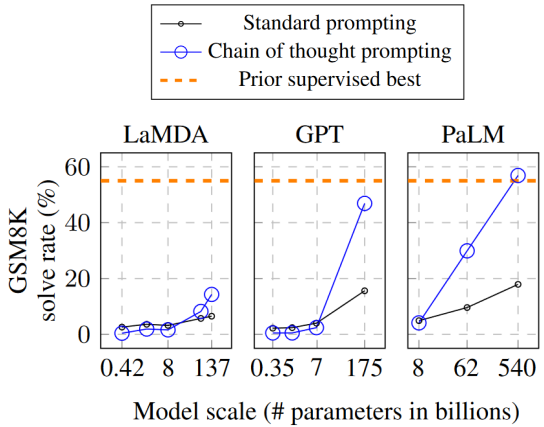
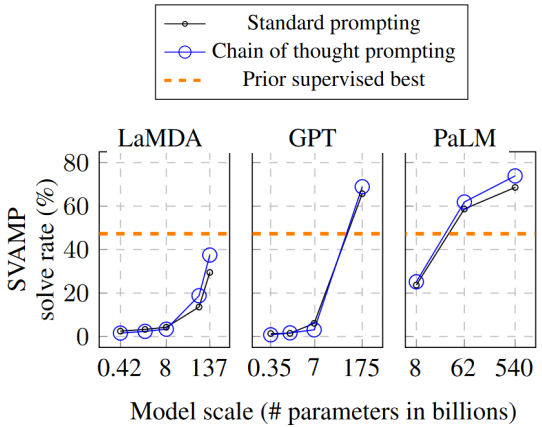
AUTOMATIC CHAIN OF THOUGHT PROMPTING IN LARGE LANGUAGE MODELS

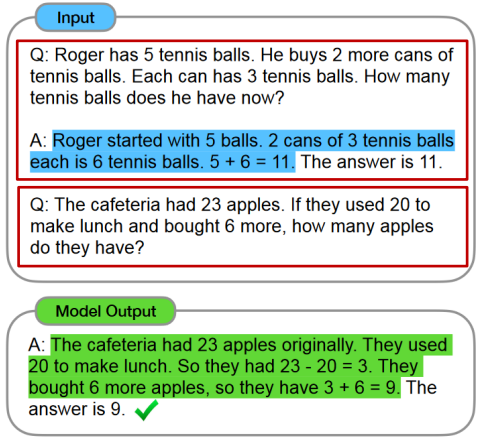
Zhuosheng Zhang[†,*] **Aston Zhang**[‡], **Mu Li**[‡], **Alex Smola**[‡]
[†]Shanghai Jiao Tong University, [‡]Amazon Web Services

Background

- Tasks: **multi-step reasoning tasks**, e.g., math word problems, commonsense reasoning, logical reasoning, etc.
- Large Language Models shows **emergent abilities** of solving challenging reasoning problems with decent prompts
- **Chain of thought (CoT)** prompting enables LLMs to achieve state of the art accuracy on math word problems
  - With only a few exemplars (e.g., eight)
  - Without gradient updates



