubey et al., 2024), we conduct experiments on 7B-Chat LLMs of different series: Qwen2.5 (Yang et al., 2024), Gemma (Team et al., 2024b,c), and Mistral (Jiang et al., 2023). The results in Table 6 exhibit a consistent superiority of the proposed ARR method over the baseline methods. This is similar to the findings in the main experiments (Table 2), solidifying the efficacy and generalizability of ARR.

Series	Method	BBH	MMLU	MMLU-Pro	Avg.
Qwen	DA	39.21	48.36	32.35	39.97
	CoT	36.66	44.91	29.26	36.94
	<b>ARR</b>	<b>40.50</b>	<b>50.34</b>	<b>39.10</b>	<b>43.31</b>
Gemma	DA	40.09	45.46	23.45	36.33
	CoT	44.39	47.17	26.20	39.25
	<b>ARR</b>	<b>45.31</b>	<b>50.73</b>	<b>26.98</b>	<b>41.01</b>
Mistral	DA	46.27	55.61	30.68	44.19
	CoT	53.42	61.16	34.73	49.77
	<b>ARR</b>	<b>53.55</b>	<b>61.49</b>	<b>35.21</b>	<b>50.08</b>

Table 6: **LLM series experiments.** The zero-shot performance (Accuracy %) of 7B-Chat models of different LLM series on multitask QA datasets.

Method	OBQA	MMLU-sub
DA	98.00	88.55
CoT	97.20	88.55
<b>ARR</b>	<b>98.20</b>	<b>88.91</b>

Table 7: **Proprietary LLM experiments.** The zero-shot performance (Accuracy %) of GPT-4o on the OpenBookQA (OBQA) dataset and a subset of MMLU.

### 6.3 Proprietary LLMs

To further validate the efficacy of our ARR method on proprietary LLMs (Hurst et al., 2024; Anthropic, 2024; Team et al., 2024a) with a humongous number of parameters and state-of-the-art (SOTA) performances, we conduct experiments using GPT-4o (Hurst et al., 2024) on OBQA (involving World Knowledge) and MMLU (involving Multitask Understanding). As MMLU is a large benchmark with 50+ subtasks, we sample 10 instances from each subtask to form an “MMLU-sub” subset with 500+ items for GPT evaluation.

As presented in Table 7, our ARR framework continues to yield performance gains even when applied to SOTA models like GPT-4o. While the improvement is not as pronounced as for smaller LLMs—likely due to the already excellent base performance—ARR still provides measurable benefits. In contrast, the CoT method either diminishes GPT-4o’s performance or offers negligible gains. This suggests that GPT-4o may already engage in CoT-like reasoning intrinsically, yet lacks the structured three-step process introduced by ARR.

### 6.4 Generation Temperatures

For reproducibility, we set the generation temperature to 0 by default, as this setting makes the generation process deterministic. However, a higher temperature brings a more diverse output, which

Temp.	Method	BBH	MMLU	MMLU-Pro	Avg.
0.0	DA	52.19	60.68	38.75	50.54
	CoT	58.40	62.08	40.10	53.53
	<b>ARR</b>	<b>59.01</b>	<b>63.51</b>	<b>42.72</b>	<b>55.08</b>
0.5	DA	50.19	59.35	36.88	48.81
	CoT	56.58	60.82	37.82	51.74
	<b>ARR</b>	<b>58.87</b>	<b>62.87</b>	<b>42.64</b>	<b>54.79</b>
1.0	DA	46.33	54.80	33.10	44.74
	CoT	51.46	55.57	33.00	46.68
	<b>ARR</b>	<b>52.90</b>	<b>56.58</b>	<b>36.73</b>	<b>48.74</b>
1.5	DA	40.84	45.03	26.85	37.57
	CoT	42.53	44.85	25.61	37.66
	<b>ARR</b>	<b>42.65</b>	<b>45.16</b>	<b>27.44</b>	<b>38.42</b>

Table 8: **Generation temperature experiments.** The zero-shot performance (Accuracy %) of the LLaMA3-8B-Chat model on multitask QA datasets using different generation temperatures (default: 0.0).

may lead to a different QA accuracy. To study the effect of this key factor, we report the QA accuracy (%) of the LLaMA3-8B-Chat model using different temperatures during the reasoning generation stage: 0.0 (default), 0.5, 1.0, and 1.5.

As shown in Table 8, our ARR method surpasses the baseline methods with different temperatures, demonstrating a strong robustness of ARR. In addition, we observe that the model generally performs better when the temperature is lower.

## 6.5 Few-shot Generation

**Few-shot Examples with Rationales.** For each subtask in a QA dataset, we randomly pick 10 examples from the training or validation set if they exists. If a subtask only has the test set, 10 test examples are held out for few-shot usage, slightly reducing the number of items for evaluation. For each raw example, we construct the CoT and ARR rationales using GPT-4o (Hurst et al., 2024). Specifically, the input prompts provided to GPT-4o match those used in the evaluation experiments under CoT/ARR settings. The model’s output is extracted as CoT/ARR rationales. In few-shot examples, these rationales, along with correct answers, are appended to the answer trigger sentence  $\phi$ . For the Direct Answer (DA) setting, few-shot examples include correct answers for in-context learning (ICL) (Brown et al., 2020; Dong et al., 2024) but exclude rationales.

**Few-shot Results.** Table 9 presents the accuracy scores (%) of the LLaMA3-Chat model on multitask QA tasks. Using different numbers of few-shot examples (1, 3, and 5), our few-shot ARR

Shot	Method	BBH	MMLU	MMLU-Pro	Avg.
0	DA	52.19	60.68	38.75	50.54
	CoT	58.40	62.08	40.10	53.53
	<b>ARR</b>	<b>59.01</b>	<b>63.51</b>	<b>42.72</b>	<b>55.08</b>
1	DA	35.68	44.80	28.62	36.37
	CoT	<b>47.39</b>	48.36	31.07	42.27
	<b>ARR</b>	47.22	<b>49.29</b>	<b>34.33</b>	<b>43.61</b>
3	DA	34.39	42.08	25.92	34.13
	CoT	<b>42.84</b>	48.21	26.69	39.25
	<b>ARR</b>	40.19	<b>49.68</b>	<b>37.04</b>	<b>42.30</b>
5	DA	34.11	41.14	25.76	33.67
	CoT	39.92	47.48	26.12	37.84
	<b>ARR</b>	<b>40.68</b>	<b>49.19</b>	<b>36.62</b>	<b>42.16</b>

Table 9: **Few-shot experiments.** The few-shot performance (Accuracy %) of the LLaMA3-8B-Chat model on multitask QA datasets using 1, 3, and 5 few-shot examples with rationales.

method outperforms the DA (i.e., ICL) and few-shot CoT (Wei et al., 2022) methods on average.

