

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

Yuwei Yin and Giuseppe Carenini

yuweiyin@cs.ubc.ca carenini@cs.ubc.ca

Department of Computer Science, Faculty of Science
University of British Columbia, on unceded Musqueam land
Vancouver, BC V6T 1Z4, Canada



THE UNIVERSITY
OF BRITISH COLUMBIA

1 Introduction

2 Our Method

3 Experiments

4 Conclusion

5 More Cases

6 References

1 Introduction

2 Our Method

3 Experiments

4 Conclusion

5 More Cases

6 References

Background

- **Large language models (LLMs)** [1, 2, 3] have been a transformative technique in Natural Language Processing (NLP) owing to their excellent text generation and conversation abilities [4, 5, 6].

Background

- **Challenging benchmarks** for language model evaluation have significantly driven LLM advancements [7], with most designed as **multiple-choice question-answering (MCQA)** tasks [8] requiring answer selection from given options [9, 10, 11].

Background

- Recent LLM benchmarks demand extensive **commonsense, world knowledge, and complex reasoning** [12, 13, 14], posing significant challenges for LLMs.
⇒ Optimizing LLM performance in QA tasks is increasingly crucial for their continued development.

Question Answering

- An example in OpenBookQA [15]:

	Fact: the summer solstice is on June 21st in the northern hemisphere.
Task Input x_i	Question: The summer solstice in the northern hemisphere is four months before (A) May (B) July (C) April (D) October

Table 1: An instance from the OpenBookQA dataset [15].

QA with LLMs (Baseline: Direct Answer)

- **Baseline** method: directly answer the question.

Task Input x_i	Fact: the summer solstice is on June 21st in the northern hemisphere. Question: The summer solstice in the northern hemisphere is four months before (A) May (B) July (C) April (D) October
Baseline ϕ_i	Answer: (B) July. Explanation: The summer solstice is the longest day of the year and marks the beginning of summer.
Reasoning r_i	It is four months before the beginning of winter. The beginning of winter is in December, so the summer solstice is in June. Therefore, the summer solstice is four months before July.

Table 2: The output by the Baseline method.

QA with LLMs (Baseline: Direct Answer)

- **Baseline** method: directly answer the question.
- **Problem**: The answer generation is only based on the task input, without reasoning or explanation.

	Fact: the summer solstice is on June 21st in the northern hemisphere.
Task Input x_t	Question: The summer solstice in the northern hemisphere is four months before (A) May (B) July (C) April (D) October
Baseline ϕ_t	Answer: (B) July. Explanation: The summer solstice is the longest day of the year and marks the beginning of summer.
Reasoning r_t	It is four months before the beginning of winter. The beginning of winter is in December, so the summer solstice is in June. Therefore, the summer solstice is four months before July.

Table 2: The output by the Baseline method.

QA with LLMs (CoT - Reasoning First)

- China-of-Thought (CoT) method: “Let’s think step by step”.
Generating step-by-step reasoning before giving the final answer.

	Fact: the summer solstice is on June 21st in the northern hemisphere.
Task Input x_i	Question: The summer solstice in the northern hemisphere is four months before (A) May (B) July (C) April (D) October
CoT ϕ_i	Answer: Let’s think step by step. The summer solstice is on June 21st. If we count four months before that, we get March 21st.
Reasoning r_i	March is the third month of the year, so the answer is (C). April is the fourth month, May is the fifth month, and October is the tenth month. None of these months is four months before June . July is the seventh month, which is five months after June, so it is not the correct answer either. The correct answer is (C).

Table 3: The output by the CoT method.

QA with LLMs (CoT - Reasoning First)

- China-of-Thought (CoT) method: “Let’s think step by step”.
Generating step-by-step reasoning before giving the final answer.
- **Advantage:** The final answer is based on the intermediate steps.
⇒ More likely to answer correctly (if the LLM reasons well)

	Fact: the summer solstice is on June 21st in the northern hemisphere.
Task Input x_i	Question: The summer solstice in the northern hemisphere is four months before (A) May (B) July (C) April (D) October
CoT ϕ_i	Answer: Let’s think step by step. The summer solstice is on June 21st. If we count four months before that, we get March 21st.
Reasoning r_i	March is the third month of the year, so the answer is (C). April is the fourth month, May is the fifth month, and October is the tenth month. None of these months is four months before June . July is the seventh month, which is five months after June, so it is not the correct answer either. The correct answer is (C).

Table 3: The output by the CoT method.

