NEURAL INFORMATION PROCESSING SYSTEMS

AutoPSV: Automated Process-Supervised Verifier

Jianqiao Lu[1], Zhiyang Dou[1], Hongru Wang[2], Zeyu Cao[3], Jianbo Dai[4],

Yingjia Wan[3], Yunlong Feng, Zhijiang Guo[3]

[1]**The University of Hong Kong** [2]The Chinese University of Hong Kong

[3]University of Cambridge [4]University of Edinburgh

TL;DR

1. AutoPSV effectively identifies **variations in model confidence** to annotate the correctness of **intermediate reasoning steps**, enabling **efficient automatic labeling for process supervision.**

2. AutoPSV significantly improves the **performance and scalability** of verification models in mathematical and commonsense reasoning tasks.

3. AutoPSV's versatility is evident in its applicability to **both labeled and unlabeled dataset settings** after completing the training process.

Background

Problem Response selection from multiple candidates for reasoning tasks

Parameterization

- q : input question
- $S_i^{(1:t)}$: i-th solution contains from 1 to t-th reasoning steps
- y_i : binary correctness label

Outcome-Supervision vs. Process-Supervision

y_i vs y_i^t

Current Process-Supervision Methods

- Human annotations: expensive
- Monte Carlo Tree Search (MCTS-based) : computationally inefficient

Motivation

Finding: Even models exceeding 70 billion parameters demonstrate suboptimal selection performance when relying solely on prompting without fine-tuning.

response generator: Mixtral-Instruct (8 x 7b)

Table 1: Performance of Mixtral-Instruct on GSM8K. All results are reported in accuracy (%).

Response Generator	**Model Size** (Parameters)	**Pass@1** (%)	**Pass@5** (%)	**Self-Consistency** (%)
Mixtral-Instruct [31]	8 x 7B (MOE)	62.55	82.31	69.06

---

selectors: Mistral-Instruct (7b), Mixtral-Instruct, Llama2-chat (70b) and Qwen (72b)

Table 2: Comparison of different selection methods across various model sizes for selecting a response from candidate responses generated by Mixtral-Instruct. All results are reported in accuracy (%).

Selector	**Model Size**	**Prompt Strategy**				
		Pairwise	Classification	Classification + CoT	Scoring	Scoring + CoT
Mistral-Instruct [32]	7B	60.73	61.18	64.82	61.49	69.75
Mixtral-Instruct [31]	8×7B	58.83	59.14	67.40	61.79	65.58
Llama2-chat [33]	70B	59.28	62.70	66.79	59.74	62.93
Qwen [34]	72B	59.14	66.64	69.52	61.86	65.88

Training Methodology

Outcome-Supervision

$$L\left(S_i^{(1:t)}, y_i\,; q\right) = \left(f_\theta\,\left(S_i^{(1:t)}\,; q\right)\, - y_i\}\right)^2$$

We firstly define $\Delta_{conf}^t = \frac{f_\theta\left(S_i^{(1:t+1)}\,; q\right) - f_\theta\left(S_i^{(1:t)}\,; q\right)}{f_\theta\left(S_i^{(1:t)}\,; q\right)}$ and

