



# Federated Prompting and Chain-of-Thought Reasoning for Improving LLMs Answering Synonymous Questions

Xiangyang Liu, Tianqi Pang, and Chenyou Fan  
South China Normal University

## ● Motivation

---



- Large language models (LLMs) have shown impressive language understanding abilities. **However, they still suffer from low accuracy on complex reasoning problems in zero-shot settings.**
- A common question might be asked frequently with same or different parameters or rephrased formulations. For example, the Chickens-and-rabbits questions can be asked with different number of heads and feet. **So Can we fully utilize those similar questions to improve the question answering of the LLMs without tuning the model parameters or infringing user privacy?**

## ● Definition of the task



Table 1: Examples of SP-questions and DP-questions. *Q1* and *Q2* are synonymous with same parameters. *Q1* and *Q3* are synonymous but with different parameters.

*Q1: "If a farmer has chickens and rabbits in a barn and there are a total of 32 heads and 100 feet, how many chickens and rabbits does the farmer have?"*

*Q2: "In a barn, there are a certain number of chickens and rabbits that have a total of 32 heads and 100 feet. how many of each animal are in the barn?"*

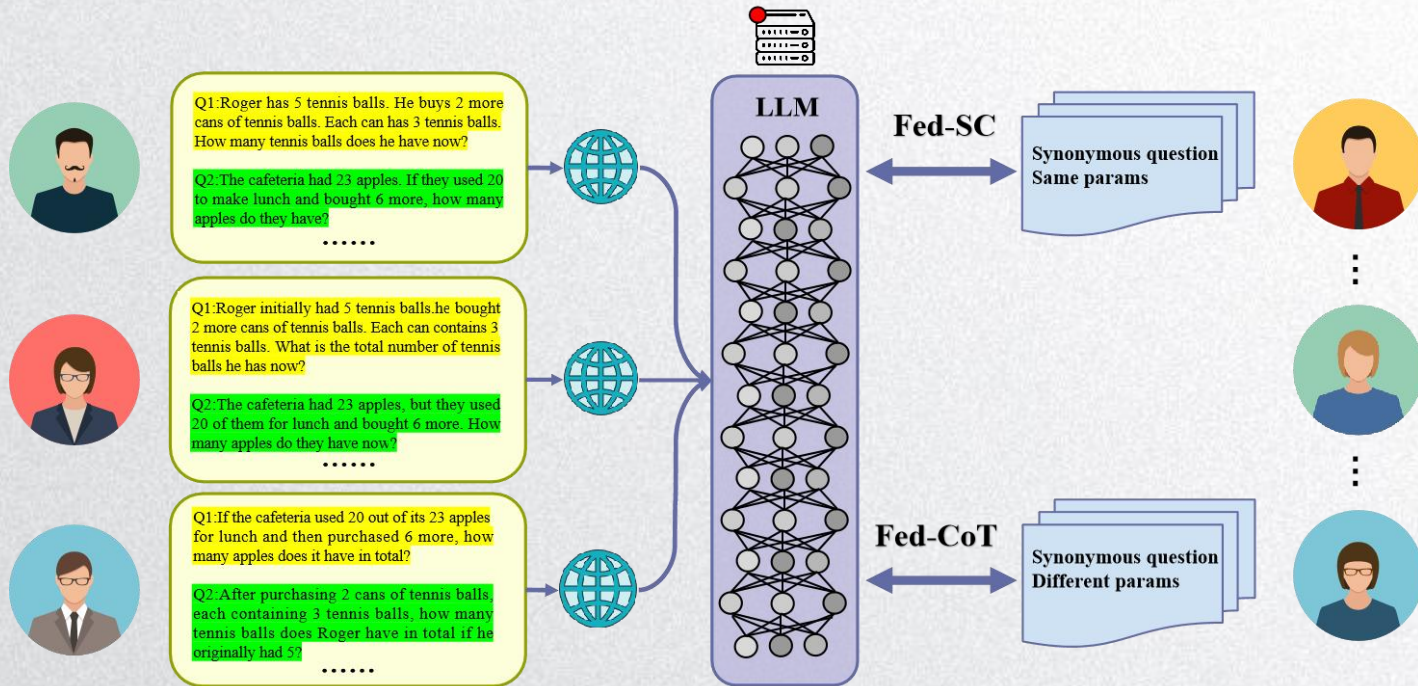
*Q3: "A farmer has a total of 20 chickens and rabbits in his barn. If the total number of legs is 56, how many chickens and how many rabbits?"*