Step-by-step Thinking in QA

- **Argument:** “Let’s think step by step” is too generic and vague.

Step-by-step Thinking in QA

- **Argument:** “Let’s think step by step” is too generic and vague.
- For QA tasks, how do we humans answer a question?

Step-by-step Thinking in QA

- **Argument:** “Let’s think step by step” is too generic and vague.
- For QA tasks, how do we humans answer a question?
- “Step by Step” → What **steps** are we thinking?

Step-by-step Thinking in QA

- **Argument:** “Let’s think step by step” is too generic and vague.
- For QA tasks, how do we humans answer a question?
- “Step by Step” → What **steps** are we thinking?
- Can we guide LLMs to answer questions with human heuristics?

1 Introduction

2 Our Method

3 Experiments

4 Conclusion

5 More Cases

6 References

ARR: Analyzing, Retrieving, and Reasoning

- For QA tasks, how do we humans answer a question?

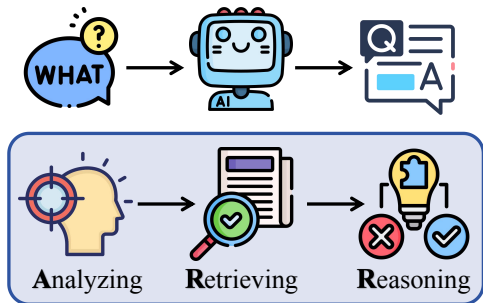
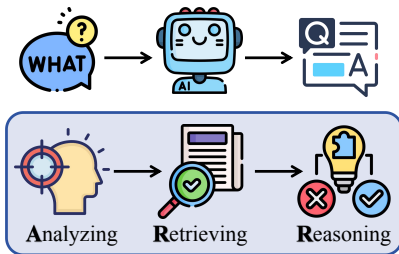


Figure 1: ARR motivation. To answer a question, we often need to **analyze** the question's intent, **retrieve** relevant information, and **reason** step by step.

ARR: Analyzing, Retrieving, and Reasoning

- “Step by Step” → What **steps** are we thinking?
 - (1) Analyze the question’s **intent** [16, 17, 18] to obtain a thorough **context understanding**, a clear **problem-solving target**, and a purposeful **planning guide**,
 - (2) Retrieve **relevant information** from context, external sources, or memory for supportive reference [19, 20], and
 - (3) Systematically apply **inductive and deductive reasoning** [21, 22, 23, 24].



ARR: Analyzing, Retrieving, and Reasoning

- Can we guide LLMs to answer questions with human heuristics?
- \Rightarrow Prompting via **Answer Trigger Sentence** Φ_i :
 - Answer: Let's analyze the intent of the question ,
find relevant information , and
answer the question with step-by-step reasoning .

QA with LLMs Workflow

- Baseline vs. CoT vs. ARR: Different Answer Trigger Sentences

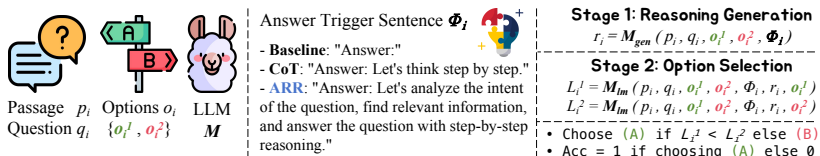


Figure 2: Question answering with LLMs. We first obtain rationale r_i by reasoning generation and then select the optimal option via evaluating the language modeling losses of different context-option combinations.

QA with LLMs (CoT - Reasoning First)

- **ARR**: Analyzing, Retrieving, and Reasoning
- Our ARR method identifies the question's intent clearly, leading to a correct reasoning path and final answer.

Task Input x_i	<p>Fact: the summer solstice is on June 21st in the northern hemisphere.</p> <p>Question: The summer solstice in the northern hemisphere is four months before (A) May (B) July (C) April (D) October</p>
ARR ϕ_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 for the month that is four months after the summer solstice.</p>
Reasoning r_i	<p>The summer solstice is on June 21st. Four months after June 21st is October 21st. Therefore, the correct answer is (D) October.</p>

Table 4: The output by our ARR method.

