
AUTOCV: Empowering Reasoning with Automated Process Labeling via Confidence Variation

Jianqiao Lu¹, Zhiyang Dou¹, Hongru Wang², Zeyu Cao³,
Jianbo Dai⁴, Yingjia Wan³, Yinya Huang⁵, Zhijiang Guo^{3†}

¹The University of Hong Kong ²The Chinese University of Hong Kong ³University of Cambridge
⁴University of Edinburgh ⁵City University of Hong Kong

jqlu@cs.hku.hk, zg283@cam.ac.uk

Abstract

In this work, we propose a novel method named **Automated Process Labeling via Confidence Variation (AUTOCV)** to enhance the reasoning capabilities of large language models (LLMs) by automatically annotating the reasoning steps. Our approach begins by training a verification model on the correctness of final answers, enabling it to generate automatic process annotations. This verification model assigns a confidence score to each reasoning step, indicating the probability of arriving at the correct final answer from that point onward. We detect relative changes in the verification’s confidence scores across reasoning steps to automatically annotate the reasoning process. This alleviates the need for numerous manual annotations or the high computational costs associated with model-induced annotation approaches. We experimentally validate that the confidence variations learned by the verification model trained on the final answer correctness can effectively identify errors in the reasoning steps. Subsequently, we demonstrate that the process annotations generated by AUTOCV can remarkably enhance the accuracy of the verification model in selecting the correct answer from multiple outputs generated by LLMs. Notably, we achieve substantial improvements across five datasets in mathematics and commonsense reasoning. The source code of AUTOCV is available at <https://github.com/rookie-joe/AUTOCV>.

1 Introduction

Large language models (LLMs) have shown impressive performance on various reasoning tasks [1–4]. Prior efforts primarily focus on specific prompting techniques, such as few-shot prompting with intermediate steps and augmented demonstrations [5–8]. While these methods have shown promise, their effectiveness is often task-specific, and designing prompts can be labor-intensive, leading to inconsistent results [9, 10]. Another approach to improve reasoning in LLMs is through instruction tuning or knowledge distillation [11–14]. These methods typically involve fine-tuning LLMs and require a large set of examples annotated with chain-of-thoughts (CoT). However, these approaches can be resource-intensive and may not always produce reliable results.

To address these challenges, verification techniques have emerged as a promising solution [15, 16]. Verification models are trained to evaluate and potentially correct the reasoning process generated by LLMs. This approach aims to mitigate the risk of relying solely on the top-1 result, which may not always be reliable [17, 18]. By reranking candidate responses, verification models can ensure higher accuracy and consistency in LLM outputs. Moreover, they provide valuable feedback for improving LLMs [19, 20] further.

[†]Corresponding Author.

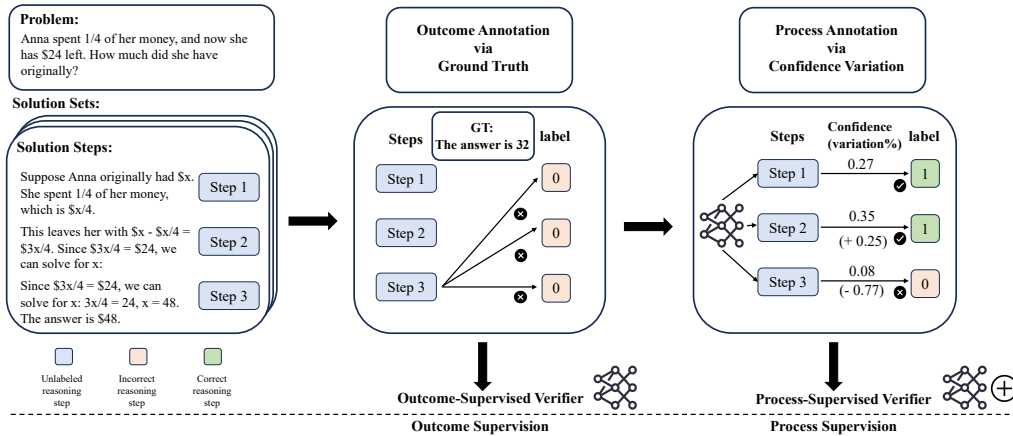


Figure 1: An overview of our framework. Our approach utilizes an outcome-supervised verifier to automatically generate process annotations for each reasoning step by detecting its own confidence variations, *without* relying on ground truth annotations. Our method efficiently produces annotations serving as process supervision during the LLM training, which sidesteps costly annotations.

Verification models generally fall into two training paradigms: outcome supervision and process supervision. In outcome supervision, the training annotations rely on the correctness of the final answer [21, 22], while in process supervision, the annotations are based on evaluations of each reasoning step [23, 19]. However, process supervision is demanding in terms of annotations. Typically, it relies on either expensive and highly skilled human evaluators [23, 16] or model-induced process annotations [18, 17] to estimate the future correctness of the current reasoning step using Monte Carlo tree search [24, 25]. In contrast, outcome supervision only requires annotations for the output, making it more economical in terms of annotation effort but less effective. That being said when answers involve multiple reasoning paths, aforementioned methods require numerous samples to ensure accurate estimations.

In this paper, we present a novel method named **Automated Process Labeling via Confidence Variation (AUTOCV)**, which enjoys both advantages of process supervision and output supervision. Our method starts by utilizing outcome supervision annotations to learn an outcome-supervised verification model. Then, this model assigns step-level scores to unannotated intermediate reasoning steps, which estimate the confidence in each reasoning step’s likelihood of leading to a correct final answer. By calculating the relative confidence variation, we could produce process annotations with significantly reduced annotation effort. With our process supervision data, we train an LLM model, resulting in a process-supervised enhanced verification model. The overall framework of our method is illustrated in Figure 1. We conduct extensive experiments across five datasets, including mathematical reasoning benchmarks and commonsense reasoning tasks. The results demonstrate that our method effectively improves the reasoning capability of the model with our highly efficient labeling scheme for process supervision, thus accomplishing two tasks with a single strategy. In summary, our contribution is as follows:

- We introduce AUTOCV, a method for automatically labeling process data to enhance the reasoning capabilities of LLMs. Our approach combines the strengths of output supervision and process supervision to annotate reasoning steps automatically.
- AUTOCV effectively identifies variations in model confidence to annotate the correctness of intermediate reasoning steps, enabling efficient automatic labeling for process supervision.
- Comprehensive experiments demonstrate that process supervision data generated by AUTOCV significantly improves the performance and scalability of verification models in mathematical and commonsense reasoning tasks, while greatly reducing the need for manual intervention and extensive computational resources.