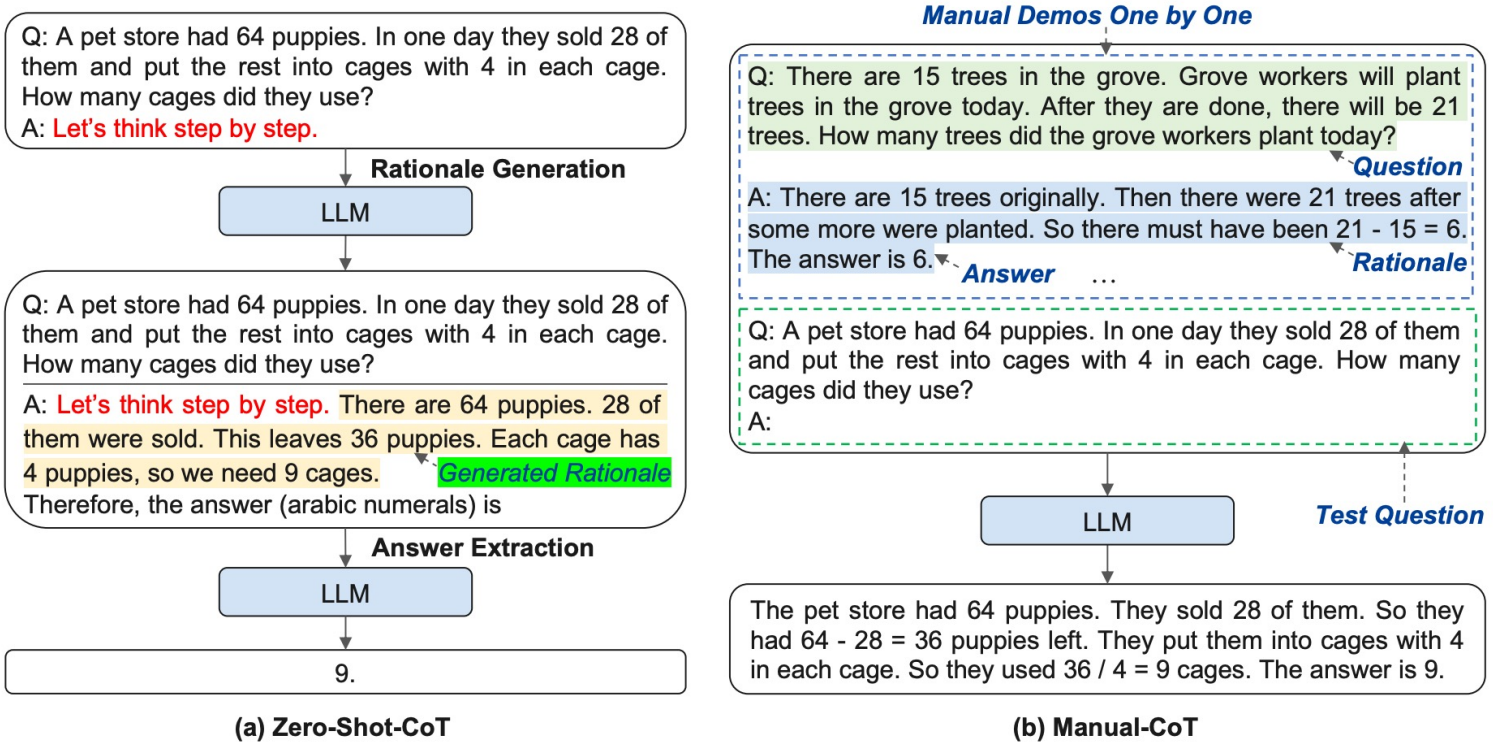




Related Work

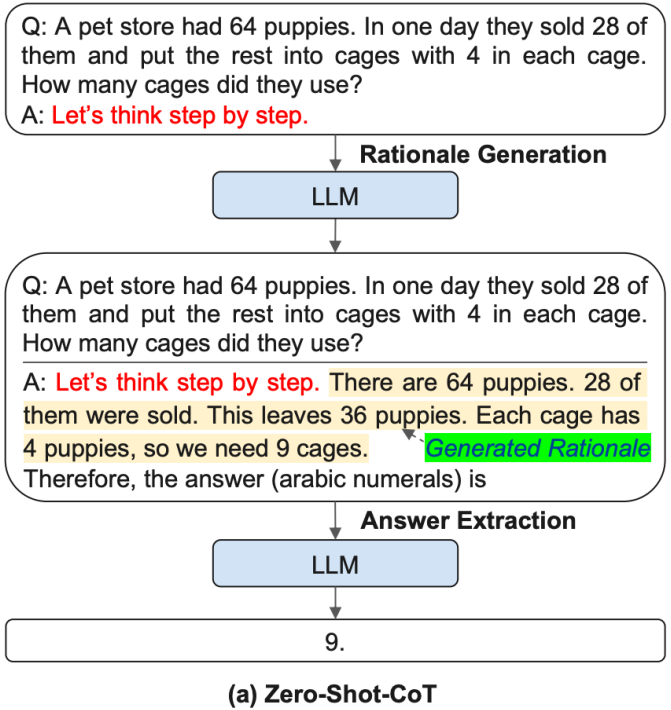
- **Zero-Shot CoT:** w/ a **trigger hint**, e.g., "let's think step by step" after the question (question + hint)
- **Manual-CoT:** w/ a few **manual-written** demonstration exemplars (question + rationale + answer)