1 Introduction

2 Our Method

3 Experiments

4 Conclusion

5 More Cases

6 References

Experimental Setup - Tasks and Datasets

Task	MCQA Dataset	Split	# Item	# Token	# Class
Reading Comprehension	BoolQ [25]	Valid	3,270	145	2
	LogiQA [10]	Test	651	192	4
Commonsense Reasoning	CSQA [26]	Valid	1,221	43	5
	SIQA [27]	Valid	1,954	51	3
World Knowledge	SciQ [28]	Test	1,000	132	4
	OBQA [15]	Test	500	55	4
	ARC [9]	Test	3,548	59	4
Multitask Understanding	BBH [13]	Test	5,281	112	2–18
	MMLU [11]	Test	13,842	108	4
	MMLU-Pro [14]	Test	12,032	186	10

Table 5: QA dataset statistics. “# Class” is the number of options m , “# Item” is the total number of data items for evaluation, and “# Token” is the average number of tokens per instance (zero-shot prompt), tokenized by the LLaMA [29] tokenizer.

Experimental Setup - Models

- Main Experiments:
 - LLaMA-3.1-8B-Instruct
- Generalizability Exp (Model Sizes):
 - LLaMA-3.2-1B-Instruct, LLaMA-3.2-3B-Instruct
- Generalizability Exp (Model Series):
 - Qwen2.5-7B-Instruct, Gemma-7b-it, Mistral-7B-Instruct-v0.3

LLM Series	Size	Type	URL
LLaMA3 [29]	8B	Chat	Link
	3B	Chat	Link
	1B	Chat	Link
Qwen2.5 [30]	7B	Chat	Link
Gemma [31, 32]	7B	Chat	Link
Mistral [33]	7B	Chat	Link

Table 6: The URL links of models and tokenizers.

Main Experiments

Method	Reading		Commonsense		World Knowledge			Multitask Understanding			Avg.
	BoolQ	LogiQA	CSQA	SIQA	SciQ	OBQA	ARC	BBH	MMLU	MMLU-Pro	
w/o Reason	77.86	35.64	50.37	47.49	91.20	69.80	64.61	50.26	45.54	29.60	56.24
Baseline	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
ARR	86.33	39.02	74.94	70.98	94.40	80.00	84.84	59.01	63.51	42.72	69.58

Table 7: Main experiments. The zero-shot performance (Accuracy %) of the LLaMA3-8B-Chat model on various multiple-choice QA datasets using different answer trigger sentences ϕ . (1) w/o Reason: directly selecting an option without rationales; (2) Baseline: ϕ is “Answer:”; (3) CoT [34]: ϕ is “Answer: Let’s think step by step.”; (4) ARR: our method that elicits intent analysis, information retrieval, and step-by-step reasoning.

Ablation Study

	A	R	R	Answer Trigger Sentence ϕ
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 8: Ablation study prompts. The answer trigger sentences ϕ used in different ARR ablation study settings.

Ablation Study

	Ablation			Reading		Commonsense		World Knowledge			Multitask Understanding			Avg.
	A	R	R	BoolQ	LogiQA	CSQA	SIQA	SciQ	OBQA	ARC	BBH	MMLU	MMLU-Pro	
①	✓	✓	✓	86.33	39.02	74.94	70.98	94.40	80.00	84.84	59.01	63.51	42.72	69.58
②	✓			86.09	38.40	75.76	70.78	94.30	86.80	85.83	57.08	63.66	42.54	70.12
③		✓		85.35	37.79	75.59	68.01	92.80	81.20	85.33	58.27	63.73	43.08	69.12
④			✓	85.87	38.86	74.53	68.01	94.50	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 9: Ablation study results. The accuracy scores (%) of the LLaMA3-8B-Chat model on various multiple-choice QA datasets using different answer trigger sentences ϕ (**Analyzing** , **Retrieving** , and **Reasoning**).

Generalizability of ARR - Model Sizes

Size	Method	BBH	MMLU	MMLU-Pro	Avg.
1B	Baseline	35.88	43.27	21.62	33.59
	CoT	36.30	41.10	22.74	33.38
	ARR	39.02	42.70	23.49	35.07
3B	Baseline	45.65	48.26	30.88	41.60
	CoT	46.89	46.80	30.03	41.24
	ARR	51.97	52.82	33.39	46.06
8B	Baseline	52.19	60.68	38.75	50.54
	CoT	58.40	62.08	40.10	53.53
	ARR	59.01	63.51	42.72	55.08

Table 10: Model size experiments. The zero-shot performance (Accuracy %) of LLaMA3-Chat models of different sizes on multiple-choice QA datasets.

Generalizability of ARR - Model Series

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

Table 11: LLM series experiments. The zero-shot performance (Accuracy %) of 7B-Chat models of different LLM series on multiple-choice QA datasets.