2 Related Works

Improving Reasoning Abilities of LLMs To enhance the reasoning capabilities of LLMs, prior research primarily focuses on specific prompting techniques [26]. Existing efforts include few-shot prompting with intermediate steps augmented demonstrations [5, 8, 27] or zero-shot prompting with specific instructions [28, 29]. Although these methods have shown promising results, their effectiveness is often constrained by their task-specific nature and the labour-intensive process of designing prompts, leading to inconsistent outcomes across different tasks [9, 10]. Another strategy to facilitate reasoning involves instruction tuning or knowledge distillation, which elicits reasoning paths from LLMs without explicit prompting [12, 13, 30, 31]. These approaches typically involve resource-intensive fine-tuning over LLMs and require a large set of examples annotated with chain-of-thoughts (CoT). Unlike methods that directly modify parameters or prompts, AUTO CV focuses on training an additional verification model to select the desired output from the original model’s output. This approach is further discussed in the context of process supervision in the following paragraph.

From Outcome to Process Supervision Recent efforts have focused on enhancing the reasoning capabilities of LLMs through the use of verifiers to select the best answer from multiple candidates. There are two main types of verifiers: the Outcome-Supervised Verifier (OSV) and the Process-Supervised Verifier (PSV). The OSV is supervised with a signal based on the final answer [21, 22], while the PSV is with detailed feedback which requires evaluating individual reasoning steps [15, 19, 16, 23]. Despite the time-consuming annotation cost, the PSV offers several advantages that make it preferable to the OSV. It can provide fine-grained feedback by pinpointing the location of errors, which is valuable for reinforcement learning and automatic correction [16, 20]. To alleviate the extensive human annotation, recent approaches [17, 18] propose a machine annotation framework using Monte Carlo Tree Search [24, 25]. This annotation process demands a lot of computing resources, potentially imposing a limitation on the usage. AUTO CV is more efficient, as it utilizes an outcome-supervised verification model to assign confidence to each reasoning step and calculate relative confidence variations, eliminating the need for additional sampling or manual labeling.

3 Method

In the following, we first introduce the problem setting in Section 3.1. Then we discuss the motivation behind why we believe it is necessary to train a verification model in Section 3.2. Finally, we describe how we transition from outcome supervision to process supervision during the training of the verification model in Section 3.3.

3.1 Problem Setting

Given an LLM as a *response generator* to produce multiple responses for a given question, we aim to have an effective method to help us select a correct response from the multiple candidates. In other words, our goal is to maximize the probability of choosing a correct solution from the multiple options. We refer to this as *response selection*.

We also outlier the notations utilized in our following training framework. q : The specific question posed to the model. $S_i^{(1:t)}$: The intermediate reasoning steps for question q up to and including step t within the i -th solution. y_i : The label of correctness for the i -th solution. \mathcal{Q} : The dataset of questions used for training.

3.2 Motivation

The methods for response selection can be categorized into two main types: models fine-tuned specifically for the choice task and those utilizing different prompting strategies.

In our exploration, we initially investigate whether existing open-source LLMs could serve as effective selection agents to evaluate model outputs and choose the correct response without fine-tuning. We choose Mixtral-Instruct [32] as the *response generator*. We employ the metric **pass@k** to evaluate model performance, defined as the scenario where at least one correct instance is within the model’s first k attempts. The evaluation results of the response from Mixtral-Instruct are detailed in Table 1. To arrive at a comprehensive and reliable conclusion, we then employ models of different sizes as

the *selector*, ranging from 7b to over 70b parameters. And we apply different prompting strategies to assess their performance on the task of response selection. The experiment results, summarized in Table 2, showcase the comparison of different prompt strategies across different model sizes tested on GSM8K test set [21]:

Table 1: Performance of Mixtral-Instruct on the GSM8K test set. All results are reported in %.

Response Generator	Model Size (Parameters)	Pass@1 (%)	Pass@5 (%)	Self-Consistency (%)
Mixtral-Instruct [32]	8 x 7B (MOE)	62.55	82.31	69.06

Table 2: Comparison of different selection methods across various model sizes (from 7B to 72B) for selecting a response from candidate responses generated by Mixtral-Instruct. All results are reported in %.

Selector	Model Size	Prompt Strategy				
		Pairwise	Classification	Classification + CoT	Scoring	Scoring + CoT
Mistral-Instruct [33]	7B	60.73	61.18	64.82	61.49	69.75
Mixtral-Instruct [32]	8 x 7B	58.83	59.14	67.40	61.79	65.58
Llama2-chat [34]	70B	59.28	62.70	66.79	59.74	62.93
Qwen [35]	72B	59.14	66.64	69.52	61.86	65.88

It can be observed that without fine-tuning and relying solely on prompting, even models with over 70 billion parameters cannot yield efficient performance in the response selection task. In the following discussion, we will introduce our approach to enhance the performance of LLMs in response selection. To this end, we will concentrate on the training and improvement of verification models for this specific task.

3.3 Training Methodology

Our proposed AUTO CV training approach combines outcome supervision with process supervision effectively. We describe each step of AUTO CV in detail below.