No gradient update



Related Work

- **Zero-Shot CoT:** w/ a **trigger hint**, e.g., "let's think step by step" after the question (question + hint)
  - **Pros:** not required to write demonstration exemplars
  - **Cons:** poor performance & expensive to search the trigger hints heuristically

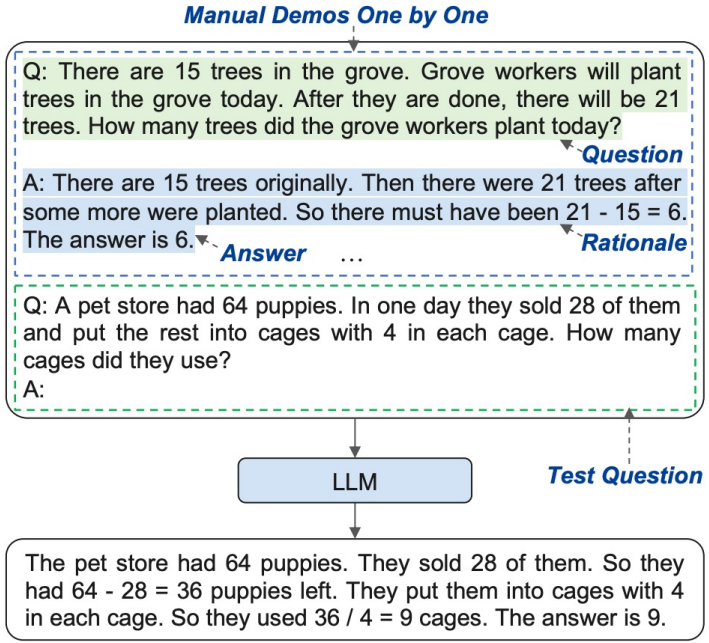


(a) Zero-Shot-CoT

No.	Template	Accuracy
1	Let's think step by step.	**78.7**
2	First, (*1)	77.3
3	Let's think about this logically.	74.5
4	Let's solve this problem by splitting it into steps. (*2)	72.2
5	Let's be realistic and think step by step.	70.8
6	Let's think like a detective step by step.	70.3
7	Let's think	57.5
8	Before we dive into the answer,	55.7
9	The answer is after the proof.	45.7
-	(Zero-shot)	17.7

Related Work

- **Manual-CoT:** w/ a few **manual-written** demonstration exemplars (question + rationale + answer)
  - **Pros**: strong performance
  - **Cons:** dependence on high-quality demonstrations written by experts



(b) Manual-CoT

	Prompting	GSM8K	SVAMP	ASDiv	AQuA	MAWPS
Prior best	N/A (finetuning)	55[a]	57.4[b]	75.3[c]	37.9[d]	88.4[e]
LaMDA 137B	Standard	6.5	29.5	40.1	25.5	43.2
	Chain of thought	14.3 (+7.8)	37.5 (+8.0)	46.6 (+6.5)	20.6 (-4.9)	57.9 (+14.7)
	+ ext. calc	17.8	42.1	53.4	20.6	69.3
GPT-3 175B (text-davinci-002)	Standard	15.6	65.7	70.3	24.8	72.7
	Chain of thought	46.9 (+31.3)	68.9 (+3.2)	71.3 (+1.0)	35.8 (+11.0)	87.1 (+14.4)
	+ ext. calc	49.6	70.3	71.1	35.8	87.5
PaLM 540B	Standard	17.9	69.4	72.1	25.2	79.2
	Chain of thought	56.9 (+39.0)	79.0 (+9.6)	73.9 (+1.8)	35.8 (+10.6)	93.3 (+14.2)
	+ ext. calc	58.6	79.8	72.6	35.8	93.5