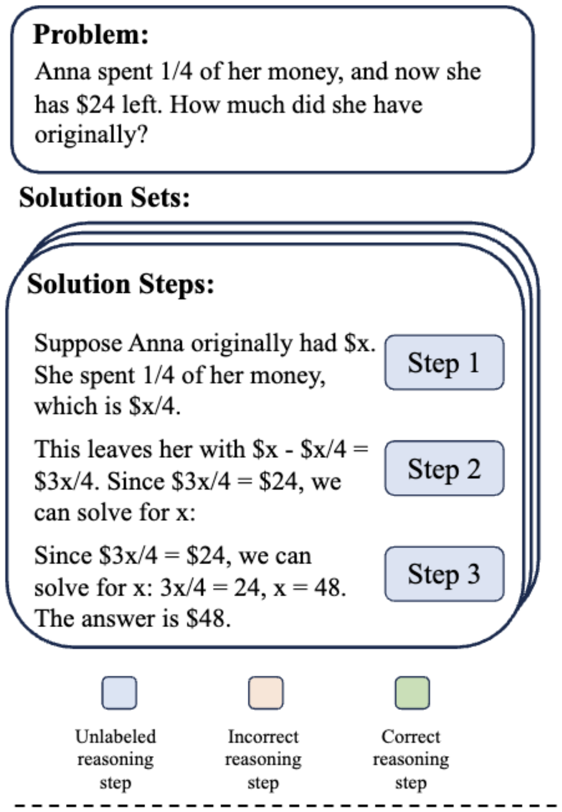
Process-Supervision

$$L\left(S_i^{(1:t)}, y_i^t\,; q\right) = \left(f_\theta\,\left(S_i^{(1:t)}\,; q\right)\, - y_i^t\}\right)^2$$

Where

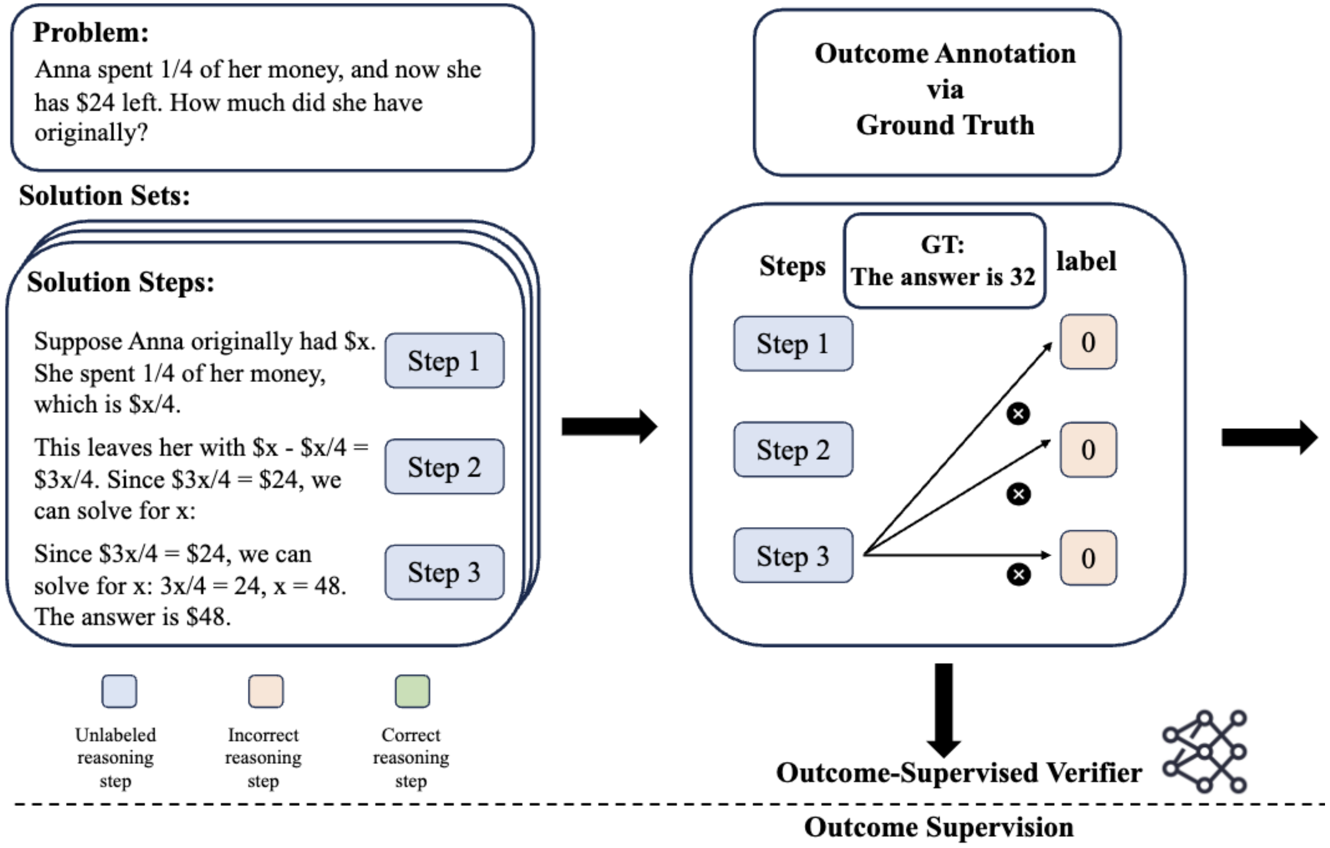
If $\Delta_{conf}^t > \theta, y_i^t = 1$, else $y_i^t = 0$

Training Methodology



Given an LLM acting as a response generator, we seek to annotate each reasoning step and perform response selection.