Outcome-Supervision Initially, we train an outcome-supervised verification model based on whether the final answer in the solution is correct or not, following [22]. We denote the outcome-supervised verification (OSV) model as $f_{\theta}(\cdot)$, where θ represents the optimized parameters. The Mean Squared Error (MSE) loss used for outcome supervision for each solution step is defined as follows:

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

The overall objective function, incorporating the entire set of training questions \mathcal{Q} , is expressed as:

$$\mathcal{L}_{\text{total}}(\mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^{m_i} \left(f_{\theta}(S_i^{(1:t)}; q) - y_i \right)^2 \quad (2)$$

Where n denotes the total number of solutions proposed for each question, and m_i denotes the total number of reasoning steps within the i -th solution. This function aims to minimize the discrepancy between the model’s output and the degree of correctness of each answer, across all questions and solution paths.

The output of this trained OSV, $f_{\theta}(q; S^{(1:t)})$, approximates the expected value of y_i , i.e., the probability of correctness from the current step t to the final answer. The following theorem [22] clarifies this claim:

Theorem 1 *For a model trained with outcome supervision, f_{θ} , characterized by optimally tuned parameters θ , the assigned score for the sequence $S^{(1:t)}$ is an estimation of the likelihood of ultimately deriving a correct answer, denoted by \hat{a} , based on the progression observed in $S^{(1:t)}$ and the pertinent question q . This is mathematically represented as:*

$$f_{\theta}(S^{(1:t)}; q) \approx p(\hat{a} | S^{(1:t)}, q)$$

Proof of this theorem can be obtained by optimizing the MSE loss defined in the overall objective function (2), with further details provided in [22].

Moreover, we calculate the relative change in the confidence score between steps, represented as:

$$\Delta_{conf}^t = \frac{f_{\theta}(S^{(1:t+1)}; q) - f_{\theta}(S^{(1:t)}; q)}{f_{\theta}(S^{(1:t)}; q)} \quad (3)$$

Δ_{conf}^t represents the relative variation in the model’s confidence score from step t to step $t + 1$. A negative value of Δ_{conf}^t signifies a reduced confidence in achieving a correct answer after incorporating information from the $(t + 1)$ -th step.

We denote the process label for the t -th step as y_i^t and θ as the variation threshold. Our automated process labelling AUTOCV follows the “first error location” strategy [15] as following:

1. If $\Delta_{conf}^t > \theta$, then $y_i^t = 1$;
2. Otherwise, $y_i^t = 0$ and $\forall t' > t, y_i^{t'} = 0$ regardless of $\Delta_{conf}^{t'}$, i.e., if any step contains an error in a reasoning problem, the subsequent steps and the final result are considered incorrect.

Process-Supervision We further fine-tune the OSV in a process-supervision manner.

$$\mathcal{L}_{proc}(S_i^{(1:t)}, y_i^t; q) = \left(f_{\theta}(S_i^{(1:t)}; q) - y_i^t \right)^2. \quad (4)$$

4 Preliminary Findings

In this section, we present our findings aiming to validate two key aspects: In Section 4.1, we present a comprehensive analysis of the OSV model, i.e., to validate that the initially trained OSV model is effective and robust. In Section 4.2, we further introduce a self-designed benchmark for process errors and calculate Δ_{conf}^t to detect these errors, i.e., to demonstrate the effectiveness and reliability of relative confidence variation in the proposed method. The validation of these two components serves as a foundation for automatic process labeling via AUTOCV, as described in Section 3.3.

4.1 Experiment on Outcome Supervised Verifier Performance

In this section, we validate the effectiveness and scalability of the OSV model. Initially, we fine-tune a pretrained language model using ground truth data from the GSM8K dataset. Then, we use this fine-tuned model to generate multiple response samples for the GSM8K training prompts. We label these samples based on the correctness of their final answers. After this, we train an OSV model using the method described in Eq. (1).

To evaluate the OSV, we measure its ability to select a sample with the correct final answer from samples generated by various LLMs, denoted as the *Response Generator*. Specifically, we assess the effectiveness of outcome supervision on two models: Phi2 (**OSV-Phi**) [36] and Mistral-7B (**OSV-Mistral**) [33]. To explore the scalability of this outcome-supervised verifier effect, we choose *response generators* of varying scales, ranging from 7B to 72B parameters, i.e., Mistral-7B-Instruct (**Mistral-7B**) [33], Mixtral $8 \times 7B$ (**Mixtral-8×7B**) [32] and Qwen-72B-Chat (**Qwen-72B**) [35]. We employ the metric *pass@k* [37] to assess the performance of those *response generators*, defined as the scenario where at least one correct instance is within the model’s first k attempts. We evaluate the OSV’s generalized selection capability across different LLM scales.

Table 3: Performance of OSV models across different configurations. The selection of trained OSV models is based on 5 outputs from *response generators*.

Response Generator	Pass@1	Pass@5	SC	OSV-Mistral	OSV-Phi
Mistral-7B	42.08	69.90	50.03	60.72	52.61
Mixtral-8×7B	62.55	82.31	69.06	74.07	69.37
Qwen-72B	77.03	91.13	81.27	85.00	84.19

The results demonstrate the effectiveness and scalability of the OSV model in selecting the correct response among multiple responses generated by different generators. Specifically, the OSV models,

trained using either Mistral or Phi, consistently outperform the self-consistency (SC) baseline across all generator configurations. The results validate the effectiveness of the OSV model in enhancing model selection strategies, particularly when applied to larger and more accurate LLM generators.

We further analyze the performance discrepancy between the two OSV models:

Performance Analysis of Different OSVs The performance disparity among the verifiers can be primarily attributed to variations in model sizes and the accuracy of their training data. Specifically, the limited selection capability of the OSV-Phi model stems from discrepancies in the accuracy of its training data. Table 4 presents a consolidated view of the model sizes alongside the accuracy metrics of their outcome supervision training data.

Table 4: Model sizes and training data accuracy for training OSVs.

Verifier	Size	Training Data	
		Quality (acc.%)	Quantity (per question)
OSV-Mistral	7B	0.9914	100
OSV-Phi	2.7B	0.9605	100

In our content, we select **OSV-Mistral** as the OSV model among other experiment settings due to its superior performance, as demonstrated in Table 3.

4.2 Detecting Hallucination During Math Reasoning

In this section, we verify the effectiveness and reliability of our method AUTOCV. Specifically, We calculate Δ_{conf}^t to identify inaccuracies in the process, as outlined in Eq. (3).

