

# Learning to Reason from Feedback at Test-Time

Yanyang Li, Michael R. Lyu, Liwei Wang\*

The Chinese University of Hong Kong

{yyli21, lyu, lwwang}@cse.cuhk.edu.hk

## Abstract

Solving complex tasks in a single attempt is challenging for large language models (LLMs). Iterative interaction with the environment and feedback is often required to achieve success, making effective feedback utilization a critical topic. Existing approaches either struggle with length generalization or rely on naive retries without leveraging prior information. In this paper, we introduce FTTT, a novel paradigm that formulates feedback utilization as an optimization problem at test time. Additionally, we propose a learnable test-time optimizer, OPTUNE, to effectively exploit feedback. Experiments on two LLMs across four reasoning datasets demonstrate that FTTT and OPTUNE achieve superior scalability and performance<sup>1</sup>.

## 1 Introduction

Leveraging external feedback from interactions with the environment during test time has emerged as a promising approach for large language models (LLMs). This includes applications such as LLM-based agents (Yao et al., 2023; Shinn et al., 2023) and, more recently, test-time scaling (Wu et al., 2024; Snell et al., 2024; Liu et al., 2025). Such methods further enhance the potential of LLMs to solve challenging tasks, e.g., Olympiad-level math problems (Guan et al., 2025) and competitive programming (OpenAI et al., 2025).

Significant progress in this area typically falls into two categories (Snell et al., 2024), as illustrated in Figure 1: sequential revision and parallel sampling. Sequential revision methods (Shinn et al., 2023; Madaan et al., 2023) incorporate previous attempts into the LLM’s context, while parallel sampling methods (Xie et al., 2023; Brown et al., 2024) generate new attempts independently of prior failures. However, both approaches have notable limitations. Sequential revision is computationally

expensive due to long context lengths and faces challenges (Muennighoff et al., 2025), such as position bias (Liu et al., 2024) and attention noise (Ye et al., 2024b). In contrast, parallel sampling, while efficient, fails to learn from previous errors (Brown et al., 2024). Unlike these paradigms, human reasoning follows a different pattern: humans store recent experiences in “fast weights” (Schmidhuber, 1992; Ba et al., 2016; Schlag et al., 2021), enabling them to neither revisit past errors explicitly nor start each attempt without any prior knowledge. Recent research suggests that the weights of neural networks could serve as a natural memory mechanism during test time (Wang et al., 2024).

Building on these observations, we propose a novel paradigm that leverages Test-Time Training (TTT) (Sun et al., 2020, 2023) to store past experiences in model weights rather than in the context. This approach bridges the gap between sequential revision and parallel sampling by indirectly incorporating knowledge into the LLM without disrupting in-context reasoning. Specifically, we introduce Feedback-based Test-Time Training (FTTT), which employs a carefully designed TTT task based on feedback enriched with self-reflection. We demonstrate that FTTT improves test-time computation scalability on two mathematical reasoning and two code generation datasets, using Llama-3.1-8B-Instruct (Dubey et al., 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023).

Inspired by advancements in learning to optimize (Chen et al., 2022), we explore training learnable test-time optimizers to yield Pareto-optimal cost-performance tradeoffs. Our proposed learnable optimizer, OPTUNE, is a lightweight neural network that predicts weight updates from the gradients of the previous attempt. Unlike traditional parameter-efficient fine-tuning (PEFT) methods, OPTUNE works on the gradient rather than the activation space. Experiments on three reasoning datasets and two different LLMs demonstrate

\*Corresponding author.

