Chain-of-Thought Reasoning In Language Models

Zhuosheng Zhang
zhangzs@sjtu.edu.cn
https://bcmi.sjtu.edu.cn/~zhangzs

What is Chain-of-Thought (CoT)?

- **Chain of thought (CoT)** prompting enables LLMs to generate **intermediate reasoning steps** before inferring an answer
  - With a few demonstrations or just a prompt sentence
  - Without gradient updates
- **Paradigm Shift of Task Format**
  - Standard Format: <input → output>
  - CoT Format: <input → rationale → output>

Q: There were 10 friends playing a video game online when 7 players quit. If each player left had 8 lives, how many lives did they have total?

A: The answer is

(Output) 80. ✗

Q: There were 10 friends playing a video game online when 7 players quit. If each player left had 8 lives, how many lives did they have total?

A: Let's think step by step.

(Output) There were 10 friends playing a video game online. This means that, at the start, there were 10 x 8 = 80 lives in total. Then, 7 players quit. This means that 7 x 8 = 56 lives were lost. Therefore, the total number of lives remaining is 80 - 56 = 24. The answer is 24. ✓

Question

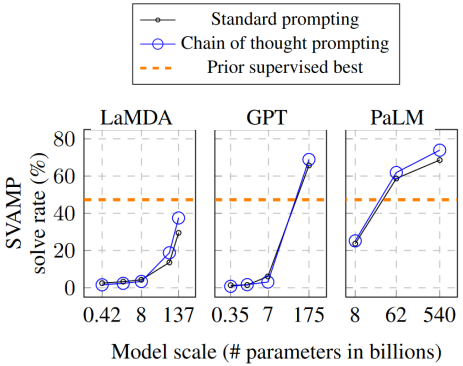
Rationale

Answer

LLMs are Strong CoT Reasoners

- Tasks: **multi-step reasoning tasks**, e.g., math word problems, commonsense reasoning, logical reasoning, etc.
- LLMs show **emergent abilities** of solving challenging reasoning problems with CoT

Model	MultiArith		GSM8K		AddSub		AQUA-RAT		SingleEq		SVAMP	
	N/A	CoT	N/A	CoT	N/A	CoT	N/A	CoT	N/A	CoT	N/A	CoT
Zero-Shot Performance												
text-davinci-002	22.7	78.7	12.5	40.7	77.0	74.7	22.4	33.5	78.7	78.7	58.8	63.7
text-davinci-003	24.2	83.7	12.6	59.5	87.3	81.3	**28.0**	40.6	82.3	86.4	64.7	73.6
ChatGPT	**30.3**	**96.0**	**14.7**	**75.4**	**89.6**	**89.9**	23.6	**47.6**	**83.1**	**91.3**	**68.1**	**82.8**
Few-Shot Performance												
UL2	5.0	10.7	4.1	4.4	18.5	18.2	20.5	23.6	18.0	20.2	10.1	12.5
LaMDA	7.6	44.9	6.5	14.3	43.0	51.9	25.5	20.6	48.8	58.7	29.5	37.5
text-davinci-002	33.8	91.7	15.6	46.9	83.3	81.3	24.8	35.8	82.7	86.6	65.7	68.9
Codex	**44.0**	**96.2**	**19.7**	**63.1**	90.9	90.9	**29.5**	**45.3**	**86.8**	**93.1**	**69.9**	76.4
PaLM	42.2	94.7	17.9	56.9	**93.9**	**91.9**	25.2	35.8	86.5	92.3	69.4	**79.0**



[1] Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M. and Yang, D., 2023. Is ChatGPT a General-Purpose Natural Language Processing Task Solver?. arXiv preprint arXiv:2302.06476.

[2] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E.H., Le, Q.V. and Zhou, D., Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems* (NeurIPS). 2022.

A Family of CoT Studies

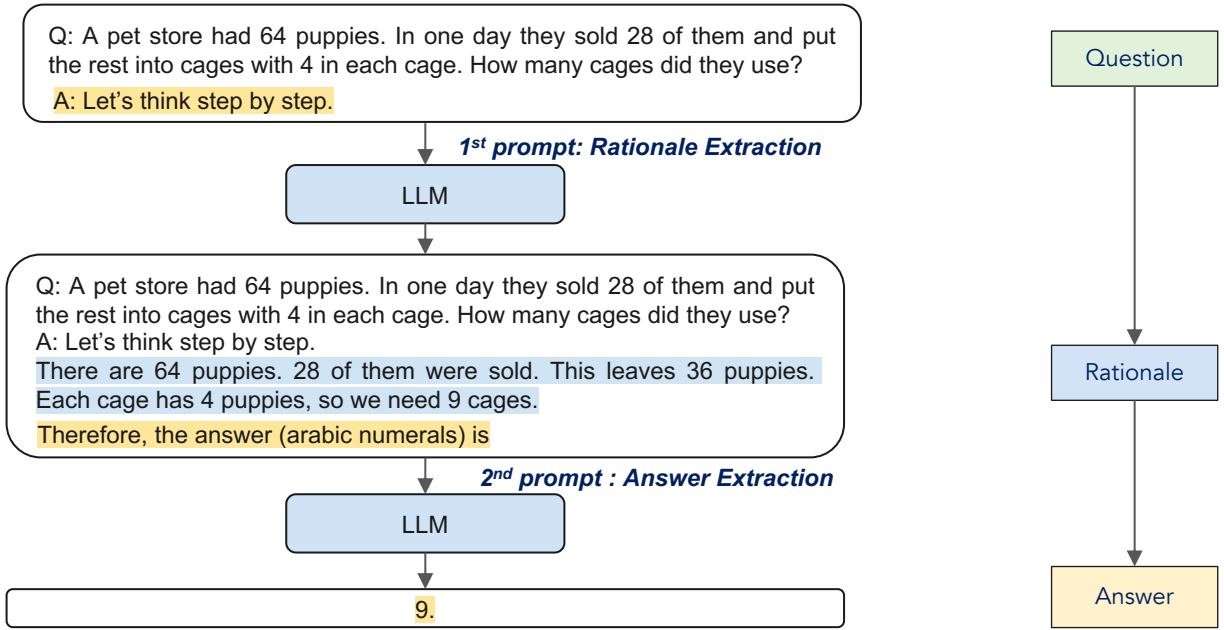
- **Key problems in existing studies**
  - Rely on **handcrafting few-shot demonstrations** for in-context learning (ICL) ⟶ **Auto-CoT**
  - Focus on the **language only modality** ⟶ **Multimodal-CoT**

Models	Mutimodal	w/o LLM	Model / Engine	Training	CoT Role	CoT Source
Zero-Shot-CoT (Kojima et al., 2022)	✗	✗	GPT-3.5 (175B)	ICL	Reasoning	Template
Few-Shot-CoT (Wei et al., 2022b)	✗	✗	PaLM (540B)	ICL	Reasoning	Hand-crafted
Self-Consistency-CoT (Wang et al., 2022a)	✗	✗	Codex (175B)	ICL	Reasoning	Hand-crafted
Least-to-Most Prompting (Zhou et al., 2022)	✗	✗	Codex (175B)	ICL	Reasoning	Hand-crafted
Retrieval-CoT (Zhang et al., 2022)	✗	✗	GPT-3.5 (175B)	ICL	Reasoning	Auto-generated
PromptPG-CoT (Lu et al., 2022b)	✗	✗	GPT-3.5 (175B)	ICL	Reasoning	Hand-crafted
Auto-CoT (Zhang et al., 2022)	✗	✗	Codex (175B)	ICL	Reasoning	Auto-generated
Complexity-CoT (Fu et al., 2022)	✗	✗	GPT-3.5 (175B)	ICL	Reasoning	Hand-crafted
Few-Shot-PoT (Chen et al., 2022)	✗	✗	GPT-3.5 (175B)	ICL	Reasoning	Hand-crafted
UnifiedQA (Lu et al., 2022a)	✗	✓	T5 (770M)	FT	Explanation	Crawled
Fine-Tuned T5 XXL (Magister et al., 2022)	✗	✗	T5 (11B)	KD	Reasoning	LLM-generated
Fine-Tune-CoT (Ho et al., 2022)	✗	✗	GPT-3 (6.7B)	KD	Reasoning	LLM-generated
Multimodal-CoT (Zhang et al., 2023b)	✓	✓	T5 (770M)	FT	Reasoning	Crawled

[1] Zhang, Z., Zhang, A., Li, M. and Smola, A. Automatic chain of thought prompting in large language models. The Eleventh International Conference on Learning Representations (ICLR). 2023.
[2] Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., and Smola, A. Multimodal Chain-of-Thought Reasoning in Language Models. arXiv preprint arXiv:2302.00923. 2023.

Typical Paradigm-1: Zero-Shot-CoT

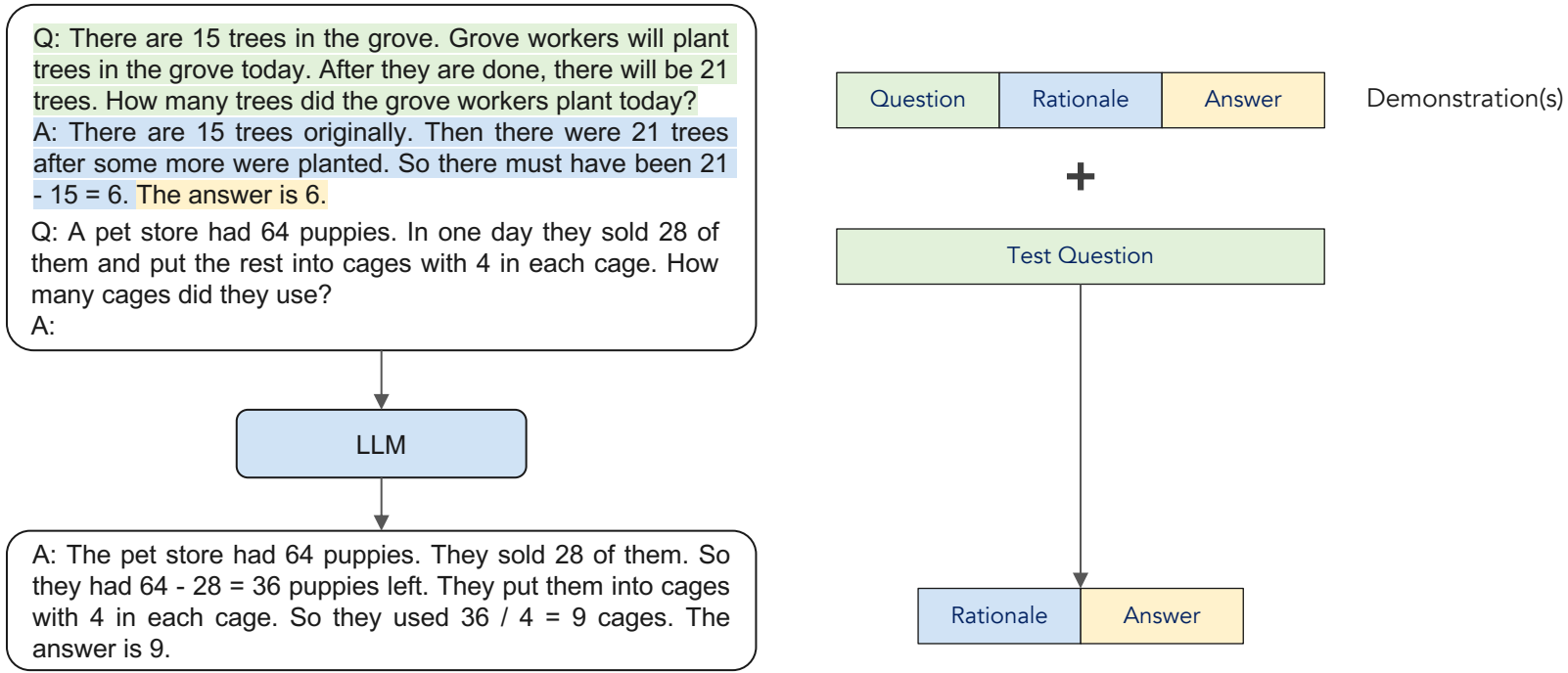
- **Zero-Shot CoT**
  - w/ a **trigger hint**, e.g., "*let's think step by step*" after the question (question + hint)



Kojima, T., Gu, S.S., Reid, M., Matsuo, Y. and Iwasawa, Y., Large Language Models are Zero-Shot Reasoners. In Advances in Neural Information Processing Systems (NeurIPS). 2022.

