

Learning to Insert [PAUSE] Tokens for Better Reasoning

Eunki Kim* Sangryul Kim* James Thorne

KAIST AI

{eunkikim, sangryul, thorne}@kaist.ac.kr

Abstract

To enhance reasoning capabilities, previous works have explored incorporating special-purpose tokens into the training process. These strategies strengthen the learning mechanism of transformer-based large language models (LLMs). Building on prior research, in which inserting dummy tokens consecutively just before reasoning steps can enhance effectiveness, we introduce a novel approach termed **Dynamic Inserting Tokens Training (DIT)**. Our method identifies positions within sequences where model confidence is lowest according to token log-likelihood. Strategically inserting [PAUSE] tokens on these positions bolsters the model’s predictive capabilities for subsequent tokens. Experimental results across diverse datasets and models, from the 2.7B model to the 8B model, demonstrate that DIT consistently outperforms traditional fine-tuning and previous token insertion methods. With this simple yet effective method, we achieve accuracy gains of up to 4.7%p on GSM8K, 3.23%p on AQUA-RAT, and pass@1 improvements of up to 3.4%p on MBPP datasets. Our work shows a model-based, dynamic approach rather than a heuristic one, thereby broadening the scope of research in reasoning¹.

1 Introduction

The primary mechanism in the text generation of recent large language models (LLMs) with the decoder-based transformer architecture (Vaswani et al., 2017; Radford and Narasimhan, 2018) involves autoregressive calculation. These models evaluate the probability of the best subsequent token, determined by the sequence of preceding tokens (Vaswani et al., 2017; Radford et al., 2018; Touvron et al., 2023). This next token prediction

*Equal contribution.

¹Code and datasets are available at <https://github.com/xfactlab/acl2025-dit>

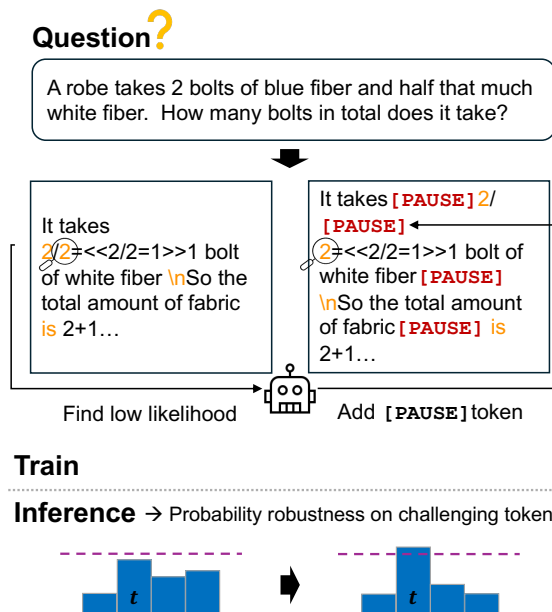


Figure 1: We propose DIT, a log-likelihood-based [PAUSE] token inserting methodology. We suggest that learning with [PAUSE] token insertion using this method will result in a probability boost for low-probability tokens during inference.

(NTP) paradigm has enabled remarkable advances in language understanding and generation, with performance scaling predictably as model size and training data increase, by scaling laws (Brown et al., 2020; Kaplan et al., 2020). However, these improvements come at the cost of substantial computational and energy demands, posing significant challenges for efficiency and sustainability (Samsi et al., 2023; Faiz et al., 2023).

To effectively adapt LLMs for specific tasks, additional reasoning steps generally support adaptation, using approaches such as the Chain of Thought (Wei et al., 2022, CoT), which can be trained as a subsequent objective (Ho et al., 2022; Shridhar et al., 2022). While effective, training decoder-only LLMs with NTP loss based on previous input tokens does not always yield optimal re-

sults (Bachmann and Nagarajan, 2024; Dziri et al., 2024). For instance, LLMs often fail on surprisingly trivial problems in mathematical reasoning tasks (Dziri et al., 2024; Ahn et al., 2024). To address such limitations, recent studies have explored mechanisms for adding dynamic computational budgets for transformer architectures. For instance, the Universal Transformer (Dehghani et al., 2018) adapts model depth based on input complexity, while speculative decoding (Leviathan et al., 2023) offloads demanding reasoning to larger expert models.

Recently, Goyal et al. (2024) have shown that inserting dummy tokens during pre-training and decoding can act as a mechanism for introducing additional computational steps. Building on prior studies (Sukhbaatar et al., 2019; Xue et al., 2023; Darcet et al., 2023), these dummy tokens strategically delay the output of decoder-only models, creating additional computational pathways in the attention layers and improving performance. However, this method relies on the stochastic inclusion of the dummy tokens during pre-training, followed by their fixed placement (depending on hand-crafted rules) during fine-tuning.

Rather than relying on heuristic methods, we propose a model-based approach for dummy token insertion, focusing solely on fine-tuning. Our novel method identifies points with low token log-likelihoods, interpreted as regions of model uncertainty, and subsequently inserts dummy tokens (denoted as [PAUSE]) immediately preceding these positions. This approach is based on previous works that set token log probability as an indicator of the model’s confidence score (Nguyen and O’Connor, 2015; Jiang et al., 2023; Kadavath et al., 2022). We conduct experiments on two reasoning tasks: mathematical reasoning and code reasoning, using three datasets (GSM8K (Cobbe et al., 2021), AQUA-RAT (Ling et al., 2017), and MBPP (Austin et al., 2021)). Our results demonstrate the broad applicability and effectiveness of the proposed method, yielding consistent performance gains. Specifically, our method shows performance improvements over existing methods, with gains of up to 4.7%p on GSM8K, 3.23%p on AQUA-RAT in accuracy, and a 3.4%p increase in pass@1 on MBPP.

We demonstrate that by incorporating [PAUSE] tokens, the model can focus more on challenging tokens with higher loss, resulting in a higher training signal for harder tokens, improving training and inference. By using dummy tokens as a contributor

to the loss term for succeeding tokens, we highlight the importance of learning difficult tokens, supported by an in-depth analysis of token probability. Furthermore, our experimental analysis reveals how [PAUSE] tokens enhance the probability robustness of outliers, solidifying their role in improving overall model performance.

To summarize, our contributions are as follows:

- We propose a novel method for dynamically inserting dummy tokens into sequences based on token likelihood (DIT).
- We demonstrate performance improvements with our approach, achieving up to 4.25%p higher accuracy on GSM8K, 3.23%p on AQUA-RAT, and up to 3.4%p improvement in pass@1 on MBPP.
- By using dummy tokens as a loss contributor, we highlight the importance of learning challenging tokens, supported by an in-depth analysis of token probability.