<sup>1</sup><https://github.com/LaVi-Lab/FTTT>

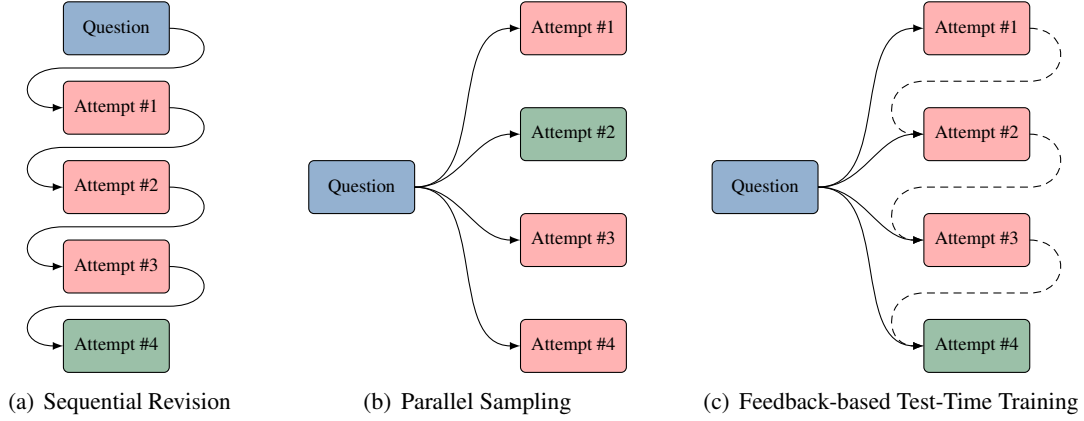


Figure 1: Comparison between sequential revision, parallel sampling, and feedback-based test-time training. ■ is the failed attempt and ■ is the successful attempt.  $\rightarrow$  indicates the LLM generation with the input on the left of the arrow and the output on the right.  $--\rightarrow$  denotes the LLM training, where the left of the arrow is the training data.

the effectiveness of OPTUNE, outperforming five widely used PEFT baselines.

## 2 Feedback-based Test-Time Training

### 2.1 The Test-Time Training Task

The problem of exploiting test-time feedback is as (Shinn et al., 2023): given a question  $Q$ , a model  $M$  attempts to solve  $Q$  within a budget of  $N$  attempts. A verifier  $V$  evaluates each attempt, such as the  $n$ -th attempt  $A_n$ , and provides feedback  $V(A_n)$ . This work focuses on binary verifiers, which determine whether  $A_n$  is correct. These verifiers are well-established, rule-based systems that are both cost-effective and efficient to evaluate.

When the model generates attempts sequentially, our goal is to enable  $M$  to learn from previous attempts to improve subsequent ones. To achieve this, we frame learning from previous attempts as a training problem: at each step  $n$ , we optimize  $M$  using  $Q$ ,  $A_n$ , and  $V(A_n)$ , aiming for  $M$  to generate a better  $A_{n+1}$ . This way internalizes the past attempts into weights for efficient inference of  $A_{n+1}$ . As a result, the sequence of attempts can be viewed as an  $N$ -step optimization process.

A key challenge is designing an effective supervised task using  $Q$ ,  $A_n$ , and  $V(A_n)$  to improve the model’s ability to solve  $Q$ . We build on the intuition that *a model capable of judging the correctness of a solution should also be able to solve the question itself*. Concretely, given  $Q$  and  $A_n$ , we train  $M$  to predict verbal feedback  $F$  that aligns with  $V(A_n)$ . This leads to our FTTT loss:

$$\mathcal{L}_{\text{FTTT}}(Q, A_n) = -\frac{1}{l_0} \log M_{n-1}(F | Q, A_n) \quad (1)$$

where  $l_0$  is the length of  $F$  and  $M_0$  denotes the raw LLM. In this work,  $F$  is set to “Your answer is incorrect.” when  $V(A_n)$  implies an incorrect  $A_n$ .

### 2.2 Self-Reflected Feedback

Since we are working with a binary verifier, the learning signal is limited at each interaction. Previous research suggests that LLMs can self-correct errors when provided with external signals (Huang et al., 2024). Inspired by this, we aim to enhance the learning signal by leveraging the model to generate silver-standard training labels via self-reflection.