Typical Paradigm-2: Few-Shot-CoT

- **Few-Shot-CoT (Manual-CoT)**
  - In-context learning method by demonstrating **step-by-step reasoning exemplars (demonstrations)**



Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E.H., Le, Q.V. and Zhou, D., Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems* (NeurIPS). 2022.

Manual-CoT is Not Scalable

- **Pros:** strong performance (by **carefully hand-crafted demonstrations**)
- **Cons:**
  - Model performance relies heavily on the **quality of the demonstrations**
  - Dependence on task-aware manual-written demonstrations (professional)

	GSM8K	SVAMP	ASDiv	MAWPS
Standard prompting	6.5 ±0.4	29.5 ±0.6	40.1 ±0.6	43.2 ±0.9
Chain of thought prompting	14.3 ±0.4	36.7 ±0.4	46.6 ±0.7	57.9 ±1.5
Ablations				
· equation only	5.4 ±0.2	35.1 ±0.4	45.9 ±0.6	50.1 ±1.0
· variable compute only	6.4 ±0.3	28.0 ±0.6	39.4 ±0.4	41.3 ±1.1
· reasoning after answer	6.1 ±0.4	30.7 ±0.9	38.6 ±0.6	43.6 ±1.0
Robustness				
· different annotator (B)	15.5 ±0.6	35.2 ±0.4	46.5 ±0.4	58.2 ±1.0
· different annotator (C)	17.6 ±1.0	37.5 ±2.0	48.7 ±0.7	60.1 ±2.0
· intentionally concise style	11.1 ±0.3	38.7 ±0.8	48.0 ±0.3	59.6 ±0.7
· exemplars from GSM8K (α)	12.6 ±0.6	32.8 ±1.1	44.1 ±0.9	53.9 ±1.1
· exemplars from GSM8K (β)	12.7 ±0.5	34.8 ±1.1	46.9 ±0.6	60.9 ±0.8
· exemplars from GSM8K (γ)	12.6 ±0.7	35.6 ±0.5	44.4 ±2.6	54.2 ±4.7

Manual-CoT is Not Scalable

- **Pros:** strong performance (w/ the best selected demos)
- **Cons:**
  - Model performance heavily relies on the **quality of the demonstrations**
  - Dependence on **task-aware manual-written demonstrations** (professional)

    Two parts: selecting ***demo questions*** and writing the ***corresponding rationales***

PROMPT FOR MATH WORD PROBLEMS

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

A: There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: There are originally 3 cars. 2 more cars arrive. 3 + 2 = 5. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?

A: Originally, Leah had 32 chocolates. Her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?

A: Jason started with 20 lollipops. Then he had 12 after giving some to Denny. So he gave Denny 20 - 12 = 8. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

A: Shawn started with 5 toys. If he got 2 toys each from his mom and dad, then that is 4 more toys. 5 + 4 = 9. The answer is 9.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?

A: There were originally 9 computers. For each of 4 days, 5 more computers were added. So 5 * 4 = 20 computers were added. 9 + 20 is 29. The answer is 29.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?

A: Michael started with 58 golf balls. After losing 23 on tuesday, he had 58 - 23 = 35. After losing 2 more, he had 35 - 2 = 33 golf balls. The answer is 33.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?

A: Olivia had 23 dollars. 5 bagels for 3 dollars each will be 5 x 3 = 15 dollars. So she has 23 - 15 dollars left. 23 - 15 is 8. The answer is 8.

PROMPT FOR CSQA

Q: What do people use to absorb extra ink from a fountain pen? Answer Choices: (a) shirt pocket (b) calligrapher's hand (c) inkwell (d) desk drawer (e) blotter

A: The answer must be an item that can absorb ink. Of the above choices, only blotters are used to absorb ink. So the answer is (e).

Q: What home entertainment equipment requires cable?
Answer Choices: (a) radio shack (b) substation (c) television (d) cabinet

A: The answer must require cable. Of the above choices, only television requires cable. So the answer is (c).

Q: The fox walked from the city into the forest, what was it looking for? Answer Choices: (a) pretty flowers (b) hen house (c) natural habitat (d) storybook

A: The answer must be something in the forest. Of the above choices, only natural habitat is in the forest. So the answer is (b).

Q: Sammy wanted to go to where the people were. Where might he go? Answer Choices: (a) populated areas (b) race track (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Of the above choices, only populated areas have a lot of people. So the answer is (a).

Q: Where do you put your grapes just before checking out? Answer Choices: (a) mouth (b) grocery cart (c)super market (d) fruit basket (e) fruit market

A: The answer should be the place where grocery items are placed before checking out. Of the above choices, grocery cart makes the most sense for holding grocery items. So the answer is (b).

Q: Google Maps and other highway and street GPS services have replaced what? Answer Choices: (a) united states (b) mexico (c) countryside (d) atlas

A: The answer must be something that used to do what Google Maps and GPS services do, which is to give directions. Of the above choices, only atlases are used to give directions. So the answer is (d).

Q: Before getting a divorce, what did the wife feel who was doing all the work? Answer Choices: (a) harder (b) anguish (c) bitterness (d) tears (e) sadness

A: The answer should be the feeling of someone getting divorced who was doing all the work. Of the above choices, the closest feeling is bitterness. So the answer is (c).

Strong & automatic?

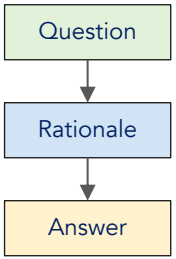
High-level Solution: Auto-CoT

- **Motivation**
  - Eliminate the need for **manually-designed input**
  - Maintain the **strong performance**
- **Key Challenges**
  - How to obtain the **representative questions** to reflect task patterns
  - How to obtain **rationales** to construct demonstrations

How to sample questions?

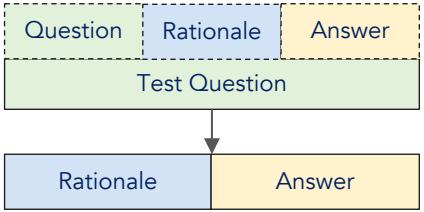
Zero-Shot-CoT

Simple w/ decent performance



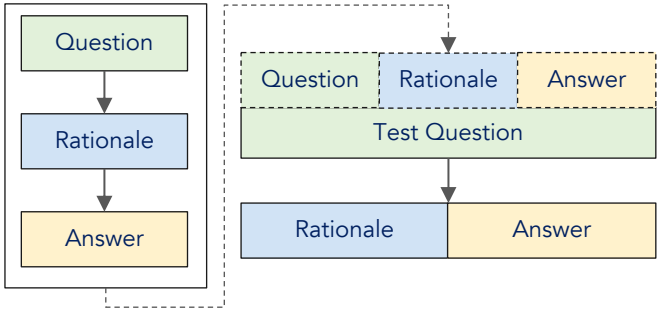
Manual-CoT

Strong but needs manual design



Auto-CoT

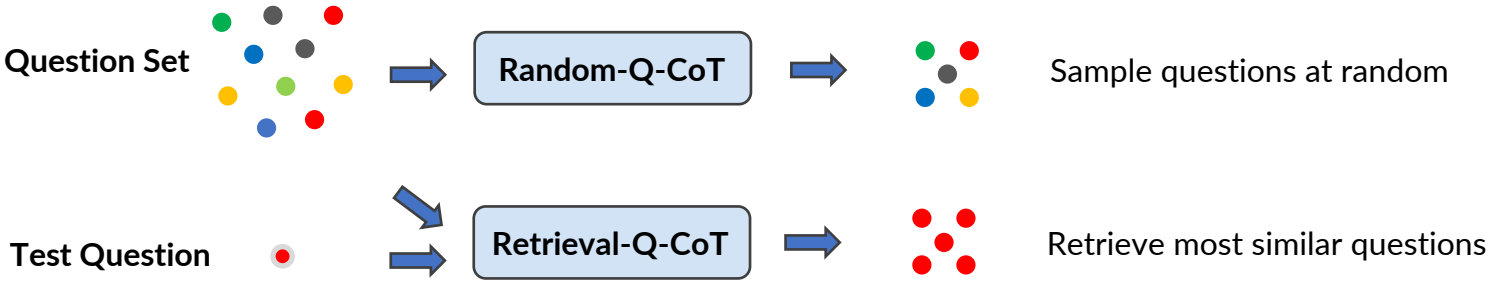
Strong and automatic?



Challenges in Automatic CoT Generation

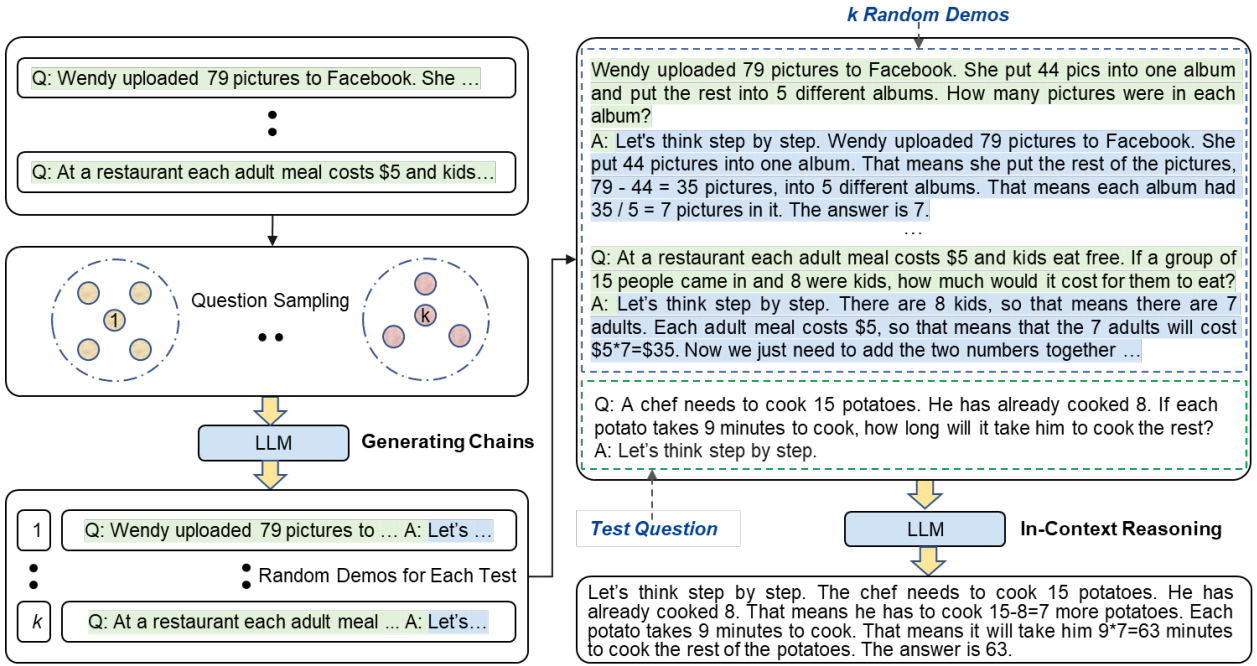
- **General Solution**
  - For each question in a test dataset, sample demo questions from the rest of the questions
  - Generate the rationale for the sampled questions by Zero-Shot-CoT