Comparison across the three few-shot settings reveals that additional examples do not necessarily enhance performance. Moreover, QA performance is lower in the few-shot experiments than in the zero-shot setting, likely because the randomly selected examples mislead the reasoning process (Zhao et al., 2021; Lu et al., 2022; Peng et al., 2024). While demonstration selection methods could mitigate this issue (Gao et al., 2021; Rubin et al., 2022; Li et al., 2023a; Wang et al., 2023b), their exploration is beyond the scope of this study.

## 7 Conclusion

In this work, we introduce ARR, an intuitive, effective, and general QA framework that effectively enhances the question-answering performance of LLMs by integrating three key steps: analyzing the question’s intent, retrieving relevant information, and reasoning step by step. Extensive experiments across diverse QA tasks demonstrate that ARR consistently outperforms baseline methods including Direct Answer (DA) and Chain-of-Thought (CoT) prompting. Ablation and case studies further validate the positive contributions of each component, with intent analysis proving particularly crucial. Furthermore, experiments on ARR prompt variations indicate that ARR remains effective regardless of the specific prompt implementation. In addition, evaluations across various model sizes, LLM series, and generation configurations confirm the effectiveness, robustness, and generalizability of the proposed ARR method.



## Limitations

Resource constraints limited our focus to open-weights LLMs with no more than 8B parameters. However, the results from model size experiments (§ 6.1) align with the scaling laws for language models (Kaplan et al., 2020), demonstrating the potential and generalizability of our ARR method when applied to larger models. Moreover, § 6.3 demonstrate that ARR continues to yield performance gains even when applied to SOTA models like GPT-4o.

As a general framework, ARR can be implemented in different ways. In this work, we realize ARR as a test-time prompting method because this is the most natural approach to elicit the power of pre-trained foundation models like LLMs. Section 5.3 shows that ARR outperforms baselines regardless of the specific prompt design, showing the efficacy of this three-step solution. Beyond prompting, the effectiveness of ARR sheds light on other approaches that may enhance LLMs, e.g., incorporating the three-step recipe of ARR (especially intent analysis) for LLM post-training.

Lastly, as mentioned in § 3.1, we only consider 10 multiple-choice question answering (MCQA) tasks, where the model is asked to answer the question by selecting an option from a list of choices. However, the ARR outputs of the Reasoning Generation stage (§ 3.2) can be used for QA tasks requiring free-form generation. As studied in recent research (Yin et al., 2025, Table 3), ARR works excellently on multiple mathematical reasoning benchmarks. Such tasks are not MCQA, and there is no “Option Selection” stage.

## Acknowledgments

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). This research was supported in part by the computational resources and services provided by Advanced Research Computing at the University of British Columbia and the Digital Research Alliance of Canada (alliancecan.ca). We would also like to thank UBC NLP Group members for their constructive feedback.

## References

- Douglas Adams. 1979. *The Hitchhiker’s Guide to the Galaxy*. Pan Books.
- Frederick Adams. 1986. [Intention and intentional action: The simple view](#). *Mind & Language*, 1(4):281–301.
- Anthropic. 2024. [The claude 3 model family: Opus, sonnet, haiku](#).
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. [Constitutional ai: Harmlessness from ai feedback](#). *arXiv preprint arXiv:2212.08073*.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, volume 33, pages 1877–1901, Virtual Event. NeurIPS.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. [Extracting training data from large language models](#). In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. [A survey on evaluation of large language models](#). *ACM Transactions on Intelligent Systems and Technology*, 15(3).
- Danqi Chen. 2018. [Neural Reading Comprehension and Beyond](#). Stanford University.
- Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023. [Universal self-consistency for large language model generation](#). *arXiv preprint arXiv:2311.17311*.
- Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024. [Teaching large language models to self-debug](#). In *The Twelfth International Conference on Learning Representations*.

- Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. [Sft memorizes, rl generalizes: A comparative study of foundation model post-training](#). *arXiv preprint arXiv:2501.17161*.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. [BoolQ: Exploring the surprising difficulty of natural yes/no questions](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.
- Herbert H Clark. 1969. [Linguistic processes in deductive reasoning](#). *Psychological review*, 76(4):387.
- 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](#). *arXiv preprint arXiv:1803.05457*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. [Training verifiers to solve math word problems](#). *arXiv preprint arXiv:2110.14168*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. [A survey on in-context learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. [Making pre-trained language models better few-shot learners](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. [Retrieval-augmented generation for large language models: A survey](#). *arXiv preprint arXiv:2312.10997*.
- Louie Giray. 2023. [Prompt engineering with chatgpt: A guide for academic writers](#). *Annals of Biomedical Engineering*, 51(12):2629–2633.
- Xinyu Guan, Li Lina Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. [rstar-math: Small llms can master math reasoning with self-evolved deep thinking](#). *arXiv preprint arXiv:2501.04519*.
- Brett K Hayes and Evan Heit. 2018. [Inductive reasoning 2.0](#). *Wiley Interdisciplinary Reviews: Cognitive Science*, 9(3):e1459.
- Evan Heit. 2000. [Properties of inductive reasoning](#). *Psychonomic bulletin & review*, 7:569–592.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. [Large language models cannot self-correct reasoning yet](#). In *The Twelfth International Conference on Learning Representations*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. [Gpt-4o system card](#). *arXiv preprint arXiv:2410.21276*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. [Mistral 7b](#). *arXiv preprint arXiv:2310.06825*.
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. [How can we know what language models know?](#) *Transactions of the Association for Computational Linguistics*, 8:423–438.
- Philip N Johnson-Laird. 1999. [Deductive reasoning](#). *Annual review of psychology*, 50(1):109–135.
- Erik Jones and Jacob Steinhardt. 2022. [Capturing failures of large language models via human cognitive biases](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 11785–11799. Curran Associates, Inc.
- Daniel Jurafsky and James H. Martin. 2025. [Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics](#),