Generalizability of ARR - Generation Temperatures

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

Table 12: Generation temperature experiments. The zero-shot performance (Accuracy %) of the LLaMA3-8B-Chat model on multiple-choice QA datasets using different generation temperatures (default: 0.0).

Generalizability of ARR - Few-shot Generation

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

Table 13: Few-shot experiments. The few-shot performance (Accuracy %) of the LLaMA3-8B-Chat model on multiple-choice QA datasets using 1, 3, and 5 few-shot examples with rationales.

1 Introduction

2 Our Method

3 Experiments

4 Conclusion

5 More Cases

6 References

Conclusion

- **Contribution 1:** This paper proposes **ARR**, an intuitive, general, and effective **zero-shot prompting method** to improve LLM performance in various **question-answering** tasks.

Conclusion

- **Contribution 1:** This paper proposes **ARR**, an intuitive, general, and effective **zero-shot prompting method** to improve LLM performance in various **question-answering** tasks.
- **Contribution 2:** Comprehensive experiments across diverse QA tasks demonstrate that **ARR consistently outperforms the Baseline and CoT** methods. Ablation and case studies further validate the positive contributions of each component.

Conclusion

- **Contribution 1:** This paper proposes **ARR**, an intuitive, general, and effective **zero-shot prompting method** to improve LLM performance in various **question-answering** tasks.
- **Contribution 2:** Comprehensive experiments across diverse QA tasks demonstrate that **ARR consistently outperforms the Baseline and CoT** methods. Ablation and case studies further validate the positive contributions of each component.
- **Contribution 3:** Additional extensive experiments on various settings solidify the **effectiveness and generalizability of ARR** across different model sizes, LLM series, and generation configurations.

Conclusion

- **Contribution 1:** This paper proposes **ARR**, an intuitive, general, and effective **zero-shot prompting method** to improve LLM performance in various **question-answering** tasks.
- **Contribution 2:** Comprehensive experiments across diverse QA tasks demonstrate that **ARR consistently outperforms the Baseline and CoT** methods. Ablation and case studies further validate the positive contributions of each component.
- **Contribution 3:** Additional extensive experiments on various settings solidify the **effectiveness and generalizability of ARR** across different model sizes, LLM series, and generation configurations.
- **Paper:** <https://huggingface.co/papers/2502.04689>
Code: <https://github.com/YuweiYin/ARR>
Current Work: Intent for planning & reasoning (Coming soon)

A stylized, calligraphic logo consisting of the letters 'Q' and 'A' intertwined in a cursive script. The 'Q' is on the left and the 'A' is on the right, with their strokes overlapping.

Thanks

1 Introduction

2 Our Method

3 Experiments

4 Conclusion

5 More Cases

6 References

More Cases - BoolQ

Task Input x_i	<p>Passage: Feather – Contour feathers are not uniformly distributed on the skin of the bird except in some groups such as the penguins, 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, pterylosis 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>Question: Do penguins have feathers arising from the epidermis?</p>
Baseline ϕ_i	<p>Answer: No, penguins have feathers arising from specific tracts of skin called pterylae. 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, pterylosis or pterylography, varies across bird families. In penguins, the feathers are uniformly distributed on the skin, but they arise from specific tracts of skin called pterylae. 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>
CoT ϕ_i	<p>Answer: Let's think step by step. 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, it also mentions that in some groups, such as penguins, ratites, and screamers, feathers are not uniformly distributed on the skin. This suggests that penguins do not have feathers arising from the epidermis.</p>
Reasoning r_i	<p>Therefore, the answer is no. Penguins do not have feathers arising from the epidermis.</p>
ARR ϕ_i	<p>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 distribution of feathers on penguins. The passage states that in most birds, feathers grow from specific tracts of skin called pterylae, but penguins are an exception. 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, the answer is yes.</p>
Reasoning r_i	

