JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

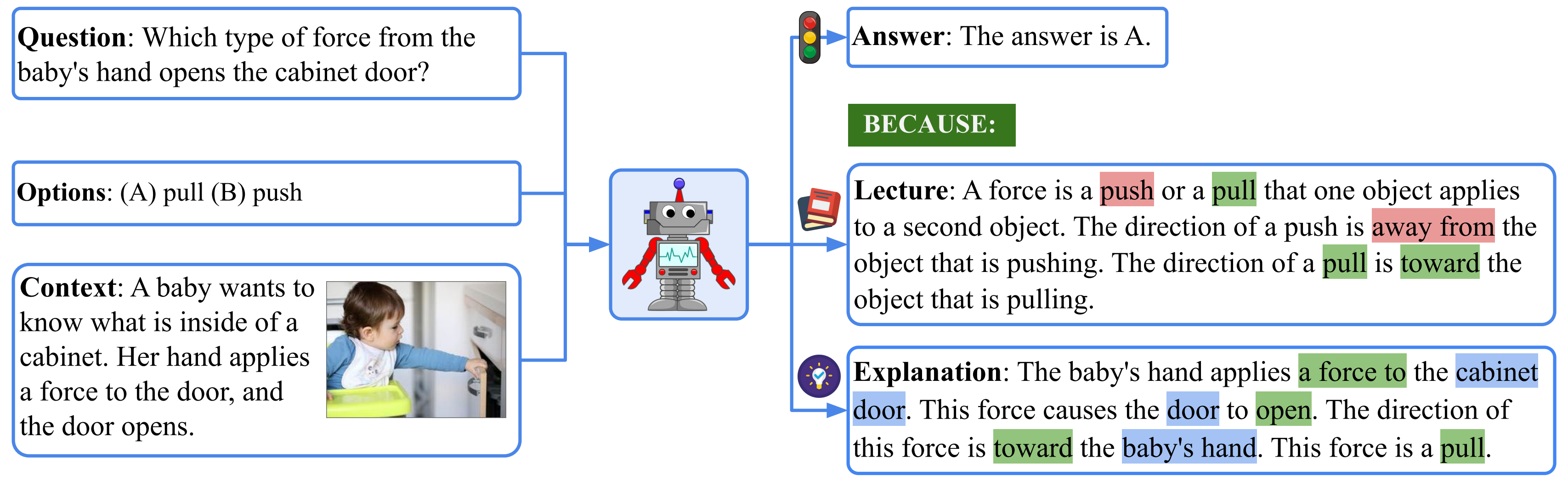
https://scienceqa.github.io

Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering

Pan Lu[1,3], Swaroop Mishra[2], Tony Xia[1], Liang Qiu[1], Kai-Wei Chang[1], Song-Chun Zhu[1], Oyvind Tafjord[3], Peter Clark[3], Ashwin Kalyan[3]

[1]University of California, Los Angeles [2]Arizona State University [3]Allen Institute for AI

Introduction



Question: Which type of force from the baby's hand opens the cabinet door?

Options: (A) pull (B) push

Context: A baby wants to know what is inside of a cabinet. Her hand applies a force to the door, and the door opens.

Answer: The answer is A.

BECAUSE:

Lecture: A force is a push or a pull that one object applies to a second object. The direction of a push is away from the object that is pushing. The direction of a pull is toward the object that is pulling.

Explanation: The baby's hand applies a force to the cabinet door. This force causes the door to open. The direction of this force is toward the baby's hand. This force is a pull.

