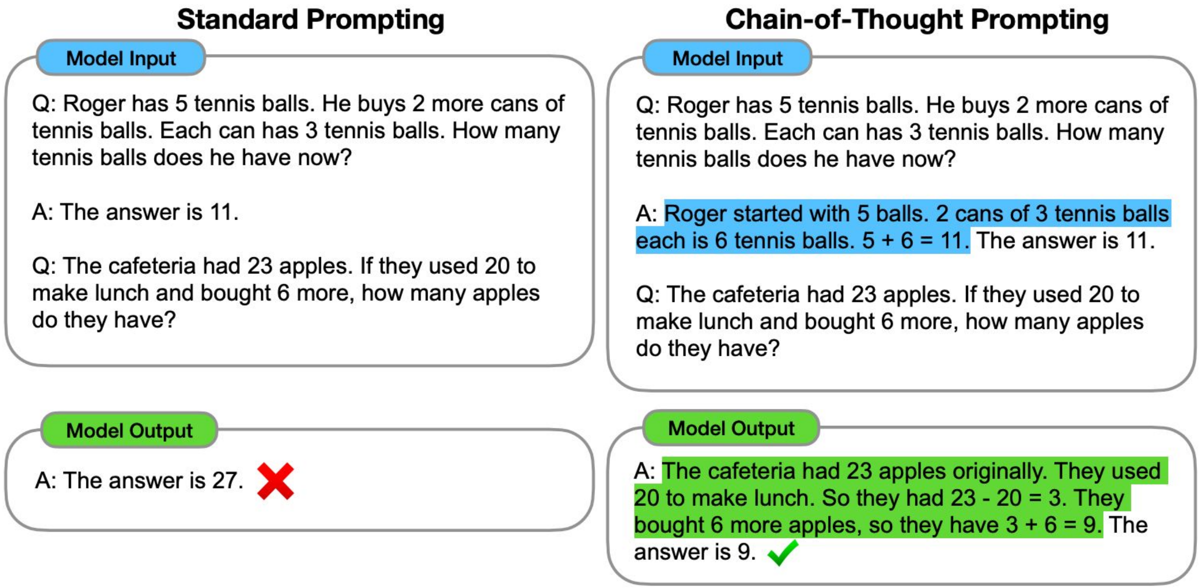
Automatic Chain of Thought Prompting in Large Language Models (2022)

Zang Zijian 522024330115

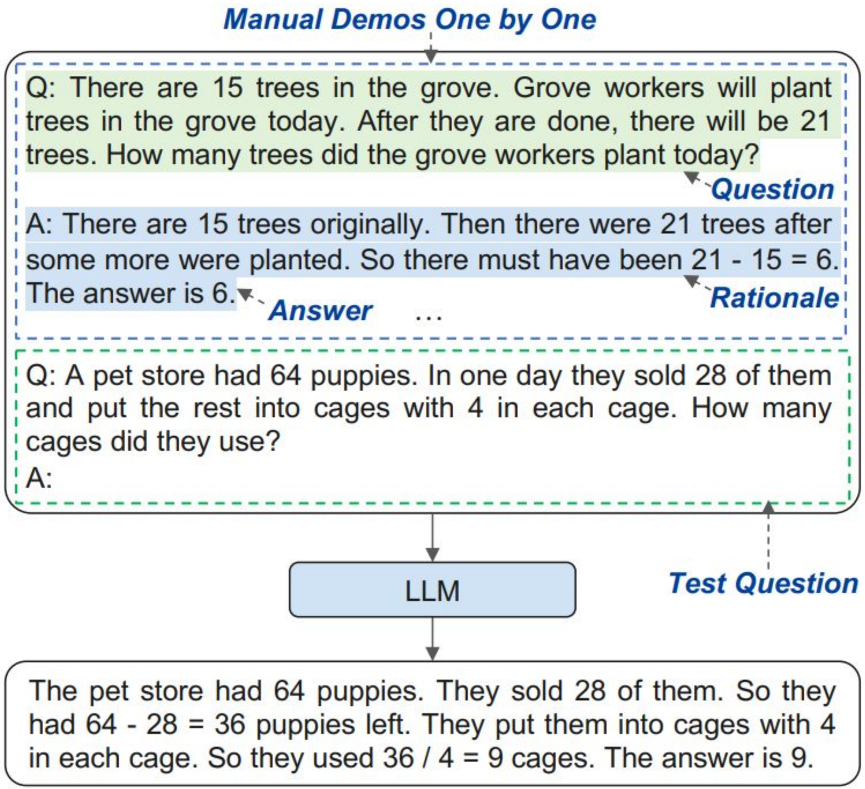
What is Chain-of-Thought



Chain of thought prompting elicits reasoning in large language models.