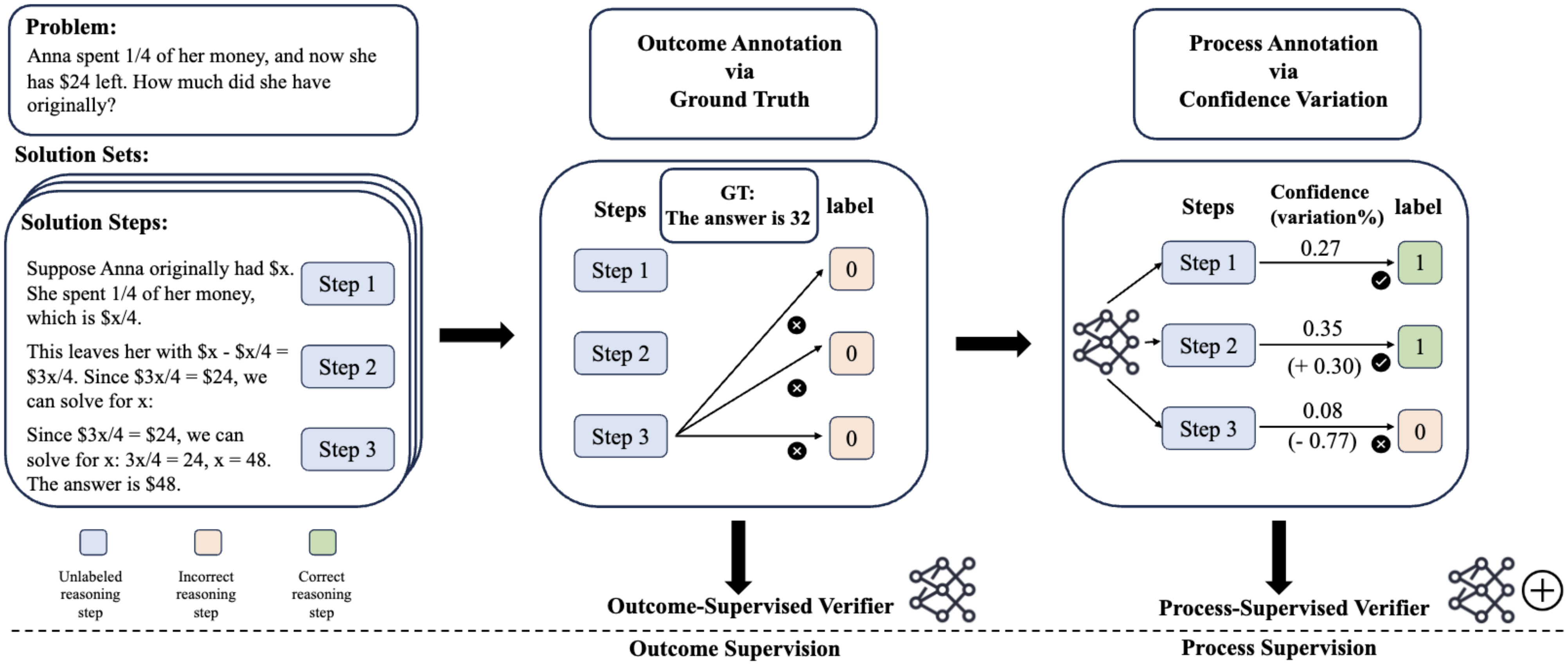
Training Methodology



Given an LLM acting as a response generator, we seek to annotate each reasoning step and perform response selection.

We train an outcome-supervised verifier based on the ground-truth answers.

Training Methodology



Given an LLM acting as a response generator, we seek to annotate each reasoning step and perform response selection.