We first sample a reflection  $R_n$  from the model given  $Q$ ,  $A_n$ ,  $F$  and the instruction  $P$  (see Appendix B for details of  $P$ ):

$$R_n \sim M_0(R | Q, A_n, F, P) \quad (2)$$

In practice, we use  $M_0$  to generate  $R_n$  to mitigate the risk of degraded self-reflection ability after training. The auxiliary loss is then defined as:

$$\mathcal{L}_{\text{aux}}(Q, A_n, R_n) = -\frac{1}{l_n} \log M_{n-1}(R_n | Q, A_n, F) \quad (3)$$

where  $l_n$  is the length of  $R_n$ . Eq. 3 can be interpreted as a sequence-level distillation loss (Kim and Rush, 2016), where knowledge from the raw model  $M_0$  is distilled into the trained model  $M_{n-1}$  to prevent overfitting. Finally, the overall loss is as:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{FTTT}} + \mathcal{L}_{\text{aux}} \quad (4)$$

Below is a training example with self-reflection, where underlined sentences are the training target:

**Algorithm 1** FEEDBACK-BASED TTT

**Require:** The question  $Q$ , the model  $M_0$ , the verifier  $V$ , the budget  $N$ , the verbal feedback  $F$ , the instruction  $P$

```

1:  $n \leftarrow 1$ 
2: while  $n \leq N$  do
3:    $A_n \sim M_{n-1}(A | Q)$ 
4:   if  $V(A_n)$  is passed then
5:     return  $A_n$ 
6:   else
7:     Compute the loss  $\mathcal{L}$  using Eq. 1
8:     if enable self-reflection then
9:        $R_n \sim M_0(R | Q, A_n, F, P)$ 
10:      Compute the loss  $\mathcal{L}_{\text{aux}}$  using Eq. 3
11:       $\mathcal{L} \leftarrow \mathcal{L} + \mathcal{L}_{\text{aux}}$ 
12:    end if
13:    Update  $M_{n-1}$  using  $\mathcal{L}$  to get  $M_n$ 
14:  end if
15:   $n \leftarrow n + 1$ 
16: end while
17: return  $A_N$ 

```

**Training Example with Self-Reflection**

**User:** Solve the following math problem . . .  
**Assistant:** . . . the final answer is: . . .  
**User:** Your answer is incorrect. Here is the summary of the mistakes in the previous solution . . .

The underlined sentence corresponds to  $F$  in Eq. 1 and the wave-underlined sentence represents  $R_n$  for Eq. 3. Algorithm 1 summarizes our FTTT.

**Discussion.** As shown in Table 1, FTTT combines the strengths of sequential revision and parallel sampling. Like sequential revision, it leverages memory (i.e., weights) to store past failed attempts, while avoiding the length generalization issues as in parallel sampling. Figure 1 highlights the advantages of FTTT from a probabilistic graphical model perspective, where both sequential revision and parallel sampling are special cases of FTTT with additional independence assumptions. In particular, sequential revision assumes that attempts form a Markov process, whereas parallel sampling treats each attempt as independent.

In terms of efficiency, FTTT is highly optimized. Since we only update  $M$  once at each attempt in practice, FTTT requires only one additional backward propagation computed in parallel for all tokens in one attempt, whose cost is negligible. The dominant overhead of FTTT is generating  $R_n$ .

Method	Self Reflection	Memory	Length Generalization
Revision (Snell et al., 2024)	✗	✓	✗
Self-Refine (Madaan et al., 2023)	✓	✓	✗
Best-of-N (Brown et al., 2024)	✗	✗	✓
Beam Search (Ow and Morton, 1988)	✗	✗	✓
Guided Beam Search (Xie et al., 2023)	✓	✗	✓
<b>FTTT (ours)</b>	✓	✓	✓

Table 1: Comparing the advantages and drawbacks of FTTT and related works.

However, FTTT is still much faster than sequential revision (see Section 4.2), as  $R_n$  is typically short.

Moreover, FTTT closely resembles vanilla TTT (Sun et al., 2020), but with additional inputs beyond  $Q$ , such as  $A_n$ ,  $V(A_n)$ , and  $R_n$ , as defined in the problem. Consequently, it inherits the same convergence guarantees as TTT.

### 3 A Learnable Test-Time Optimizer

#### 3.1 The Learning to Optimize Problem

Although FTTT achieves success (see Section 4.2), it simply applies the gradients of the feedback received so far to update the weights. This raises the question: can we design a better test-time optimizer that more effectively exploits feedback?