- and *Speech Recognition with Language Models*, 3rd edition. Prentice Hall PTR. Online manuscript released January 12, 2025.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). *arXiv preprint arXiv:2001.08361*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. 2024. [Tulu 3: Pushing frontiers in open language model post-training](#). *arXiv preprint arXiv:2411.15124*.
- Kuang-Huei Lee, Ian Fischer, Yueh-Hua Wu, Dave Marwood, Shumeet Baluja, Dale Schuurmans, and Xinyun Chen. 2025. [Evolving deeper llm thinking](#). *arXiv preprint arXiv:2501.09891*.
- Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023a. [Unified demonstration retriever for in-context learning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.
- Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023b. [Making language models better reasoners with step-aware verifier](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *The Twelfth International Conference on Learning Representations*.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. [Deepseek-v3 technical report](#). *arXiv preprint arXiv:2412.19437*.
- Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. [Logiqa: A challenge dataset for machine reading comprehension with logical reasoning](#). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3622–3628. International Joint Conferences on Artificial Intelligence Organization. Main track.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. [Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing](#). *ACM Computing Surveys*, 55(9):195:1–195:35.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. [Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-refine: Iterative refinement with self-feedback](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.
- Alfred R Mele. 1989. [Intention, belief, and intentional action](#). *American Philosophical Quarterly*, 26(1):19–30.
- Alfred R Mele and Paul K Moser. 1994. [Intentional action](#). *Nous*, 28(1):39–68.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. [Can a suit of armor conduct electricity? a new dataset for open book question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.
- Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. [Recent advances in natural language processing via large pre-trained language models: A survey](#). *ACM Computing Surveys*, 56(2):1–40.
- Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. [Large language models: A survey](#). *arXiv preprint arXiv:2402.06196*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong



- Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. [The fineweb datasets: Decanting the web for the finest text data at scale](#). *arXiv preprint arXiv:2406.17557*.
- Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. [Revisiting demonstration selection strategies in in-context learning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9090–9101, Bangkok, Thailand. Association for Computational Linguistics.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. [Reasoning with language model prompting: A survey](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. [Improving language understanding by generative pre-training](#). *OpenAI blog*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Joshua Robinson and David Wingate. 2023. [Leveraging large language models for multiple choice question answering](#). In *The Eleventh International Conference on Learning Representations*.
- Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. [Learning to retrieve prompts for in-context learning](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.
- Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. [A systematic survey of prompt engineering in large language models: Techniques and applications](#). *arXiv preprint arXiv:2402.07927*.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. [Social iqa: Commonsense reasoning about social interactions](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.
- Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2024. [Rewarding progress: Scaling automated process verifiers for llm reasoning](#). *arXiv preprint arXiv:2410.08146*.
- Claude Elwood Shannon. 1948. [A mathematical theory of communication](#). *The Bell System Technical Journal*, 27(3):379–423.
- Claude Elwood Shannon. 1951. [Prediction and entropy of printed english](#). *Bell System Technical Journal*, 30(1):50–64.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Yu Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *arXiv preprint arXiv:2402.03300*.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. [Large language models can be easily distracted by irrelevant context](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.
- Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. [Detecting pretraining data from large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxin Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hananeh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse



- Dodge, and Kyle Lo. 2024. [Dolma: an open corpus of three trillion tokens for language model pretraining research](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. [Conceptnet 5.5: An open multilingual graph of general knowledge](#). In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *Transactions on Machine Learning Research*.
- Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, et al. 2023. [A survey of reasoning with foundation models](#). *arXiv preprint arXiv:2312.11562*.
- Richard S. Sutton and Andrew G. Barto. 2018. [Reinforcement Learning: An Introduction](#). A Bradford Book, Cambridge, MA, USA.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. [Challenging big-bench tasks and whether chain-of-thought can solve them](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. [CommonsenseQA: A question answering challenge targeting commonsense knowledge](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024a. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#). *arXiv preprint arXiv:2403.05530*.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024b. [Gemma: Open models based on gemini research and technology](#). *arXiv preprint arXiv:2403.08295*.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024c. [Gemma 2: Improving open language models at a practical size](#). *arXiv preprint arXiv:2408.00118*.
- Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. 2024. [LLMs cannot find reasoning errors, but can correct them given the error location](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13894–13908, Bangkok, Thailand. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Huaijie Wang, Shibo Hao, Hanze Dong, Shenao Zhang, Yilin Bao, Ziran Yang, and Yi Wu. 2024a. [Offline reinforcement learning for llm multi-step reasoning](#). *arXiv preprint arXiv:2412.16145*.
- Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. [Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.
- Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023b. [Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Xuezhi Wang and Denny Zhou. 2024. [Chain-of-thought reasoning without prompting](#). *arXiv preprint arXiv:2402.10200*.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024b. [Mmlu-pro: A more robust and challenging multi-task language understanding benchmark](#). *arXiv preprint arXiv:2406.01574*.
- Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, et al. 2024. [Redpajama: an open dataset for](#)

- training large language models. *arXiv preprint arXiv:2411.12372*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- 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*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.
- Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. [A prompt pattern catalog to enhance prompt engineering with chatgpt](#). *arXiv preprint arXiv:2302.11382*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Fengli Xu, Qian Yue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. 2025. [Towards large reasoning models: A survey of reinforced reasoning with large language models](#). *arXiv preprint arXiv:2501.09686*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of thoughts: Deliberate problem solving with large language models](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2024. [Large language models as analogical reasoners](#). In *The Twelfth International Conference on Learning Representations*.
- Yuwei Yin, EunJeong Hwang, and Giuseppe Carenini. 2025. [Swi: Speaking with intent in large language models](#). *arXiv preprint arXiv:2503.21544*.
- Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. [Making retrieval-augmented language models robust to irrelevant context](#). In *The Twelfth International Conference on Learning Representations*.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. [A survey of large language models](#). *arXiv preprint arXiv:2303.18223*.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. [Calibrate before use: Improving few-shot performance of language models](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023a. [Least-to-most prompting enables complex reasoning in large language models](#). In *The Eleventh International Conference on Learning Representations*.
- Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023b. [Large language models are human-level prompt engineers](#). In *The Eleventh International Conference on Learning Representations*.