Simple Methods: Manual CoT

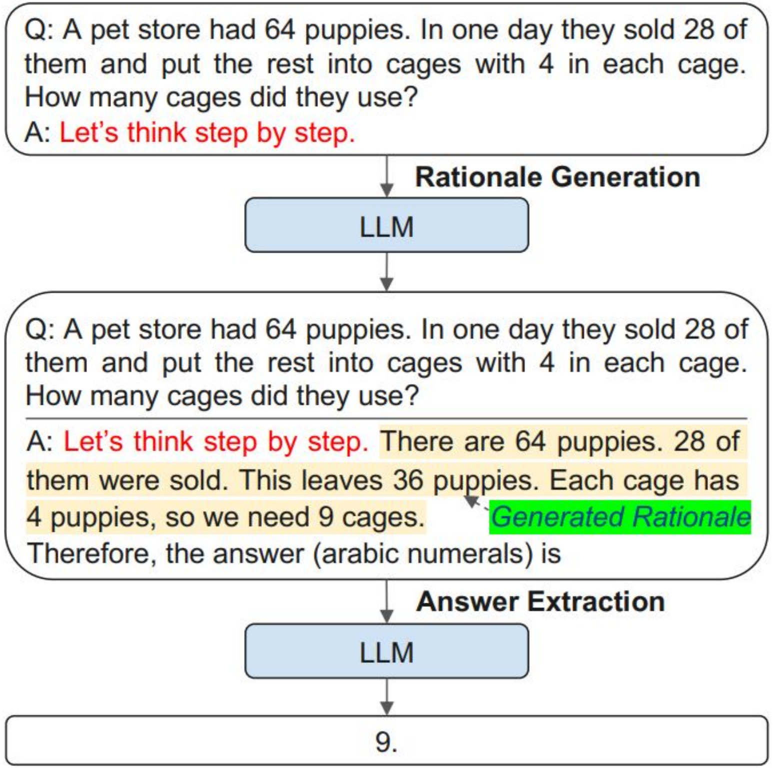
- Few-shot Prompting
- Ilustrate the intermediate steps
- In-Context Learning



Simple Methods: ZeroShot CoT

Magic words:

Let’s think step by step.



Auto CoT: Powered Manual CoT

Manual CoT generally performs better than ZeroShot, but:

1. It hinges on the hand-drafting, involves:
   a. designing question
   b. designing reasoning chain
2. Different tasks require different ways of demonstration
   a. arithmetic
   b. commonsense reasoning
   c. …