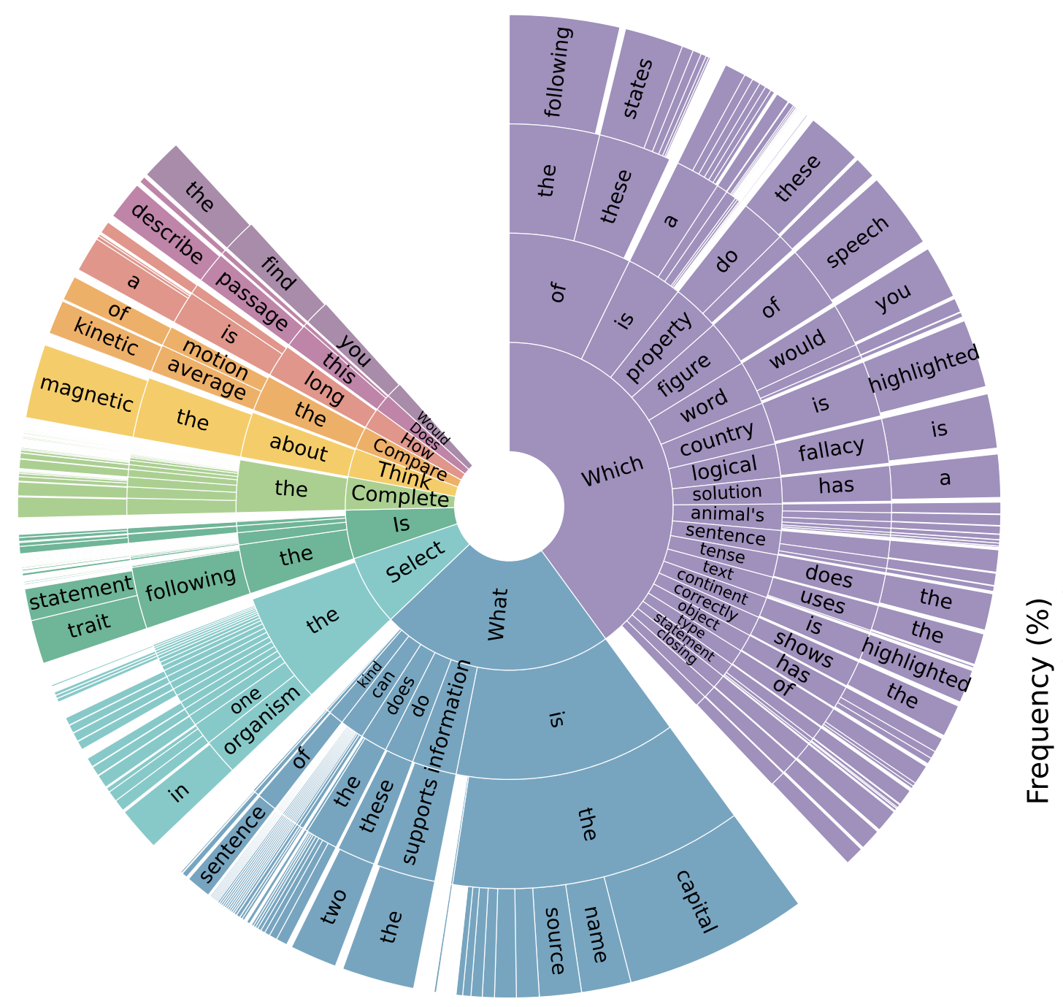
- We propose **Science Question Answering (ScienceQA)**, a new dataset that contains **21,208** MC questions with multimodal contexts from the science curriculum. ScienceQA is the first large-scale multimodal science dataset that annotates lectures and explanations
- We show that **Chain of Thought (CoT)** benefits large language models in both few-shot and fine-tuning settings by improving model performance and reliability via generating explanations