In Section 4.2.1, we introduce the concept of *Process Calculation Hallucination* (PCH) and establish a preliminary benchmark. In Section 4.2.2, we assess the performance of calculating Δ_{conf}^t to detect these PCH against our established benchmark.

4.2.1 Process Calculation Hallucination

We outline a method for identifying instances of PCH, which we define as occurrences where the numerical values on either side of an equals sign within a mathematical expression do not align. This misalignment indicates a breakdown in logical reasoning, categorizing the instance as a hallucination in the context of mathematical problem-solving. This process establishes a benchmark for PCH detection with more details in Appendix E.

Process Calculation Hallucination Detection To identify hallucinations in mathematical reasoning, we monitor the relative confidence variations between the intermediate steps as defined in Eq. (3). If $\Delta_{conf}^t \leq \theta$, the corresponding step would be viewed as “incorrect”. Samples of detection are presented in Figure 2 of Appendix E.

4.2.2 Quantitative Results

We employ three metrics for a thorough evaluation of process calculation hallucination detection:

Precision, the ratio of samples accurately identified as true positives to the total number of positive predictions detected by the OSV model via confidence variations. In our setting, this metric calculates the proportion of samples that have correct final answers but exhibit hallucinatory errors during the reasoning process, relative to the total examples detected by the OSV model.

Recall, which determines the proportion of samples with hallucinatory computational errors that the OSV model successfully identifies through confidence variations;

F1-score, which gauges the verifier’s overall efficacy. In Table 5, we explore how different threshold (θ) values affect the accuracy, recall, and F1-score for process calculation hallucination detection.

The results in Table 5 demonstrate that our method using confidence variation effectively detects calculation errors across threshold values from - 0.5 to - 0.9. As the threshold becomes more negative

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

Metric	Threshold (θ) Value				
	- 0.5	- 0.6	- 0.7	- 0.8	- 0.9
Precision	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

(stricter for labeling errors), the precision increases, indicating higher precision in identifying true errors. However, the recall decreases, meaning fewer actual errors are caught. Importantly, the F1-score, balancing precision and recall, remains relatively stable across thresholds. This demonstrates that our method strikes a good balance between detecting real errors and minimizing incorrect flagging of valid calculations. Overall, our detection method is effective and robust, performing well over a range of thresholds without significantly compromising overall detection quality.

We find that setting $\theta = -0.5$ in our detection methods helps maintain a balance between precision and recall, which can ensure a balanced distribution of labeled “incorrect” and “correct” responses.

4.3 Validation and Foundation for AUTO CV

By validating these two points, we establish an experimental basis for automating process annotations using our proposed method, AUTO CV. Meanwhile, Theorem 1 provides the theoretical groundwork for leveraging the OSV model to estimate the likelihood of arriving at a correct final answer from an intermediate reasoning step onward. These combined theoretical and empirical insights lay a solid foundation for applying AUTO CV to construct process-supervision training data for the main experiments that follow.

5 Experiment

In this section, we first introduce the experimental setup in a subsection, which includes the response generator LLMs and evaluation settings in Section 5.1. We then present the main result of our process supervision-enhanced verification model on both mathematical and commonsense reasoning benchmarks, as described in Section 5.2.

5.1 Experimental Setup

Models: We selected three different LLMs of varying sizes, ranging from 7 billion to over 70 billion parameters, to serve as the *response generator*. Specifically, we used Mistral-Instruct-7B (**Mistral-Instruct**), Mixtral-8x7B-Instruct-v0.1 (**Mixtral-Instruct**), and Qwen-72B-Chat (**Qwen**). Note that all of these are instruction fine-tuning versions.

Datasets: We conduct experiments on five datasets. For mathematical reasoning, we include GSM8K [21], a dataset containing math word problems requiring multi-step reasoning, and MATH [38], composed of high school-level competition problems covering a range of math subjects. For commonsense reasoning, we use HellaSwag [39], a dataset for physically situated commonsense reasoning, Winogrande [40], fill-in-the-blank problems requiring commonsense pronoun resolution and ANLI [41], a dataset for natural language understanding and reasoning.

Evaluation: We follow evaluation metrics in [42] to ensure consistency across all benchmarks. To ensure more reliable $\text{pass}@k$ results, we utilized the estimation method described in [43]. We generated n samples per task, where n is greater than k , and evaluated the number of correct samples that passed unit tests. We then calculated the unbiased estimator for $\text{pass}@k$. For the self-consistency (**Self-Cons.**) and verifier’s selection results, we randomly choose k out of n samples and conducted separate calculations. The results are reported with an accuracy of ± 0.1 at a 95% confidence level. Further details are provided in Appendix F.2.

5.2 Enhanced LLMs Reasoning via Process Supervision

We report experimental results on both mathematical and commonsense reasoning tasks across five datasets to showcase the efficacy and scalability of our proposed approach as described in Section 5.2.

Specifically, we calculate Δ_{conf}^t based on the model confidence from OSV and set the threshold $\theta = -0.5$ to annotate process-supervision training data autonomously. We further leverage this process-supervision training data to fine-tune the OSV model continually, denoted as OSV + PSV.

For a comprehensive evaluation of our framework, we demonstrate its impact on both mathematics reasoning (GSM8K and MATH datasets) and commonsense reasoning (HellaSwag, Winogrande, and ANLI datasets). We note that Pass@5 represents the upper limit of performance on these benchmarks. We compare the performance of three models: Self-Consistency (**Self-Cons.**), outcome-supervised verifier (**OSV**), and process-supervised enhanced verifier (**OSV + PSV**).

