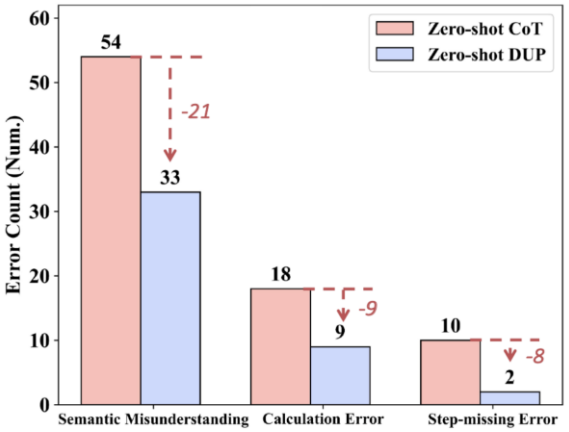
Achieving >97% on GSM8K: Deeply Understanding the Problems Makes LLMs Better Solvers for Math Word Problems

Qihuang ZHONG, Kang WANG, Ziyang XU, Liang DING, Juhua Liu, Bo DU

Frontiers of Computer Science, DOI: 10.1007/s11704-025-41102-z

Problems & Ideas

- Problems of conventional Chain-of-thought (CoT) approaches:
  - CoT usually suffers from three pitfalls: *semantic misunderstanding errors, calculation errors, and step-missing errors*.
  - Prior studies involve addressing the calculation errors and step-missing errors, but **neglect the semantic misunderstanding errors**, which is the major factor limiting the reasoning performance of LLMs.
- Ideas: The core of our method is to encourage the LLMs to deeply understand the problems and extract the key problem-solving information used for better reasoning.



Error analysis of GSM8K problems with incorrect answers returned by zero-shot CoT and our DUP method using GPT-3.5 LLM.

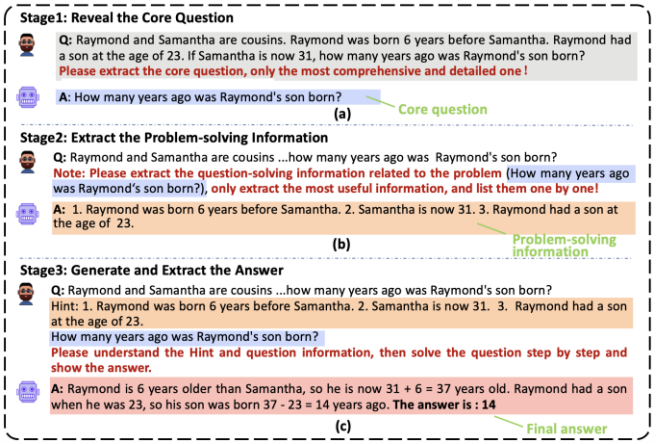


Illustration of our DUP strategy, containing three-stage processes: 1) reveal the core question, 2) extract the problem-solving information, and 3) generate and extract the answer.

Main Contributions

- Contributions:
  - We reveal the underlying causes of semantic misunderstanding errors, and propose a simple yet effective approach (DUP) to effectively address the semantic misunderstanding and boost LLMs' math reasoning ability;
  - DUP is easy-to-implement and plug-and-play. It can be easily applied to various LLMs;
  - Extensive experiments on various reasoning benchmarks show that DUP outperforms the other counterparts by a large margin.

Model	Method	Arithmetic Reasoning							Score	
		SVAMP	GSM8K	AddSub	MultiArith	AQuA	SingleEq	MathQA	*Avg.*	Δ
Performance of Zero-shot Methods										
GPT-3.5-Turbo	Zero-shot CoT	79.3	78.9	85.8	95.3	53.0	93.5	63.7	78.5	-
	Least-to-Most	80.9	77.5	91.3	95.5	57.4	93.5	66.0	80.3	+1.8
	Zero-shot PS+	80.7	79.3	86.5	92.0	55.9	93.0	66.8	79.2	+0.7
	R^3 prompting	81.6	79.4	92.0	95.5	59.4	94.6	64.4	81.0	+2.5
	DUP (Ours)	**82.5**	**82.3**	**92.1**	**97.8**	**60.2**	**94.9**	68.1	**82.6**	**+4.1**
GPT-4	Zero-shot CoT	90.4	94.6	92.4	97.8	72.8	95.0	82.1	89.3	-
	Least-to-Most	90.3	92.1	92.1	97.1	71.6	95.0	82.4	88.7	-0.6
	Zero-shot PS+	92.6	94.3	93.1	**98.1**	75.5	95.3	83.4	90.3	+1.0
	R^3 prompting	93.4	92.1	94.5	**98.1**	75.1	94.8	81.7	90.0	+0.7
	DUP (Ours)	**94.2**	**97.1**	**95.1**	**98.1**	**77.1**	**96.0**	84.1	**91.7**	**+2.4**
Performance of Few-shot Methods										
GPT-3.5-Turbo	Manual-CoT	78.5	81.6	90.6	95.6	55.9	94.2	64.2	80.1	+1.6
	Auto-CoT	82.9	80.2	89.9	99.0	54.3	94.6	64.8	80.8	+2.3

Performance comparison (%) of different CoT methods on various arithmetic reasoning benchmarks.