- **How?**



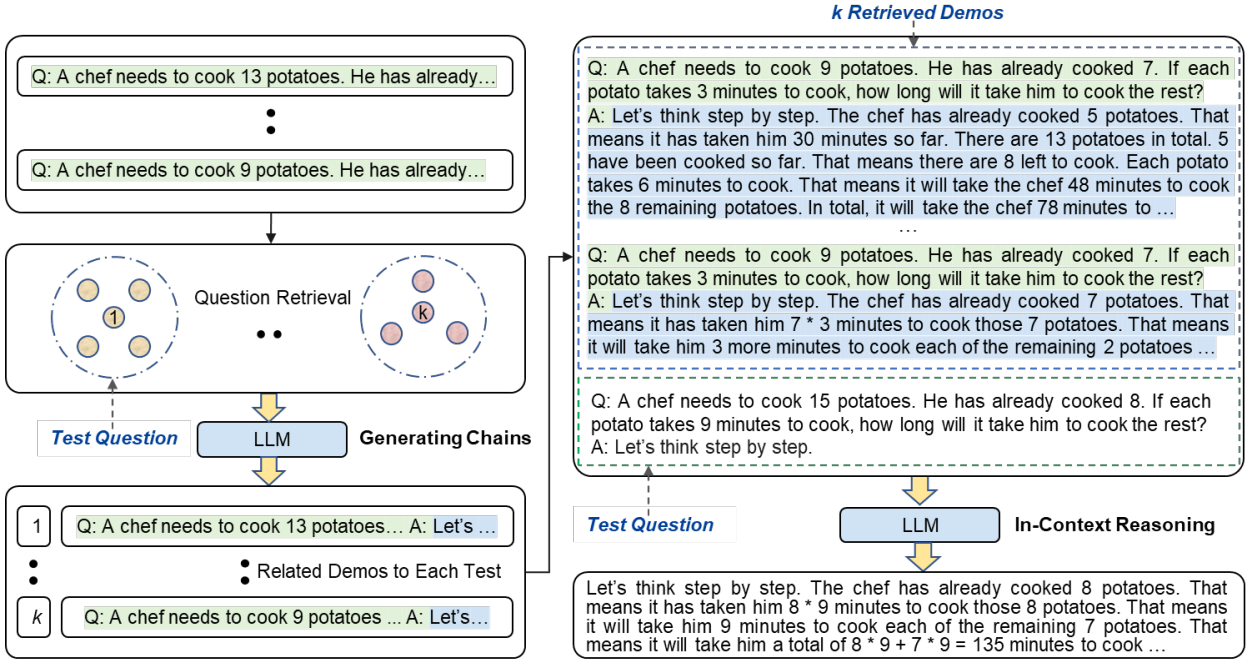
Possible Solution-1: Random-Q-CoT

- **Random-Q-CoT**
  - Randomly sampling q_i^{demo} (i = 1, … , k) from a set of questions
  - Generate the rationale for the sampled questions by Zero-Shot-CoT



Possible Solution-2: Retrieval-Q-CoT

- **Retrieval-Q-CoT**
  - For each question q^{test} in a test dataset, sample demo questions q_i^{demo} (i = 1, … , k) from the rest of the questions
  - Generate the rationale for the sampled questions by Zero-Shot-CoT



Preliminary Experiments

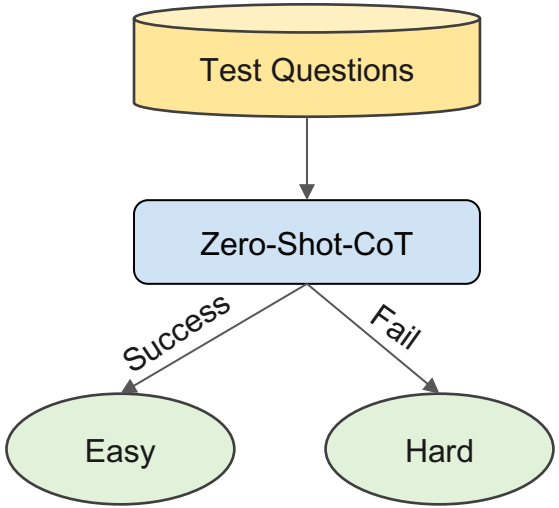
- **Settings**
  - **Engine:** GPT-3.5 (text-davinci-002)
  - **Dataset:** MultiArith, GSM8K, AQuA

Method	MultiArith	GSM8K	AQuA
Zero-Shot-CoT	78.7	40.7	33.5
Manual-CoT	**91.7**	46.9	35.8
Random-Q-CoT	87.1±1.8	47.3±0.5†	36.4±2.2†
Retrieval-Q-CoT	82.4±0.5	48.4±0.6†	39.6±2.4†

- **Findings**
  - With **generated rationales** (MultiArith): Retrieval-Q-CoT is **worst**

Retrieval-Q-CoT **fails with incorrect reasoning chains** by Zero-Shot-CoT

Error Amplification in Retrieval-Q-CoT



Retrieval-Q-CoT	9.7%	46.9%
Random-Q-CoT	9.4%	25.8%

Amplification of errors
through retrieval of erroneous chains

Case Study of Error Amplification in Retrieval-Q-CoT

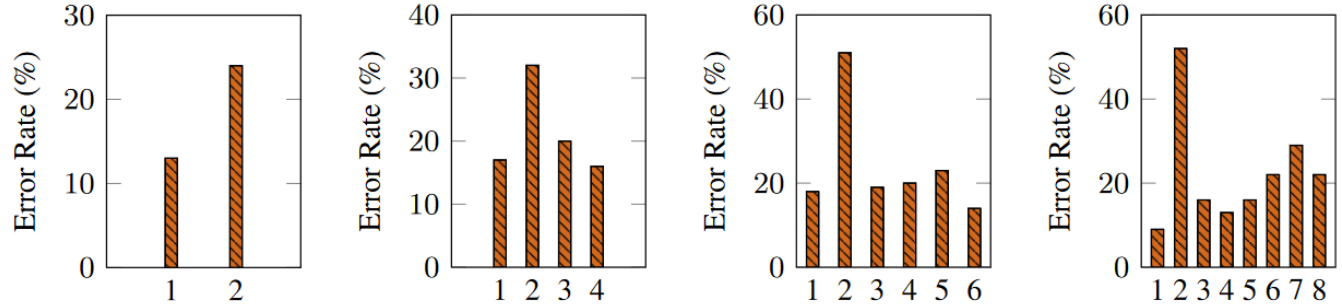
- Retrieval-Q-CoT fails by misunderstanding the meaning of **"the rest"**

Retrieval-Q-CoT	Random-Q-CoT
Q: A chef needs to cook 9 potatoes. He has already cooked 7. If each potato takes 3 minutes to cook, how long will it take him to cook **the rest**? **A:** Let's think step by step. The chef has already cooked 7 potatoes. That means it has taken him 7 * 3 minutes to cook those 7 potatoes. That means it will take him 3 more minutes to cook each of the remaining 2 potatoes. That means it will take him a total of 7 * 3 + 2 * 3 = 21 minutes to cook all 9 potatoes. The answer is 21. ✗ **Q:** A chef needs to cook 13 potatoes. He has already cooked 5. If each potato takes 6 minutes to cook, how long will it take him to cook **the rest**? **A:** Let's think step by step. The chef has already cooked 5 potatoes. That means it has taken him 30 minutes so far. There are 13 potatoes in total. 5 have been cooked so far. That means there are 8 left to cook. Each potato takes 6 minutes to cook. That means it will take the chef 48 minutes to cook the 8 remaining potatoes. In total, it will take the chef 78 minutes to cook all 13 potatoes. The answer is 78. ✗	**Q:** Wendy uploaded 79 pictures to Facebook. She put 44 pics into one album and put the rest into 5 different albums. How many pictures were in each album? **A:** Let's think step by step. Wendy uploaded 79 pictures to Facebook. She put 44 pictures into one album. That means she put the rest of the pictures, 79 - 44 = 35 pictures, into 5 different albums. That means each album had 35 / 5 = 7 pictures in it. The answer is 7. ✓ **Q:** At a restaurant each adult meal costs $5 and kids eat free. If a group of 15 people came in and 8 were kids, how much would it cost for the group to eat? **A:** Let's think step by step. There are 8 kids, so that means there are 7 adults. Each adult meal costs $5, so that means that the 7 adults will cost $5*7=$35. Now we just need to add the two numbers together. $35+8=43$. So it would cost $43 for the group to eat. The answer is 43. ✗
Q: A chef needs to cook 15 potatoes. He has already cooked 8. If each potato takes 9 minutes to cook, how long will it take him to cook **the rest**? **A:** Let's think step by step. The chef has already cooked 8 potatoes. That means it has taken him 8 * 9 minutes to cook those 8 potatoes. That means it will take him 9 minutes to cook each of the remaining 7 potatoes. That means it will take him a total of 8 * 9 + 7 * 9 = 135 minutes to cook all 15 potatoes. The answer is 135. ✗	**Q:** A chef needs to cook 15 potatoes. He has already cooked 8. If each potato takes 9 minutes to cook, how long will it take him to cook **the rest**? **A:** Let's think step by step. The chef needs to cook 15 potatoes. He has already cooked 8. That means he has to cook 15-8=7 more potatoes. Each potato takes 9 minutes to cook. That means it will take him 9*7=63 minutes to cook **the rest** of the potatoes. The answer is 63. ✓

How to detect?

Frequent-Error Cluster

- Clustering: use **k-means** to partition all test questions into **k clusters**
- We find **frequent-error cluster(s)**

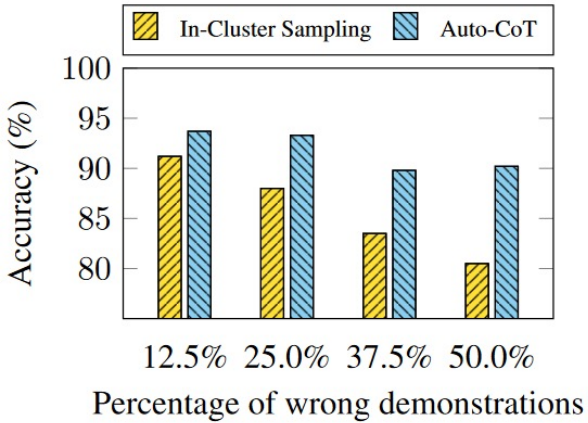


(a) Clusters Num. = 2 (b) Clusters Num. = 4 (c) Clusters Num. = 6 (d) Clusters Num. = 8

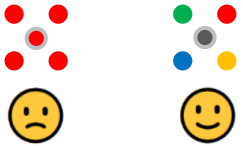
- **Diversity**: higher chance to obtain **good demonstrations** that is not too heavily perturbed (extreme case: 1/8 mistakes)

Mitigation through Diversity

- **A small portion of errors** will not harm reasoning performance



More alternative skills for solving target questions



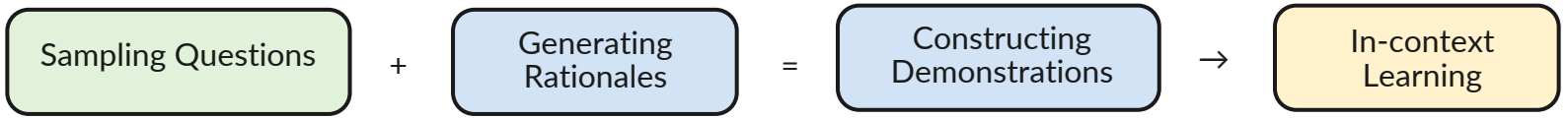
Auto-CoT: Design

- **Principle -> Feasibility**
  - **Questions**: cover the **typical patterns** of the dataset

    -> sample the representative questions via **clustering**
  - **Rationales**: reflect **step-by-step reasoning** processes