2 [PAUSE] Token

Goyal et al. (2024) introduce a method for pre-training models with dummy tokens (denoted as the <pause> tokens) that enables arbitrary allocation of additional computation at run-time through the rule-based insertion of the <pause> tokens at inference. This approach involves using $(K + M)$ vectors instead of just the K existing vectors when generating the $(K + 1)^{th}$ token by adding M <pause> tokens to each dataset instance. In their study, the authors introduce a method for incorporating <pause> tokens into a pre-training sequence $y_{1:N}$. They insert multiple <pause> tokens at random positions represented as M_{pt} , throughout the sequence to get a <pause>-injected sequence, $\tilde{y}_{1:N+M_{PT}}$. The model is trained using the next token prediction (NTP) loss while disregarding the loss associated with predicting the <pause> tokens. The positions where the next token is a <pause> token are denoted by $S_{ignore} = \{k : \tilde{y}_{k+1} = \text{<pause>}\}$. The pre-training loss for a decoder-only language model f_θ is formulated as follows:

$$\begin{aligned} & \mathcal{L}_{PausePT}(f_\theta, \tilde{y}_{1:N+M_{PT}}) \\ &= \sum_{k=1, k \notin S_{ignore}}^{N+M_{PT}-1} \mathcal{L}_{CE}(\tilde{y}_{k+1}, f_\theta(\tilde{y}_{1:k})), \end{aligned} \quad (1)$$

where \mathcal{L}_{CE} denotes the cross-entropy loss. For fine-tuning, given a prefix $y_{1:N}$ annotated with target

$\mathbf{t}_{1:T}$, they append a series of <pause> tokens (denoted as M_{ft}) to the end of the prefix, resulting in a <pause>-appended prefix $\tilde{y}_{1:N+M_{ft}}$. Using this <pause>-appended prefix, the model is fine-tuned with the NTP loss on the target $\mathbf{t}_{1:T}$, minimizing the loss as follows:

$$\sum_{k=0}^{T-1} \mathcal{L}_{CE}(\mathbf{t}_{k+1}, f_{\theta}([\tilde{y}_{1:N+M_{ft}}, \mathbf{t}_{1:k}])), \quad (2)$$

where $[\cdot]$ denotes the concatenation operation. Authors suggest that inserting <pause> tokens during pre-training and their subsequent utilization during fine-tuning for downstream tasks can create additional computational pathways within the attention layers and the associated MLP layers. This approach has demonstrated significant performance enhancements across various downstream tasks.

However, they show that merely using <pause> tokens for fine-tuning without its prior insertion during pre-training does not result in substantial performance gains. Additionally, in reasoning tasks such as GSM8K, the performance deteriorates when the <pause> token insertion is exclusively implemented during the fine-tuning stage. To address these limitations, we propose a targeted approach for the effective placement of dummy tokens (denoted as [PAUSE]) in reasoning tasks and clarify their role in model training. Previous studies have primarily focused on the pre-training phase, which is resource-intensive. Instead, our research demonstrates that significant performance gains can be achieved by strategically inserting [PAUSE] tokens during the fine-tuning phase. Unlike the arbitrary insertion of [PAUSE] tokens during pre-training, our method identifies optimal positions during fine-tuning, enhancing performance even in models not pre-trained with them.

3 Methods

3.1 Training with [PAUSE] Token

To incorporate the [PAUSE] token during fine-tuning, we first determine the positions of effective insertion. Unlike previous research, our novel method activates [PAUSE] token insertion during the training phase, based on token log-likelihood rather than heuristic or rule-based approaches. Since the inference of the language model starts from the prefix (*e.g.*, prompt, question) and predicts the target (*e.g.*, output), we strategically insert [PAUSE] tokens within the target sequence, which is our primary focus. By executing

a forward pass of the model, we compute the log-likelihoods of each token and pinpoint positions where the log-likelihood is lowest in the target sequence. Subsequently, we train the model using this modified dataset incorporated [PAUSE] tokens. At this point, we provide a hyperparameter M_{DIT} to tune the number of [PAUSE] tokens to insert per data instance. The pipeline is described in Figure 2, with an example data.

Consider a sequence $Y_{1:N} = \{y_1, y_2, \dots, y_N\}$ that the language model is learning. In the traditional fine-tuning scenario (SFT), the loss for a token y_{k+1} , which ranks within the top M_{DIT} lowest log-likelihoods at time step k , is given by:

$$\mathcal{L}_{SFT}(f_{\theta}, y_{k+1}) = \mathcal{L}_{CE}(y_{k+1}, f_{\theta}(y_{1:k})), \quad (3)$$

where \mathcal{L}_{CE} denotes the cross-entropy loss. The overall loss for fine-tuning on the sequence $Y_{1:N}$ is given by:

$$\mathcal{L}_{SFT}(f_{\theta}, Y_{1:N}) = \sum_{k=1}^{N-1} \mathcal{L}_{CE}(y_{k+1}, f_{\theta}(y_{1:k})). \quad (4)$$

Conversely, in the DIT (**D**ynamic **I**nserting **T**okens **T**raining) scenario, where the [PAUSE] tokens are inserted, the loss becomes:

$$\mathcal{L}_{DIT}(f_{\theta}, \tilde{y}_{k+1}) = \mathcal{L}_{CE}(\tilde{y}_{k+1}, f_{\theta}([\tilde{y}_{1:k}, [\text{PAUSE}]])), \quad (5)$$

where $\tilde{Y}_{1:N+M_{DIT}} = \{\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_{N+M_{DIT}}\}$ denotes pause-injected sequences with M_{DIT} [PAUSE] tokens inserted. The overall loss for DIT on the sequence $\tilde{Y}_{1:N+M_{DIT}}$ is given by:

$$\begin{aligned} & \mathcal{L}_{DIT}(f_{\theta}, \tilde{Y}_{1:N+M_{DIT}}) \\ &= \sum_{k=1, k \notin S_{ignore}}^{N+M_{DIT}-1} \mathcal{L}_{CE}(\tilde{y}_{k+1}, f_{\theta}(\tilde{y}_{1:k})). \end{aligned} \quad (6)$$

This is similar to equation 1, but in the DIT scenario, S_{ignore} is determined based on token log-likelihood rather than random selection.

3.2 Heuristic Findings

We provide justifications and an overview of preliminary experiments that elucidate the rationale behind the development of DIT. In our prior experiments, we leverage a fine-tuned model to pinpoint optimal positions for inserting [PAUSE] tokens. During the token generation for a given input, we calculate the log probabilities of each token and

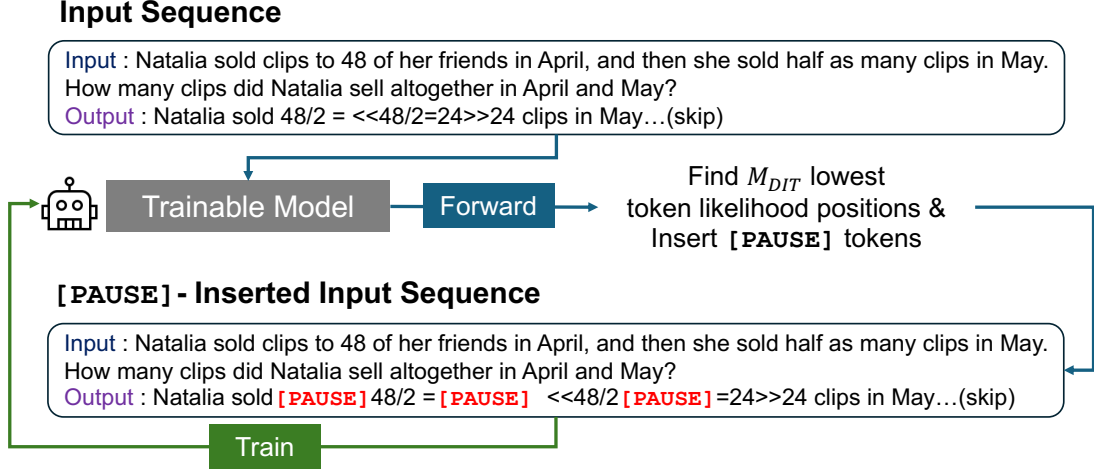


Figure 2: We illustrate the DIT training process, which is divided into two steps. In the first step, we perform a forward pass to identify M_{DIT} positions with low log-likelihood in the model’s input. In the second step, we insert [PAUSE] tokens at those positions and train the model using the modified data.