	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

Challenges

- ❑ Competitive performance requires **manual annotation** of the **complex step-by-step reasoning chains** (demonstrations)
- ❑ Model performance heavily relies on the **quality of the demonstrations**
- ❑ **Expensive** to evaluate the effectiveness of the written demonstrations heuristically

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).

Challenges

- ❑ Require **manual annotation** of the **complex step-by-step reasoning chains** (demonstrations)
- ❑ Model performance heavily relies on the **quality of the demonstrations**
- ❑ **Expensive** to evaluate the effectiveness of the written demonstrations heuristically

	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

a) Few-Shot CoT

No.	Template	Accuracy
1	Let's think step by step.	**78.7**
2	First, (*1)	77.3
3	Let's think about this logically.	74.5
4	Let's solve this problem by splitting it into steps. (*2)	72.2
5	Let's be realistic and think step by step.	70.8
6	Let's think like a detective step by step.	70.3
7	Let's think	57.5
8	Before we dive into the answer,	55.7
9	The answer is after the proof.	45.7
-	(Zero-shot)	17.7

b) Zero-Shot CoT

	Zero-shot	Few-shot-CoT †	Zero-shot-CoT	Few-shot-CoT
AQUA-RAT	22.4	31.9	33.5	39.0
MultiArith	17.7	27.0	78.7	88.2

c) Manual exemplars from other tasks (CSQA)

Challenges

- Require **manual annotation** of the **complex step-by-step reasoning chains** (demonstrations)
- Model performance heavily relies on the **quality of the demonstrations**
- **Expensive** to evaluate the effectiveness of the written demonstrations heuristically

	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

MODEL	USAGE
Ada	$0.0040 / 1K tokens
Babbage	$0.0050 / 1K tokens
Curie	$0.0200 / 1K tokens
Davinci	$0.2000 / 1K tokens

Each exp: 600 test examples -> ≈$12

Motivation

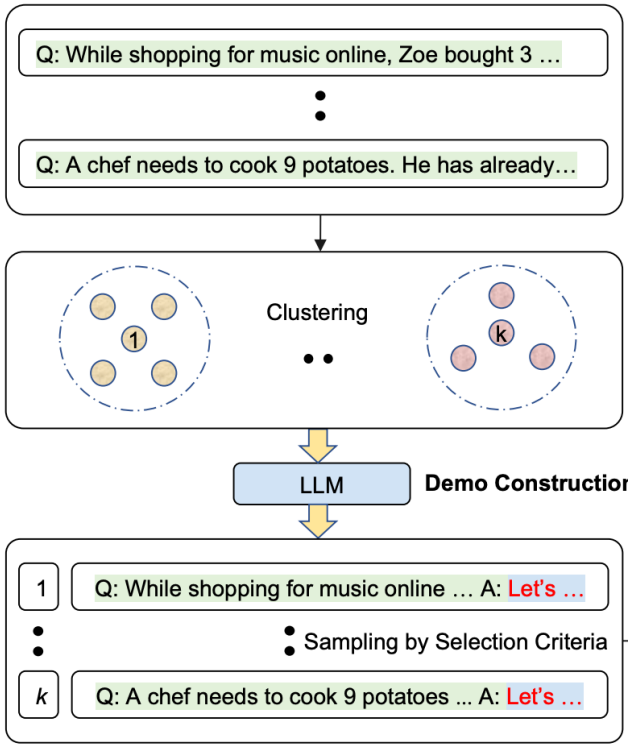
- **Challenges**
  - Require **manual annotation** of the complex **step-by-step reasoning chains** (demonstrations)
  - Model performance heavily relies on the **quality of the demonstrations**
  - **Expensive** to evaluate the effectiveness of the written demonstrations heuristically
- **Goals**
  - Get rid of human annotations and innovate the research line of **automatic CoT**
  - Figure out what makes **good demonstrations** and how to obtain them