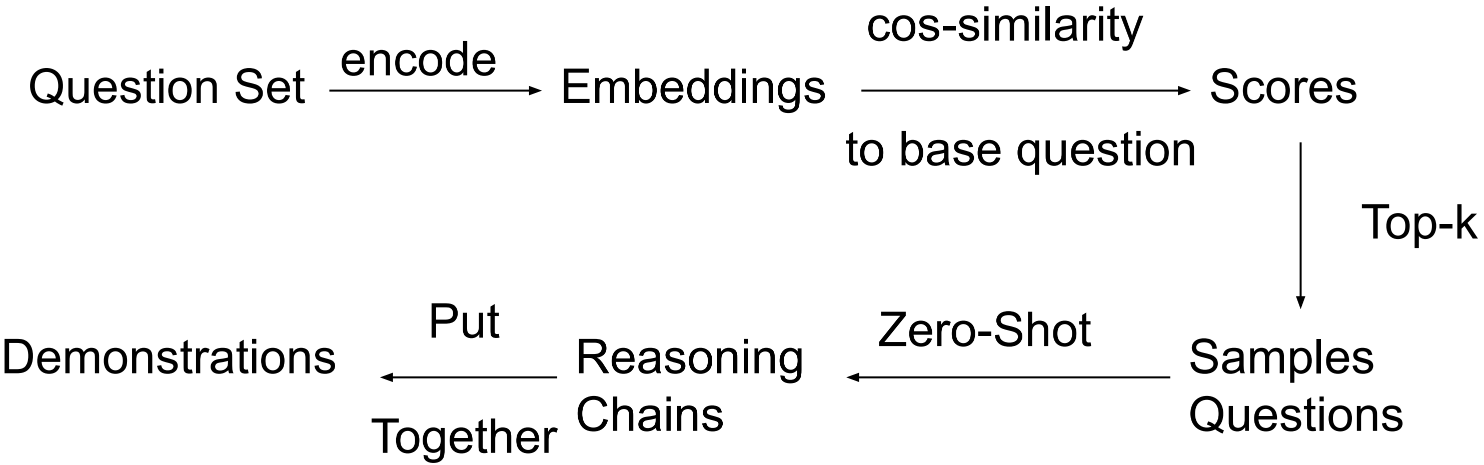
Key Challenge: Automatically constructing demonstrations

Fact: demonstrations written by different annotators brings up 28.2% accuracy disparity

Two Parts: Question, Reasoning Chain

Select Questions: The Trivial Way

similarity-based retrieval methods:



Challenge

Wrong!

Table 1: Accuracy (%) of different sampling methods. Symbol † indicates using training sets with annotated reasoning chains.

Method	MultiArith	GSM8K	AQuA
Zero-Shot-CoT	78.7	40.7	33.5
Manual-CoT	**91.7**	46.9	35.8†
Random-Q-CoT	86.2	47.6†	36.2†
Retrieval-Q-CoT	82.8	**48.0**†	**39.7**†

with human-annotation it works — but its nontrivial

Why ? Assumption

Fact:
Zero-Shot CoT may lead to incorrect reasoning chains for difficult problems.

Assumption: For challenging problems:

- Similarity-based sampling gathers similar hard problems and produce large number of incorrect reasoning chain.
- Random sampling collects a diverse range of problems, alloing some problems to be solved, which aids further reasoning.

Why ? Experiment

To put it simply:
They collect all the problems that Zero-Shot cannot solve and test them using these two CoT method.

Random sampling greatly outperforms Retrieval sampling.

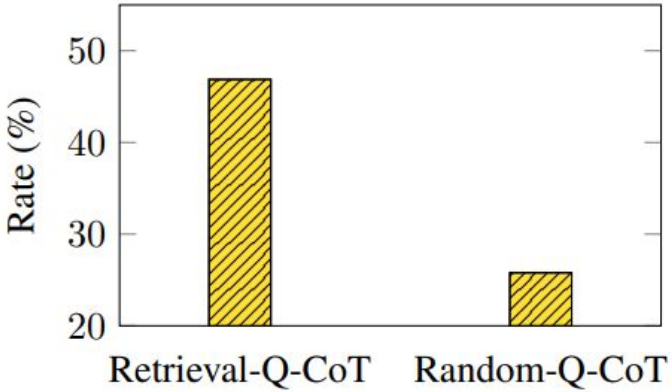


Figure 2: Unresolving Rate.

Inspiration: Errors Frequently Fall into the Same Cluster

- k-means to partition all test questions into 8 clusters
- cluster 2 is extremely difficult to solve