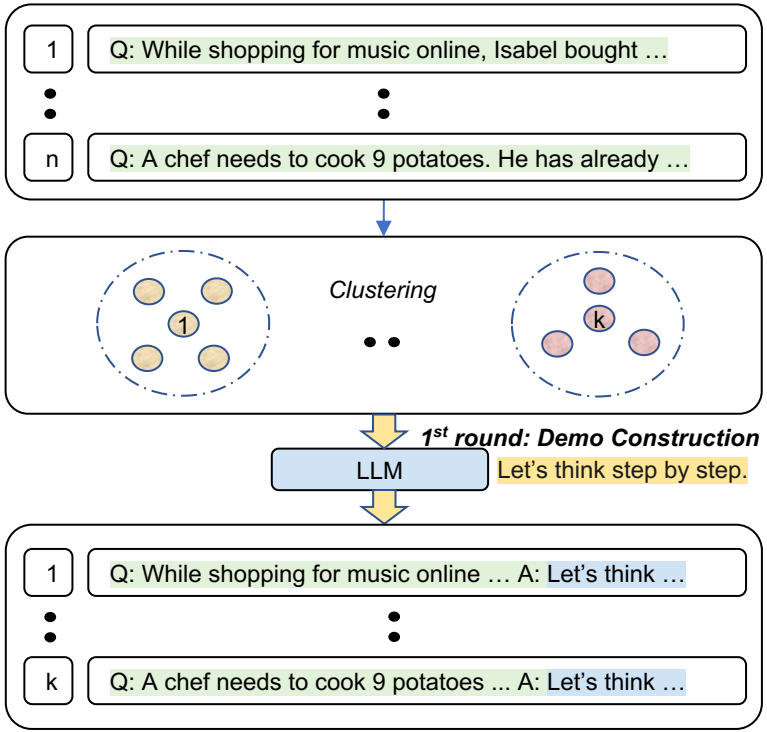
    -> generate rationales by **pre-existing zero-shot CoT** abilities of LLMs

Sampling Criteria	
Diverse	√
Similar	×
Random	×

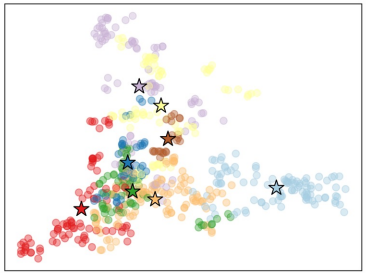


Auto-CoT: Methodology

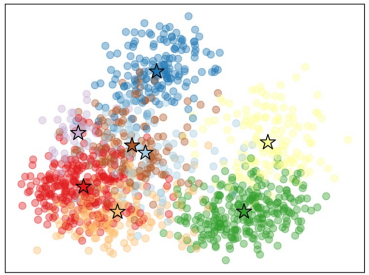
- ❑ **Step-1: Zero-shot Demo Construction**
- ❑ Step-2: Automatic In-context Reasoning



1. **Encoding**: Encode each question with **Sentence-BERT**.
2. **Clustering**: Use K-means to cluster the embeddings into ***k* clusters**.
3. **Sampling**: Select the question **closest to the cluster center** from each cluster.



MultiArith

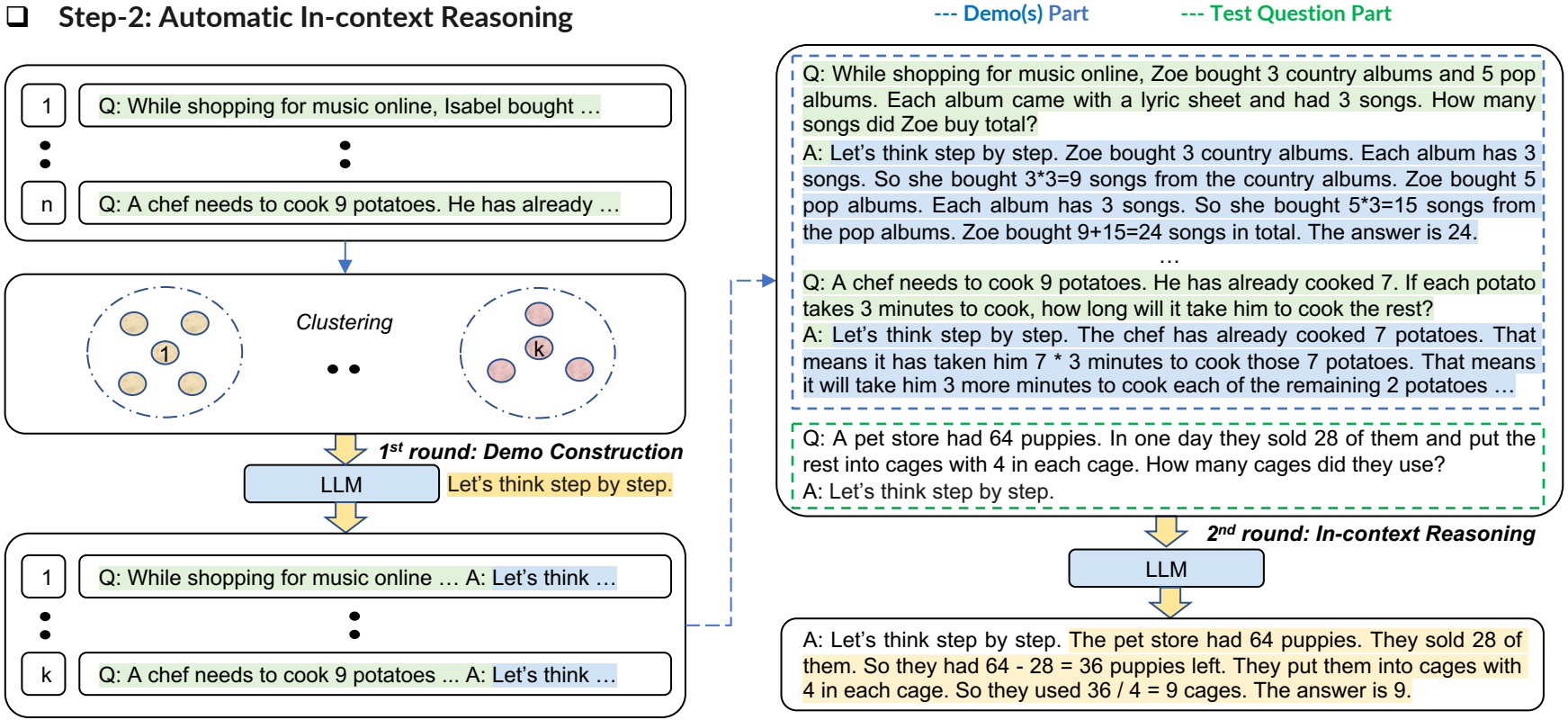


GSM8K

* k is the number of our desired demonstrations

Auto-CoT: Methodology

- ❑ Step-1: Zero-shot Demo Construction
- ❑ **Step-2: Automatic In-context Reasoning**



Experimental Settings

- **Datasets**
  - Our method is evaluated on 10 public benchmark datasets
  - Cover arithmetic, commonsense, and logical reasoning tasks
- **Backbone Model: GPT-3.5 (175B Text-davinci-002)**

Dataset	Number of samples	Average words	Answer Format	Licence
MultiArith	600	31.8	Number	Unspecified
AddSub	395	31.5	Number	Unspecified
GSM8K	1319	46.9	Number	MIT License
AQUA	254	51.9	Multiple choice	Apache-2.0
SingleEq	508	27.4	Number	No License
SVAMP	1000	31.8	Number	MIT License
CSQA	1221	27.8	Multiple choice	Unspecified
StrategyQA	2290	9.6	Yes or No	Apache-2.0
Last Letters	500	15.0	String	Unspecified
Coin Flip	500	37.0	Yes or No	Unspecified

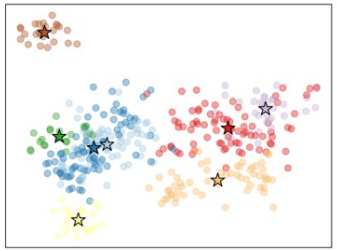
Main Results

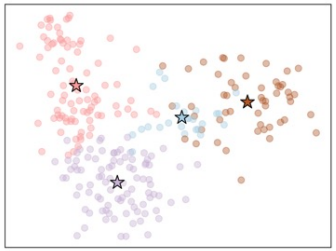
- Auto-CoT method **substantially outperforms** the Zero-Shot-CoT and Manual-CoT baselines
- Auto-CoT is **robust towards randomness**

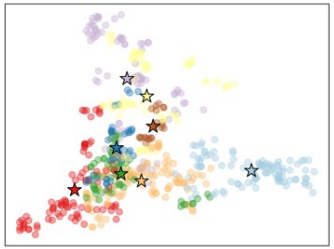
Model	*Arithmetic*						*Commonsense*		*Symbolic*	
	MultiArith	GSM8K	AddSub	AQuA	SingleEq	SVAMP	CSQA	Strategy	Letter	Coin
Zero-Shot	22.7	12.5	77.0	22.4	78.7	58.8	72.6	54.3	0.2	53.8
Zero-Shot-CoT	78.7	40.7	74.7	33.5	78.7	63.7	64.6	54.8	57.6	91.4
Few-Shot	33.8	15.6	83.3	24.8	82.7	65.7	**79.5**	**65.9**	0.2	57.2
Manual-CoT	91.7	46.9	81.3	35.8	86.6	68.9	73.5	65.4	59.0	97.2
Random-Q-CoT	$87.1_{\pm 1.8}$	$40.4_{\pm 0.4}$	$82.7_{\pm 1.3}$	$31.5_{\pm 1.1}$	$81.5_{\pm 0.3}$	$66.7_{\pm 1.8}$	$71.9_{\pm 0.2}$	$58.0_{\pm 0.1}$	$58.2_{\pm 0.3}$	$95.9_{\pm 0.1}$
Auto-CoT	$\mathbf{92.0}^{\Uparrow}_{\pm 1.7}$	$\mathbf{47.9}^{\Uparrow}_{\pm 3.7}$	$\mathbf{84.8}^{\uparrow}_{\pm 2.9}$	$\mathbf{36.5}^{\Uparrow}_{\pm 2.2}$	$\mathbf{87.0}^{\Uparrow}_{\pm 1.2}$	$\mathbf{69.5}^{\uparrow}_{\pm 2.2}$	$74.4^{\uparrow}_{\pm 2.5}$	$65.4^{\Uparrow}_{\pm 0.4}$	$\mathbf{59.7}^{\uparrow}_{\pm 3.2}$	$\mathbf{99.9}^{\Uparrow}_{\pm 0.1}$

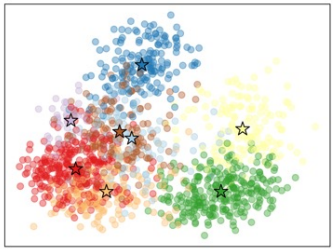
Visualization of Demonstration Clustering

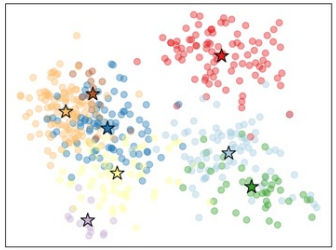
- The number of clusters = num. of desired demos = num. of few-shot demos in Few-Shot CoT.
- The clustered demonstrations are likely to represent generic themes of the datasets.