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-7B	69.90	50.03	61.18	61.41	7.7	1.64	5.10	5.30
Mixtral-8×7B	82.30	69.06	74.91	76.04	22.80	10.66	15.2	16.92
Qwen-72B	91.13	81.27	84.91	85.15	56.10	40.10	38.94	39.36

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-7B	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-8×7B	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-72B	95.28	85.44	93.08	93.99	88.63	72.21	80.34	79.32	82.4	63.8	69.1	71.4

Mathematics Reasoning: As shown in Table 6, the process-supervised enhanced verifier demonstrates superior performance over the outcome-supervised verifier and Self-Consistency models for all evaluated response generators on GSM8K. For the MATH benchmark, the process-supervised enhanced verifier outperforms the other two approaches for Mistral-Instruct and Mixtral-Instruct, but it is slightly less effective than the Self-Consistency model when applied to Qwen-72b.

Commonsense Reasoning: According to Table 7, OSV + PSV again leads to the best results among the three methods for each response generator tested on HellaSwag. For Winogrande, Mistral-Instruct paired with OSV + PSV achieves the highest performance, whereas, for Mixtral-Instruct and Qwen-72b, the original OSV without process supervision has a marginal advantage. When looking at the results of the ANLI benchmark, OSV + PSV is the highest-performing method for Mistral-Instruct and Mixtral-Instruct. Despite this, for Qwen-72b, the OSV model alone falls slightly behind the integrated OSV + PSV.

In summary, our process-supervision enhanced OSV model (**OSV + PSV**) consistently matches or improves upon the OSV and Self-Consistency baselines across most benchmarks and response generator models, demonstrating the effectiveness of the automatic process annotation technique in enhancing the verifier’s capabilities for different reasoning tasks.

6 Analysis

In Section 6.1, we compare our process annotation method, AUTO CV with two other model-induced annotation methods to showcase the effectiveness and efficiency of our proposed approach. In Section 6.2, we validate the data quality constructed via AUTO CV as described in Section 5.2.

6.1 Different Process Labelling Strategy

Aside from the labeling method defined by Eq. (3) in AUTO CV, another labeling strategy is the Monte Carlo sampling estimation (MCS), as described in [17, 18]. To better demonstrate the effect of our method, we make a comparison with the approaches proposed by [17, 18], and conduct experiments on mathematical benchmarks. The results are shown in Table 8. We follow the experimental settings

described in their work to ensure a fair comparison. More implementation details are provided in Appendix F.3.

Table 8: Comparison of process labeling methods’ performance across different response generators on GSM8K and MATH datasets. The table evaluates the Pass@5, Self-Consistency (Self-Cons.), and response selection performance of models fine-tuned using process annotations labeled by MCS and our proposed method (AUTOCV).

Response Generator	GSM8K				MATH			
	Pass@5	Self-Cons.	Process (MCS)	Process (AUTOCV)	Pass@5	Self-Cons.	Process (MCS)	Process (AUTOCV)
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-72b	91.13	81.27	82.17	82.83	56.10	40.10	36.88	37.10

Table 9: Comparison of annotation costs between MCS and AUTOCV for process labeling on the GSM8K and MATH datasets. Annotation cost represents the processed tokens into a model when generating process annotations, encompassing both input and output tokens.

Dataset	#Questions	#Steps(Avg.)	#Solution Statistical			Annotation Cost	
			#Steps(Overall)	#Tokens(Avg.)	#Tokens(Overall)	Process (MCS)	Process (AUTOCV)
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

The experimental results shown in Table 8 suggest that our proposed method for process labeling, which relies on detecting changes in model confidence, performs competitively with the MCS method from [18, 17]. In some cases, our method even outperforms the MCS method, especially on the more challenging MATH benchmark.

It is important to note that, as mentioned in the analysis and shown in Table 9, our method is computationally more efficient than the MCS method, as it does not require generating costly multiple samples to label each reasoning step. This computational efficiency advantage of our method could be particularly beneficial for large-scale applications or scenarios with limited computational resources.

In summary, the experimental results and analysis demonstrate that our proposed process labeling method offers a promising alternative to the MCS method, providing competitive or better performance while being computationally more efficient. Moreover, further research can benefit from the advantages of both approaches, potentially achieving better overall performance and efficiency.

6.2 Outcome-Supervised Verification vs. Process-Supervised Verification

We apply the OSV to relabel the process-supervised training data as in Section 5.2. We then retrain a new model using this relabeled data. This experiment highlights the performance gap between outcome-supervised and process-supervised training.

Table 10: Experimental results showing the performance of OSV models across different configurations tested on GSM8K test sets.

Response Generator	Pass@1	Pass@5	SC	OSV	PSV
Mistral-Instruct	42.08	69.90	50.03	60.72	59.14
Mixtral-Instruct	62.55	82.31	69.06	74.07	71.39
Qwen-72b	77.03	91.13	81.27	85.00	83.70

The experimental results in Table 10 reveal that retraining the model with process supervision from AUTOCV still yields better performance than self-consistency across three different response generators. We also noticed a small performance gap between the PSV and OSV. It is worth noting that our PSV was trained using data from the OSV. The small performance gap between the PSV and OSV models demonstrates that the relabeled process-supervised training method successfully inherits information from the outcome-supervised model without requiring ground truth annotations. This ablation study further provides quality assurance for automatic process labeling via AUTOCV.

7 Conclusion

In this paper, we propose a novel method for automated process labeling via confidence variation (AUTOCV) to enhance the reasoning capabilities of LLMs by detecting relative changes in model confidence. It combines the strengths of output supervision and process supervision to annotate reasoning steps automatically. Extensive experiments demonstrate that AUTOCV significantly enhances the precision and scalability of the verifier models in various reasoning tasks, ranging from mathematical to commonsense reasoning. AUTOCV could considerably enhance existing LLMs’ performance while drastically reducing the need for intensive computation and manual intervention.

For future work, we aim to utilize the automatically constructed PSV to supervise the generator using step-wise proximal policy optimization, to enhance the accuracy of the generator’s output during greedy decoding without the need for subsequent reranking. This avenue of research could lead to even more advancements in the capabilities of LLMs and their application in reasoning tasks. The limitations and broader impact of the paper are discussed in Appendix A and B.