More Cases - LogiQA

Task Input x_i	<p>Passage: 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 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.</p> <p>Question: If Haoyun stands to the east of Jiushang Station, 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>
Baseline ϕ_i	<p>Answer: (C) Jiushang Station is adjacent to Haoyun Station and west of Haoyun Station; (Explanation) Since Jiushang Station is adjacent to Yinling Station, and Yinling Station is adjacent to Haoyun Station,</p>
Reasoning r_i	<p>Jiushang Station is adjacent to Haoyun Station. Since Jiushang Station is east of Yinling Station, and Yinling Station is west of Haoyun Station, Jiushang Station is west of Haoyun Station. Therefore, Jiushang Station is adjacent to Haoyun Station and west of Haoyun Station.</p>
CoT ϕ_i	<p>Answer: Let's think step by step. If Haoyun stands to the east of Jiushang Station, then Haoyun Station is west of Jiushang Station. Since Jiushang Station is adjacent to Yinling Station, Haoyun Station is adjacent to Yinling Station. Since Fuyi Station is east of Haoyun Station, Fuyi Station is east of Yinling Station. Since Huyao Station is west of Fuyi Station, Huyao Station is west of Yinling Station. Since Jiushang Station is adjacent to Yinling Station, Huyao Station is adjacent to Jiushang Station ...</p>
ARR ϕ_i	<p>Answer: Let's analyze the intent of the question, find relevant information, and answer the question with step-by-step reasoning. The question asks us to determine the correct relationship between the stations based on the given information. 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 Fuyi Station is between Haoyun Station and Huyao Station. Additionally, we know that Jiushang Station is adjacent to Yinling Station. The question also states that Haoyun stands to the east of Jiushang Station. This implies that Haoyun Station is to the east of Jiushang Station. Option (B) states that Haoyun Station is adjacent to Fuyi Station and west of Fuyi Station. This is consistent with the given information that Fuyi Station is east of Haoyun Station. Therefore, the correct answer is option (B)</p>

More Cases - Commonsense QA

Task Input x_i	<p>Question: What do people aim to do at work? (A) complete job (B) learn from each other (C) kill animals (D) wear hats (E) talk to each other</p>
Baseline ϕ_i	<p>Answer: (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 it doesn't necessarily imply learning or personal growth. Option (C) "kill animals" is unrelated to the typical goals of work in most industries.</p>
Reasoning r_i	<p>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) "learn from each other" is a goal that aligns with personal and professional development, 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. The final answer is: B</p>
CoT ϕ_i	<p>Answer: Let's think step by step. (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 people do not go to work to complete a job. They go to work to do a job. 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. They go to work to do a job. The best answer is B.</p>
Reasoning r_i	<p>Choice (A) is incorrect because people do not go to work to complete a job. They go to work to do a job. 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. They go to work to do a job. The best answer is B.</p>
ARR ϕ_i	<p>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 people may learn from each other at work, but that is not the main purpose of going to work. 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. The best answer is A.</p>
Reasoning r_i	<p>Choice (B) is incorrect because people may learn from each other at work, but that is not the main purpose of going to work. 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. The best answer is A.</p>

More Cases - SocialIQA

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

More Cases - SciQ

Task Input x_i	Question: 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
Baseline ϕ_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).
CoT ϕ_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.
Reasoning r_i	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.
ARR ϕ_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,
Reasoning r_i	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.

More Cases - OpenBookQA

Task Input x_i	<p>Fact: force causes the speed of an object to decrease.</p> <p>Question: The amount of friction and the speed of an object have what kind of relationship? (A) inverse (B) reverse (C) direct (D) equal</p>
Baseline ϕ_i	<p>Answer: (C) Direct</p>
Reasoning r_i	<p>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.</p>
CoT ϕ_i	<p>Answer: Let's think step by step.</p> <p>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.</p> <p>Step 2: Determine the type of relationship</p>
Reasoning r_i	<p>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.</p> <p>Step 3: Choose the correct answer Based on the understanding of the relationship between friction and speed, the correct answer is (C) direct.</p>
ARR ϕ_i	<p>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.</p>
Reasoning r_i	<p>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.</p>

More Cases - ARC

Task Input x_i	<p>Question: 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?</p> <p>(A) Volcanism also characterizes the region. (B) Earthquake centers in the region occur at shallow depths. (C) The region shows extensive faulting of sediments. (D) Equal crust densities are found on opposite sides of the region.</p>
Baseline ϕ_i	<p>Answer: The best answer is C. Transform plate boundaries are characterized by horizontal motion and faulting, but the faulting is not necessarily extensive. 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.</p>
Reasoning r_i	
CoT ϕ_i	<p>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.</p>
Reasoning r_i	
ARR ϕ_i	<p>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.</p>
Reasoning r_i	<p>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.</p>