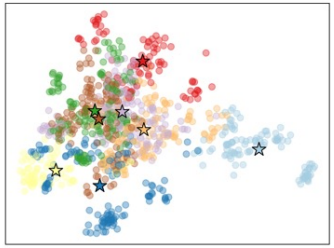

AddSub

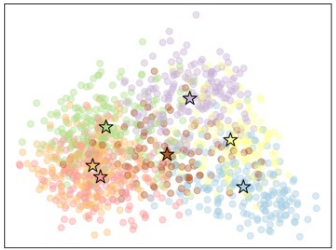

AQUA

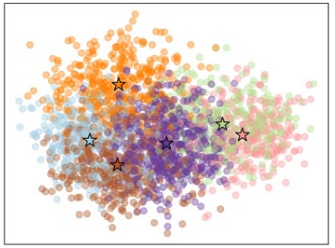

MultiArith


GSM8K


SingleEq


SVAMP

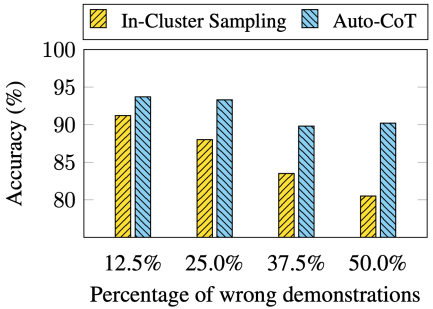

CSQA

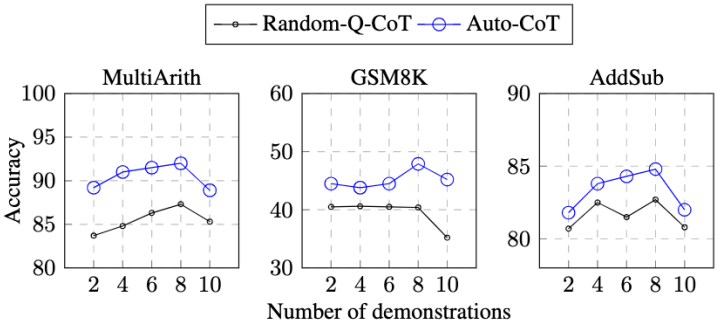

Strategy

Analysis: Different Methods for Obtaining Demonstrations

- Demonstrations are better if they are **closer to each cluster centers**
- Auto-CoT **tolerates** incorrect rationales
- Our method is **robust** against k-means

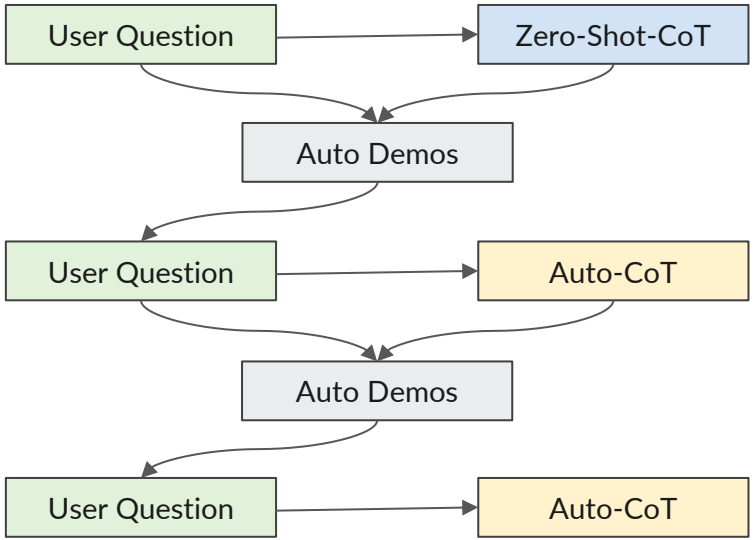
	Method	MultiArith
	Auto-CoT	93.7
↑	In-Cluster Min Dist	93.7
	In-Cluster Random	89.2
	In-Cluster Max Dist	88.7

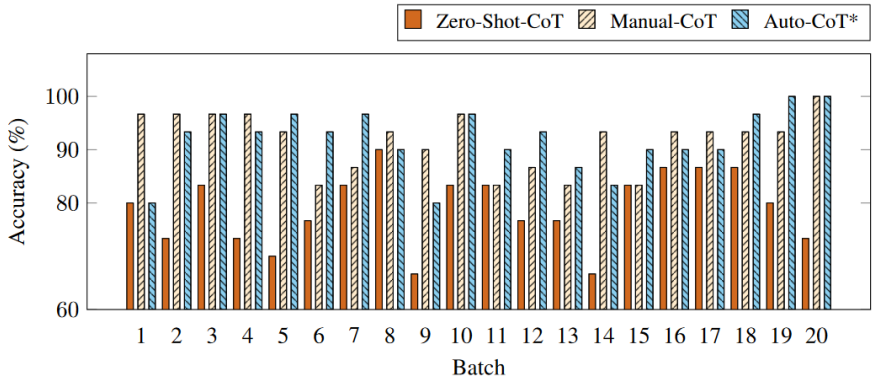




Beyond Auto-CoT: Evolution with Streaming Queries

- Assume we **do not have a full test set**
- Consider a case where **questions arrive in small batches** of, say m=30 questions at a time



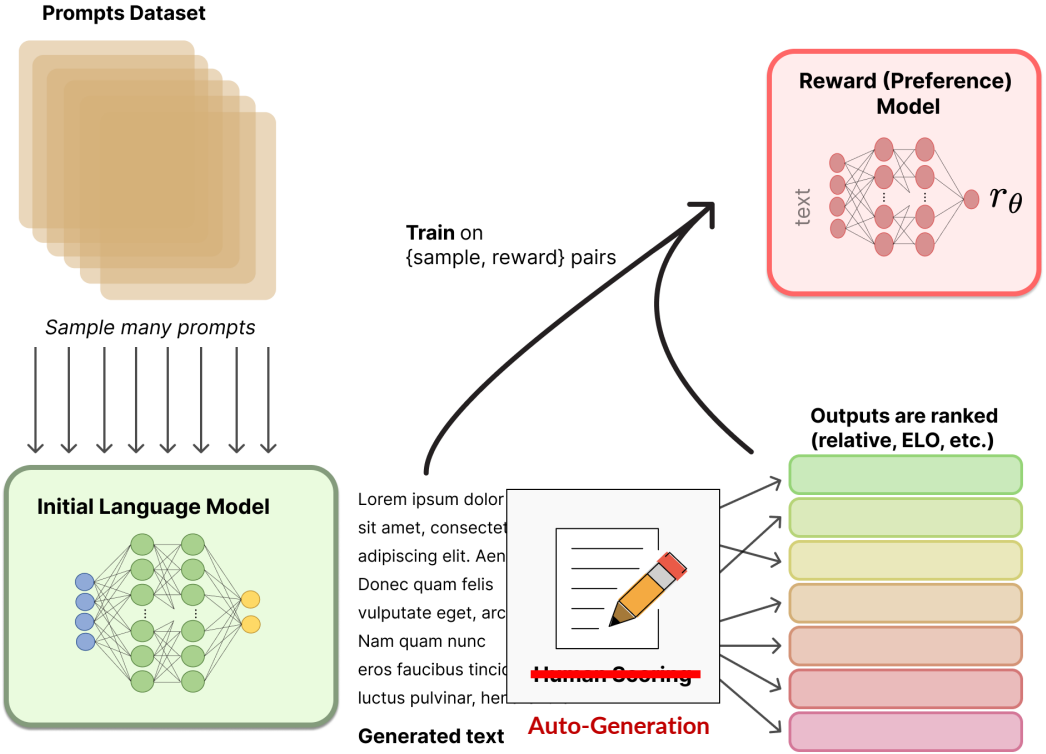


- For batch 1, Auto-CoT* and Zero-Shot-CoT obtain equal accuracy
- From batch 2, Auto-CoT* performs even better than Manual-CoT

Enhancing zero-shot reasoning in an **automatic few-shot** manner!

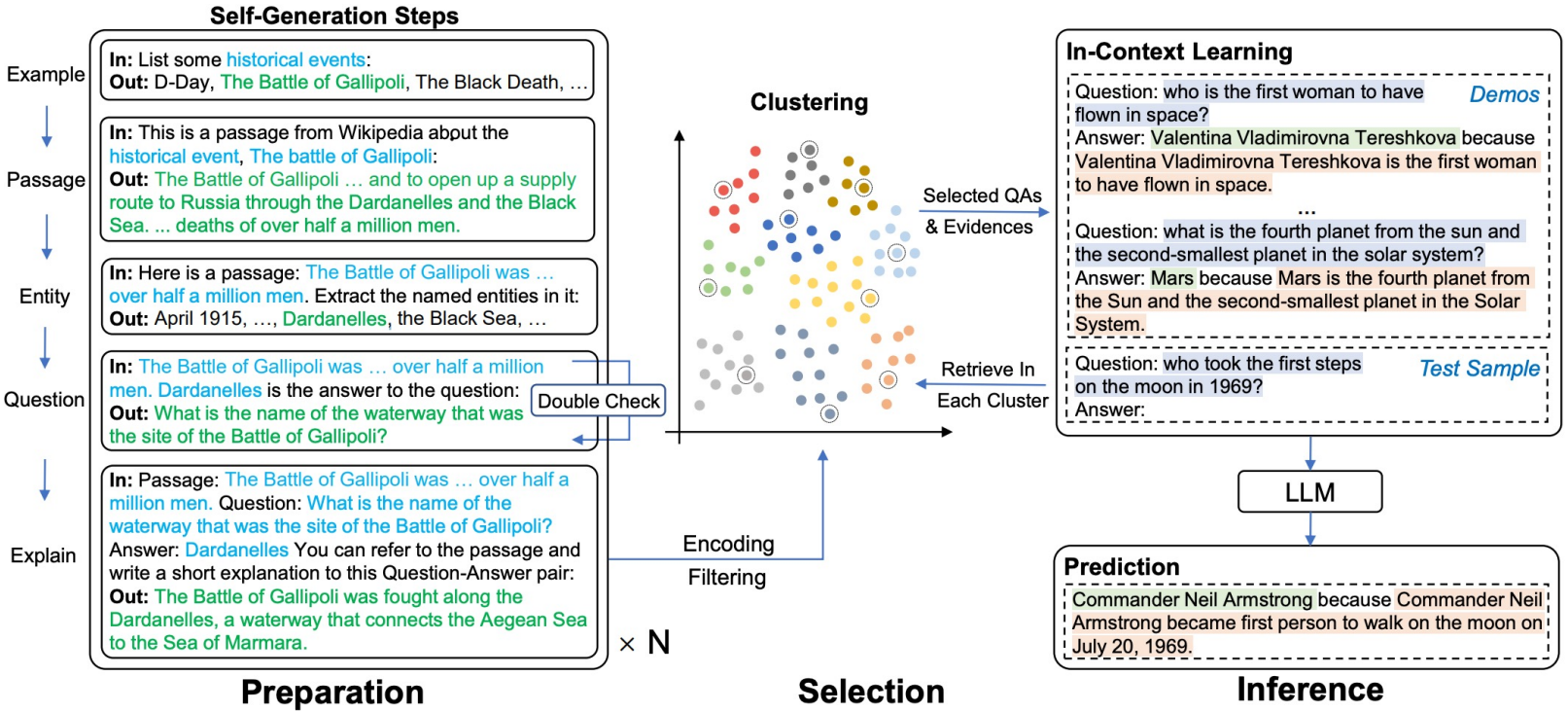
Beyond Auto-CoT: RF with AI Feedback

- Reinforcement learning with AI Feedback



Beyond Auto-CoT: RF with AI Feedback

- ❑ **Free from training data and external knowledge corpus for ODQA**



Li, J., Zhang, Z. and Zhao, H. Self-Prompting Large Language Models for Open-Domain QA. arXiv preprint arXiv:2212.08635. 2022.