### SP-questions and DP-questions

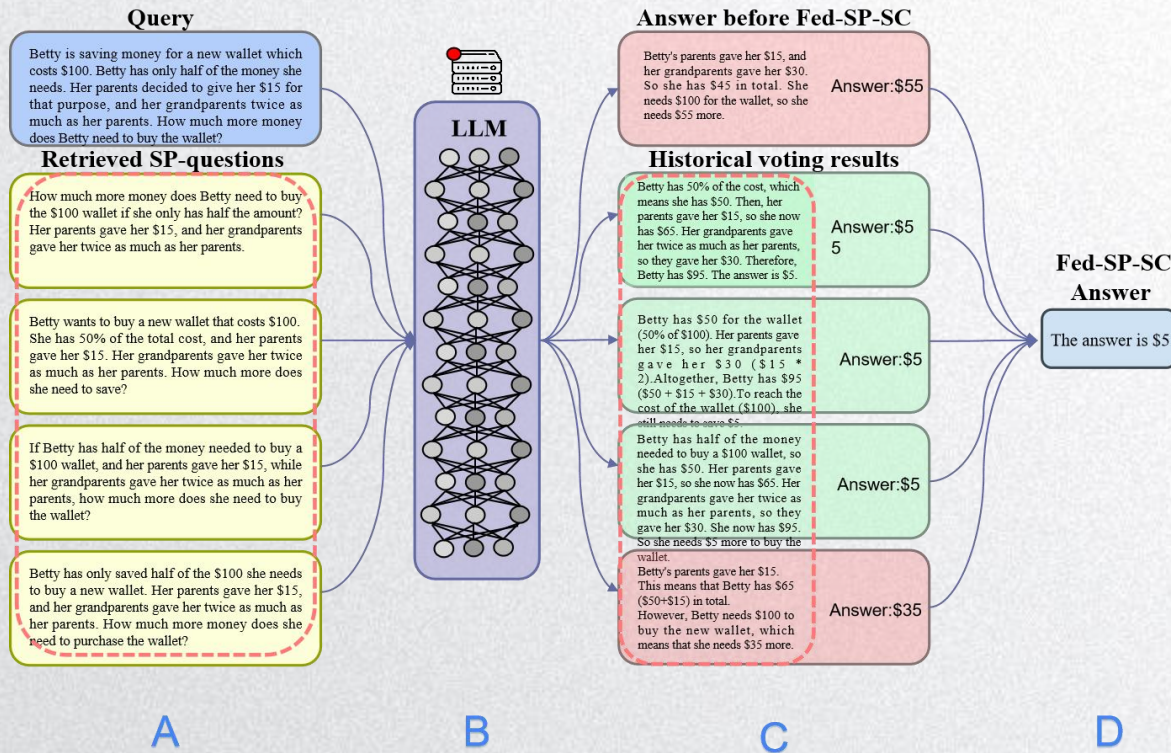
We consider improving the reasoning capacity of LLMs by better understanding crowd-sourced similar questions. We propose two typical scenarios: synonymous questions with same parameters or different parameters.

When a user sends a QA request to the LLM and the LLM tries to answer with a collected question database.

# Overview



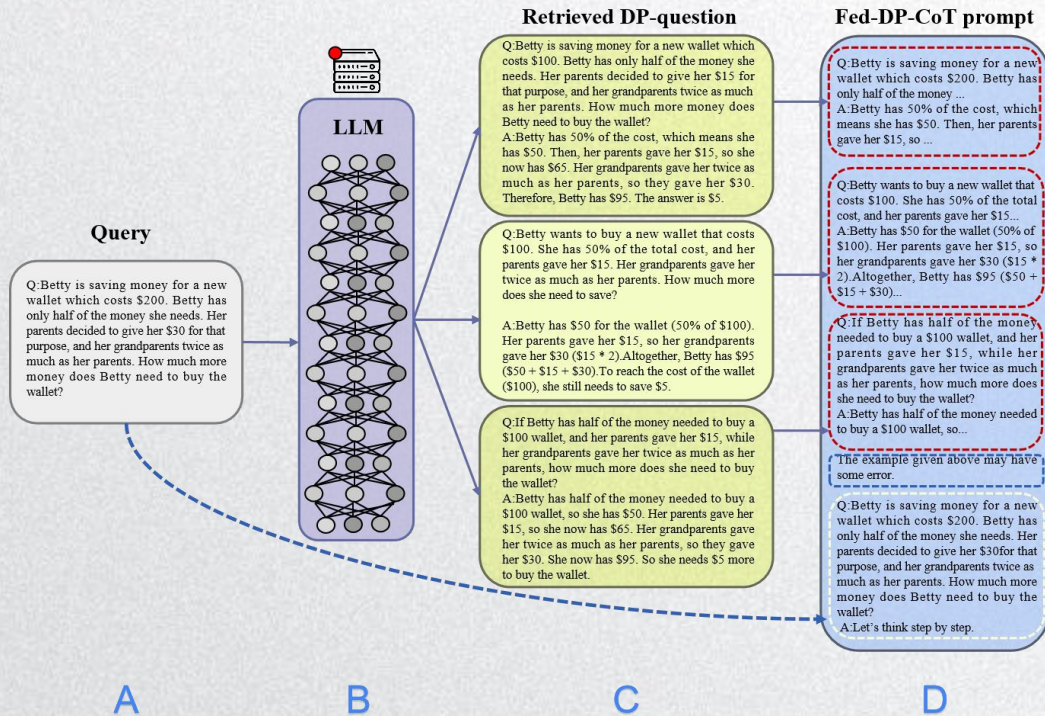
We federate synonymous questions from distributed users and apply self-consistency and Chain-of-Thought techniques to improve answering from central LLMs.