Motivated by learning to optimize (Chen et al., 2022), we adopt a neural network as the test-time optimizer. Concretely, this learnable test-time optimizer is formulated as  $f_\theta(Q, \{A_i, V(A_i)\}_{i=1}^n)$ , which predicts updates for all LLM weights based on the previous  $n$  attempts, and  $\theta$  is the optimizer parameter. However, this direct formulation leads to prohibitively large networks due to high-dimensional input and output spaces. For a maximum number of  $m$  tokens per attempt and an  $l$ -layer LLM, the input space grows to  $n \times m \times l$ , even when updating only a scalar weight at each layer (we exclude the token count of  $Q$ , as it is significantly smaller than  $m$ ). Since updates for all weight matrices across all layers are predicted jointly, the dimensionality of the input and output spaces becomes unmanageable. We therefore simplify  $f_\theta$  by introducing the following assumptions:

**(A1): Markov Property:** The latest attempt captures all relevant information from previous attempts.

**(A2): Independent Update:** The optimizer predicts updates for each weight matrix independently, similar to conventional optimizers (Kingma and Ba, 2015).

A1 eliminates the dependency on  $n$  and A2 enables updates to be predicted independently for

each weight, significantly reducing the size of the output space. The learnable test-time optimizer now becomes  $\tilde{\nabla}_{W_i} = f_{\theta_{W_i}}(Q, A_n, V(A_n))$ , where it predicts the update for the weight  $W_i$  in the  $i$ -th layer based on  $Q$  and the latest attempt  $A_n$ .

To train all  $f$ , we define the following loss:

$$\mathcal{L}_{\text{meta}} = -\log M(\hat{A} | Q, \{W + \tilde{\nabla}_W | \forall W \in \mathcal{W}\}) \quad (5)$$

where  $\hat{A}$  is the ground-truth for  $Q$  and  $\mathcal{W}$  is the set of LLM weights. Here we consider only a one-step optimization problem, as a multi-step version involves the non-differentiable generation of  $A$ . Eq. 5 encourages  $f$  to predict updates that increase the likelihood of generating the correct answer after applying these updates.

### 3.2 A Parameter-Efficient Architecture in The Gradient Space

Given the limited learning signal at test time, we design the learnable optimizer to be parameter-efficient to alleviate overfitting. However, the input and output spaces of  $f_{\theta_{W_i}}(Q, A_n, V(A_n))$  are large due to their lengths, making even a simple linear projection parameter-intensive. Additionally,  $V(A_n)$  may be heterogeneous to  $Q$  and  $A_n$ , e.g., a scalar, posing challenges for modeling.

Inspired by the success of FTTT in Section 2 and recent works (Mitchell et al., 2022; Wang et al., 2024), we propose a parameter-efficient architecture in the gradient space as the learnable optimizer.

**Gradient-based Input Compression.** Instead of directly inputting  $Q$ ,  $A_n$ , and  $V(A_n)$ , we first project them into the gradient space, since recent work suggests that long context can be effectively compressed by gradients (Wang et al., 2024). This way reduces the token count  $m$  in  $A_n$  to a constant and unifies the spaces of  $Q$ ,  $A_n$ , and  $V(A_n)$  to ease the modeling. Besides, the gradient encodes the knowledge about  $W_i$ , enabling efficient modeling of  $f_{\theta_{W_i}}$ , which no longer needs to learn such knowledge with massive resources as in the black-box optimization (Chen et al., 2017). To compress  $Q$  and  $A_n$ , we use the next token prediction loss, while for  $V(A_n)$ , we include  $\mathcal{L}_{\text{FTTT}}$  in Eq. 1. The final loss for compressing the optimizer input is:

$$\mathcal{L}_{\text{compress}} = -\frac{1}{m} \log M(A_n | Q) + \mathcal{L}_{\text{FTTT}} \quad (6)$$

The input of  $f_{\theta_{W_i}}$  to predict the update of  $W_i$  now is the gradient  $\nabla_{W_i}$  of  $\mathcal{L}_{\text{compress}}$  w.r.t.  $W_i$ . Consequently,  $f_{\theta_{W_i}}$  receives a fixed-size tensor as input rather than a variable-length sequence.