We train an outcome-supervised verifier based on the ground-truth answers.

We then train a process-supervised verifier to annotate steps via confidence variation.

Preliminary Findings

1. Good Performance of Outcome-Supervised Verifier for Response Selection Task

Table 3: Performance of OSV models across different configurations.

Response Generator	Pass@1	Pass@5	SC	OSV (Mistral)	OSV (Phi)
Mistral-Instruct	42.08	69.90	50.03	60.72	52.61
Mixtral-Instruct	62.55	82.31	69.06	74.07	69.37
Qwen	77.03	91.13	81.27	85.00	84.19

2. High Efficiency of Δ_{conf}^{t} for Detecting Calculation Error During Math Reasoning

Table 5: Process Calculation Error Detection Performance with Varying Threshold (θ) Values.

Metric	Threshold (θ) Value				
	0.5	0.6	0.7	0.8	0.9
Prec.	0.85	0.88	0.91	0.93	0.94
Recall	0.90	0.89	0.86	0.83	0.80
F1-Score	0.88	0.89	0.88	0.88	0.86

Experiment: Main Results

Mathematics Reasoning

Table 6: Results on mathematics benchmarks.

Response Generator	GSM8K				MATH			
	Pass@5	Self-Cons.	OSV	OSV + PSV	Pass@5	Self-Cons.	OSV	OSV + PSV
Mistral-Instruct	69.90	50.03	61.18	**61.41**	7.7	1.64	5.10	**5.30**
Mixtral-Instruct	82.30	69.06	74.91	**76.04**	22.80	10.66	15.2	**16.92**
Qwen	91.13	81.27	84.91	**85.15**	56.10	**40.10**	38.94	39.36

Commonsense Reasoning

Table 7: Results on commonsense reasoning benchmarks.

Response Generator	HellaSwag				Winogrande				ANLI			
	Pass@5	Self-Cons.	OSV	OSV + PSV	Pass@5	Self-Cons.	OSV	OSV + PSV	Pass@5	Self-Cons.	OSV	OSV + PSV
Mistral-Instruct	76.84	40.30	73.81	**74.45**	91.16	58.64	79.16	**79.98**	73.4	45.6	59.8	**59.3**
Mixtral-Instruct	84.05	73.67	82.83	**83.62**	79.16	68.75	73.40	**73.88**	68.4	59.0	62.9	**64.0**
Qwen	95.28	85.44	93.08	**93.99**	88.63	72.21	**80.34**	79.32	82.4	63.8	69.1	**71.4**

Experiment: Analysis

Performance in Labeled Settings

Performance Comparison

Response Generator	GSM8K Pass@5	GSM8K Self-Cons.	GSM8K Process (MCTS)	GSM8K Process (AUTOPSV)	MATH Pass@5	MATH Self-Cons.	MATH Process (MCTS)	MATH Process (AUTOPSV)
Mistral-Instruct	69.90	50.03	54.13	55.32	7.7	1.64	3.3	3.24
Mixtral-Instruct	82.30	69.06	72.36	72.12	22.80	10.66	12.18	12.54
Qwen	91.13	81.27	82.17	82.83	56.10	40.10	36.88	37.10

Annotation Cost Comparison

Dataset	#Questions	#Solution Statistical				Annotation Cost	
		#Steps(Avg.)	#Steps(Overall)	#Tokens(Avg.)	#Tokens(Overall)	Process (MCTS)	Process (AUTOPSV)
GSM8K	7,473	4.47	334,358	126	9,379,258	2,808	127
MATH	7,498	16.00	1,200,177	272	1,621,515,894	21,626	273

Performance in Unlabeled Settings

Further Performance Improvement

Response Generator	Pass@5	Self-Cons.	OSV (GSM8K)	MCTS (GSM8K)	OSV+PSV (GSM8K)	OSV+PSV (GSM8K+WizardLM)
Mistral-Instruct	69.90	50.03	61.18	60.82	61.41	63.11
Mixtral-Instruct	82.30	69.06	74.91	75.10	76.04	78.15
Qwen	91.13	81.27	84.91	84.85	85.15	86.77

Thanks!