The illustration of performing Fed-SP-SC for answering synonymous SP-questions.



- (A  $\rightarrow$  B): When receiving the users' query, the LLM retrieves synonymous SP-questions from the centralized database.
- (B  $\rightarrow$  C): The LLM generates the answers with zero-shot prompting for the query and combines the retrieved SP-questions' answers from the database for a majority vote to ensure self-consistency.
- (C  $\rightarrow$  D): The most voted answer is returned to the user as the best answer. The database could store the query and answer pair back to the database, caching for later retrieval. This procedure can grow the database quickly by gathering distributed user queries.



The illustration of performing Fed-DP-CoT.



- (A  $\rightarrow$  B): DP-questions are firstly retrieved from the centralized user query database.
- (B  $\rightarrow$  C): The system selects SP-questions with consistent answers after applying Fed-SP-SC as DP-questions and retrieves consistent answers as “pseudo-labels” .
- (C  $\rightarrow$  D): The system concatenates questions and pseudo-labels, adds a pseudo-label disclaimer such as “The examples may have errors.” after, and finally appends the original user query to form the complete CoT prompt.





## ● Experimental Settings

---

- Benchmark: GSM8K, SVAMP (Subset)
- Language Model: GPT3 (Openai API)
- Baseline: We compare our methods with Zero-Shot-CoT, which refers to adding “Let’s think step by step.” to prompt as a composite prompt, such as “[Question] A: Let’s think step by step.”
- We use the prompt “Rephrase in 4 ways: [question]” to generate synonymous questions in the experiment. In practice, we use GPT-3 and GPT-3.5 to generate synonymous questions. In the experiment, we selected OpenAI text-davinci-003 for the GSM8K dataset and text-davinci-002 for the SVAMP dataset so that the accuracy neither too high nor too low.

# ● Result 1 : Fed-SP-SC



**Table 2.** Fed-SP-SC results

Data\Method	Zero-Shot-CoT	Fed-SP-SC (GPT-3 Gen.)	Fed-SP-SC (GPT-3.5 Gen.)
GSM8K	52.5%	62.7%	70.6%
SVAMP	77.2%	86.3%	91.1%

- **Fed-SP-SC can improve answering accuracy of LLMs** by federating multiple SP-questions through self-consistency.
- **Fed-SP-SC(GPT-3.5 Gen.) performs best on the GSM8K and SVAMP datasets**, improved the performance by 17.5% and 14% on the GSM8K and SVAMP datasets, respectively.
- **The quality of the synonymous questions can affect the accuracy significantly**, as seen in the larger improvement from the synonymous questions generated by GPT-3.5 compared to GPT-3.