## A Experiment Details

### A.1 Dataset Details

All QA datasets used in this work are loaded from Hugging Face datasets<sup>5</sup>. Table 10 lists the URL link of each dataset.

QA Datasets	URL
BoolQ (Clark et al., 2019)	<a href="#">Link</a>
LogiQA (Liu et al., 2020)	<a href="#">Link</a>
CSQA (Talmor et al., 2019)	<a href="#">Link</a>
SIQA (Sap et al., 2019)	<a href="#">Link</a>
SciQ (Welbl et al., 2017)	<a href="#">Link</a>
OBQA (Mihaylov et al., 2018)	<a href="#">Link</a>
ARC (Clark et al., 2018)	<a href="#">Link</a>
BBH (Suzgun et al., 2023)	<a href="#">Link</a>
MMLU (Hendrycks et al., 2021)	<a href="#">Link</a>
MMLU-Pro (Wang et al., 2024b)	<a href="#">Link</a>

Table 10: The URL links of adopted QA datasets.

### A.2 Model Details

As mentioned in § 4.2, we mainly employ LLaMA3-8B-Chat (Dubey et al., 2024), an instruction-following LLM with 8 billion model parameters, for most experiments. In generalizability experiments (§ 6), we also explore LLaMA3-Chat models of different sizes in § 6.1 and 7B-Chat models of different LLM series in § 6.2, i.e., Qwen2.5 (Yang et al., 2024), Gemma (Team et al., 2024b,c), and Mistral (Jiang et al., 2023). Table 11 lists the URL link of each model and tokenizer provided by Hugging Face Transformers (Wolf et al., 2020).<sup>6</sup>

LLM Series	Size	Type	URL
LLaMA3 (Dubey et al., 2024)	8B	Chat	<a href="#">Link</a>
	3B	Chat	<a href="#">Link</a>
	1B	Chat	<a href="#">Link</a>
Qwen2.5 (Yang et al., 2024)	7B	Chat	<a href="#">Link</a>
Gemma (Team et al., 2024b,c)	7B	Chat	<a href="#">Link</a>
Mistral (Jiang et al., 2023)	7B	Chat	<a href="#">Link</a>

Table 11: The URL links of models and tokenizers.

### A.3 LLM Generation Details

For each experimental setting, the model needs to perform reasoning generation and option selection

<sup>5</sup>Data source: <https://huggingface.co/datasets>

<sup>6</sup>Model source: <https://huggingface.co/models>

sessions on every QA dataset. For each running session, all experiments are conducted on a single NVIDIA V100 GPU with 32GB memory except the few-shot experiments in § 6.5, which use a single NVIDIA A100 GPU with 40GB memory since the input length is much longer considering the few-shot examples with rationales. To avoid out-of-memory issue, all the models are loaded in a half-precision (float16) mode, and the generation batch size is 1. The input sequence is not truncated since we do not want to lose the context information or the answer trigger sentence  $\phi$ , but we set the maximum number of newly generated tokens as 512 during reasoning generation.

### A.4 Reproducibility

For the reproducibility of this work, we set the generation temperature as 0 by default and disable token sampling for deterministic generation. In addition, we pre-set the random seed for all random modules at the beginning of each experiment session. By an unofficial tradition<sup>7</sup>, we set 42 as the random seed and do not tune the value. To validate the reproducibility, we ran the main experiments twice and obtained the same results as shown in Table 2. Our source code is available on GitHub: <https://github.com/YuweiYin/ARR>

### A.5 Experimental Cost

In the reasoning generation stage, the total computational cost is approximately 8,000 GPU hours on NVIDIA V100 clusters (about 333 days) and 1,300 hours on A100 clusters (about 54 days). We only use V100 clusters for option selection, and the overall running time is approximately 560 hours (about 23 days). The expense for GPT-4o API calls in § 6.3 and § 6.5 is below US\$100.

## B ARR Prompt Variants

As mentioned in § 5.3, we conduct experiments on different ARR prompt variants to show that ARR works effectively irrespective of specific prompt design. The original ARR prompt is paraphrased into five different versions by GPT-4o (Hurst et al., 2024). The full experimental results on 10 QA datasets are presented in Table 12.

V1: “Answer: Let’s identify the question’s intent, gather the necessary information, and then work through a logical, step-by-step solution.”

<sup>7</sup> “The answer to the ultimate question of life, the universe, and everything is forty-two.” (Adams, 1979)

Method	Reading		Commonsense		World Knowledge			Multitask Understanding			Avg.
	BoolQ	LogiQA	CSQA	SIQA	SciQ	OBQA	ARC	BBH	MMLU	MMLU-Pro	
w/o RG	77.86	35.64	50.37	47.49	91.20	69.80	64.61	50.26	45.54	29.60	56.24
DA	84.16	35.79	72.97	69.55	85.90	72.20	82.59	52.19	60.68	38.75	65.48
CoT	84.65	38.10	73.71	68.12	93.70	78.20	84.31	58.40	62.08	40.10	68.14
<b>ARR</b>	<b>86.33</b>	39.02	74.94	<b>70.98</b>	<b>94.40</b>	80.00	84.84	59.01	63.51	42.72	69.58
V1	84.40	36.56	75.51	68.63	93.10	83.20	84.05	61.58	63.45	42.96	69.34
V2	85.14	37.63	76.82	70.68	93.90	82.80	85.90	60.66	<b>65.53</b>	<b>44.91</b>	70.40
V3	84.68	38.71	75.76	69.34	93.70	83.40	85.92	59.59	65.05	44.31	70.05
V4	84.40	<b>39.94</b>	<b>77.31</b>	68.78	93.90	<b>84.00</b>	<b>87.01</b>	<b>63.08</b>	65.42	44.38	<b>70.82</b>
V5	84.22	38.10	76.25	69.34	93.40	81.40	85.36	61.02	64.61	44.12	69.78
Vars AVG	84.57	38.19	76.33	69.35	93.60	82.96	85.65	61.19	64.81	44.14	70.08
Vars STD	$\pm 0.32$	$\pm 1.12$	$\pm 0.66$	$\pm 0.72$	$\pm 0.31$	$\pm 0.87$	$\pm 0.96$	$\pm 1.15$	$\pm 0.75$	$\pm 0.64$	$\pm 0.75$

Table 12: **Main experiments and ARR prompt variation experiments.** “ARR” is the original prompt design, and “V1”–“V5” are five paraphrased prompt variants. “Vars AVG” and “Vars STD” denote the average and standard deviation of the accuracy score on each QA dataset, respectively.