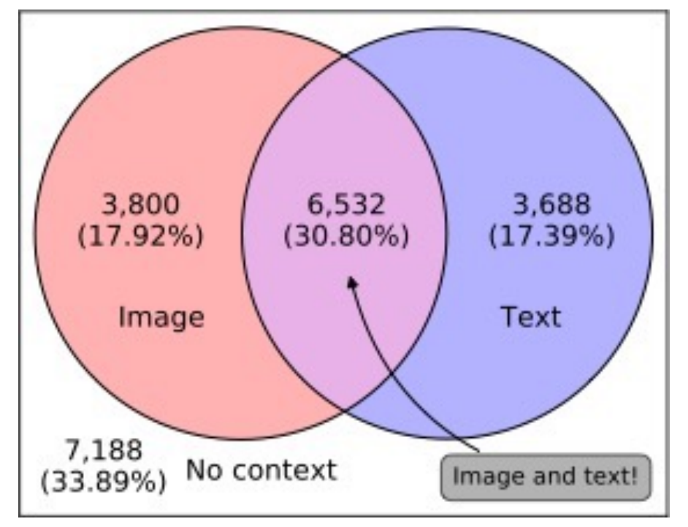
The ScienceQA Dataset

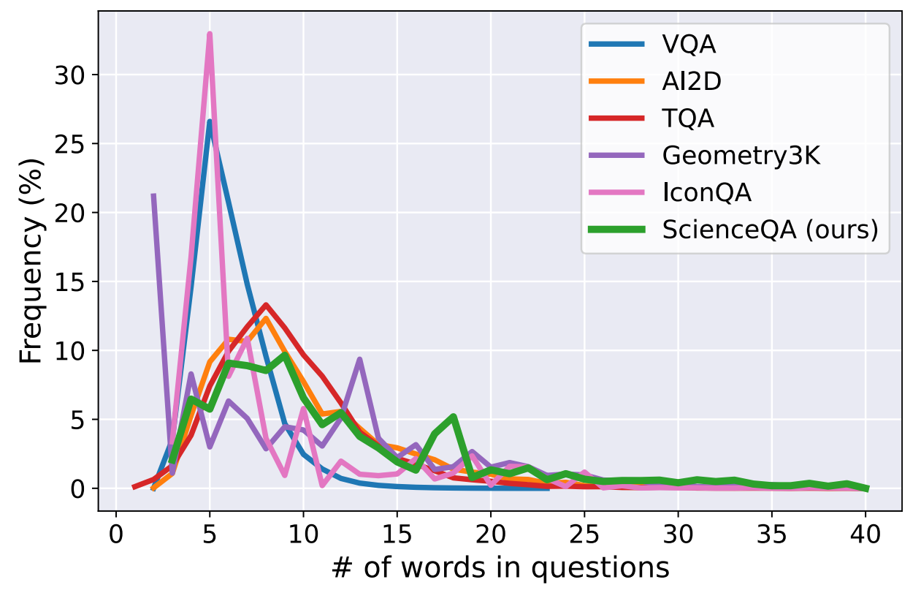
Main statistics

Statistic	Number
Total questions	21,208
Questions with text context	10,220 (48.2%)
Questions with image context	10,332 (48.7%)
* Image of natural format	≈2,960 (14.0%)
* Image of diagram format	≈7,372 (34.8%)
Questions with both contexts	6,532 (30.8%)
Questions without any context	7,188 (33.9%)
Questions with a lecture	17,798 (83.9%)
Questions with a explanation	19,202 (90.5%)
Different questions	9,122
Different lectures	261
Topic classes	26
Category classes	127
Skill classes	379
Average question length	12.11
Average choice length	4.40
Average lecture length	125.06
Average explanation length	47.66

Question distribution







- ScienceQA is collected from elementary and high school science curricula and contains **21,208** examples
- **48.7%** have an image context, **48.2%** have a text context, and **30.8%** have both
- **83.9%** are annotated with a lecture, and **91.3%** with an explanation

Comparisons with existing VQA and Science datasets

	#Q	#I	AvgQ	MaxQ	Grades	Science subjects	Contexts	Images	Lecture	Explanation
Geometry3K [30]	3,002	2,342	10.1	46	6-12	natural (geometry)	image	diagram	✗	✗
AI2D [16]	4,563	4,903	9.8	64	1-6	natural	image	diagram	✗	✗
FOODWEBS [23]	≈5,000	≈5,00	-	-	8	natural (foodweb only)	image	diagram	✗	✗
ARC [5]	7,787	0	**20.4**	128	3-9	natural	✗	✗	✗	✗
TQA [17]	**26,260**	3,455	9.2	57	6-8	natural	image, text	diagram	✓	✗
IconQA [34]	107,439	96,817	8.4	73	PreK-3	math	visual	diagram	✗	✗
WorldTree [12]	1,680	0	-	-	3-5	natural	✗	✗	✗	✓
OpenBookQA [36]	5,957	0	10.6	68	1-6	natural	✗	✗	✗	✓
QASC [19]	9,980	0	8.0	25	1-9	natural	✗	✗	✗	✓
SCIENCEQA (ours)	21,208	**10,332**	12.1	**141**	**1-12**	natural, social, language	image, text	natural, diagram	✓	✓

- ScienceQA is much larger than most existing datasets and different from in various aspects