pinpoint the positions of the lowest tokens. Drawing on prior research that underscores log probability as a crucial indicator of a model’s confidence and its knowledge gaps (Jiang et al., 2023; Kada-vath et al., 2022), we identify strategic positions for inserting [PAUSE] tokens to enhance downstream task performance.

Interestingly, our observations reveal that log probabilities at specific points within each task tend to be low across experimental datasets. Training our models by inserting [PAUSE] tokens before these low-probability tokens in the training datasets results in noticeable performance improvements. This finding suggests that the model’s reasoning or generation confidence diminishes according to certain reasoning patterns, and the [PAUSE] token aids in learning these challenging tokens. By inserting [PAUSE] tokens at these critical locations, we aim to validate their effectiveness through changes in downstream metrics. However, this finding is based on heuristic and task-specific methods for determining the positions to insert [PAUSE] tokens. From this point, we report on automated methods to identify optimal positions for [PAUSE] token insertion and train the model accordingly. We report the results of these experiments in Appendix B.

4 Experiments

4.1 Comparison against other works

Our method is dynamically inserting [PAUSE] tokens based on the calculated token likelihood, rather than relying on statically predetermined posi-

tions. We validate our method through comparison against the following methods and baselines:

- **Dynamic Inserting Tokens Training (DIT):** Our method (§3.1) dynamically inserts a [PAUSE] token in front of any token that is among the bottom M_{DIT} tokens in terms of the likelihood of the generated output token.
- **Supervised Fine-Tune (SFT):** A model that is fine-tuned using only the given downstream data without inserting any tokens. This serves as the baseline experiment for all comparisons in our study.
- **Random (RAN):** Insert [PAUSE] tokens at random positions between output tokens for a given input sequence in the training set.
- **Append (APPD):** Appending [PAUSE] tokens at the end of the input in the given training set has been recognized as the best performing one in previous studies (Goyal et al., 2024). During inference, we also append [PAUSE] tokens at the end of the prefix, following established practices.
- **After All Words (AAW):** Herel and Mikolov (2024) suggests inserting so-called “thinking tokens(<T>)” after each word in a sentence to introduce extra computational steps within the hidden layers, resulting in improved perplexity. To compare our method, we utilize this idea with the same setting, but we replace the thinking tokens with [PAUSE] tokens.

Base Model	Method	Dataset		
		GSM8K	AQUA-RAT	MBPP
Phi-2	SFT	51.63	40.15	<u>14.0</u>
	RAN	<u>54.20</u>	32.03	0.0
	APPD	52.38	<u>41.43</u>	10.4
	AAW	39.04	17.03	5.4
	DIT(ours)	56.33	43.38	17.4
Phi-3 Mini	SFT	<u>78.84</u>	62.46	32.2
	RAN	75.43	48.00	17.0
	APPD	79.45	63.22	<u>29.6</u>
	AAW	73.84	19.83	22.4
	DIT(ours)	78.62	<u>62.82</u>	32.2
Llama 3 8B	SFT	65.04	<u>70.71</u>	<u>12.6</u>
	RAN	60.95	54.92	9.4
	APPD	63.83	72.18	11.0
	AAW	64.06	19.25	9.2
	DIT(ours)	66.86	69.96	14.0

Table 1: **Impact of Dynamic Inserting Tokens Training (DIT) on Various Datasets.** In this table, we showcase the performance comparison across three base models—Phi-2, Phi-3 Mini, and Llama 3 8B—evaluated on three distinct datasets: GSM8K (accuracy), AQUA-RAT (accuracy), and MBPP (pass@1). The methods evaluated include Supervised Fine-Tune (SFT), Random (RAN), Append (APPD), After All Words (AAW), and our proposed DIT (ours). For each method, the best score is indicated in **bold**, and the second-best score is underlined.

4.2 Downstream Datasets

We conduct experiments utilizing datasets designed to assess the reasoning and generative capabilities of the models. These include two renowned mathematical reasoning datasets and one code reasoning dataset. The specific processing and settings are detailed in Appendix A.

Math Reasoning GSM8K (Cobbe et al., 2021) consists of 8.5K diverse grade school math word problems written in linguistic expression containing calculation annotations with special tokens(*e.g.*, ‘<’, ‘>’) where arithmetic calculation occurs. **AQUA-RAT** (Ling et al., 2017) is a dataset that includes a correct answer and a rationale, a natural language description of the solution for each given math problem, among five possible options. For both datasets, the metric is accuracy, which is determined by whether the final answer, generated through the rationales, is correct or incorrect.

Code Reasoning MBPP (Austin et al., 2021) is a dataset consisting of an instruction prompt, the corresponding Python function and module to be generated accordingly, and three test cases along with code to verify its functionality. The metric is pass@1, which requires the model to generate one

piece of code from the instruction that passes all provided test cases.

4.3 Models

In this experiment, we leverage well-known open-source large language models, specifically Phi-2 (Microsoft, 2023), which comprises 2.7B parameters. Additionally, we include the latest models, Phi-3 Mini (3.8B parameters) (Abdin et al., 2024) and Llama 3 (8B parameters) (AI@Meta, 2024), as baseline models. This diverse setup allows us to systematically compare the impact of the [PAUSE] token across various models with differing levels of pre-training and parameter counts. Details of experiment settings are described in Appendix C.

5 Results and Analysis

5.1 Impact of [PAUSE] Token

The experimental results, summarized in Table 1, clearly demonstrate the efficacy of the [PAUSE] token insertion, especially using the DIT method. Notably, DIT consistently shows optimal performance across multiple datasets and model configurations.

For the Phi-2 model, DIT outperforms all other methods, achieving performance improvements of 4.7%p on GSM8K, 3.23%p on AQUA-RAT, and

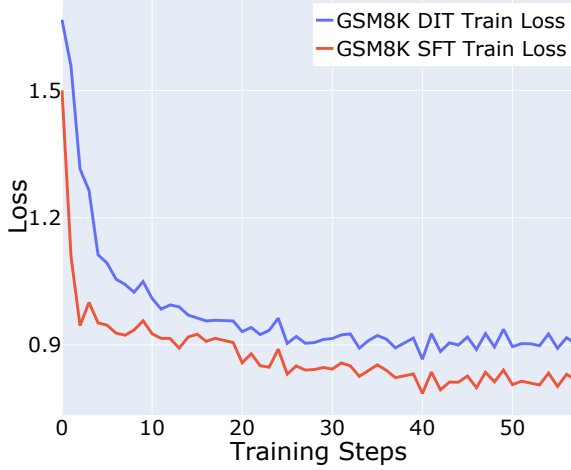


Figure 3: **Train Loss Comparison between SFT and DIT (ours) on GSM8K Dataset.** The DIT approach results in higher training loss compared to SFT, even when excluding losses associated with [PAUSE] tokens.