V2: “Answer: We’ll begin by examining what the question is asking, then retrieve any relevant details, and finally provide a well-reasoned answer step by step.”

V3: “Answer: First, we’ll interpret the purpose behind the question, collect supporting information, and proceed to solve it methodically.”

V4: “Answer: Let’s break this down by understanding the goal of the question, pulling in the required data, and then reasoning through the answer in a clear sequence.”

V5: “Answer: To approach this, we’ll clarify the question’s intent, locate pertinent information, and then build our answer using structured, logical reasoning.”

## C Case Study

As mentioned in § 5.1, this section presents case studies (Table 13–22) from our main experiments (§ 5, Table 2) to provide insights into how the ARR method enhances reasoning and improves QA accuracy. We present one case for each QA dataset introduced in § 4.1.



<b>Task Input</b> $x_i$	<p><b>Passage:</b> Feather – Contour feathers are <b>not uniformly distributed</b> on the skin of the bird <b>except in some groups</b> such as the <b>penguins</b>, ratites and screamers. In most birds the feathers grow from specific tracts of skin called pterylae; between the pterylae there are regions which are free of feathers called apterylae (or apteria). Filoplumes and down may arise from the apterylae. The arrangement of these feather tracts, pterylousis or pterylography, varies across bird families and has been used in the past as a means for determining the evolutionary relationships of bird families.</p> <p><b>Question:</b> Do penguins have feathers arising from the epidermis?</p>
<b>DA</b> $\phi_i$	<p>Answer:</p> <p><b>No</b>, penguins have feathers arising from specific tracts of skin called <b>pterylae</b>.</p> <p>The pterylae are the areas of skin where the feathers grow. The pterylae are not uniformly distributed on the skin of the penguin. The feathers grow from the pterylae, and the areas between the pterylae are free of feathers and are called apterylae.</p>
Reasoning $r_i$	<p>The arrangement of these feather tracts, pterylousis or pterylography, varies across bird families. <b>In penguins, the feathers are uniformly distributed on the skin, but they arise from specific tracts of skin called pterylae.</b></p> <p>The feathers do not arise from the epidermis. The epidermis is the outermost layer of the skin, and in most birds, the feathers grow from the dermal layer of the skin, not the epidermis.</p>
<b>CoT</b> $\phi_i$	<p>Answer: Let's think step by step.</p> <p>Penguins are a group of birds that have feathers uniformly distributed on their skin. The passage states that in most birds, feathers grow from specific tracts of skin called pterylae, and between these tracts, there are regions called apterylae that are free of feathers. However, <b>it also mentions that in some groups, such as penguins, ratites, and screamers, feathers are not uniformly distributed on the skin.</b> This suggests that penguins do not have feathers arising from the epidermis. Therefore, <b>the answer is no.</b> Penguins do not have feathers arising from the epidermis.</p>
<b>ARR</b> $\phi_i$	<p>Answer: Let's analyze the intent of the question, find relevant information, and answer the question with step-by-step reasoning.</p> <p>The question asks about the distribution of feathers on penguins. The passage states that in most birds, feathers grow from specific tracts of skin called pterylae, but <b>penguins are an exception.</b> The passage does not mention the epidermis, but it does mention that in penguins, feathers are uniformly distributed on the skin. This suggests that feathers in penguins may arise from the epidermis, as there are no specific tracts of skin called pterylae. Therefore, <b>the answer is yes.</b></p>
Reasoning $r_i$	

Table 13: **Case study.** An instance from the **Boolq** dataset, where our ARR method correctly answers the question but the baseline methods fail.

<b>Task Input</b> $x_i$	<p><b>Passage:</b> There are five stations set up from west to east along a section of the high-speed railway along the river. It is known that <b>Fuyi Station is east of Haoyun Station and West of Huyao Station, and is adjacent to Huyao Station; Jiushang Station is adjacent to Yinling Station.</b></p> <p><b>Question:</b> If <b>Haoyun stands to the east of Jiushang Station</b>, you can get</p> <p>(A) Yinling Station is adjacent to Haoyun Station and west of Haoyun Station;          (B) Haoyun Station is adjacent to Fuyi Station and west of Fuyi Station;          (C) Jiushang Station is adjacent to Haoyun Station and west of Haoyun Station;          (D) Yinling Station is adjacent to Fuyi Station and west of Fuyi Station;</p>
<b>DA</b> $\phi_i$	<p>Answer:</p> <p><b>(C)</b> Jiushang Station is adjacent to Haoyun Station and west of Haoyun Station;</p> <p>(Explanation) Since Jiushang Station is adjacent to Yinling Station, and Yinling Station is adjacent to Haoyun Station,</p>
Reasoning $r_i$	<p><b>Jiushang Station is adjacent to Haoyun Station.</b> Since <b>Jiushang Station is east of Yinling Station</b>, and Yinling Station is west of Haoyun Station, Jiushang Station is west of Haoyun Station.</p> <p>Therefore, <b>Jiushang Station is adjacent to Haoyun Station and west of Haoyun Station.</b></p>
<b>CoT</b> $\phi_i$	<p>Answer: Let's think step by step.</p> <p>If Haoyun stands to the east of Jiushang Station, then <b>Haoyun Station is west of Jiushang Station.</b></p> <p>Since Jiushang Station is adjacent to Yinling Station, <b>Haoyun Station is adjacent to Yinling Station.</b></p>
Reasoning $r_i$	<p>Since Fuyi Station is east of Haoyun Station, Fuyi Station is east of Yinling Station.</p> <p>Since Huyao Station is west of Fuyi Station, Huyao Station is west of Yinling Station.</p> <p>Since Jiushang Station is adjacent to Yinling Station, Huyao Station is adjacent to Jiushang Station ...</p>
<b>ARR</b> $\phi_i$	<p>Answer: Let's analyze the intent of the question, find relevant information, and answer the question with step-by-step reasoning.</p> <p>The question asks us to determine the correct relationship between the stations based on the given information.</p> <p>We are told that Fuyi Station is east of Haoyun Station and west of Huyao Station, and is adjacent to Huyao Station. This means that <b>Fuyi Station is between Haoyun Station and Huyao Station.</b></p> <p>Additionally, we know that Jiushang Station is adjacent to Yinling Station.</p>
Reasoning $r_i$	<p>The question also states that Haoyun stands to the east of Jiushang Station.</p> <p>This implies that <b>Haoyun Station is to the east of Jiushang Station.</b></p> <p>Option (B) states that Haoyun Station is adjacent to Fuyi Station and west of Fuyi Station.</p> <p>This is <b>consistent with the given information</b> that Fuyi Station is east of Haoyun Station.</p> <p>Therefore, the correct <b>answer is option (B)</b></p>