Summary: Auto-CoT

- **Contributions**
  - **An automatic CoT method for prompting LLMs**
  - **State-of-the-art results** using the public GPT-3.5 model in the single model setting
- **Insights**
  - LLMs are able to perform complex reasoning with **self-generated demonstrations**
  - LLMs **tolerate** incorrect rationales generated by zero-shot learning
- **Sources**
  - Paper: https://arxiv.org/abs/2210.03493 (ICLR 2023)
  - Code: https://github.com/amazon-science/auto-cot

bilibili

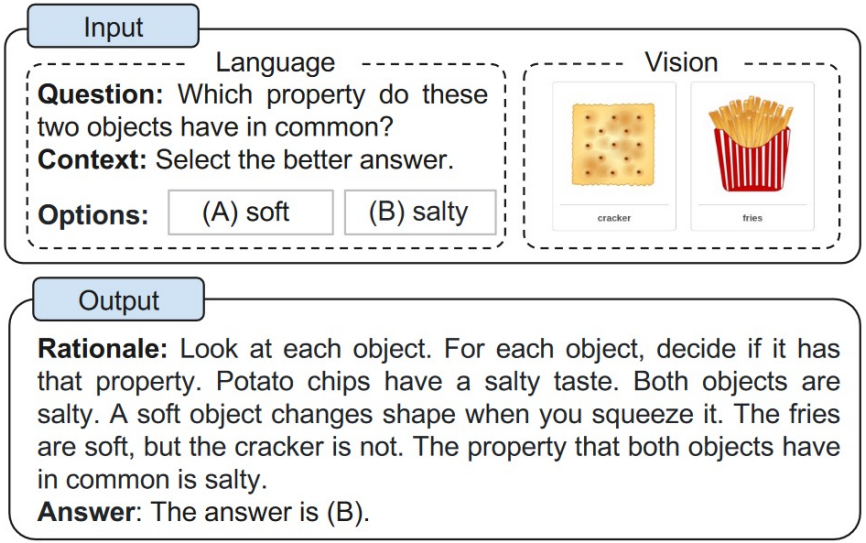
Hands-on learning CoT

Welcome to the world of Multimodal-CoT

Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., and Smola, A.
Multimodal Chain-of-Thought Reasoning in Language Models. arXiv preprint arXiv:2302.00923. 2023.

Background

- **Imagine reading a textbook with no figures or tables**
  - Our ability to knowledge acquisition is greatly strengthened by jointly **modeling diverse data modalities**
  - Existing studies related to CoT reasoning are largely isolated in the **language modality only**

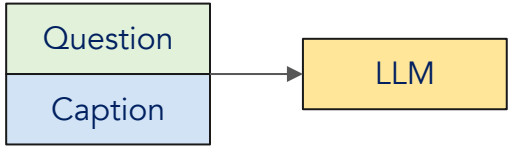


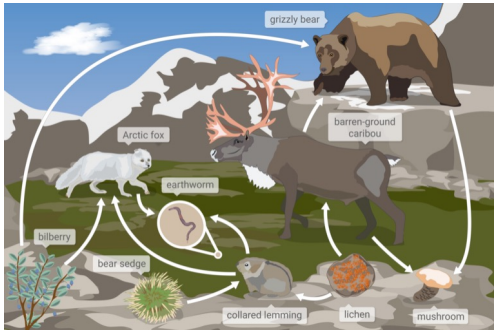
- Two ways to elicit Multimodal-CoT reasoning
  - Prompting LLMs
  - Fine-tuning small models

with less than 1 billion parameters (1B-models)

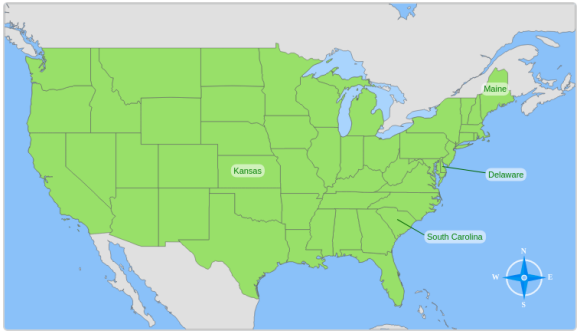
Approach-1: Prompting LLMs

- **Transform the input of different modalities into one modality**
  - Extract the caption of an image by **a captioning model**
  - **Concatenate** the caption with the original language input
- **Mistakes and information loss in the captioning process**





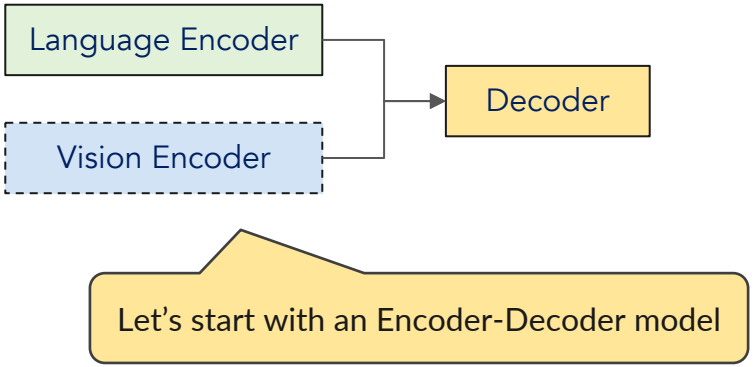
A painting of a horse and a cow



An aerial view of a painting of a forest

Approach-2: Fine-tuning Small Models

- **Fine-tune smaller language models (LMs) by fusing multimodal features**
  - allows the flexibility of adjusting model architectures to incorporate multimodal features



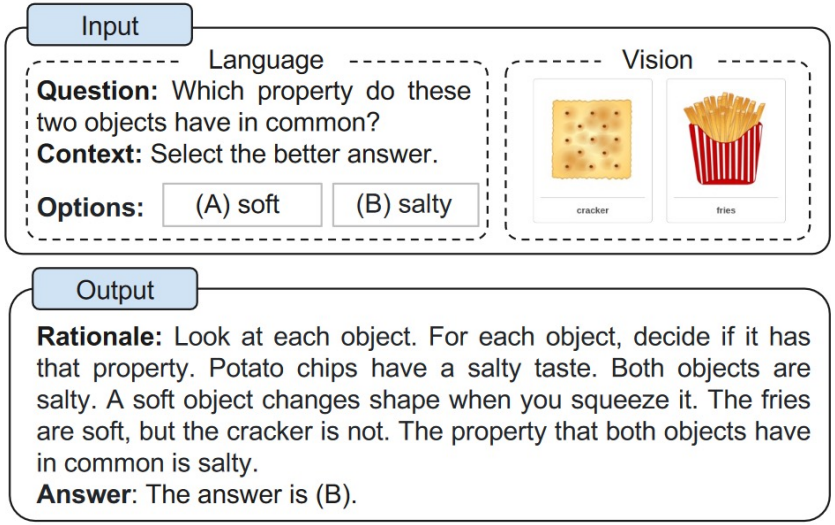
Challenge of Multimodal-CoT

- **Understanding the role of CoT**

Method	Format	Accuracy
No-CoT	QCM→A	80.40
Reasoning	QCM→RA	67.86
Explanation	QCM→AR	69.77

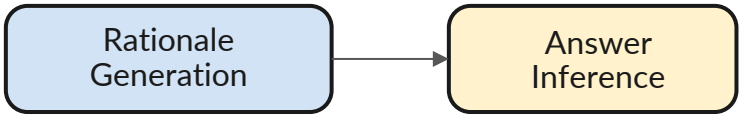
- Surprisingly, **a 12.54% accuracy decrease** (80.40 -> 67.86%) if the model predicts rationales before answers (QCM->RA)

Generated rationales might not contribute to answer inference



Misleading by Hallucinated Rationales

- To dive into how the rationales affect the answer prediction
  - Separate the CoT problem into **two stages**



- The **generated rationale** in the two-stage framework does not improve answer accuracy

Table 3. Two-stage setting of (i) rationale generation (RougeL) and (ii) answer inference (Accuracy).

Method	(i) QCM→ R	(ii) QCMR→ A
Two-Stage Framework	91.76	70.53
w/ Captions	91.85	71.12
w/ Vision Features	96.97	84.91

Misleading by Hallucinated Rationales

- **Case studies**: 50 error cases
  - generate **hallucinated rationales** that mislead the answer inference (64%)

Problem

Question: Will these magnets attract or repel each other?
Context: Two magnets are placed as shown. Hint: Magnets that attract pull together. Magnets that repel push apart.
Options: (A) attract (B) repel

Gold Rationale: Magnets can pull or push on each other without touching. When magnets attract, they pull together. When magnets repel, they push apart. Whether a magnet attracts or repels other magnets depends on the positions of its poles, or ends. Every magnet has two poles, called north and south. Here are some examples of magnets. The north pole of each magnet is marked N, and the south pole is marked S. If different poles are closest to each other, the magnets attract. The magnets in the pair below attract. If the same poles are closest to each other, the magnets repel. The magnets in both pairs below repel. Will these magnets attract or repel? To find out, look at which poles are closest to each other. The north pole of one magnet is closest to the south pole of the other magnet. Poles that are different attract. So, these magnets will attract each other.
Answer: The answer is (A).

Baseline

Generated Rationale: Magnets can pull or push on each other without touching. When magnets attract, they pull together. When magnets repel, they push apart. Whether a magnet attracts or repels other magnets depends on the positions of its poles, or ends. Every magnet has two poles, called north and south. Here are some examples of magnets. The north pole of each magnet is marked N, and the south pole is marked S. If different poles are closest to each other, the magnets attract. The magnets in the pair below attract. If the same poles are closest to each other, the magnets repel. The magnets in both pairs below repel. Will these magnets attract or repel? To find out, look at which poles are closest to each other. The south pole of one magnet is closest to the south pole of the other magnet. Poles that are the same repel. So, these magnets will repel each other.
Answer: The answer is (B).

+ Vision Features

Generated Rationale: Magnets can pull or push on each other without touching. When magnets attract, they pull together. When magnets repel, they push apart. Whether a magnet attracts or repels other magnets depends on the positions of its poles, or ends. Every magnet has two poles, called north and south. Here are some examples of magnets. The north pole of each magnet is marked N, and the south pole is marked S. If different poles are closest to each other, the magnets attract. The magnets in the pair below attract. If the same poles are closest to each other, the magnets repel. The magnets in both pairs below repel. Will these magnets attract or repel? To find out, look at which poles are closest to each other. The north pole of one magnet is closest to the south pole of the other magnet. Poles that are different attract. So, these magnets will attract each other.
Answer: The answer is (A).

Lack of Information Results in Hallucinated

Will these magnets attract or repel each other?

Will these magnets attract or repel each other?

Baseline

$X_1 = X_2 \rightarrow \{Y_1, Y_2\}$

The **south pole** of one magnet is closest to the **south pole** of the other magnet.

The **north pole** of one magnet is closest to the **south pole** of the other magnet.

Will these magnets attract or repel each other?

Will these magnets attract or repel each other?

+ Vision

$X_1 \rightarrow Y_1$
$X_2 \rightarrow Y_2$

The **south pole** of one magnet is closest to the **south pole** of the other magnet.

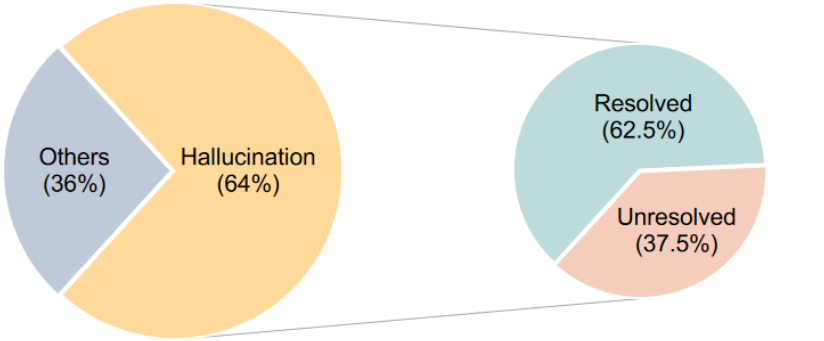
The **north pole** of one magnet is closest to the **south pole** of the other magnet.