Domain Diversity of ScienceQA

Biology
Genes to traits
Classification
Adaptations
Traits and heredity
Ecosystems
Classification
Scientific names
Heredity
Ecological interactions
Cells
Plants
Animals
Plant reproduction

Earth Science
Weather and climate
Rocks and minerals
Astronomy
Fossils
Earth events
Plate tectonics

Physics
Materials
Magnets
Velocity and forces
Force and motion
Particle motion and energy
Heat and thermal energy
States of matter
Kinetic and potential energy
Mixture

Chemistry
Solutions
Physical and chemical change
Atoms and molecules
Chemical reactions

Engineering
Designing experiments
Engineering practices

Units and Measurement
Weather and climate

Geography
State capitals
Geography
Maps
Oceania: geography
Physical Geography
The Americas: geography
Oceans and continents
Cities
States

History
Colonial America
English colonies in North America
The American Revolution

World History
Greece
Ancient Mesopotamia
World religions
American history
Medieval Asia

Civics
Social skills
Government
The Constitution

Economics
Basic economic principles
Supply and demand
Banking and finance

Global Studies
Society and environment

Writing Strategies
Supporting arguments
Sentences, fragments, and run-ons
Word usage and nuance
Creative techniques
Audience, purpose, and tone
Pronouns and antecedents
Persuasive strategies
Editing and revising
Visual elements
Opinion writing

Vocabulary
Categories
Shades of meaning
Comprehension strategies
Context clues

Grammar
Sentences and fragments
Phrases and clauses

Figurative Language
Literary devices

Verbs
Verb tense

Capitalization
Formatting

Punctuation
Fragments

Phonology
Rhyming

Reference
Research skills

🟦 Nature Science
🟥 Social Science
🟩 Language Science

3 subjects

26 topics

127 categories

379 skills

Methods and Main Results

Baselines and GPT-3 (CoT)

Question: $\texttt{question}: I_i^{ques}$
Options: (A) $\texttt{option}: I_{i1}^{opt}$ (B) $\texttt{option}: I_{i2}^{opt}$ (C) $\texttt{option}: I_{i3}^{opt}$
Context: $\texttt{context}: I_i^{cont}$
Answer: The answer is $\texttt{answer}: I_i^{a}$. BECAUSE: $\texttt{lecture}: I_i^{lect}$ $\texttt{explanation}: I_i^{exp}$

Question: $\texttt{question}: I_t^{ques}$
Options: (A) $\texttt{option}: I_{t1}^{opt}$ (B) $\texttt{option}: I_{t2}^{opt}$ (C) $\texttt{option}: I_{t3}^{opt}$ (D) $\texttt{option}: I_{t4}^{opt}$
Context: $\texttt{context}: I_t^{cont}$
Answer:

We establish various baselines on ScienceQA
- Heuristic baselines: random chance and human performance
- Fine-tuning VQA models
- Pre-trained and fine-tuning UnifiedQA
- Zero-shot and few-shot GPT-3

To mimic the multi-hop reasoning process, we further build GPT-3 via **chain-of-thought (CoT) prompting** to generate the answer followed by the lecture and explanation