3.40%p on MBPP compared to SFT. These results indicate that our approach of inserting [PAUSE] tokens at strategically identified positions significantly enhances performance. Similarly, with the Phi-3 Mini model, DIT achieves a 0.36%p performance improvement on AQUA-RAT compared to SFT. In addition, DIT outperforms other token-insertion methodologies on MBPP. This further validates the effectiveness of our method across different model sizes and pre-training levels. For the Llama 3 8B model, DIT achieves performance improvements of 1.82%p on GSM8K and 1.4%p on MBPP compared to the baseline. For Llama 3 on AQUA-RAT and MBPP, the performance of DIT slightly decreases compared to SFT. This may be due to the application of instruction format, not generally used for Llama 3, and the generalization of hyperparameters (*e.g.*, learning rates and epochs) for fine-tuning, which may not be optimal.

To summarize, across the nine experimental settings comparing our method with the baseline SFT and other [PAUSE] token-inserting methods, DIT demonstrated superior performance in six cases, achieving the highest scores. Additionally, in one case, DIT secured the second-best performance, further emphasizing the effectiveness and robustness of our approach. Although it does not surpass all other methods on every dataset, it remains highly competitive, underscoring the robustness of the [PAUSE] token strategy. In conclusion, DIT consistently demonstrates superior performance across various datasets, affirming that strategically inserting [PAUSE] tokens can significantly enhance model efficacy in reasoning tasks.

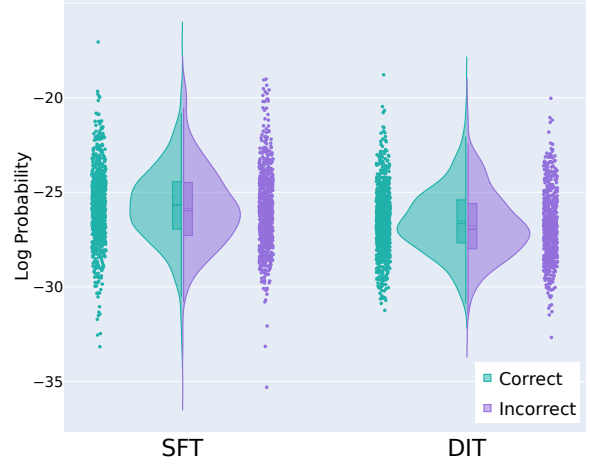


Figure 4: **Comparison of the lowest 20 log probabilities for tokens between SFT and DIT (ours) on the GSM8K dataset.** The violin plots illustrate that Dynamic Insertion of [PAUSE] Tokens (DIT) results in higher log probabilities for tokens initially ranked with low probabilities, indicating enhanced model confidence in predicting challenging tokens.

5.2 Role of [PAUSE] Token in Training

We investigate the impact of inserting [PAUSE] tokens into sequences during language model training. As shown in Figure 3, our results indicate that irrespective of the model and dataset, the insertion of [PAUSE] tokens leads to an increase in losses. The observed increase in losses, where $\mathcal{L}_{DIT}(f_\theta, \tilde{Y}_{1:N+M_{DIT}})$ exceeds $\mathcal{L}_{SFT}(f_\theta, Y_{1:N})$ suggests that the presence of the [PAUSE] token complicates the model’s ability to predict \tilde{y}_{t+1} based on the preceding tokens up to step t .

By strategically inserting a [PAUSE] token at a position where the log-likelihood of \tilde{y}_{t+1} is already among the lowest in the sequence, the log-likelihood further decreases, increasing in the loss term. This method effectively heightens the loss for tokens that are challenging to predict based on previous tokens, thereby encouraging the model trained more robustly. Previous studies have highlighted that dynamically adjusting the loss term for challenging instances can mitigate performance degradation (Jiang et al., 2020; Lin et al., 2017; Fernando and Tsokos, 2021; Wang et al., 2022). Consequently, DIT induces a more significant shift in the model’s parameters compared to standard fine-tuning, enhancing the model’s fit to the dataset. This improved alignment likely accounts for the observed performance improvements.

5.3 [PAUSE] token with token probability

In this section, we delve into the impact of [PAUSE] token insertion on the performance of downstream tasks, focusing on token log probability. By strategically placing [PAUSE] tokens where model confidence is low during training, we aim to bolster the model’s certainty in selecting subsequent tokens, thereby enhancing performance in downstream tasks. This expectation is substantiated by significant performance improvements across various domains, as demonstrated in Table 1 and §5.1. To understand the changes during inference, we contrast SFT and DIT methodologies using the GSM8K dataset, depicted in Figure 4. Here, the y-axis represents the log probability, with the x-axis contrasting SFT and DIT. Each violin graph showcases correct predictions in cyan on the left and incorrect predictions in violet on the right, with actual log probability values displayed on both sides. These values represent the averages of the lowest 20 log probabilities for tokens in each test dataset, facilitating observation of alterations in low log probability tokens.

As shown in Figure 4, DIT exhibits a shortened long tail compared to the SFT method, indicative of heightened log probability. Moreover, the data appears more clustered around the mean, signaling reduced variance. Although overall log probabilities for each data point may slightly decrease due to [PAUSE] token insertion, the probabilities of previously low-confidence tokens experience an upsurge. Essentially, DIT training, facilitated by [PAUSE] token insertion before low log-likelihood tokens, enables the model to prioritize learning challenging tokens. Consequently, while the overall average probability may decrease due to task-irrelevant [PAUSE] tokens, we can see an increase in the average probability of outliers, ultimately leading to performance enhancement.

5.4 Analysis for the Loss Calibration

Here, we examine the impact of [PAUSE] tokens on loss calibration, a key factor in ensuring stable learning, especially in models dealing with complex or imbalanced data (Mukhoti et al., 2020; Tao et al., 2023). In our approach, the [PAUSE] token insertion plays a critical role in this process. These tokens act as a temporal regularization, inducing the model to slow down and pay closer attention to challenging tokens that otherwise might be overlooked.

This is similar to the principles behind the focal loss (Lin et al., 2017), where the loss function is adjusted to emphasize challenging examples of those with low likelihood while down-weighting easier ones. Focal loss adds relatively more weight for misclassified instances by dynamically scaling cross-entropy (CE) loss depending on model confidence. Whereas the conventional CE objective is akin to $-\log p_t$, FL scales the objective as $-(1 - p_t)^\gamma \log p_t$. This focal loss objective has been adapted for reducing bias in tasks such as natural language inference (NLI) by dynamically reducing instance weighting based on undesirable properties during training (Mahabadi et al., 2020).