References

- [1] OpenAI. GPT-3.5 Turbo, 2023.
- [2] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.
- [3] Mistral AI. Au large, 2023.
- [4] Anthropic. Introducing the next generation of claude, 2023.
- [5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.
- [6] Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
- [7] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- [8] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
- [9] Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.
- [10] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
- [11] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu,

- Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022.
- [12] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of GPT-4. *CoRR*, abs/2306.02707, 2023.
- [13] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need. *CoRR*, abs/2306.11644, 2023.
- [14] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *CoRR*, abs/2308.09583, 2023.
- [15] Jonathan Uesato, Nate Kushman, Ramana Kumar, H. Francis Song, Noah Y. Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *CoRR*, abs/2211.14275, 2022.
- [16] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. *CoRR*, abs/2305.20050, 2023.
- [17] Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *CoRR*, abs/2312.08935, 2023.
- [18] Zihan Wang, Yunxuan Li, Yuexin Wu, Liangchen Luo, Le Hou, Hongkun Yu, and Jingbo Shang. Multi-step problem solving through a verifier: An empirical analysis on model-induced process supervision. *CoRR*, abs/2402.02658, 2024.
- [19] Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making language models better reasoners with step-aware verifier. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5315–5333. Association for Computational Linguistics, 2023.
- [20] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- [21] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021.
- [22] Fei Yu, Anningzhe Gao, and Benyou Wang. Outcome-supervised verifiers for planning in thematical reasoning. *CoRR*, abs/2311.09724, 2023.
- [23] Qianli Ma, Haotian Zhou, Tingkai Liu, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. Let’s reward step by step: Step-level reward model as the navigators for reasoning. *CoRR*, abs/2310.10080, 2023.
- [24] Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In H. Jaap van den Herik, Paolo Ciancarini, and H. H. L. M. Donkers, editors, *Computers and Games, 5th International Conference, CG 2006, Turin, Italy, May 29-31, 2006. Revised Papers*, volume 4630 of *Lecture Notes in Computer Science*, pages 72–83. Springer, 2006.

- [25] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.
- [26] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.
- [27] Jing Xiong, Zixuan Li, Chuanyang Zheng, Zhijiang Guo, Yichun Yin, Enze Xie, Zhicheng Yang, Qingxing Cao, Haiming Wang, Xiongwei Han, et al. Dq-lore: Dual queries with low rank approximation re-ranking for in-context learning. *arXiv preprint arXiv:2310.02954*, 2023.
- [28] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.
- [29] Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. Large language models as analogical reasoners. *CoRR*, abs/2310.01714, 2023.
- [30] Jianqiao Lu, Wanjun Zhong, Wenyong Huang, Yufei Wang, Fei Mi, Baojun Wang, Weichao Wang, Lifeng Shang, and Qun Liu. SELF: language-driven self-evolution for large language model. *CoRR*, abs/2310.00533, 2023.
- [31] Jianqiao Lu, Wanjun Zhong, Yufei Wang, Zhijiang Guo, Qi Zhu, Wenyong Huang, Yanlin Wang, Fei Mi, Baojun Wang, Yasheng Wang, et al. Yoda: Teacher-student progressive learning for language models. *arXiv preprint arXiv:2401.15670*, 2024.
- [32] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. Mixtral of experts. *CoRR*, abs/2401.04088, 2024.
- [33] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023.
- [34] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aur lien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023.

- [35] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [36] Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. Llava-phi: Efficient multi-modal assistant with small language model. *CoRR*, abs/2401.02330, 2024.
- [37] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021.
- [38] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *CoRR*, abs/2103.03874, 2021.
- [39] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics, 2019.
- [40] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press, 2020.
- [41] Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. Avoiding the hypothesis-only bias in natural language inference via ensemble adversarial training. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8281–8291. Association for Computational Linguistics, 2020.
- [42] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023.
- [43] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain,

- William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021.
- [44] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.
- [45] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073, 2022.
- [46] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023.
- [47] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *CoRR*, abs/2308.03188, 2023.
- [48] Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, and André F. T. Martins. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *CoRR*, abs/2305.00955, 2023.
- [49] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- [50] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

- [51] Yuxuan Yao, Han Wu, Zhijiang Guo, Biyan Zhou, Jiahui Gao, Sichun Luo, Hanxu Hou, Xiaojin Fu, and Linqi Song. Learning from correctness without prompting makes llm efficient reasoner. *arXiv preprint arXiv:2403.19094*, 2024.
- [52] Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. Re3: Generating longer stories with recursive reprompting and revision. *CoRR*, abs/2210.06774, 2022.
- [53] Sarah Pan, Vladislav Lialin, Sherin Muckatira, and Anna Rumshisky. Let’s reinforce step by step. *CoRR*, abs/2311.05821, 2023.
- [54] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujie Yang, Nan Duan, and Weizhu Chen. CRITIC: large language models can self-correct with tool-interactive critiquing. *CoRR*, abs/2305.11738, 2023.
- [55] Yiannis Charalambous, Norbert Tihanyi, Ridhi Jain, Youcheng Sun, Mohamed Amine Ferrag, and Lucas C. Cordeiro. A new era in software security: Towards self-healing software via large language models and formal verification. *CoRR*, abs/2305.14752, 2023.
- [56] Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: researching and revising what language models say, using language models. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16477–16508. Association for Computational Linguistics, 2023.
- [57] Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. Improving language models via plug-and-play retrieval feedback. *CoRR*, abs/2305.14002, 2023.
- [58] Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1266–1279. Association for Computational Linguistics, 2022.
- [59] Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. Generating sequences by learning to self-correct. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
- [60] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. *CoRR*, abs/2303.17651, 2023.
- [61] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- [62] Alec Helbling, Mansi Phute, Matthew Hull, and Duen Horng Chau. LLM self defense: By self examination, llms know they are being tricked. *CoRR*, abs/2308.07308, 2023.
- [63] Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Qizhe Xie. Self-evaluation guided beam search for reasoning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

- [64] Hao Yan, Saurabh Srivastava, Yintao Tai, Sida I. Wang, Wen-tau Yih, and Ziyu Yao. Learning to simulate natural language feedback for interactive semantic parsing. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3149–3170. Association for Computational Linguistics, 2023.
- [65] Seonghyeon Ye, Yongrae Jo, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, and Minjoon Seo. Selfee: Iterative self-revising llm empowered by self-feedback generation. Blog post, May 2023.
- [66] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *CoRR*, abs/2310.01798, 2023.