Results on ScienceQA

Model	Learning	Format	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
Random chance	-	M→A	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67	39.83
Q only [1]	train set	Q→A	41.34	27.22	47.00	41.79	35.15	44.60	39.28	40.87	39.85
C_I only [1]	train set	C_I→A	41.34	29.25	45.45	42.33	36.09	42.93	39.21	41.07	39.87
Q+M only [1]	train set	QM→A	52.66	51.86	60.18	55.57	50.37	57.42	52.53	57.88	54.44
Q+C_T+M only [1]	train set	QC_TM→A	57.28	49.04	61.36	60.46	52.80	58.82	54.44	60.51	56.61
Q+C_I+M only [1]	train set	QC_IM→A	58.97	53.77	60.45	62.85	54.49	57.63	56.72	61.04	58.26
MCAN [54]	train set	QCM→A	56.08	46.23	58.09	59.43	51.17	55.40	51.65	59.72	54.54
Top-Down [1]	train set	QCM→A	59.50	54.33	61.82	62.90	54.88	59.79	57.27	62.16	59.02
BAN [20]	train set	QCM→A	60.88	46.57	66.64	62.61	52.60	65.51	56.83	63.94	59.37
DFAF [8]	train set	QCM→A	64.03	48.82	63.55	65.88	54.49	64.11	57.12	67.17	60.72
ViLT [21]	train set	QCM→A	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90	61.14
Patch-TRM [34]	train set	QCM→A	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50	61.42
VisualBERT [25, 26]	train set	QCM→A	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92	61.87
UnifiedQA$_{SMALL}$ [47]	zero-shot	QCM→A	47.78	40.49	46.00	50.24	44.12	44.39	45.56	46.21	45.79
UnifiedQA$_{BASE}$ [47]	zero-shot	QCM→A	50.13	44.54	48.18	53.08	48.09	46.69	47.58	50.03	48.46
UnifiedQA$_{SMALL}$ [47]	train set	QCM→A	53.77	58.04	61.09	52.10	51.51	61.46	58.22	53.59	56.57
UnifiedQA$_{BASE}$ [47]	train set	QCM→A	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00	70.12
UnifiedQA$_{BASE}$ (CoT)	train set	QCM→AE	70.60	74.02	78.36	65.69	64.80	81.53	75.48	69.48	73.33$_{3.21\uparrow}$
UnifiedQA$_{BASE}$ (CoT)	train set	QCM→ALE	71.00	76.04	78.91	66.42	66.53	**81.81**	77.06	68.82	74.11$_{3.99\uparrow}$
GPT-3 [4]	zero-shot	QCM→A	75.04	66.59	78.00	74.24	65.74	79.58	76.36	**69.87**	74.04
GPT-3 [4]	2-shot	QCM→A	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3 (CoT)	2-shot	QCM→AE	**76.60**	65.92	77.55	**75.51**	66.09	79.58	**78.49**	67.63	74.61$_{0.64\uparrow}$
GPT-3 (CoT)	2-shot	QCM→ALE	75.44	**70.87**	**78.09**	74.68	**67.43**	79.93	78.23	69.68	**75.17**$_{1.20\uparrow}$
Human	-	QCM→A	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40

- **Blind studies** show that all input components provide critical information for question answering
- Current **VQA** models are not well generalized to ScienceQA
- The fine-tuned **UnifiedQA** model can benefit from CoT
- **GPT-3** shows its power in both zero-shot and few-shot settings
- **2-shot GPT-3 (CoT)** achieves a SOTA accuracy of 75.17%
- **Humans** perform much better