Within the context of DIT, the [PAUSE] token is previously unseen and will be associated with a high loss. Given that $\mathcal{L}_{CE}(\tilde{y}_{k+1}, f_\theta([\tilde{y}_{1:k}, [\text{PAUSE}]]) > \mathcal{L}_{CE}(\tilde{y}_{k+1}, f_\theta(\tilde{y}_{1:k}))$, inserting [PAUSE] tokens with DIT dynamically increases the loss for low-confidence tokens. This elevated loss signal prompts the model to focus more attention on these uncertain tokens, enhancing its ability to learn from challenging contexts. As a result, the models become more robust and show improved performance in downstream tasks, as demonstrated empirically in Section 5.2.

The effect of [PAUSE] tokens can be likened to the NoiseBoost (Wu et al., 2024) method, where perturbations are introduced to balance attention across different modalities. Just as NoiseBoost prevents over-reliance on specific inputs by distributing attention more evenly, [PAUSE] tokens help recalibrate the model’s focus during training. This recalibration stabilizes the learning process by preventing the model from becoming overly confident in easy predictions or neglecting challenging tokens, thereby enhancing generalization.

6 Ablations

In this section, we conduct an ablation study on the number of [PAUSE] tokens and the consecutive insertion lengths, which we consider important in our experimental setup. Using the GSM8K dataset, which shows the most significant performance improvement over the baseline setting SFT, we perform the ablation study on the DIT model based on the Phi-2 model.

The number of [PAUSE] token As detailed in Appendix C, we insert a total of five [PAUSE] tokens during training. A critical consideration is

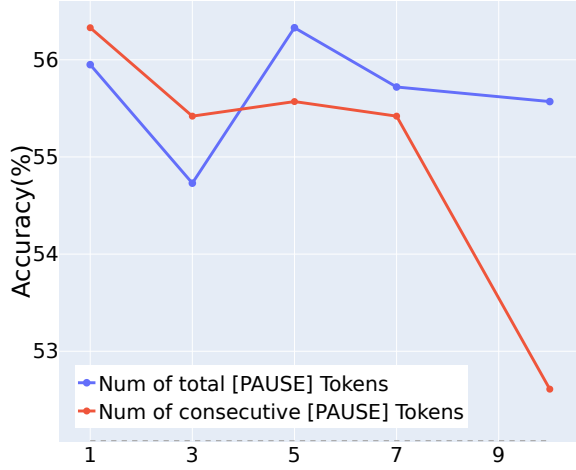


Figure 5: **Impact of varying the number of [PAUSE] tokens on model performance.** The graph shows that inserting five [PAUSE] tokens achieves optimal performance improvement for the Phi-2 model on the GSM8K dataset.

how many low log-likelihood tokens should have [PAUSE] tokens inserted before them to effectively modify their likelihood values. In addition to our initial setting of five tokens, we experimented with inserting 1, 3, 7, and 10 [PAUSE] tokens under the same conditions.

The results, depicted in the blue graph in Figure 5, indicate that our choice of five [PAUSE] tokens is optimal, yielding the highest performance improvement for the Phi-2 model base using the GSM8K dataset. The graph also shows no consistent pattern suggesting that a higher or lower number of [PAUSE] tokens is inherently better. Notably, across all configurations, Dynamic Insertion of [PAUSE] Tokens (DIT) consistently achieves superior scores compared to other methods.

The role of the [PAUSE] token is pivotal: strategically inserting [PAUSE] tokens in front of tokens with a likelihood below a certain threshold can aid in learning challenging tokens. However, as evidenced by the performance degradation of AAW and RAN in Table 1, inserting [PAUSE] tokens when token log-likelihood is above a certain threshold can potentially disrupt maximizing likelihood or introduce noise. Hence, identifying the optimal number of [PAUSE] tokens to insert is crucial.

Length of [PAUSE] token blocks In Goyal et al. (2024)’s work, a series of [PAUSE] tokens is created by inserting n consecutive [PAUSE] tokens, forming token blocks. In contrast, our approach involves inserting a single [PAUSE] token at each designated position. To assess the impact on per-

formance, we experiment with inserting multiple consecutive [PAUSE] tokens at the optimal five positions identified previously. Specifically, we insert 3, 5, 7, and 10 consecutive [PAUSE] tokens at each position. For instance, inserting three consecutive [PAUSE] tokens at each of the five positions results in a total of fifteen [PAUSE] tokens. As shown in Figure 5, the red graph indicates that using a single [PAUSE] token consecutively, as in our approach, is most effective. The performance decreases as the number of consecutive [PAUSE] tokens increases, proportionally increasing the total number of tokens.

This performance degradation is likely due to the increased delay in the calculation of the next token caused by inserting consecutive [PAUSE] tokens. This finding contrasts with Goyal et al. (2024)’s results, underscoring the significance of the consecutive number of [PAUSE] tokens inserted, which critically impacts performance. Our results suggest that strategically placed single [PAUSE] tokens can significantly improve performance, highlighting the importance of precise placement over the quantity of inserted tokens.

7 Related Works

Token Insertion Recent decoder-based language models (Radford et al., 2018; Microsoft, 2023; Abidin et al., 2024; AI@Meta, 2024) generate the upcoming token based on the probabilities of previous tokens. This characteristic has been extensively explored in various studies, which have proposed incorporating special tokens to modify the token log-likelihood values. For instance, some have introduced controllable codes in language models (Keskar et al., 2019), while others have explored the removal of tokens during processing in a vision transformer (Rao et al., 2021). In addition, the concept of memory transformers has emerged, where dummy tokens are prefixed to token vectors to enable additional trainable memory for storing representations (Burtsev et al., 2020; Bulatov et al., 2022). Despite these advancements, there remains an opportunity for further research on the strategic insertion and removal of tokens in transformer models.

Model Confidence Training of LLMs using the transformer’s decoder ensures that during inference, generated tokens are based on the highest log probability tokens or methods like beam search to find optimal probabilities within a range (Germann,

2003; Freitag and Al-Onaizan, 2017). Hence, when the model generates tokens, a low log probability value for a specific token indicates that the probability differences among token candidates are relatively small. This means that the model is more likely to generate tokens with higher uncertainty. Prior works indicate that this low probability considered as low confidence of the model can indicate a lack of knowledge in model (Dong et al., 2018; Jiang et al., 2023; Kadavath et al., 2022; Varshney et al., 2022). Our study, which aims to bolster model confidence on a token-by-token basis, builds upon these foundational insights.

8 Conclusions

Previous research (Goyal et al., 2024) has underscored the effectiveness of the dummy token in enhancing performance by creating additional computational pathways in attention layers and delaying output generation. Building on this, our study strategically employs [PAUSE] tokens to improve reasoning tasks. We introduced a novel approach where [PAUSE] tokens are dynamically inserted based on token log-likelihood, rather than following fixed rules or relying on randomness. This method targets low-confidence areas, enhancing model stability during training and inference. Our findings demonstrate that inserting [PAUSE] tokens at low log-likelihood positions significantly boosts the model’s ability to handle challenging tokens with higher losses and improve the probability robustness of outlier tokens during inference.