Problem-1: Find the appropriate questions

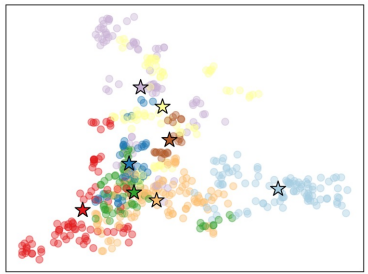
Problem-2: Generate decent reasoning chains

Auto-CoT

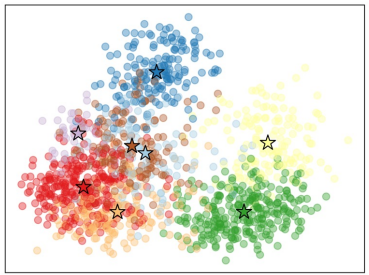
- **Questions**: cover the typical patterns of the dataset -> sample the representative questions via clustering
- **Rationales:** reflect the step-by-step reasoning process -> fetch the intermediate rationales via Zero-Shot CoT



1. **Encoding**: Encode each question with **Sentence-Transformer**.
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

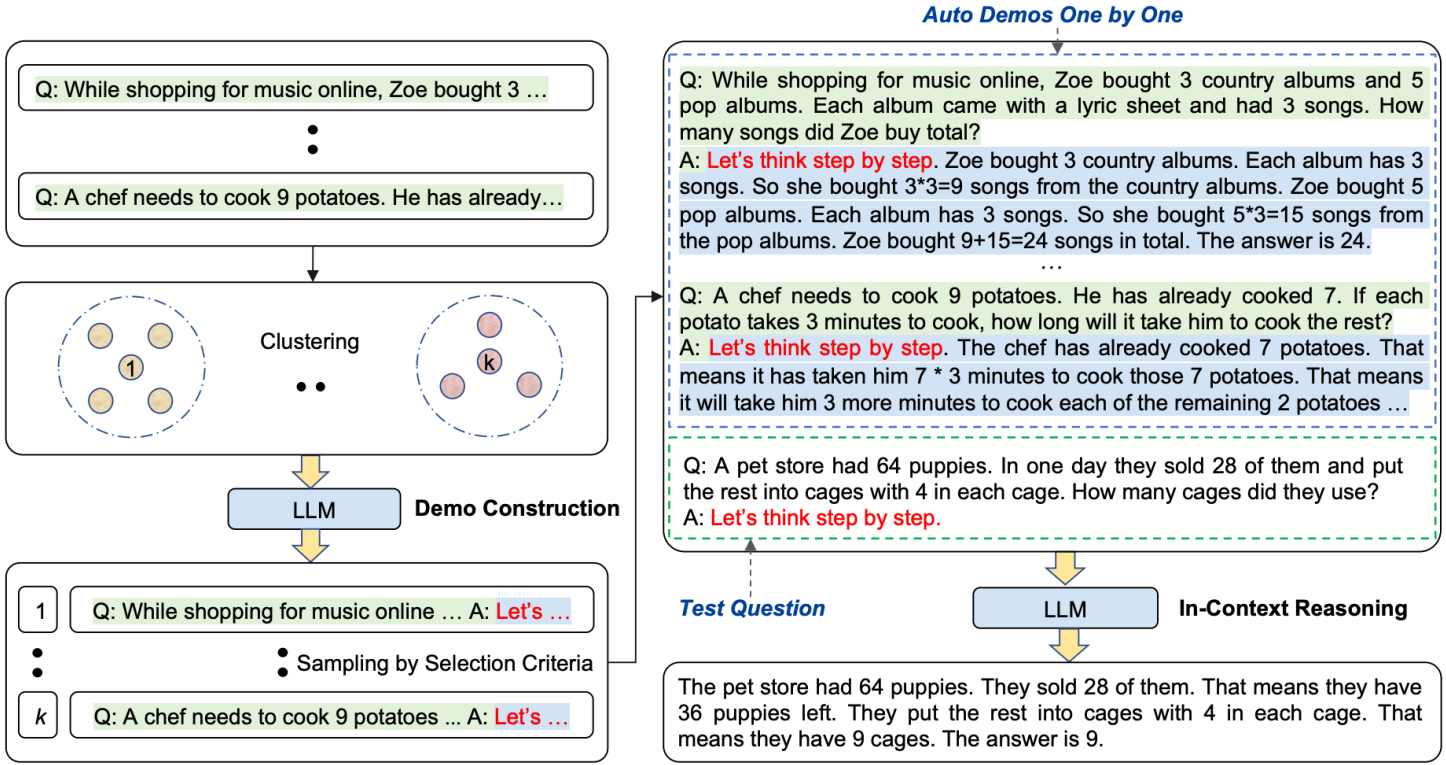


GSM8K

* k is the number of our desired demonstrations

Auto-CoT

- **Questions**: cover the typical patterns of the dataset -> sample the representative questions via clustering
- **Rationales:** reflect the step-by-step reasoning process -> fetch the intermediate rationales via Zero-Shot CoT



Experimental Settings

Datasets:

1. Our method is evaluated on 10 public benchmark datasets
2. Cover arithmetic, commonsense, and logical reasoning tasks

Backbone Model: GPT-3 (*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