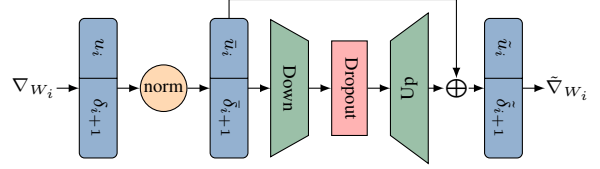


Figure 2: The model architecture of OPTUNE.

**Gradient Decomposition.** Although  $f_{\theta_{W_i}}$  operates on a smaller space after compression, the dimensionality of the gradient space remains large for direct processing. We utilize the observation that  $\nabla_{W_i} \in \mathbb{R}^{d \times d}$  (assuming  $W_i \in \mathbb{R}^{d \times d}$ ) can be decomposed into two vectors to further reduce the dimensionalities (Mitchell et al., 2022): the input to a linear projection with weight  $W_i$ ,  $u_i \in \mathbb{R}^{d \times 1}$ , and the gradient of  $\mathcal{L}_{\text{compress}}$  w.r.t. the output of the same linear projection,  $\delta_{i+1} \in \mathbb{R}^{d \times 1}$ . In this framework,  $f_{\theta_{W_i}}$  takes the decomposed  $u_i$  and  $\delta_{i+1}$  as its input and predicts  $\tilde{u}_i$  and  $\tilde{\delta}_{i+1}$ . The update is then reconstructed as  $\tilde{\nabla}_{W_i} = \tilde{\delta}_{i+1} \tilde{u}_i^T$ . This approach reduces the dimension from  $d^2$  to  $2d$ .

**Model Architecture.** The architecture of  $f_{\theta_{W_i}}(u_i, \delta_{i+1})$ , named OPTUNE, is shown in Figure 2 and defined as follows:

$$[\bar{u}_i, \bar{\delta}_{i+1}] = \text{Norm}([u_i, \delta_{i+1}]) \quad (7)$$

$$h_i = \theta_2 \text{Dropout}(\theta_1 [\bar{u}_i, \bar{\delta}_{i+1}]) \quad (8)$$

$$[\tilde{u}_i, \tilde{\delta}_{i+1}] = h_i + [\bar{u}_i, \bar{\delta}_{i+1}] \quad (9)$$

where  $\theta_1 \in \mathbb{R}^{r \times 2d}$  and  $\theta_2 \in \mathbb{R}^{2d \times r}$  are the optimizer parameters with  $r \ll d$ .  $[\cdot]$  denotes the vector concatenation. Norm normalizes  $u_i$  and  $\delta_{i+1}$  to have zero mean and unit variance separately. Dropout is the dropout regularization (Srivastava et al., 2014). In practice,  $\theta_1$  and  $\theta_2$  are shared across all weights with the same shape. We design OPTUNE this way as rank reduction and dropout help mitigate the overfitting risk, which is severe in our scenario due to the sparse, binary feedback. OPTUNE is similar to the Bottleneck Adapter (Houlsby et al., 2019), with the key difference that its input is gradients and its output is the weight update. As such, OPTUNE can also be regarded as a specialized PEFT technique tailored to reasoning.