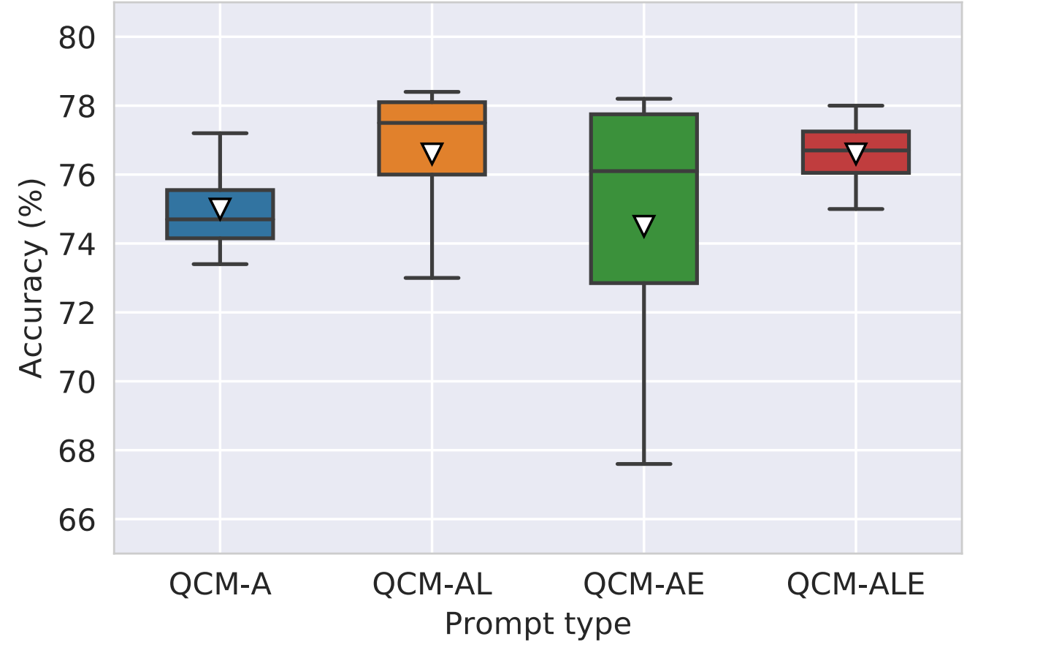
Evaluations of generated explanations

Model	Format	BLEU-1	BLEU-4	ROUGE-L	Similarity	Relevant	Correct	Complete	Gold
UnifiedQA$_{BASE}$ (CoT)	QCM→ALE	**0.397**	**0.370**	**0.714**	**0.811**	80.4%	76.6%	76.1%	56.9%
GPT-3 (CoT)	QCM→AE	0.234	0.048	0.351	0.561	76.9%	73.0%	70.5%	52.5%
GPT-3 (CoT)	QCM→ALE	0.192	0.052	0.323	0.595	**88.5%**	**78.8%**	**84.5%**	**65.2%**

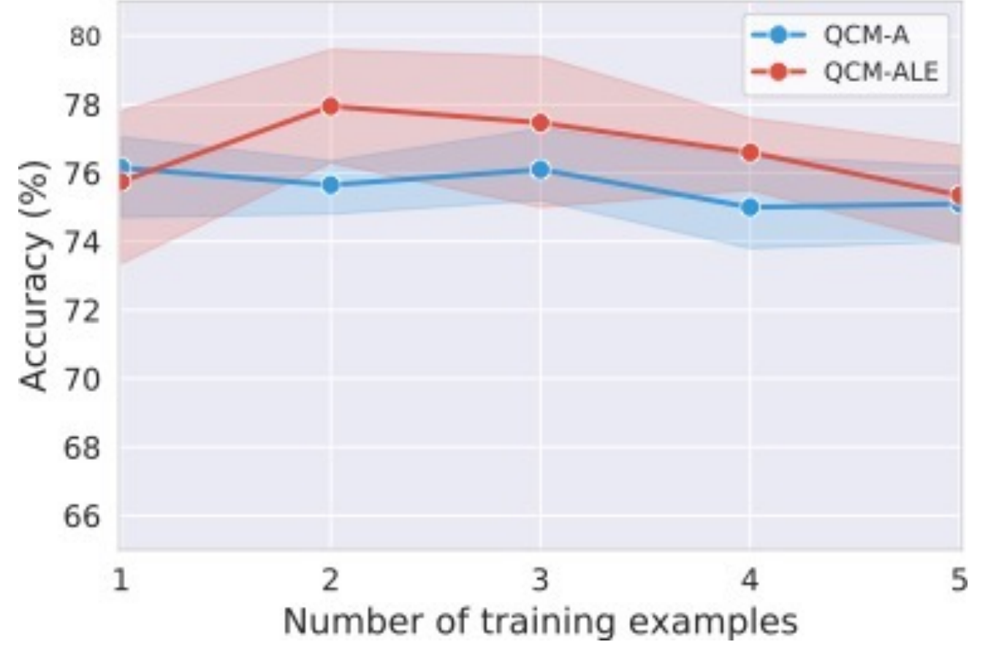
- Fine-tuning models (UnifiedQA) generate **similar** explanations to training data
- The explanations generated by GPT-3 (CoT) conform best to **human judgment**

Abalation Study

Different prompt types



QCM-ALE performs the best



2-shot examples perform the best

Prompt type	Sampling	Acc. (%)
QCML*→A	Random	73.59
QCML*→AE	Random	74.32
QCME*→A	Random	94.03$_{18.86\uparrow}$
QCMLE*→A	Random	**94.13**$_{18.96\uparrow}$
QCM→ALE	Random	75.17