Multimodality Contributes to Effective Rationales

- **Solutions**
  - **Image captioning**
  - **Vision features** (i.e., DETR)

Table 3. Two-stage setting of (i) rationale generation (RougeL) and (ii) answer inference (Accuracy).

Method	(i) QCM→ R	(ii) QCMR→ A
Two-Stage Framework	91.76	70.53
w/ Captions	91.85	71.12
w/ Vision Features	96.97	84.91

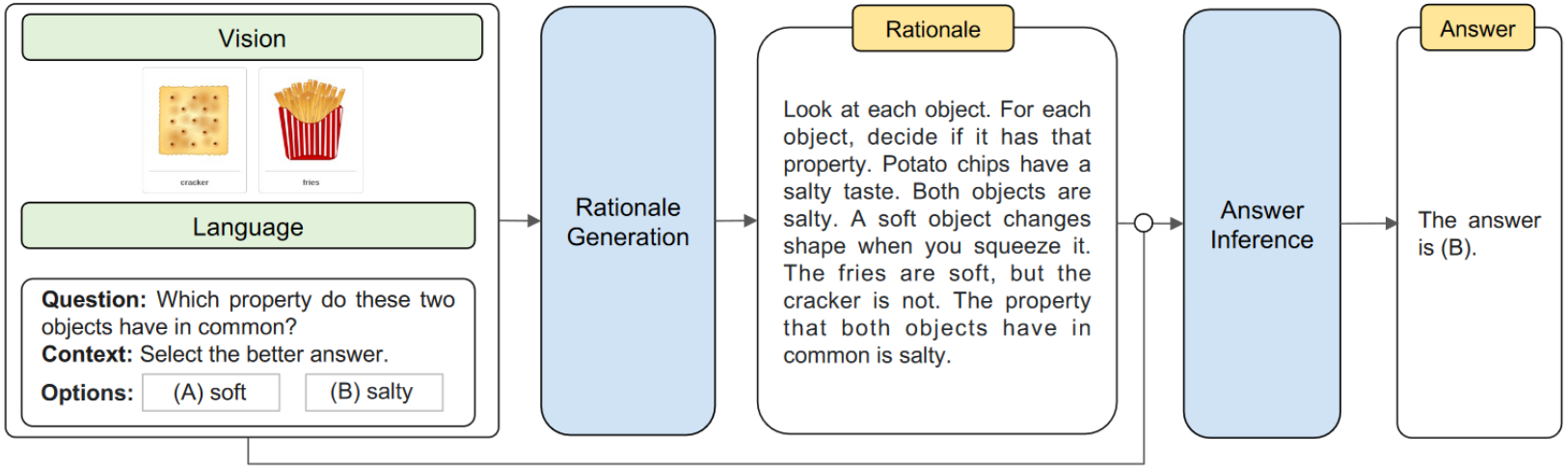


(a) ratio of hallucination mistakes (b) correction rate w/ vision features

- **Findings**
  - **Hallucination is mitigated** (solve rate: 62.5%)
  - **Vision features** are **indeed beneficial** for generating effective rationales
  - The **two-stage method** (QCMR→A) achieves better performance than one-stage methods

Multimodal-CoT: Overview

- **Hypothesis**: due to a lack of necessary vision contexts for performing effective Multimodal-CoT
- **Two stages**
  - share the same model architecture but differ in the input X and output Y



Multimodal-CoT: Architecture

- **Objective**: Given the language input $X_{\text{language}} \in \{X^1_{\text{language}}, X^2_{\text{language}}\}$ vision input X_{vision}, compute the probability of generating target text Y (either the rationale or the answer) by

$$p(Y|X_{\text{language}}, X_{\text{vision}}) = \prod_{i=1}^{N} p_\theta \left(Y_i \mid X_{\text{language}}, X_{\text{vision}}, Y_{<i}\right)$$

- **Model**
  - **Encoding**

$$\begin{aligned} H_{\text{language}} &= \text{LanguageEncoder}(X_{\text{language}}), \\ H_{\text{vision}} &= W_h \cdot \text{VisionExtractor}(X_{\text{vision}}), \end{aligned}$$

  - **Interaction**

$$\begin{aligned} H^{\text{attn}}_{\text{vision}} &= \text{Softmax}(\frac{QK^\top}{\sqrt{d_k}})V, \\ \lambda &= \text{Sigmoid}(W_l H_{\text{language}} + W_v H^{\text{attn}}_{\text{vision}}), \\ H_{\text{fuse}} &= (1-\lambda) \cdot H_{\text{language}} + \lambda \cdot H^{\text{attn}}_{\text{vision}}, \end{aligned}$$

Algorithm 1 Multimodal-CoT

Input: Language input X^1_{language}, vision input X_{vision}
Output: Generated rationale R, inferred answer A

1: Construct the input $X = \{X_{\text{language}}, X_{\text{vision}}\}$
2: Generate rationale $R = F(X)$ using the model $F(\cdot)$
3: Append the rationale R to the original language input $X^2_{\text{language}} = X^1_{\text{language}} \circ R$.
4: Construct new input $X' = \{X^2_{\text{language}}, X_{\text{vision}}\}$
5: Infer the answer A by conditioning on the new input, $A = F(X')$.
6: **procedure** F(X)
7: Encode the language and vision inputs H_{language} and H_{vision}, respectively
8: Build the interaction between language and vision features by attention $H^{\text{attn}}_{\text{vision}}$
9: Fuse H_{language} and $H^{\text{attn}}_{\text{vision}}$ by a gated fusion mechanism to have H_{fuse}
10: Feed H_{fuse} to the decoder to obtain the target prediction Y
11: **return** Y
12: **end procedure**

- **Decoding**: the fused output is fed into the Transformer decoder to predict the target Y

Experimental Settings

- **ScienceQA**
  - 21k multimodal questions with domain diversity across 3 subjects, 26 topics, 127 categories, and 379 skills
  - The benchmark dataset is split into training, validation, and test splits with 12726, 4241, and 4241 examples

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

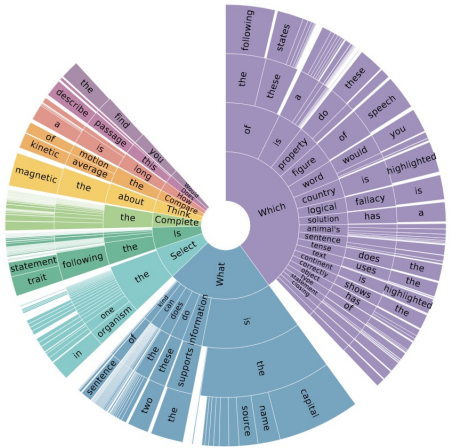
Verbs
Verb tense

Capitalization
Formatting

Punctuation
Fragments

Phonology
Rhyming

Reference
Research skills



- **Backbone Models**
  - UnifiedQA (default)
  - FlanT5

Main Results

- Mutimodal-CoT **outperforms previous SoTA** (GPT-3.5) by 16.51% and surpasses human performance
- Using **image features is more effective** compared with existing UnifiedQA and GPT-3.5 that **leverage image captions**

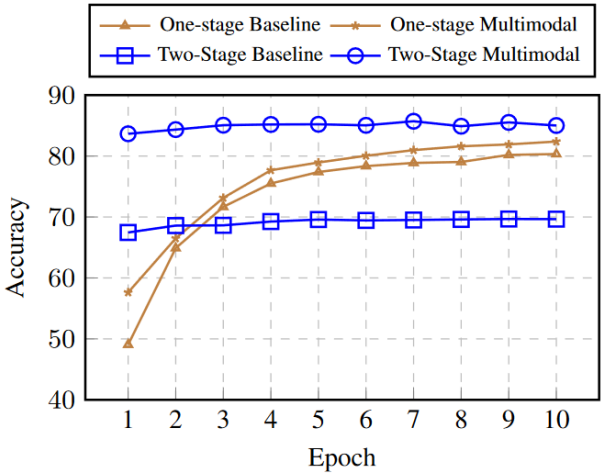
Model	Size	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
Human	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
MCAN (Yu et al., 2019)	95M	56.08	46.23	58.09	59.43	51.17	55.40	51.65	59.72	54.54
Top-Down (Anderson et al., 2018)	70M	59.50	54.33	61.82	62.90	54.88	59.79	57.27	62.16	59.02
BAN (Kim et al., 2018)	112M	60.88	46.57	66.64	62.61	52.60	65.51	56.83	63.94	59.37
DFAF (Gao et al., 2019)	74M	64.03	48.82	63.55	65.88	54.49	64.11	57.12	67.17	60.72
ViLT (Kim et al., 2021)	113M	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90	61.14
Patch-TRM (Lu et al., 2021)	90M	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50	61.42
VisualBERT (Li et al., 2019)	111M	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92	61.87
UnifiedQA$_{\text{Base}}$ (Khashabi et al., 2020)	223M	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00	70.12
UnifiedQA$_{\text{Base}}$ w/ CoT (Lu et al., 2022a)	223M	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82	74.11
GPT-3.5 (Chen et al., 2020)	175B	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT (Lu et al., 2022a)	175B	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
Mutimodal-CoT$_{\text{Base}}$	223M	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
Mutimodal-CoT$_{\text{Large}}$	738M	**95.91**	**82.00**	**90.82**	**95.26**	**88.80**	**92.89**	**92.44**	**90.31**	**91.68**

Analysis

- Both **two-stage framework** and **vision features** help

Model	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
Multimodal-CoT	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
w/o Two-Stage Framework	80.99	87.40	81.91	80.25	78.83	83.62	82.78	82.20	82.57
w/o Vision Features	71.09	70.75	69.18	71.16	65.84	71.57	71.00	69.68	70.53

- **Multimodality boosts convergence**



- **Using vision features generally achieves better performance**

Method	One-stage	Two-Stage
w/ CLIP	81.21	84.81
w/ DETR	82.57	84.91
w/ ResNet	80.97	84.77

- **General effectiveness across backbone models**

Method	Size	Language Only	Mutimodal-CoT
UnifiedQA$_{\text{Base}}$	223M	80.40	84.91
UnifiedQA$_{\text{Large}}$	738M	83.60	91.68
FLAN-T5$_{\text{Base}}$	248M	83.42	85.85
FLAN-T5$_{\text{Large}}$	783M	85.19	93.02

Case Studies

- Randomly picked up **50 samples whose answers were correct** and **50 samples whose answers were incorrect**

Answer	CoT Category	Percentage (%)
Correct	CoT is correct	90
	CoT is incorrect	10
Incorrect	Commonsense Mistake	82
	Logical Mistake	12
	CoT is correct	6

Case Studies: correct answers

- **Correct samples** (i.e., whose answers are correct) contain a certain amount of **incorrect chain-of-thought** (10%)
- The model is **robust to some extent** — it can predict the correct answer by ignoring incorrect rationales

Problem

Question: Can Leopardus wiedii cells make their own food?

Context: This organism is Leopardus wiedii. It is a member of the animal kingdom. Leopardus wiedii is commonly called a margay. Margays live in the forest in Central and South America. Margays are nocturnal. A nocturnal animal sleeps most of the day and is awake at night.

Options: (A) no (B) yes

Vision

Rationale : In the past, scientists classified living organisms into two groups: plants and animals. Over the past 300 years, scientists have discovered many more types of organisms. Today, many scientists classify organisms into six broad groups, called kingdoms. Organisms in each kingdom have specific traits. The table below shows some traits used to describe each kingdom. | Bacteria | Archaea | Protists | Fungi | Animals | Plants. How many cells do they have? | one | one | one or many | one or many | many | many Do their cells have a nucleus? | no | no | yes | yes | yes | yes Can their cells make food? | some species can | some species can | some species can | no | no | yes Leopardus wiedii is an animal. Animal cells cannot make their own food. Animals get their food by digesting other organisms.

Answer: The answer is (A).