## 4 Experiments

### 4.1 Setup

**Datasets.** We evaluate both baselines and our method on math and coding reasoning tasks: (a)

Method	MATH	GSM8K	MBPP	HumanEval	Avg.
Llama-3.1-8B-Instruct					
Revision (Snell et al., 2024)	0.2960	0.4570	0.2991	0.3889	0.3603
Beam Search (Ow and Morton, 1988)	0.2560	0.4842	0.1869	0.2407	0.2920
Self-Consistency (Wang et al., 2023)	0.3000 <sub>0.0057</sub>	0.4525 <sub>0.0111</sub>	0.1371 <sub>0.0180</sub>	0.1235 <sub>0.0531</sub>	0.2533
Self-Refine (Madaan et al., 2023)	0.4693 <sub>0.0207</sub>	0.7828 <sub>0.0064</sub>	0.2305 <sub>0.0088</sub>	0.2963 <sub>0.0800</sub>	0.4447
Best-of-N (Brown et al., 2024)	0.6427 <sub>0.0154</sub>	0.8069 <sub>0.0043</sub>	0.5452 <sub>0.0154</sub>	0.6728 <sub>0.0087</sub>	0.6669
FTTT	0.6707 <sub>0.0222</sub>	<b>0.8100</b> <sub>0.0037</sub>	<b>0.5607</b> <sub>0.0212</sub>	<b>0.6852</b> <sub>0.0302</sub>	0.6817
+ w/o Self-Reflected Feedback	<b>0.6720</b> <sub>0.0113</sub>	0.8054 <sub>0.0133</sub>	0.5405 <sub>0.0117</sub>	0.6790 <sub>0.0087</sub>	0.6742
Mistral-7B-Instruct-v0.3					
Revision (Snell et al., 2024)	0.0497	0.1686	0.1351	0.1000	0.1134
Beam Search (Ow and Morton, 1988)	0.1783	0.4537	0.1318	0.1600	0.2310
Self-Consistency (Wang et al., 2023)	0.1618 <sub>0.0011</sub>	0.3549 <sub>0.0052</sub>	0.1243 <sub>0.0135</sub>	0.1033 <sub>0.0236</sub>	0.1861
Self-Refine (Madaan et al., 2023)	0.1287 <sub>0.0139</sub>	0.3312 <sub>0.0078</sub>	0.2347 <sub>0.0091</sub>	0.3533 <sub>0.0309</sub>	0.2620
Best-of-N (Brown et al., 2024)	0.4688 <sub>0.0138</sub>	0.7807 <sub>0.0055</sub>	<b>0.4962</b> <sub>0.0171</sub>	0.6500 <sub>0.0141</sub>	0.5989
FTTT	0.4733 <sub>0.0087</sub>	0.7820 <sub>0.0045</sub>	<b>0.4962</b> <sub>0.0015</sub>	0.6633 <sub>0.0125</sub>	0.6037
+ w/o Self-Reflected Feedback	<b>0.4876</b> <sub>0.0133</sub>	<b>0.7858</b> <sub>0.0021</sub>	0.4941 <sub>0.0040</sub>	<b>0.6833</b> <sub>0.0205</sub>	0.6127

Table 2: Experimental results on four datasets with a budget of 32. For stochastic algorithms, we report the mean of three runs with different random seeds and standard deviation in the subscript. **Bold** entries are the best results, and underlined entries are the second-best results.

Mathematical reasoning: MATH (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021), using the test split from Lightman et al. (2024) for MATH. (b) Code generation: MBPP (Austin et al., 2021) and HumanEval (Chen et al., 2021). For all datasets, we report results on subsets where models fail with greedy decoding. We use Exact Match as the evaluation metric as well as the verifier for math tasks and Pass@1 for code.

**Models.** We conduct experiments with Llama-3.1-8B-Instruct (Dubey et al., 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). We evaluate both models with zero-shot prompting and follow the official instructions when evaluating the model on each dataset<sup>2</sup>.

**Baselines.** We compare FTTT with the following test-time scaling methods:

- **Beam Search** (Ow and Morton, 1988) is a popular search algorithm that approximates the most confident model prediction.
- **Self-Consistency** (Wang et al., 2023) samples multiple predictions and selects the most frequent answer.
- **Best-of-N** (Brown et al., 2024) samples  $N$  predictions independently and picks the best

one based on external feedback.

- **Revision** (Snell et al., 2024) iteratively refines answers by conditioning the model on previous attempts. We pick the best answer according to feedback.
- **Self-Refine** (Madaan et al., 2023) alternates between self-critique and refinement. We select the best solution based on feedback.

For sampling-based methods, we use nucleus sampling (Holtzman et al., 2020) with a temperature of 0.6 and  $p = 0.95$ , following Brown et al. (2024). All methods are allocated a budget of 32. For FTTT, we fine-tune the model with LoRA (Hu et al., 2022), using a rank of 4 and a dropout ratio of 0.05. We use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of  $1e-5$ , except for Mistral-7B-Instruct-v0.3 in coding tasks, where we use  $2e-5$ .