A Limitations

AUTOCV is a promising solution for enhancing the reasoning capabilities of LLMs. However, it is important to acknowledge several potential limitations of the method. Firstly, while AUTOCV aims to reduce the need for manual intervention, there is still a risk of inaccurate annotations. The relative confidence variation used to produce process annotations is an estimation and may not always accurately represent the actual correctness of a reasoning step. This could compromise the quality of the annotations and, in turn, the effectiveness of the method. Secondly, the success of AUTOCV is heavily dependent on the performance of the verifier. If the model is not accurate enough in its step-level scores, the quality of the process annotations generated by AUTOCV could be compromised. Thirdly, AUTOCV is specifically designed to improve the reasoning capabilities of LLMs. Therefore, its applicability may be limited to tasks that involve complex multi-step reasoning. It is unclear how well the method will scale or generalize to other tasks or domains that do not involve intensive reasoning. This is an important consideration for future research and development of the method.

B Broader Impact

Positive Societal Impacts The proposed AUTOCV has the potential to bring about several positive societal impacts. By enhancing the reasoning capabilities of LLMs, AUTOCV can lead to more accurate and reliable information, which in turn can support better decision-making in various sectors, including healthcare, education, and finance. Moreover, AUTOCV combines the strengths of output supervision and process supervision to automatically annotate reasoning steps, significantly reducing the time, effort, and cost associated with manual annotation. This makes the process of training LLMs more efficient and accessible. Additionally, the process supervision data generated by AUTOCV can improve the performance and scalability of verification models, allowing for the development of more complex and sophisticated LLMs capable of handling a wider range of tasks and applications.

Positive Societal Impacts However, AUTOCV also presents several potential negative societal impacts. The automation of the annotation process could lead to job displacement for individuals currently employed in this role. There is also a risk that AUTOCV and the enhanced LLMs could be misused, for instance, to spread misinformation or manipulate public opinion. The increased reliance on LLMs for decision-making could potentially result in a decrease in critical thinking and problem-solving skills among individuals. Furthermore, the use of LLMs in various sectors could lead to privacy and security issues, as these models often require large amounts of data for training.

C More Related Work

Learning From Feedback Improving LLMs through learning from feedback has become a prevalent strategy, notably through reinforcement learning from human feedback, which seeks to align LLMs with human values by refining their outputs based on feedback [44–46]. However, this method faces challenges such as high costs due to manual labor and a lack of real-time feedback capabilities [47, 48]. An alternative strategy involves using self-correcting LLMs, which rely on automated feedback to iteratively adapt and understand the consequences of their actions without

heavy reliance on human intervention. This feedback can be derived from inside sources such as the model itself [49, 50] or generation logits [51], and outside sources such as other models [52, 53], tools [54, 55], knowledge bases [56, 57], or evaluation metrics [58, 59].

External feedback leverages external perspectives to identify errors and verify factual accuracy, offering insights that may not be recognized by the LLM alone. Conversely, feedback can also be internally generated, where the LLM evaluates and refines its output iteratively until a desired quality is achieved [60–63]. This self-improvement mechanism is particularly valuable in scenarios where external feedback is scarce or restricted [64, 65]. However, recent effort [66] suggests that LLMs struggle to independently identify and correct errors through self-generated prompts.

D Vanilla Evaluation Methods Description

Classification: For this method, the evaluator is presented with multiple answers for a given question and is required to choose the best answer among them. The selection is made based on the evaluator’s judgment of which answer most accurately addresses the question or provides the most relevant information.

Given the following question: '[question]', and these five answers:

1. [answer]
2. [answer]
3. [answer]
4. [answer]
5. [answer]

Which answer is the best? Please provide the number of the best answer. You should strictly follow the output format requirements and not output any other content.

Example: Answer [number] is better. Let’s Begin!

Classification + COT: In Classification + COT, the evaluator must not only identify the best answer but also analyze and compare all provided answers before making their decision. This method requires a deeper examination of the content and context of each answer to determine its quality and relevance to the question.

Given the following question: '[question]', and these five answers:

1. [answer]
2. [answer]
3. [answer]
4. [answer]
5. [answer]

Which answer is the best? Please analyze and compare the provided answers and then identify the number of the best answers. You should strictly follow the output format requirements and not output any other content.

Example: Comparison and Analysis: [analysis]. Best answer: [number]. Let’s Begin!

Scoring: In the Scoring method, the evaluator assigns a numerical score to each answer based on its quality or relevance to the given question. The scores typically range from 1 to 10, with 10 representing the highest quality or most relevant answer.

Given the following question: '[question]', please score these five answers on a scale from 1 to 10, where 10 is the best:

1. [answer]
2. [answer]
3. [answer]
4. [answer]
5. [answer]

Please provide a score for each answer. You should strictly follow the output format requirements and not output any other content.

Example: Answer i: [score]. Let's Begin!

Scoring + COT: Similar to Scoring, Scoring + COT also involves assigning numerical scores to each answer. However, in Scoring_cot, the evaluator is required to provide an analysis of each answer before assigning a score. This analysis informs the scoring process and ensures a more informed evaluation of each answer.

Given the following question: '[question]', please score these five answers on a scale from 1 to 10, where 10 is the best:

1. [answer]
2. [answer]
3. [answer]
4. [answer]
5. [answer]

Please analyze each answer, and then provide a score for each answer. You should strictly follow the output format requirements and not output any other content.

Example: Answer i: analysis: [analysis]. score: [score]. Let's Begin!

Pairwise Comparison: In this method, the evaluator is presented with pairs of answers for a given question and is tasked with determining which answer is better in each pair. The evaluator compares the content or quality of each answer and selects the one they deem superior. The Pairwise Comparison method differs from other evaluation methods in that it evaluates two candidate answers at a time and chooses the winner to proceed to the next comparison with the next candidate answer. For a set of n candidates, this method conducts $n-1$ pairwise comparisons. To mitigate the potential order preference bias exhibited by LLMs, we adopt a method similar to [30] which shuffles the positions of two answers during prompting. This ensures a fair evaluation process by eliminating any bias toward the position of the answers.