## ● Result 2 : Fed-DP-CoT

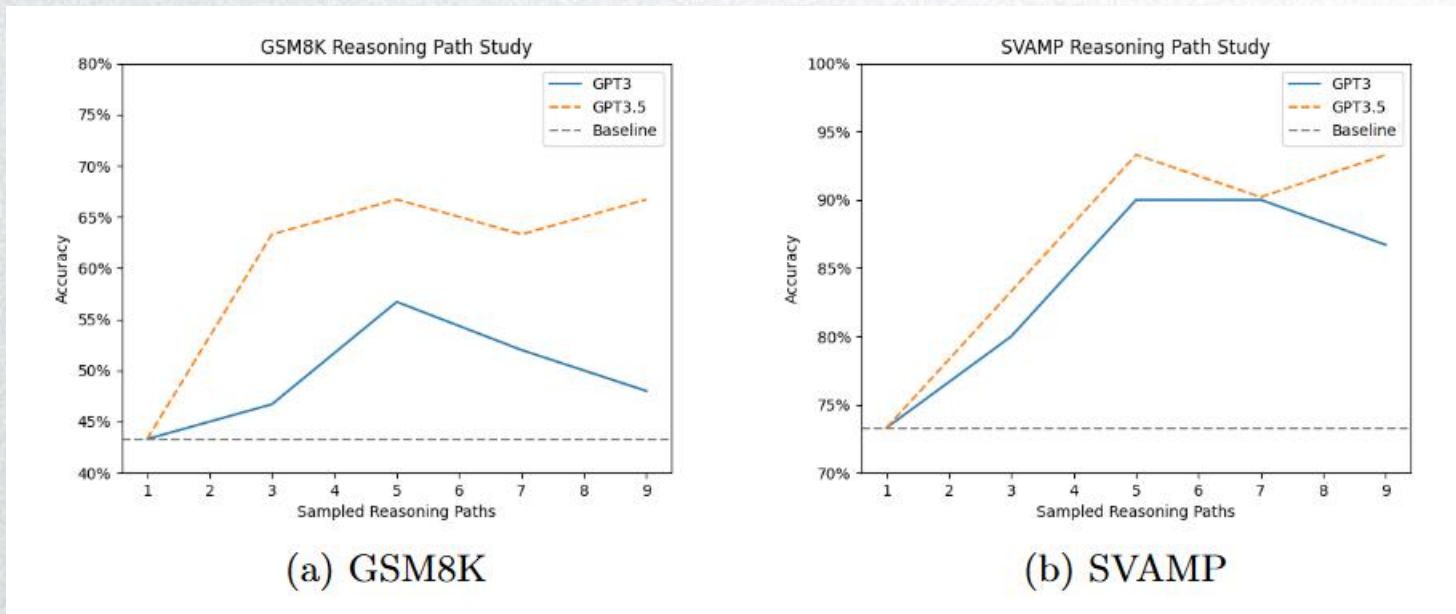


**Table 3.** Fed-DP-CoT results.

Setting\Method	Zero-Shot-CoT	Fed-DP-CoT (GPT-3 Gen.)	Fed-DP-CoT (GPT-3.5 Gen.)
GSM8K	48.3%	59.2%	62.5%
SVAMP	76.5%	82.4%	85.7%

- **Fed-DP-CoT can improve the performance.** Compared to Zero-Shot-CoT, CoT Prompt(GPT-3 Gen.) and CoT Prompt(GPT-3.5 Gen.) improve by approximately 10.9%-14.2% and 6.6%-10% respectively on the datasets GSM8K and SVAMP.
- **Fed-SP-SC performs better than Fed-DP-CoT.** The results of Fed-SP-SC (GPT3 Gen.) and Fed-SP-SC (GPT-3.5 Gen.) on the GSM8K and SVAMP datasets are both higher than Fed-DP-CoT (GPT-3 Gen.) and CoT Prompt (GPT-3.5 Gen.), with an approximate improvement of 5%.
- **Less performance difference between GPT-3 Gen. and GPT-3.5 Gen.** compared to Fed-SP-SC experiment.

# ● Ablation Studies 1



We study the effect of using different number of reasoning paths for Fed-SP-SC to apply self-consistency. In the line chart, as the number of sampled reasoning paths increases from one to five, the accuracy rate gradually increases. **However, when the number of synonymous questions exceeds five, the accuracy of the model starts to decrease.**

## ● Ablation Studies 2



Table 5: GSM8K disclaimer ablation.

Method\Setting	Zero-shot -CoT	Fed-DP-CoT (GPT-3 Gen.)	Fed-DP-CoT (GPT-3.5 Gen.)
w/o disclaimer	48.3%	57.7%	60%
w/ disclaimer	NA	59.2%	62.5%

We investigate **whether the disclaimer is effective** of correcting noisy CoTs in this ablation experiment. We observe that the use of a simple disclaimer can potentially improve the accuracy of LLMs by approximately 2% for the Fed-DP-CoT task.

We postulate that the improvement in accuracy may be attributed to the fact that the disclaimer prompts LLMs to be careful of the pseudo-labels and self-examine the reasoning steps.

## ● Conclusion

---



- We investigate the potential benefits of employing synonymous queries from distributed users to enhance question-answering of LLMs.
- We abstract two scenarios: SP- and DP-questions. We propose Fed-SP-SC and Fed-DP-CoT to solve the problems in these two cases.
- Moving forward, future research may investigate more realistic systems that can efficiently retrieve federated questions while also improving CoT correctness to further advance reasoning capabilities.



Thank for listening!

Email: [2022024952@m.scnu.edu.cn](mailto:2022024952@m.scnu.edu.cn)