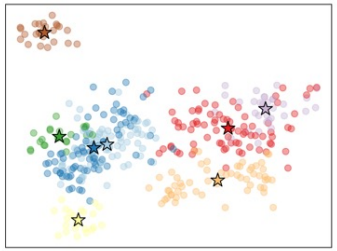
Main Results

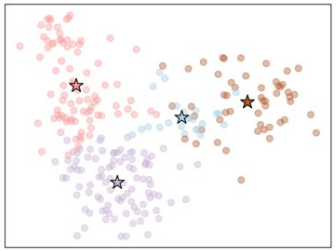
1. Auto-CoT method substantially outperforms the Zero-Shot-CoT and Manual-CoT baselines
2. Competitive Performance based on the public GPT-3 text-davinci-002 (single model)
3. Auto-CoT is robust towards randomness

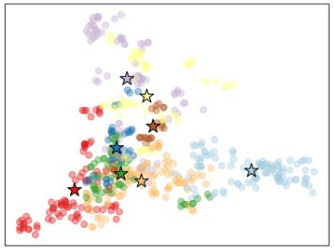
Model	*Arithmetic*						*Commonsense*		*Symbolic*	
	MultiArith	GSM8K	AddSub	AQuA	SingleEq	SVAMP	CSQA	Strategy	Letter	Coin
Zero-Shot	22.7	12.5	77.0	22.4	78.7	58.8	72.6	54.3	0.2	53.8
Zero-Shot-CoT	78.7	40.7	74.7	33.5	78.7	63.7	64.6	54.8	57.6	91.4
Few-Shot	33.8	15.6	83.3	24.8	82.7	65.7	**79.5**	**65.9**	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	$\mathbf{92.0}^{\Uparrow}_{\pm 1.7}$	$\mathbf{47.9}^{\Uparrow}_{\pm 3.7}$	$\mathbf{84.8}^{\uparrow}_{\pm 2.9}$	$\mathbf{36.5}^{\Uparrow}_{\pm 2.2}$	$\mathbf{87.0}^{\Uparrow}_{\pm 1.2}$	$\mathbf{69.5}^{\uparrow}_{\pm 2.2}$	$74.4^{\uparrow}_{\pm 2.5}$	$65.4^{\Uparrow}_{\pm 0.4}$	$\mathbf{59.7}^{\uparrow}_{\pm 3.2}$	$\mathbf{99.9}^{\Uparrow}_{\pm 0.1}$