Limitations

Negative Impact of [PAUSE] token During our experiments, we have observed not only the positive outcomes of inserting [PAUSE] tokens but also some negative effects. While DIT generally demonstrates improvements, there are instances where certain datasets show worse performance compared to straightforward fine-tuning (SFT). Notably, in the Phi-2 GSM8k dataset, where DIT exhibits significant enhancement over SFT, the improvement stems largely from correcting errors introduced by SFT rather than the reverse. Nevertheless, there are cases in the test set where DIT introduces errors absent in the SFT model. These observations prompt further investigation into unintended consequences, including the potential for [PAUSE] tokens to introduce token noise. Addressing these nuances will require additional studies to refine the deploy-

ment and optimization of [PAUSE] tokens in future model architectures.

Interpretations Still Obscured Within a Black Box

In the original work by Goyal et al. (2024), they mention that it is unclear what changes their method would bring about in practice and offer their own interpretation that a “wider” computational pathway induced by the delay could have a positive effect, leading to performance improvements. Our research focuses on optimizing the insertion points of [PAUSE] tokens and automating this process rather than relying on heuristic methods. Therefore, we partially rely on the original interpretation to explain why this approach is effective. From the model’s perspective, it is imperative to achieve a more profound comprehension of the effects of the [PAUSE] token.

Ethics Statement

In our experiments throughout the paper, we used three models with publicly available weights. Using our method, there is a possibility of obtaining unintended experimental outputs. Additionally, we conduct our experiments using research data available in public repositories. Detailed information about this can be found in the Appendix. It is important to note that all training models and data, as well as our work using them, may produce incorrect outputs if used with improper intent. Also, We acknowledge the use of the LLM, GPT-4o in writing the paper, for grammar and contextual corrections for better expression. Ultimately, all authors rewrite the final version manuscript in their own words.

Acknowledgments

This work was partly supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)) and Institute for Information & communications Technology Promotion(IITP) grants funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI).

References

Marah Abidin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla,

- Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.
- AI@Meta. 2024. [Llama 3 model card](#).
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.
- Gregor Bachmann and Vaishnavh Nagarajan. 2024. The pitfalls of next-token prediction. *arXiv preprint arXiv:2403.06963*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. 2022. Recurrent memory transformer. *Advances in Neural Information Processing Systems*, 35:11079–11091.
- Mikhail S Burtsev, Yuri Kuratov, Anton Peganov, and Grigory V Sapunov. 2020. Memory transformer. *arXiv preprint arXiv:2006.11527*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Tri Dao. 2024. [Flashattention-2: Faster attention with better parallelism and work partitioning](#). In *The Twelfth International Conference on Learning Representations*.
- Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2023. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*.
- Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. 2018. Universal transformers. *arXiv preprint arXiv:1807.03819*.
- Li Dong, Chris Quirk, and Mirella Lapata. 2018. [Confidence modeling for neural semantic parsing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 743–753, Melbourne, Australia. Association for Computational Linguistics.
- Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. 2024. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36.
- Ahmad Faiz, Sotaro Kaneda, Ruhan Wang, Rita Chukwunyere Osi, Prateek Sharma, Fan Chen, and Lei Jiang. 2023. Llmcarbon: Modeling the end-to-end carbon footprint of large language models. In *The Twelfth International Conference on Learning Representations*.
- K Ruwani M Fernando and Chris P Tsokos. 2021. Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2940–2951.
- Markus Freitag and Yaser Al-Onaizan. 2017. [Beam search strategies for neural machine translation](#). In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics.
- Ulrich Germann. 2003. [Greedy decoding for statistical machine translation in almost linear time](#). In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 72–79.
- Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2024. [Think before you speak: Training language models with pause tokens](#). In *The Twelfth International Conference on Learning Representations*.
- David Herel and Tomas Mikolov. 2024. Thinking tokens for language modeling. *arXiv preprint arXiv:2405.08644*.
- Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.
- Shaojie Jiang, Thomas Wolf, Christof Monz, and Maarten de Rijke. 2020. Tldr: token loss dynamic reweighting for reducing repetitive utterance generation. *arXiv preprint arXiv:2003.11963*.
- Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. [Active retrieval augmented generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.
- Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pages 19274–19286. PMLR.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.
- Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. [Program induction by rationale generation: Learning to solve and explain algebraic word problems](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.
- Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. [End-to-end bias mitigation by modelling biases in corpora](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.
- Microsoft. 2023. [Phi-2: The surprising power of small language models](#). [Blog post].
- Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. 2020. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299.
- Khanh Nguyen and Brendan O’Connor. 2015. [Posterior calibration and exploratory analysis for natural language processing models](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1587–1598, Lisbon, Portugal. Association for Computational Linguistics.
- Alec Radford and Karthik Narasimhan. 2018. [Improving language understanding by generative pre-training](#).
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.
- Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949.
- Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadeppally. 2023. From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–9. IEEE.
- Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2022. Distilling reasoning capabilities into smaller language models. *arXiv preprint arXiv:2212.00193*.
- Sainbayar Sukhbaatar, Edouard Grave, Guillaume Lample, Herve Jegou, and Armand Joulin. 2019. Augmenting self-attention with persistent memory. *arXiv preprint arXiv:1907.01470*.
- Linwei Tao, Minjing Dong, and Chang Xu. 2023. Dual focal loss for calibration. In *International Conference on Machine Learning*, pages 33833–33849. PMLR.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Neeraj Varshney, Man Luo, and Chitta Baral. 2022. Can open-domain qa reader utilize external knowledge efficiently like humans? *arXiv preprint arXiv:2211.12707*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Rose E Wang, Esin Durmus, Noah Goodman, and Tatsunori Hashimoto. 2022. Language modeling via stochastic processes. *arXiv preprint arXiv:2203.11370*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Kai Wu, Boyuan Jiang, Zhengkai Jiang, Qingdong He, Donghao Luo, Shengzhi Wang, Qingwen Liu, and Chengjie Wang. 2024. Noiseboost: Alleviating hallucination with noise perturbation for multimodal large language models. *arXiv preprint arXiv:2405.20081*.
- Fuzhao Xue, Valerii Likhoshesterov, Anurag Arnab, Neil Houlsby, Mostafa Dehghani, and Yang You. 2023. Adaptive computation with elastic input sequence. In *International Conference on Machine Learning*, pages 38971–38988. PMLR.

Dataset	GSM8K	AQUA-RAT	MBPP
split	number of datasets		
train	5,610	70,104	282
validation	-	2,254	90
test	1,319	2,254	500

Table 2: These are the specific numbers of data involved in the entire experiments.

A Details of Datasets

We download datasets from Hugging Face repository²³⁴. Information regarding the splits and numbers of the data we used is provided in Table 2. For the AQUA-RAT, the original dataset contains over 98k instances, with relatively small validation and test sets of around 200 each. For more efficient experimentation, we sample this data to approximately 74k instances. And due to the limited number of validation and test samples for AQUA-RAT (254 each), we supplement these sets by randomly extracting 2000 samples from the training set. Additionally, to ensure accurate evaluation of the model, we modify the AQUA-RAT options by appending four hashtags (####) and the correct alphabet at the end of the data, similar to GSM8K, to facilitate learning.