More Cases - BBH

Task Input x_i	<p>Question: 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.</p> <p>The peaches are more expensive than the apples. The mangoes are the cheapest.</p> <p>(A) The peaches are the second-most expensive (B) The mangoes are the second-most expensive (C) The apples are the second-most expensive</p>
Baseline ϕ_i	<p>Answer:</p> <p>The peaches are the second-most expensive. The mangoes are the cheapest. The apples are the least expensive of the three fruits.</p>
Reasoning r_i	<p>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 ...</p>
CoT ϕ_i	<p>Answer: Let's think step by step.</p> <p>The peaches are more expensive than the apples. The mangoes are the cheapest. Therefore, the mangoes are the least expensive of the three fruits.</p>
Reasoning r_i	<p>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.</p>
ARR ϕ_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 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.</p>
Reasoning r_i	<p>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.</p>

More Cases - MMLU

Task Input x_i	<p>Question: 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</p>
Baseline ϕ_i	<p>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.</p>
Reasoning r_i	<p>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.</p>
CoT ϕ_i	<p>Answer: Let's think step by step.</p>
Reasoning r_i	<p>(B) The predicate logic translation of the sentence is Tcd, where Txy means x teaches y. The best answer is B.</p>
ARR ϕ_i	<p>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.</p>
Reasoning r_i	<p>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.</p>

More Cases - MMLU-Pro

Task Input x_i	<p>Question: Which type of research methods are designed to elicit responses to predetermined, standardized questions from many respondents?</p> <p>(A) Non-probability. (B) Cross-sectional. (C) Qualitative. (D) Ethnographic. (E) Longitudinal. (F) Experimental. (G) Probability. (H) Observational. (I) Case Study. (J) Quantitative.</p>
Baseline ϕ_i	<p>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.</p>
Reasoning r_i	
CoT ϕ_i	<p>Answer: Let's think step by step. (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.</p>
Reasoning r_i	
ARR ϕ_i	<p>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. The correct answer is (J). The other choices are incorrect because they are not quantitative research methods.</p>
Reasoning r_i	

1 Introduction

2 Our Method

3 Experiments

4 Conclusion

5 More Cases

6 References

- [6] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al.
Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.
arXiv preprint arXiv:2403.05530, 2024.
- [7] 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.
A survey on evaluation of large language models.
ACM Transactions on Intelligent Systems and Technology, 15(3), March 2024.

- [8] Joshua Robinson and David Wingate.
Leveraging large language models for multiple choice question answering.
In The Eleventh International Conference on Learning Representations, 2023.
- [9] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord.
Think you have solved question answering? try arc, the ai2 reasoning challenge.
arXiv preprint arXiv:1803.05457, 2018.

- [10] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang.

Logiqa: A challenge dataset for machine reading comprehension with logical reasoning.

In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3622–3628. International Joint Conferences on Artificial Intelligence Organization, 7 2020.

Main track.

- [11] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt.

Measuring massive multitask language understanding.

In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

OpenReview.net, 2021.

- [12] 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. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.
- [13] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics.

[14] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al.

Mmlu-pro: A more robust and challenging multi-task language understanding benchmark.

arXiv preprint arXiv:2406.01574, 2024.

[15] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal.

Can a suit of armor conduct electricity? a new dataset for open book question answering.

In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium, October–November 2018. Association for Computational Linguistics.

- [16] Frederick Adams.
Intention and intentional action: The simple view.
Mind & Language, 1(4):281–301, 1986.
- [17] Alfred R Mele.
Intention, belief, and intentional action.
American Philosophical Quarterly, 26(1):19–30, 1989.
- [18] Alfred R Mele and Paul K Moser.
Intentional action.
Nous, 28(1):39–68, 1994.

[19] Erik Jones and Jacob Steinhardt.

Capturing failures of large language models via human cognitive biases.

In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 11785–11799. Curran Associates, Inc., 2022.

[20] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou.

Large language models can be easily distracted by irrelevant context.

In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine*

[26] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant.

CommonsenseQA: A question answering challenge targeting commonsense knowledge.

In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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, June 2019. Association for Computational Linguistics.

[27] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi.

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, November 2019.

Association for Computational Linguistics.

[28] Johannes Welbl, Nelson F. Liu, and Matt Gardner.

Crowdsourcing multiple choice science questions.

In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin, editors, *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark, September 2017.

Association for Computational Linguistics.

- [29] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al.
The llama 3 herd of models.
arXiv preprint arXiv:2407.21783, 2024.
- [30] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al.
Qwen2.5 technical report.
arXiv preprint arXiv:2412.15115, 2024.

- [31] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology.

arXiv preprint arXiv:2403.08295, 2024.

- [32] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118, 2024.*