As OPTUNE is a specialized PEFT method, we compare it with the following PEFT approaches: **Adapter** (Houlsby et al., 2019), **(IA)**<sup>3</sup> (Liu et al., 2022), **LoRA** (Hu et al., 2022) and **LN-Tuning** (Zhao et al., 2024). We also include **full fine-tuning** that updates all LLM weights. For OPTUNE,  $r = 16$  and the dropout ratio is 0.1. We only apply OPTUNE to the query and value projections in the last two layers of the LLM. Detailed configurations are in Appendix A.

<sup>2</sup><https://huggingface.co/datasets/meta-llama/Llama-3.1-8B-Instruct-evals>

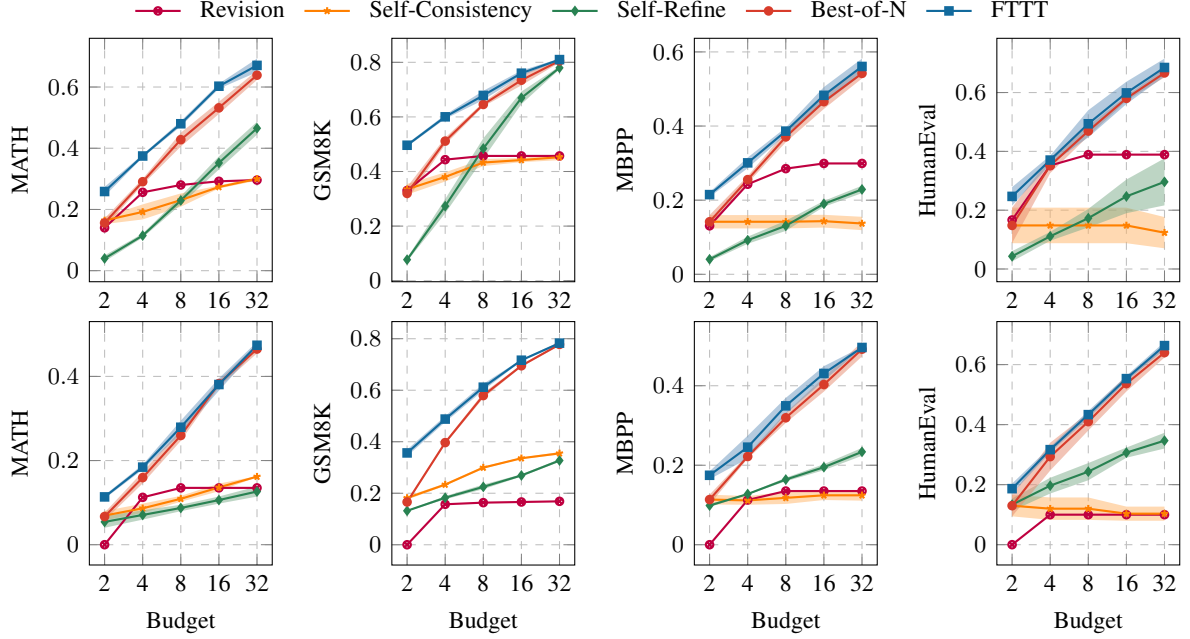


Figure 3: The scaling trends of different methods under varying budgets. The colored area around the line denotes the standard deviation. The first row is the results of Llama-3.1-8B-Instruct and the second row is Mistral-7B-Instruct-v0.3.

## 4.2 Test-Time Scaling Results

Table 2 compares FTTT with various baselines across four reasoning datasets. FTTT, both with and without self-reflected feedback, outperforms conventional test-time scaling methods on average. This success is partially explained by the findings of Ye et al. (2024a), which show that training with error-correction data enhances reasoning capabilities and models barely retry during inference. FTTT is also efficient. For instance, the inference time of Llama-3.1-8B-Instruct on GSM8K with a budget of 32 is 3 GPU hours for the best parallel sampling method (Best-of-N) and 20 GPU hours for the best sequential revision method (Self-Refine). In contrast, FTTT achieves inference times of approximately 3 GPU hours without self-reflected feedback and 4 GPU hours with self-reflected feedback.

Notably, self-reflected feedback does not always improve results. Its effectiveness appears to depend on the LLM’s self-reflection ability. To test this, we computed the Spearman rank correlation between FTTT and Self-Refine, a self-reflection-based algorithm. The Spearman coefficient ( $r = 0.8656$ ,  $p \leq 0.05$ ) indicates a strong positive correlation, supporting our hypothesis. We also observe that Self-Consistency performs poorly on code tasks because sampled code snippets rarely match exactly,

making majority voting akin to random selection.