B A Priori Experiments

In this section, we describe the preliminary experiments conducted prior to our main experiments. Our primary experiments focus on the GSM8K dataset. Initially, we have fine-tuned a model on GSM8K to generate predictions on the test set while also calculating the log probabilities of each token. We have observed that certain parts of the data such as the start and end positions of equations and the points where sentences change consistently exhibit low log probabilities. By identifying these positions, we specify the rule-based approach for inserting [PAUSE] tokens. Subsequently, we insert [PAUSE] tokens at these annotated positions and conduct fine-tuning. The results of these experiments are presented in Table 3.

Glossary of Terms for A Priori Experiments

- Math Front GSM8K (**MF-G**) : Position [PAUSE] token between equal sign(‘=’) in math equation and start sign of calculator annotations(‘«’) at GSM8K dataset.
- Math Behind GSM8K (**MB-G**) : Position [PAUSE] token behind of end sign of calculator annotations(‘»’) at GSM8K dataset.
- Sentence Behind (**SB**) : Position [PAUSE] token right after the sentence in reasoning steps ends.

Base Model	Method	GSM8K
Phi-2	SFT	52.08
	MF-G	<u>53.90</u>
	MB-G	52.53
	SB	53.82
	DIT(learnable)	47.15
	DIT(ours)	56.33

Table 3: Results of comparing the heuristic methods for finding locations to insert the [PAUSE] tokens and DIT. For each method, the best score is indicated in **bold**, and the second-best score is underlined.

²<https://huggingface.co/datasets/openai/gsm8k>

³https://huggingface.co/datasets/deepmind/aqua_rat

⁴<https://huggingface.co/datasets/google-research-datasets/mbpp>

At this point, we consider the [PAUSE] token as a learnable token included in the loss function and proceed with our experiments accordingly. All methods in Table 3, except for DIT (ours), are conducted using this approach. Additionally, we insert ten [PAUSE] tokens at each position, while inserting five [PAUSE] tokens in DIT (learnable) and DIT (ours). However, through additional experiments, we have discovered that excluding the [PAUSE] tokens from the loss calculation is optimal, thereby strongly assigning it the role of a dummy token results in our final experimental setup (DIT) exhibiting probability robustness and improves performance across the overall dataset and models.

C Experimental Settings

We conduct experiments on the GSM8K, AQUA-RAT, and MBPP tasks. For each type of [PAUSE] setting, we insert five [PAUSE] tokens into the sequence ($M_{DIT} = 5$), with the exception of the original fine-tuning baseline (SFT). The losses associated with predicting [PAUSE] tokens are excluded in all methods. We use learning rates of $1e-5$ for GSM8K, and $3e-5$ for both the MBPP and AQUA-RAT datasets. Although performance varies with different learning rates, the overall trends remain consistent. Each task is trained for three epochs, using the model with the lowest evaluation loss for inference. Fine-tuning is performed using two RTX 6000 GPUs for GSM8K and MBPP, and two A100 GPUs for AQUA-RAT, with a training and test batch size of 4.

D Training Complexity Notation

Let F denote the computational cost of a single forward pass, and B denote the computational cost of a single backward pass (including parameter updates).

Standard SFT In standard supervised fine-tuning (SFT), each iteration performs one forward pass and one backward pass, yielding an approximate cost of

$$\text{Cost}_{\text{SFT}} \approx F + B.$$

DIT Approach In our DIT approach, we first perform an extra forward pass to identify low log-likelihood positions for inserting [PAUSE] tokens, and then perform a second forward pass together with a backward pass. Concretely, as shown in Figure 2,

1. Forward pass(blue line): identify tokens/positions with low log-likelihood. In this step, a single forward pass is performed.
2. Training after token insertion(green line): train on the augmented input containing [PAUSE] tokens. In this step, a single forward pass followed by a backward pass is performed.

Thus, each DIT iteration incurs a cost of

$$\text{Cost}_{\text{DIT}} \approx 2F + B.$$

The primary cost difference between the standard SFT approach and the proposed DIT method lies in the additional forward pass introduced by DIT. This extra computation serves as a minor cost paid for improved performance. Notably, forward passes are generally less computationally expensive than backward passes and offer greater room for optimization (Dao, 2024), suggesting that the performance gain can outweigh the added cost.

E Examples of Methods

We provide examples of the dataset with the [PAUSE] token used during training and inference to illustrate the application of each method. From Table 4 to Table 8, we extract actual examples from each dataset to visualize how and where the [PAUSE] tokens are inserted.

E.1 DIT

Dataset	Example
GSM8K	<p>Instruct: Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers’ market daily for \$2 per fresh duck egg. How much in dollars does she make every day at the farmers’ market?</p> <p>Output:Jan[PAUSE]et sells[PAUSE] 16 - 3 - 4 = «16-3-4=9»9 duck eggs a day.[PAUSE]</p> <p>She makes $9 * 2 = \\$\langle 9*2=18 \rangle 18$ every day at the farmer’s market.[PAUSE]</p> <p>[PAUSE]####[PAUSE] 18</p>
AQUA-RAT	<p>Instruct:Question : If $A = x^4 + y^4 = 100$, then the greatest possible value of x is between</p> <p>Options : A)0 and 3 B)3 and 6 C)6 and 9 D)9 and 12 E)12 and 15</p> <p>Output:My attempt:</p> <p>If[PAUSE] $A = x^4 + y^4 = 100$, then the greatest possible value of x would be when y is minimum.</p> <p>Let[PAUSE] y^4 be 0. Now $x^4 = 100$. x should be definitely greater than 3 but less than 4. The only option that fits this range is B</p> <p>[PAUSE]Hence answer is –[PAUSE]</p> <p>b) 3 and 6.[PAUSE] #### B</p>
MBPP	<p>Instruct:Write a python function to find minimum sum of factors of a given number.</p> <p>Output:</p> <pre>def find_Min_Sum(num): [PAUSE] sum = 0 i = 2 while(i * i <= num): [PAUSE] while(num % i == 0): sum += i [PAUSE] num /= i [PAUSE] i += 1 sum += num [PAUSE] return sum</pre>

Table 4: These are examples of [PAUSE] token insertion in each dataset using the DIT method.

E.2 SFT

Dataset	Example
GSM8K	<p>Instruct: Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers’ market daily for \$2 per fresh duck egg. How much in dollars does she make every day at the farmers’ market?</p> <p>Output: Janet sells $16 - 3 - 4 = \ll 16 - 3 - 4 = 9 \gg 9$ duck eggs a day. She makes $9 * 2 = \\$\ll 9 * 2 = 18 \gg 18$ every day at the farmer’s market. #### 18</p>
AQUA-RAT	<p>Instruct: Question : If $A = x^4 + y^4 = 100$, then the greatest possible value of x is between</p> <p>Options : A)0 and 3 B)3 and 6 C)6 and 9 D)9 and 12 E)12 and 15</p> <p>Output: My attempt: If $A = x^4 + y^4 = 100$, then the greatest possible value of x would be when y is minimum. Let y^4 be 0. Now $x^4 = 100$. x should be definitely greater than 3 but less than 4. The only option that fits this range is B Hence answer is – b) 3 and 6. #### B</p>
MBPP	<p>Instruct: Write a python function to find minimum sum of factors of a given number.</p> <p>Output:</p> <pre>def find_Min_Sum(num): sum = 0 i = 2 while(i * i <= num): while(num % i == 0): sum += i num /= i i += 1 sum += num return sum</pre>

Table 5: These are examples of [PAUSE] token insertion in each dataset using the SFT method.