Given the following question: '[question]', compare each pair of answers and decide which one is better:

Compare 1. [answer1] with 2. [answer2]

For the comparison, indicate the better answer with its number. You should strictly follow the output format requirements and not output any other content.

Example: Answer i is better. Let's Begin!

By employing these different evaluation methods, we aim to comprehensively assess the quality and relevance of the answers generated by our models for various questions. Each method offers a unique perspective and contributes to a more thorough evaluation process.

E PCH benchmark

In our approach, we employ the LLaMA2-chat model (LLaMA) to generate steps for mathematical reasoning. Through the use of regular expressions, we isolate steps that involve numerical calculations. These steps, specifically the expressions to the left of the "=" sign, are then evaluated using Python's *eval* function to verify their correctness against the results on the right-hand side. We denote it as "PCH Detection" and exhibit the results in Table 11.

Table 11: Performance of the LLaMA model on GSM8K training data, evaluated through few-shot cot prompting strategy.

Model	Pass@5 (%)	Self-Consistency (%)	PCH Detection (%)
LLaMA	0.4791	0.2881	0.1824

Instances that contain expressions that cannot be evaluated due to unsolvability or incorrect formatting (for example, "x + 1 + 2 = 4") are categorized as non-computational and excluded from the ground truth data. This approach ensures that our ground truth reflects only the model’s computational errors during reasoning, avoiding any overstatement of its computational accuracy.

It’s worth noting that while organizations like DeepMind [15] have undertaken similar annotation tasks with numerous human labelers, referring to them as "trace errors," our method employs Python’s ‘eval’ function for automated labeling. This strategy provides a scalable and efficient way to approximate trace errors.

To further illustrate how we calculate Δ_{conf}^t to identify PCH error, consider the following example. As shown in Figure 2, the highlighted errors in red signify hallucinations during the reasoning process. This is detected by the model’s confidence decreasing significantly.

Question:

Randy just turned 12 and started playing the piano. His friend Sheila told him about the 10,000– hour rule which says, after 10,000 hours of practice, you become an expert or master in your field. If Randy wants to become a piano expert before he is 20, how many hours a day will he need to practice if he practices every day, Monday – Friday, and takes two weeks off for vacation each year?

Solution:

If Randy wants to become a piano expert before he is 20, he will need to practice for 10,000 hours in total. Since he is 12 now, that means he will need to practice for $10,000/20 = 500$ hours a year. [0.2]

If he practices every day, Monday – Friday, that means he will need to practice for $500*52 = 500*52=260$ 260 weeks [– 0.6].

Since he takes two weeks off for vacation each year, that means he will need to practice for $260/52 = 5$ hours a day [0.1].

The answer is 5.

Figure 2: **A data example during hallucination detection.** We apply the OSV model to detect calculation hallucinations by calculating the confidence variation at each step, denoted inside the brackets [] for the step containing calculations. The ground truth location of PCH is marked in red.

F Experiment Details

F.1 Training Hyperparameters

Our experiments were conducted in a computing environment equipped with 8 NVIDIA A100 GPUs, each having 40GB of memory. All models are fine-tuned in a full-parameter setting.

We employed the AdamW optimizer for model training over 1 epochs, with a batch size of 512. The learning rate was set at $2e \times 10^{-6}$, incorporating a 3% learning rate warmup period. Below, we present a comprehensive overview of the training hyperparameters utilized. These parameters were

consistently applied across training both process-supervised and outcome-supervised methods in our experiments in Table 12.

Table 12: Verifier Training Hyperparameters.

Hyperparameter	Global Batch Size	LR	Epo.	Max Length	Weight Decay	Warmup Ratio
Value	512	2×10^{-6}	1	2048	0	0.03

Before we train a verifier, it is essential to first train a supervised fine-tuning model (SFT). This SFT model enables us to generate responses and label outcome supervision data by checking whether the final answer being correct or not. The training parameters for the SFT model are outlined in Table 13. For further training details regarding OSV, refer to [22].

Table 13: SFT Training Hyperparameters.

Hyperparameter	Global Batch Size	LR	Epo.	Max Length	Weight Decay	Warmup Ratio
Value	128	5×10^{-6}	2	2048	0	0.03

F.2 Generation Settings

For the generation of results using the "greedy" strategy, we set the temperature parameter to 0.0 and 0.7 for the "pass@k" strategy. To present unbiased results for "pass@k", we follow the calculation method outlined in [43]. Specifically, we generate $n = 20$ samples for each instance, evaluate the number of correct samples passing unit tests, and then calculate the unbiased estimator for pass@k. We repeat the above experiments 5 times to obtain the 95% confidence intervals reported in the Section 5.1

F.3 Process-Supervision Labelling Strategy

We detail the implementation process for both our proposed method (Process Ours) and the comparative method (Process MCS). Initially, we train a generator through standard supervised fine-tuning using training datasets from GSM8K and MATH, comprising approximately 15,000 instances.

Following this, we utilize the trained generator to produce 10 samples for each unlabeled prompt within the training datasets. These samples are then labeled as either 0 or 1 based on the accuracy of the final answer obtained.

For Process Ours, we subsequently train an output-supervised verifier using the total of $15,000 * 10$ generated samples. These samples are then relabeled using a relative confidence change criterion (refer to Eq. 3), followed by retraining a Process Ours-supervised verifier.

In the case of Process MCS, each of the $15,000 * 10$ samples is divided into reasoning paths. For each incomplete reasoning path, we employ the generator to generate eight complete reasoning paths. The accuracy of the final answer is then computed for each complete path, and the fraction of correct paths is utilized to relabel the samples. Finally, a Process MCS-supervised verifier is retrained.

Note: We do not sample 50 times for each training problem as we do in the main results, since $50 * 15,000$ data points for MCS labeling would take approximately 10-12 days.