Table 14: **Case study.** An instance from the **LogiQA** dataset, where our ARR method correctly answers the question but the baseline methods fail. **Analysis:** the correct station order can be Jiushang-Yinling-Haoyun-Fuyi-Huyao or Yinling-Jiushang-Haoyun-Fuyi-Huyao.

<b>Task Input</b> $x_i$	<b>Question:</b> What do people aim to do at work? <b>(A)</b> complete job <b>(B)</b> learn from each other <b>(C)</b> kill animals <b>(D)</b> wear hats <b>(E)</b> talk to each other
<b>DA</b> $\phi_i$	Answer: <b>(B)</b> learn from each other. The best answer is B. Step 1: Understand the context of the question: The question is asking about the purpose or goal of people at work. Step 2: Analyze the options given Option (A) "complete job" is a task-oriented answer, but <b>it doesn't necessarily imply learning or personal growth.</b> Option (C) "kill animals" is unrelated to the typical goals of work in most industries.
Reasoning $r_i$	Option (D) "wear hats" is a superficial aspect of work attire and not a goal. Option (E) "talk to each other" is a common aspect of work but doesn't necessarily imply learning. Step 3: Identify the most appropriate answer: Option (B) <b>"learn from each other" is a goal that aligns with personal and professional development,</b> which is a common aim in many workplaces. It suggests a focus on growth, improvement, and collaboration, which are key aspects of many work environments. <b>The final answer is: B</b>
<b>CoT</b> $\phi_i$	Answer: Let's think step by step. <b>(B)</b> People go to work to earn a living, but they also go to learn from each other and to be part of a community. Choice (A) is incorrect because <b>people do not go to work to complete a job. They go to work to do a job.</b> Choice (C) is incorrect because people do not go to work to kill animals.
Reasoning $r_i$	Choice (D) is incorrect because people do not go to work to wear hats. Choice (E) is incorrect because people do not go to work to talk to each other. They go to work to do a job. <b>The best answer is B.</b>
<b>ARR</b> $\phi_i$	Answer: Let's analyze the intent of the question, find relevant information, and answer the question with step-by-step reasoning. The question asks what people aim to do at work. The relevant information is that people go to work to earn a living. The answer is that people aim to complete their jobs at work. Choice (B) is incorrect because <b>people may learn from each other at work, but that is not the main purpose of going to work.</b>
Reasoning $r_i$	Choice (C) is incorrect because people do not go to work to kill animals. Choice (D) is incorrect because people do not go to work to wear hats. Choice (E) is incorrect because people do not go to work to talk to each other. <b>The best answer is A.</b>

Table 15: **Case study.** An instance from the **Commonsense QA** dataset, where our ARR method correctly answers the question but the baseline methods fail.

<b>Task Input</b> $x_i$	<b>Context:</b> Alex grew closer to their <b>significant other</b> after they vacationed together. <b>Question:</b> How would Alex feel afterwards? <b>(A)</b> in love <b>(B)</b> happy <b>(C)</b> back from vacation
<b>DA</b> $\phi_i$	Answer: <b>(B) happy</b> Reasoning Skill: Empathy And Perspective Taking This question requires the test-taker to put themselves in Alex's shoes and understand how they would feel after a positive experience like vacationing with their significant other.
Reasoning $r_i$	The correct answer, "happy", is a natural emotional response to a pleasant experience. The other options are not directly related to the emotional state of Alex after the vacation. "In love" is a more intense and long-term feeling, "back from vacation" is a factual statement rather than an emotional response, and "happy" is the most appropriate answer given the context.
<b>CoT</b> $\phi_i$	Answer: Let's think step by step. Alex grew closer to their significant other after they vacationed together. This implies that the vacation strengthened their relationship. Therefore, Alex would likely feel happy after the vacation.
Reasoning $r_i$	Reasoning skill: Empathy And Perspective Taking. This question requires the reader to put themselves in Alex's shoes and understand how they would feel after a positive experience. It also requires the reader to consider the context of the situation and how it would impact Alex's emotions. <b>The best answer is (B)</b>
<b>ARR</b> $\phi_i$	Answer: Let's analyze the intent of the question, find relevant information, and answer the question with step-by-step reasoning. Step 1: Identify the context of the question. The context is about Alex growing closer to their significant other after a vacation. Step 2: Determine the emotional state of Alex after the vacation. The question asks how Alex would feel afterwards. Step 3: Analyze the relationship between Alex and their significant other. The question implies that the vacation strengthened their bond.
Reasoning $r_i$	Step 4: Determine the emotional state of Alex based on the context. Given that the vacation brought them closer together, it is likely that <b>Alex would feel happy or in love.</b> Step 5: Choose <b>the most appropriate answer.</b> Between "happy" and "in love," <b>"in love" is a more intense emotion</b> that aligns with the context of <b>growing closer to someone.</b> <b>The final answer is: (A) in love.</b>

Table 16: **Case study.** An instance from the **SocialIQA** dataset, where our ARR method correctly answers the question but the baseline methods fail. **Analysis:** Option B is not wrong, but Option A is more appropriate as they are in a romantic relationship ("significant other").