Ground truth LE in the input performs pretty well

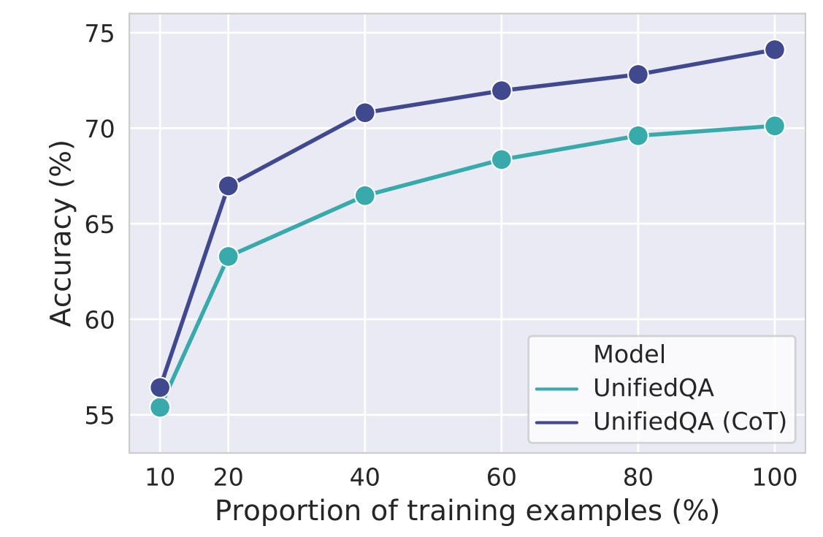
Prompt type	Sampling	Acc. (%)
QCM→LA	Random	60.6
QCM→EA	Random	56.0
QCM→LEA	Random	55.4
QCM→ELA	Random	51.5
QCM→ALE	Random	**73.6**

LE after A leads to the performance drop

Dynamic sampling

Prompt type	Sampling	Acc. (%)
QCM→ALE	Dynamic (same topic)	75.15
QCM→ALE	Dynamic (same category)	74.58
QCM→ALE	Dynamic (same skill)	75.10

- Select the in-context examples to prompt with the **same class** as the test sample instead of random sampling
- This kind of dynamic sampling does not work



UnifiedQA (CoT) learns efficiently with fewer training data

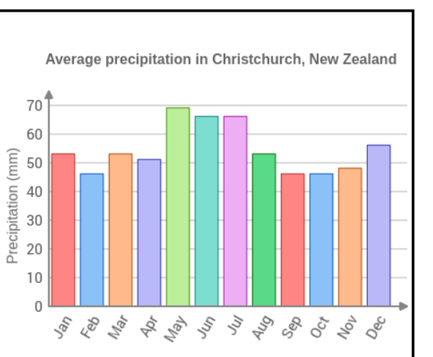
Case Study

Successful example from GPT-3 CoT

Question: What do these two changes have in common?\n
bleaching clothes\n
a piece of apple turning brown\n
Options: (A) Both are only physical changes. (B) Both are caused by cooling. (C) Both are chemical changes. (D) Both are caused by heating.\n
Answer: **The answer is C. BECAUSE:** A **physical change** is a change to a substance that does not change what the substance is. A **physical change** only changes the substance's physical properties, such as its size, shape, or color.\nA **chemical change** is a change to a substance that changes what the substance is. A **chemical change** changes the substance's chemical properties, such as its chemical structure.\n**The changes that you listed are both chemical changes.**\nWhen you **bleach clothes**, the chemical structure of the fabric changes. This change makes the fabric less likely to hold color.\n When **a piece of fruit turns brown**, the chemical structure of the fruit changes. This change makes the fruit taste different.

Failure cases from GPT-3 CoT

Question: Which month is the wettest on average in Christchurch?
Context: Use the graph to answer the question below. A series of photographs showing a variety of colorful objects.



Options: (A) April (B) December (C) May
Answer: **The answer is B. BECAUSE:** The graph shows the average monthly rainfall in Christchurch from 1971 to 2000.\nThe wettest month on average is December, with an average rainfall of 154.8 mm.\nThe second wettest month is May, with an average rainfall of 153.1 mm.\nThe third wettest month is April, with an average rainfall of 151.4 mm.

Question: Is the following trait inherited or acquired?
Marshall has five fingers on each hand.
Context: Hint: Marshall was born with five fingers on each hand.
Options: (A) acquired (B) inherited
Answer: **The answer is B. BECAUSE:** Marshall was born with five fingers on each hand. This trait was inherited from Marshall's parents.

Project Page
https://scienceqa.github.io

- Fail to understand **multimodal** inputs and lack **domain** knowledge to predict correct answers
- Generate the wrong chain of thought with **irrelevant**, **incorrect**, or **incomplete** information