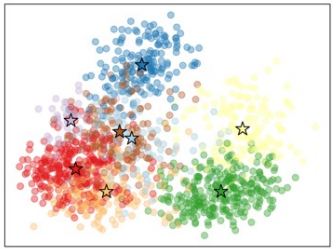
Visualization of Demonstration Clustering

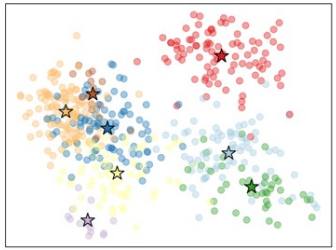
1. The number of clusters = num. of desired demos = num. of few-shot demos in Few-Shot CoT.
2. The clustered demonstrations are likely to represent generic themes of the datasets.

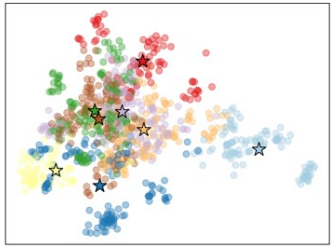

AddSub

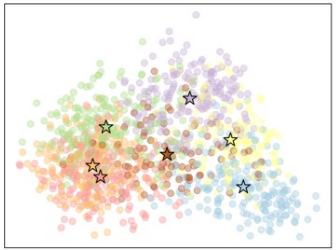

AQUA

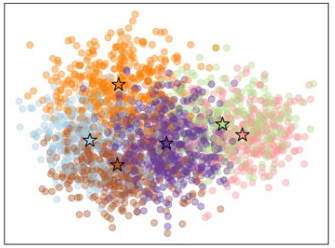

MultiArith


GSM8K


SingleEq


SVAMP

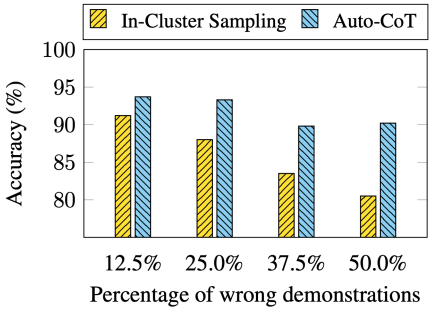

CSQA

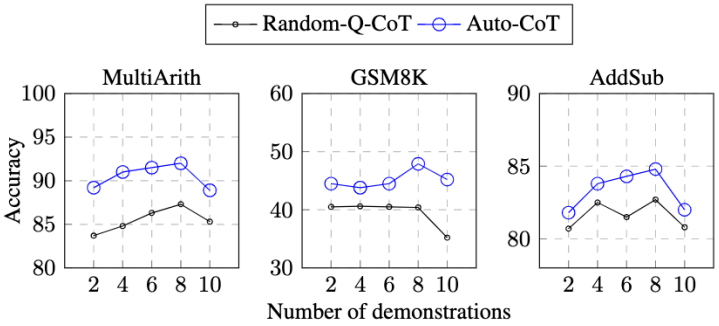

Strategy

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





Summary: Large Language Models Are Automatic Chain of Thought Reasoners

- **Problem**
  - **Chain of thought (CoT)** prompting for large language models (LLMs)
- **Goals**
  - Get rid of human annotations and innovate the research line of **automatic CoT**
  - Figure out what makes **good demonstrations** and how to obtain them
- **Contributions**
  - **A complete automatic CoT** method that outperforms few-shot CoT methods that rely human expert annotations
  - **State-of-the-art results** using the public GPT-3 model in the single model setting
  - Discloses the potential of **automatically constructing effective demonstrations** using public LLMs
- **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
  - Code: https://github.com/amazon-science/auto-cot

Thanks & QA

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