Prediction

Rationale: Leopard the past, scientists classified living organisms into two groups: plants and animals. Over the past 300 years, scientists have discovered many more types of organisms. Today, many scientists classify organisms into six broad groups, called kingdoms. Organisms in each kingdom have specific traits. The table below shows some traits used to describe each kingdom. | Bacteria | Archaea | Protists | Fungi | Animals | Plants. How many cells do they have? | one | one | one or many | one or many | many | many. Do their cells have a nucleus? | no | no | yes | yes | yes | yesnCan their cells make food? | some species can | some species can | some species can | no | no | yes Leopardus wiedii is an animal. Animal cells cannot make their own food. Animals cannot their food by digesting other organisms.

Answer: The answer is (A).

Case Studies: incorrect answers

- **Commonsense mistake** in the CoT is the most frequent error type (88%)
- The model often makes commonsense mistakes when answering the questions requires commonsense knowledge, e.g., **understand maps** and **counting numbers** in the images, and **utilizing the alphabet**

Problem

Question: Which ocean is highlighted?

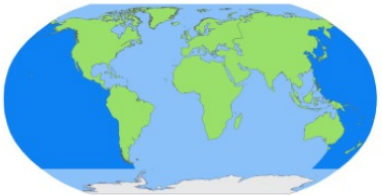
Context: N/A

Options: (A) the Atlantic Ocean (B) the Indian Ocean (C) the Pacific Ocean (D) the Southern Ocean

Rationale : Oceans are huge bodies of salt water. The world has five oceans. All of the oceans are connected, making one world ocean. This is the Pacific Ocean.

Answer: The answer is (C).

Vision



Prediction

Rationale: Oceans are huge bodies of salt water. The world has five oceans. All of the oceans are connected, making one world ocean. This is the Indian Ocean.

Answer: The answer is (B).

Case Studies: incorrect answers

- The other type of mistake is a **logical mistake** (12%), with contradictions in the reasoning chains

Problem

Question: Is this a run-on sentence? Right-handedness is controlled by genetics, interestingly enough, the same genes are also responsible for the brain becoming more specialized at certain tasks.
Context: N/A
Options: (A) yes (B) no

Vision: N/A

Rationale : A sentence is a group of words that expresses a complete thought. The band I'm in has been rehearsing daily because we have a concert in two weeks. A run-on sentence is formed when two sentences are run together, joined by just a comma or by no punctuation at all. If only a comma is used, the run-on is called a comma splice … The band I'm in has been rehearsing daily since we have a concert in two weeks. This is a run-on sentence. It is a comma splice formed from two sentences run together, joined by just a comma. Right-handedness is controlled by genetics, interestingly enough, the same genes are also responsible for the brain becoming more specialized at certain tasks.
Answer: The answer is (A).

Prediction

Rationale: : A sentence is a group of words that expresses a complete thought. The band I'm in has been rehearsing daily because we have a concert in two weeks. A run-on sentence is formed when two sentences are run together, joined by just a comma or by no punctuation at all. If only a comma is used, the run-on is called a comma splice … The band I'm in has been rehearsing daily since we have a concert in two weeks. This is not a run-on sentence. It is formed a comma splice formed from two sentences run together, joined without just a comma. Right-handedness is controlled by genetics, interestingly enough, the same genes are also responsible for the brain becoming more specialized at certain tasks.
Answer: The answer is (B).

Case Studies: incorrect answers

- There are cases with incorrect answers while their **CoT are correct** (6%) but might not be necessarily related to answers

Problem

Question: Which trait did Palaeopython have? Select the trait you can observe on the fossil.

Context: This picture shows a fossil of an ancient animal called Palaeopython. Palaeopython lived in trees and could grow more than six feet long.

Vision

Options: (A) black stripes on its skin (B) large fins on its body (C) a long, thin body

Rationale : The way an organism looks or acts is called a trait. Scientists use fossils to learn more about the traits of ancient organisms. Fossils can preserve the remains of body parts and activities. A fossil of a body part, such as a tail or a wing, can tell you what an organism looked like. A fossil of an organism's activities, such as a burrow or a footprint, can tell you about the organism's behavior. Here are three examples of fossils and the traits that you can observe from them: This is a fossil of an animal. This fossil tells you that the animal had a spiral-shaped shell. This is a fossil of a plant. This fossil tells you that the plant had small leaves arranged in a branched pattern. This is a fossil of an animal's footprint. This fossil tells you that the animal could walk on land. An organism's fossil may not show all of the organism's traits. This is because most body parts are destroyed during fossil formation. When an organism's body turns into a fossil, only a few body parts are usually preserved.
Answer: The answer is (C).

Prediction

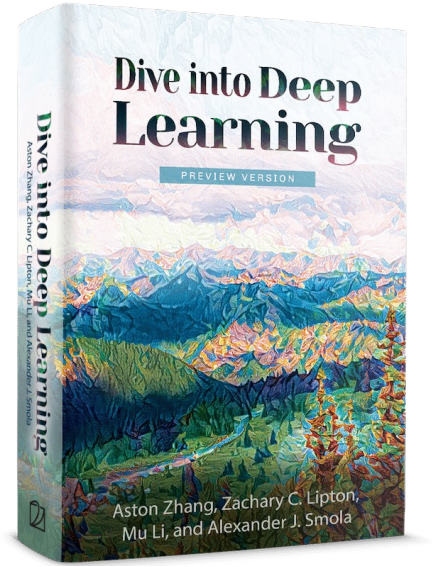
Rationale : The way an organism looks or acts is called a trait. Scientists use fossils to learn more about the traits of ancient organisms. Fossils can preserve the remains of body parts and activities. A fossil of a body part, such as a tail or a wing, can tell you what an organism looked like. A fossil of an organism's activities, such as a burrow or a footprint, can tell you about the organism's behavior. Here are three examples of fossils and the traits that you can observe from them: This is a fossil of an animal. This fossil tells you that the animal had a spiral-shaped shell. This is a fossil of a plant. This fossil tells you that the plant had small leaves arranged in a branched pattern. This is a fossil of an animal's footprint. This fossil tells you that the animal could walk on land. An organism's fossil may not show all of the organism's traits. This is because most body parts are destroyed during fossil formation. When an organism's body turns into a fossil, only a few body parts are usually preserved.
Answer: The answer is (B).

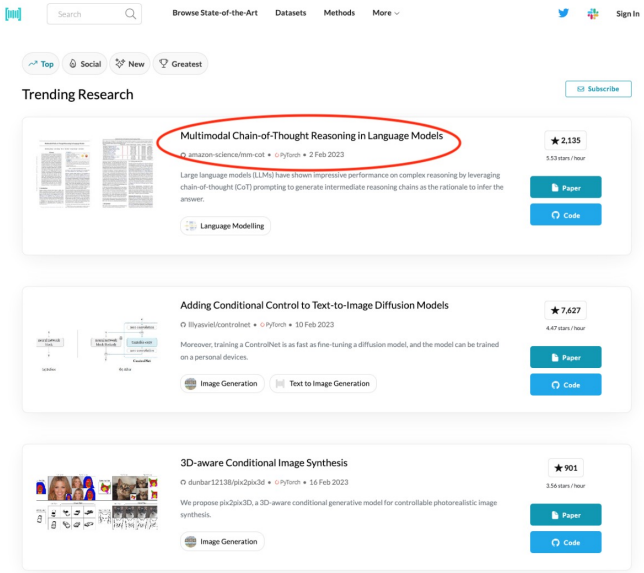
Summary: Multimodal-CoT

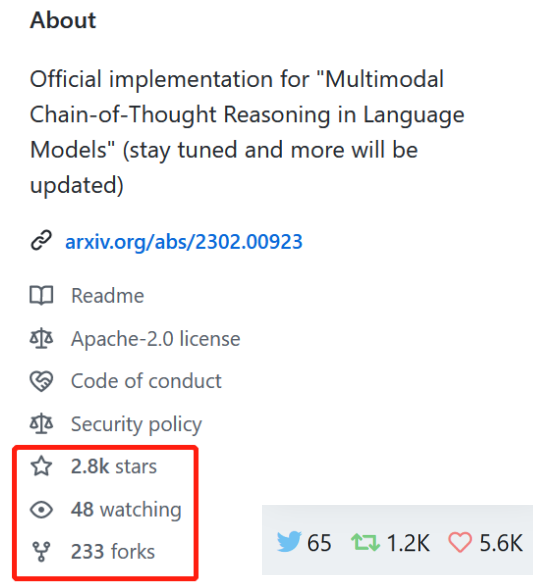
- **Problem**
  - **Multimodal chain of thought reasoning** for large language models (LLMs)
- **Key Points**
  - Multimodal-CoT: use **vision features** to generate **more effective rationales**
  - Our method **surpasses previous SoTA by 16%** in accuracy on the ScienceQA benchmark
  - Error analysis: potential to leverage **more effective vision features**, inject **commonsense knowledge**, and apply **filtering mechanisms**
- **Sources**
  - Paper: https://arxiv.org/abs/2302.00923
  - Code: https://github.com/amazon-science/mm-cot

Broad Impact

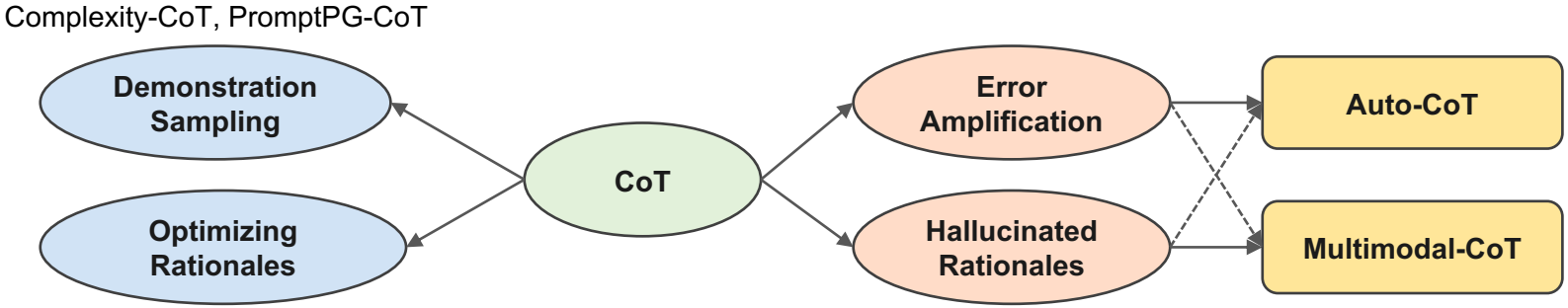
- Both Auto-CoT and Multimodal-CoT have been featured in ***Dive into Deep Learning***
  - Adopted at 400 universities from 60 countries
- Multimodal-CoT becomes a **Top Trending Research** in paperwithcode







Discussion



- Zhang, Z., Zhang, A., Li, M. and Smola, A. Automatic chain of thought prompting in large language models. The Eleventh International Conference on Learning Representations (ICLR). 2023.
  - Paper: https://arxiv.org/abs/2210.03493
  - Code: https://github.com/amazon-science/auto-cot ☆ 270 27
- Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., and Smola, A. Multimodal Chain-of-Thought Reasoning in Language Models. arXiv preprint arXiv:2302.00923. 2023.
  - Paper: https://arxiv.org/abs/2302.00923
  - Code: https://github.com/amazon-science/mm-cot ☆ 2.8k 233

Open Questions

- **Philosophy**: How does CoT capability emerge in LLMs?
  - How to make small models CoT reasoners, too?
- **Technique**: How does ICL/CoT affect the answer inference?
  - How to avoid incorrect rationales?
  - How to fix the mistakes in the rationales?
- **Application**: How would CoT techniques empower general tasks?
  - For Open-domain QA
  - For summarization
  - ...

Thanks & QA

Zhuosheng Zhang
zhangzs@sjtu.edu.cn
https://bcmi.sjtu.edu.cn/~zhangzs