E.3 RAN

Dataset	Example
GSM8K	<p>Instruct: Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers’ market daily for \$2 per fresh duck egg. How much in dollars does she make every day at the farmers’ market?</p> <p>Output: Janet sells 16 - 3 - 4 [PAUSE] = [PAUSE] «16-3-4=9» [PAUSE] > 9 duck eggs a [PAUSE] day.</p> <p>She makes $9 * 2 = \\$\ll 9 * 2 = 18 \gg 18$ every day at the farmer’s market.</p> <p>#### 18</p>
AQUA-RAT	<p>Instruct: Question : If $A = x^4 + y^4 = 100$, then the greatest possible value of x is between</p> <p>Options : A) 0 and 3 B) 3 and 6 C) 6 and 9 D) 9 and 12 E) 12 and 15</p> <p>Output: M[PAUSE]y attempt:</p> <p>If $A = x^4 + y^4 = 100$, then the greatest possible value of [PAUSE]x would be when y is minimum.</p> <p>Let y^4 be 0. Now [PAUSE]$x^4 = 100$. x sho[PAUSE]uld be definitely greater than 3 but less than 4. The only optio[PAUSE]n that fits this range is B</p> <p>Hence answer is –</p> <p>b) 3 and 6. #### B</p>
MBPP	<p>Instruct: Write a python function to find minimum sum of factors of a given number.</p> <p>Output:</p> <pre>def find_Min_Sum(num)[PAUSE]: [PAUSE] sum = 0 i = 2 while(i * [PAUSE]i [PAUSE]<= num): while(num % i == 0): sum += i num /= i i += 1 sum += num[PAUSE] return sum</pre>

Table 6: These are examples of [PAUSE] token insertion in each dataset using the RAN method.

E.4 APPD

Dataset	Example
GSM8K	<p>Instruct: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for \$2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?</p> <p>Output: [PAUSE][PAUSE][PAUSE][PAUSE][PAUSE] Janet sells 16 - 3 - 4 = «16-3-4=9»9 duck eggs a day. She makes $9 * 2 = \\$\langle 9*2=18 \rangle 18$ every day at the farmer's market. #### 18</p>
AQUA-RAT	<p>Instruct:Question : If $A = x^4 + y^4 = 100$, then the greatest possible value of x is between</p> <p>Options : A)0 and 3 B)3 and 6 C)6 and 9 D)9 and 12 E)12 and 15</p> <p>Output: [PAUSE][PAUSE][PAUSE][PAUSE][PAUSE] My attempt: If $A = x^4 + y^4 = 100$, then the greatest possible value of x would be when y is minimum. Let y^4 be 0. Now $x^4 = 100$. x should be definitely greater than 3 but less than 4. The only option that fits this range is B Hence answer is – b) 3 and 6. #### B</p>
MBPP	<p>Instruct:Write a python function to find minimum sum of factors of a given number.</p> <p>Output: [PAUSE][PAUSE][PAUSE][PAUSE][PAUSE]</p> <pre>def find_Min_Sum(num): sum = 0 i = 2 while(i * i <= num): while(num % i == 0): sum += i num /= i i += 1 sum += num return sum</pre>

Table 7: These are examples of [PAUSE] token insertion in each dataset using the APPD method.

E.5 AAW

Dataset	Example
GSM8K	<p>Instruct: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for \$2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?</p> <p>Output: Janet [PAUSE] sells [PAUSE] 16 [PAUSE] - [PAUSE] 3 [PAUSE] - [PAUSE] 4 [PAUSE] = [PAUSE] «16-3-4=9» 9 [PAUSE] duck [PAUSE] eggs [PAUSE] a [PAUSE] day. [PAUSE]</p> <p>She [PAUSE] makes [PAUSE] 9 [PAUSE] * [PAUSE] 2 [PAUSE] = [PAUSE] \$«9*2=18» 18 [PAUSE] every [PAUSE] day [PAUSE] at [PAUSE] the [PAUSE] farmer's [PAUSE] market. [PAUSE]</p> <p>#### [PAUSE] 18</p>
AQUA-RAT	<p>Instruct: Question : If $A = x^4 + y^4 = 100$, then the greatest possible value of x is between</p> <p>Options : A)0 and 3 B)3 and 6 C)6 and 9 D)9 and 12 E)12 and 15</p> <p>Output: My [PAUSE] attempt: [PAUSE]</p> <p>If [PAUSE] $A = x^4 + y^4 = 100$, [PAUSE] then [PAUSE] the [PAUSE] greatest [PAUSE] possible [PAUSE] value [PAUSE] of [PAUSE] x [PAUSE] would [PAUSE] be [PAUSE] when [PAUSE] y [PAUSE] is [PAUSE] minimum. [PAUSE]</p> <p>Let [PAUSE] y^4 [PAUSE] be [PAUSE] 0. [PAUSE] Now [PAUSE] x^4 [PAUSE] = [PAUSE] 100. [PAUSE] x [PAUSE] should [PAUSE] be [PAUSE] definitely [PAUSE] greater [PAUSE] than [PAUSE] 3 [PAUSE] but [PAUSE] less [PAUSE] than [PAUSE] 4. [PAUSE] The [PAUSE] only [PAUSE] option [PAUSE] that [PAUSE] fits [PAUSE] this [PAUSE] range [PAUSE] is [PAUSE] B [PAUSE]</p> <p>Hence [PAUSE] answer [PAUSE] is [PAUSE] - [PAUSE]</p> <p>b) [PAUSE] 3 [PAUSE] and [PAUSE] 6. [PAUSE] #### [PAUSE] B</p>
MBPP	<p>Instruct: Write a python function to find minimum sum of factors of a given number.</p> <p>Output:</p> <pre>def [PAUSE] find_Min_Sum(num): [PAUSE] sum [PAUSE] = [PAUSE] 0 [PAUSE] i [PAUSE] = [PAUSE] 2 [PAUSE] while (i [PAUSE] * [PAUSE] i [PAUSE] <= [PAUSE] num): [PAUSE] while (num [PAUSE] % [PAUSE] i [PAUSE] == [PAUSE] 0): [PAUSE] sum [PAUSE] += [PAUSE] i [PAUSE] num [PAUSE] /= [PAUSE] i [PAUSE] i [PAUSE] += [PAUSE] 1 [PAUSE] sum [PAUSE] += [PAUSE] num [PAUSE] return [PAUSE] sum</pre>

Table 8: These are examples of [PAUSE] token insertion in each dataset using the AAW method.