<b>Task Input</b> $x_i$	<b>Question:</b> Rupture can cause fragments of what to travel via the bloodstream and become lodged in other arteries? (A) red blood cells (B) white blood cells (C) plaque (D) enamel
<b>DA</b> $\phi_i$	Answer: Rupture of an atherosclerotic plaque can cause fragments of the plaque to travel via the bloodstream and become lodged in other arteries. This is known as an embolus.
Reasoning $r_i$	Choice (B) is incorrect because white blood cells are not typically associated with atherosclerosis. Choice (C) is incorrect because plaque is the substance that ruptures. Choice (D) is incorrect because enamel is the hard, outer layer of the teeth. The best answer is (A).
<b>CoT</b> $\phi_i$	Answer: Let's think step by step. The question is asking about the consequences of a rupture, which implies a physical injury or damage to a structure. The correct answer should be related to the physical structure that can be damaged and cause fragments to travel through the bloodstream. Option (A) red blood cells are a type of cell that carries oxygen throughout the body via the bloodstream, and they can be damaged or ruptured in the event of a physical injury or damage to a blood vessel. Option (B) white blood cells are part of the immune system and help fight infections, but they are not a physical structure that can rupture and cause fragments to travel through the bloodstream. Option (C) plaque is a sticky film of bacteria that can cause dental problems, but it's not a physical structure that can rupture and cause fragments to travel through the bloodstream. Option (D) enamel is a hard, outer layer of teeth, which is not directly related to the bloodstream or arteries. Therefore, the correct answer is (A) red blood cells.
<b>ARR</b> $\phi_i$	Answer: Let's analyze the intent of the question, find relevant information, and answer the question with step-by-step reasoning. The question is asking about the consequences of a rupture, which is likely referring to a rupture of an aneurysm or an atherosclerotic plaque. In the context of aneurysms or atherosclerosis, a rupture can lead to the release of fragments into the bloodstream. These fragments can then travel to other parts of the body and become lodged in other arteries, potentially causing blockages or other complications. Considering the options provided, the most relevant choice is (C) plaque, as atherosclerotic plaque is a common cause of arterial blockages and can rupture, releasing fragments into the bloodstream. The other options are not directly related to the consequences of a rupture in the context of aneurysms or atherosclerosis.

Table 17: **Case study.** An instance from the SciQ dataset, where our ARR method correctly answers the question but the baseline methods fail.

<b>Task Input</b> $x_i$	<b>Fact:</b> force causes the speed of an object to decrease. <b>Question:</b> The amount of friction and the speed of an object have what kind of relationship? (A) inverse (B) reverse (C) direct (D) equal
<b>DA</b> $\phi_i$	Answer: (C) Direct
Reasoning $r_i$	Explanation: The more friction an object has, the slower it will move. The less friction an object has, the faster it will move. This is a direct relationship. The amount of friction and the speed of an object are directly related.
<b>CoT</b> $\phi_i$	Answer: Let's think step by step. Step 1: Understand the relationship between friction and speed The relationship between friction and speed is that as friction increases, the speed of an object decreases. Step 2: Determine the type of relationship Since an increase in friction leads to a decrease in speed, this is a direct relationship. A direct relationship means that as one variable increases, the other variable also increases. Step 3: Choose the correct answer Based on the understanding of the relationship between friction and speed, the correct answer is (C) direct.
<b>ARR</b> $\phi_i$	Answer: Let's analyze the intent of the question, find relevant information, and answer the question with step-by-step reasoning. The question asks about the relationship between the amount of friction and the speed of an object. The relevant information is that force causes the speed of an object to decrease. Reasoning $r_i$ The intent of the question is to determine the relationship between the amount of friction and the speed of an object. The correct answer is (A) inverse. The amount of friction and the speed of an object have an inverse relationship. The more friction, the slower the object moves. The less friction, the faster the object moves.

Table 18: **Case study.** An instance from the OpenBookQA dataset, where our ARR method correctly answers the question but the baseline methods fail. **Analysis:** The CoT method provides a correct fact about the direct relationship (“A direct relationship means that . . .”), but wrongly infers that “this is a direct relationship” from “an increase in friction leads to a decrease in speed.” This is because the wrong statement is made without conditioning on the correct fact that is presented after the statement (due to the autoregressive generation). In contrast, the ARR method performs reasoning based on sufficient context after analyzing the intent and finding relevant information.

Task Input $x_i$	<b>Question:</b> A scientist maps a long region in which earthquakes originate and determines this region is a transform plate boundary. Which evidence would cause the scientist to reevaluate this determination? <b>(A)</b> Volcanism also characterizes the region. <b>(B)</b> Earthquake centers in the region occur at shallow depths. <b>(C)</b> The region shows extensive faulting of sediments. <b>(D)</b> Equal crust densities are found on opposite sides of the region.	
	DA $\phi_i$	Answer: The best answer is C. Transform plate boundaries are characterized by horizontal motion and faulting, but the faulting is not necessarily extensive.
Reasoning $r_i$		Volcanism is characteristic of divergent and convergent plate boundaries, not transform plate boundaries. Shallow earthquake foci are characteristic of transform plate boundaries. Equal crust densities on opposite sides of the region would indicate that the region is a transform plate boundary. Therefore, (A), (B), and (D) are incorrect. (C) is correct because extensive faulting is characteristic of transform plate boundaries.
Reasoning $r_i$	CoT $\phi_i$	Answer: Let's think step by step. Transform plate boundaries are characterized by horizontal motion of the Earth's crust, which is evidenced by faulting. Volcanism is characteristic of divergent and convergent plate boundaries, not transform plate boundaries. Therefore, (A) is incorrect. Earthquakes at transform plate boundaries can occur at any depth, so (B) is incorrect. (C) is correct because faulting is a characteristic of transform plate boundaries. (D) is incorrect because transform plate boundaries are characterized by horizontal motion of the Earth's crust, not by equal crust densities.
	ARR $\phi_i$	Answer: Let's analyze the intent of the question, find relevant information, and answer the question with step-by-step reasoning. The question asks for evidence that would cause a scientist to reevaluate the determination that a region is a transform plate boundary. A transform plate boundary is a region where two plates slide past one another horizontally. The evidence that would cause a scientist to reevaluate this determination would be evidence that is characteristic of a different type of plate boundary. A divergent plate boundary is a region where two plates move apart from one another, and a convergent plate boundary is a region where two plates move toward one another. Volcanism is characteristic of a divergent plate boundary, so (A) is correct.