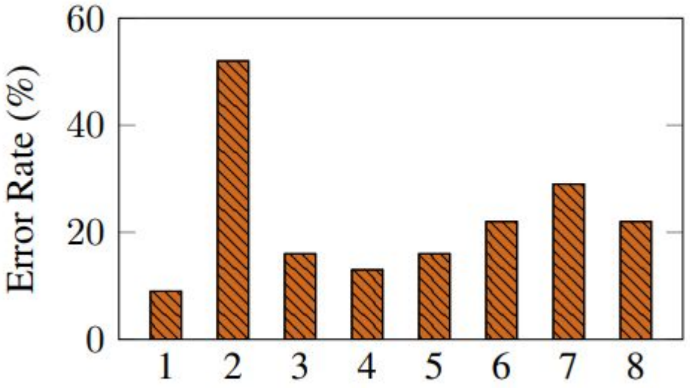
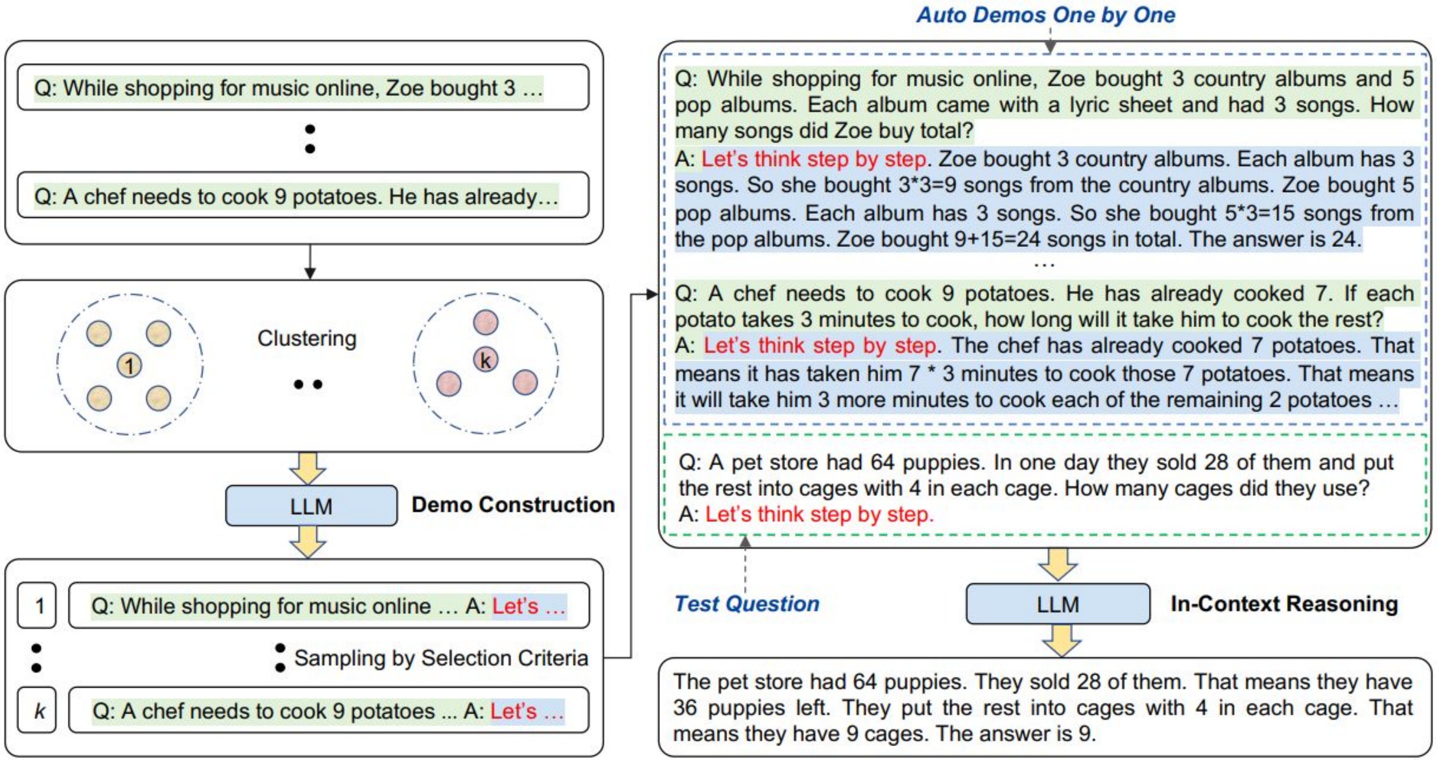


Figure 3: Clusters of similar questions.

Inspiration: Diversity May Mitigate Misleading by Similarity

- a small portion of mistakes would not harm the overall reasoning performance
- different clusters reflect diverse semantics of the questions
- diverse demonstrations seem to cover more alternative skills for solving target questions

Inspiration: Diversity May Mitigate Misleading by Similarity



Thank you