Table 19: **Case study.** An instance from the **ARC** dataset, where our ARR method correctly answers the question but the baseline methods fail.

Task Input $x_i$	<b>Question:</b> The following paragraphs each describe a set of three objects arranged in a fixed order. The statements are logically consistent within each paragraph. A fruit stand sells three fruits: peaches, mangoes, and apples. The peaches are more expensive than the apples. The mangoes are the cheapest. <b>(A)</b> The peaches are the second-most expensive <b>(B)</b> The mangoes are the second-most expensive <b>(C)</b> The apples are the second-most expensive	
	DA $\phi_i$	Answer: The peaches are the second-most expensive. The mangoes are the cheapest. The apples are the least expensive of the three fruits.
Reasoning $r_i$		The peaches are more expensive than the apples. The mangoes are the cheapest. The peaches are the second-most expensive. The mangoes are the cheapest. The apples are the least expensive of the three fruits. The peaches are more expensive than the apples. The mangoes are the cheapest . . .
Reasoning $r_i$	CoT $\phi_i$	Answer: Let's think step by step. The peaches are more expensive than the apples. The mangoes are the cheapest. Therefore, the mangoes are the least expensive of the three fruits. The peaches are more expensive than the apples, so the peaches are the second-most expensive of the three fruits. Choice (B) is incorrect. The mangoes are the cheapest of the three fruits. Choice (C) is incorrect. The apples are less expensive than the peaches. The best answer is A.
	ARR $\phi_i$	Answer: Let's analyze the intent of the question, find relevant information, and answer the question with step-by-step reasoning. The question asks for the second-most expensive fruit. The peaches are more expensive than the apples, and the mangoes are the cheapest. Therefore, the second-most expensive fruit is the apples. Choice A is incorrect. The peaches are more expensive than the apples, and the mangoes are the cheapest. Therefore, the peaches are the most expensive fruit, not the second-most expensive. Choice B is incorrect. The mangoes are the cheapest fruit.

Table 20: **Case study.** An instance from the **BBH** dataset, where our ARR method correctly answers the question but the baseline methods fail. **Analysis:** The DA (Direct Answer) method rushes to conclusions and then tries to present an explanation. However, the false statement confuses the whole reasoning process and makes the model caught in an infinite self-doubting loop.



<b>Task Input</b> $x_i$		<b>Question:</b> Select the best translation into predicate logic. David teaches Chris. (c: Chris; d: David; Txy: x teaches y) (A) Tdc (B) Tcd (C) Tcc (D) dTc
<b>DA</b> $\phi_i$	Answer:	(B) The translation of the sentence into predicate logic is Tcd, where Txy means x teaches y. The subject of the sentence is David, the object is Chris, and the predicate is teaches.
Reasoning $r_i$		The correct translation is Tcd, not Tdc, because the subject comes before the predicate in a predicate logic statement. Choice (C) is incorrect because it reverses the subject and object. Choice (D) is incorrect because it reverses the subject and predicate.
<b>CoT</b> $\phi_i$	Answer:	Let's think step by step.
Reasoning $r_i$		(B) The predicate logic translation of the sentence is Tcd, where Txy means x teaches y. The best answer is B.
<b>ARR</b> $\phi_i$	Answer:	Let's analyze the intent of the question, find relevant information, and answer the question with step-by-step reasoning. The question asks for the best translation into predicate logic. The relevant information is the statement "David teaches Chris." The statement is in the active voice, so the subject is the teacher, and the object is the student. The predicate logic translation is Tdc, where Txy means x teaches y, c is Chris, and d is David.
Reasoning $r_i$		Choice (B) is incorrect. This is the translation of "Chris teaches David." Choice (C) is incorrect. This is the translation of "Chris teaches Chris." Choice (D) is incorrect. This is the translation of "David is taught by Chris." The best answer is A.

Table 21: **Case study.** An instance from the MMLU dataset, where our ARR method correctly answers the question but the baseline methods fail. **Analysis:** The DA (Direct Answer) and CoT methods jump to the wrong conclusions confidently, while the ARR method understands the question first and analyzes each option correctly.

<b>Task Input</b> $x_i$		<b>Question:</b> Which type of research methods are designed to elicit responses to predetermined, standardized questions from many respondents? (A) Non-probability. (B) Cross-sectional. (C) Qualitative. (D) Ethnographic. (E) Longitudinal. (F) Experimental. (G) Probability. (H) Observational. (I) Case Study. (J) Quantitative.
<b>DA</b> $\phi_i$	Answer:	The best answer is G. Probability sampling is a type of research method that is designed to elicit responses to predetermined, standardized questions from many respondents. Probability sampling is a type of sampling method that is used to select a sample of participants from a larger population. Probability sampling methods include simple random sampling, stratified sampling, and cluster sampling.
Reasoning $r_i$		
<b>CoT</b> $\phi_i$	Answer:	Let's think step by step.
Reasoning $r_i$		(G) Probability sampling is a method of selecting a sample from a population in which every member of the population has an equal chance of being selected. Probability sampling is used in research to ensure that the sample is representative of the population.
<b>ARR</b> $\phi_i$	Answer:	Let's analyze the intent of the question, find relevant information, and answer the question with step-by-step reasoning. The question asks about research methods that are designed to elicit responses to predetermined, standardized questions from many respondents. This is a clue that the research method is quantitative. Quantitative research methods are designed to elicit responses to predetermined, standardized questions from many respondents.
Reasoning $r_i$		The correct answer is (J). The other choices are incorrect because they are not quantitative research methods.

Table 22: **Case study.** An instance from the MMLU-Pro dataset, where our ARR method correctly answers the question but the baseline methods fail. **Analysis:** The ARR method grasps the question's intent and implications before answering.