Figure 3 illustrates performance for FTTT and baselines under varying budgets. FTTT consistently outperforms baselines, with greater gains under constrained budgets. In contrast, Revision and Self-Consistency do not scale well. Revision struggles with long-context reasoning due to length generalization issues (Li et al., 2024), while Self-Consistency fails to leverage feedback, often discarding correct answers during majority voting due to long-tailed distributions of correct answers (Brown et al., 2024).

## 4.3 Fine-Tuning Results

We present the results of PEFT baselines and OPTUNE with a budget of 32 in Table 3. Best-of-N is applied to PEFT baselines to exploit test-time feedback. HumanEval is excluded as it lacks a training set. Table 3 highlights the effectiveness of OPTUNE, outperforming all PEFT baselines by at least 2.58% on average. OPTUNE is also parameter-efficient, with 439K trainable parameters that are comparable to the most lightweight PEFT method (LN-Tuning, 266K parameters), while surpassing the best PEFT method (LoRA, 1.7M parameters) with an order of magnitude fewer parameters. However, OPTUNE shows suboptimal performance on MATH for Mistral-7B-Instruct-v0.3, which is consistent

Method	#Param.	MATH	GSM8K	MBPP	Avg.
Llama-3.1-8B-Instruct					
Adapter (Houlsby et al., 2019)	134M	0.5933 <sub>0.0151</sub>	0.7979 <sub>0.0056</sub>	0.2632 <sub>0.0058</sub>	0.5515
(IA) <sup>3</sup> (Liu et al., 2022)	524K	0.6187 <sub>0.0105</sub>	0.8929 <sub>0.0107</sub>	0.5685 <sub>0.0022</sub>	0.6934
LoRA (Hu et al., 2022)	1.7M	0.6387 <sub>0.0136</sub>	0.9186 <sub>0.0037</sub>	0.5639 <sub>0.0242</sub>	0.7071
LN-Tuning (Zhao et al., 2024)	266K	0.6280 <sub>0.0113</sub>	0.8899 <sub>0.0056</sub>	0.5748 <sub>0.0175</sub>	0.6976
Full Fine-Tuning	8B	0.6027 <sub>0.0136</sub>	0.7722 <sub>0.0056</sub>	0.4034 <sub>0.0096</sub>	0.5928
OPTUNE	439K	<b>0.7013</b> <sub>0.0050</sub>	<b>0.9246</b> <sub>0.0056</sub>	<b>0.6184</b> <sub>0.0159</sub>	0.7481
Mistral-7B-Instruct-v0.3					
Adapter (Houlsby et al., 2019)	134M	<b>0.5418</b> <sub>0.0111</sub>	0.8264 <sub>0.0021</sub>	0.2763 <sub>0.0076</sub>	0.5482
(IA) <sup>3</sup> (Liu et al., 2022)	524K	0.5041 <sub>0.0056</sub>	0.8686 <sub>0.0060</sub>	0.4914 <sub>0.0185</sub>	0.6214
LoRA (Hu et al., 2022)	1.7M	0.5117 <sub>0.0091</sub>	0.8686 <sub>0.0016</sub>	0.4968 <sub>0.0046</sub>	0.6257
LN-Tuning (Zhao et al., 2024)	266K	0.4357 <sub>0.0115</sub>	0.8259 <sub>0.0051</sub>	0.4065 <sub>0.0095</sub>	0.5560
Full Fine-Tuning	7B	0.5388 <sub>0.0157</sub>	0.7355 <sub>0.0016</sub>	0.2548 <sub>0.0095</sub>	0.5097
OPTUNE	439K	0.4891 <sub>0.0111</sub>	<b>0.9003</b> <sub>0.0039</sub>	<b>0.5194</b> <sub>0.0070</sub>	0.6363

Table 3: Fine-tuning results on four datasets with a budget of 32. #Param. denotes the number of trainable parameters. We report the mean of three runs with different random seeds and standard deviation in the subscript. **Bold** entries are the best results.

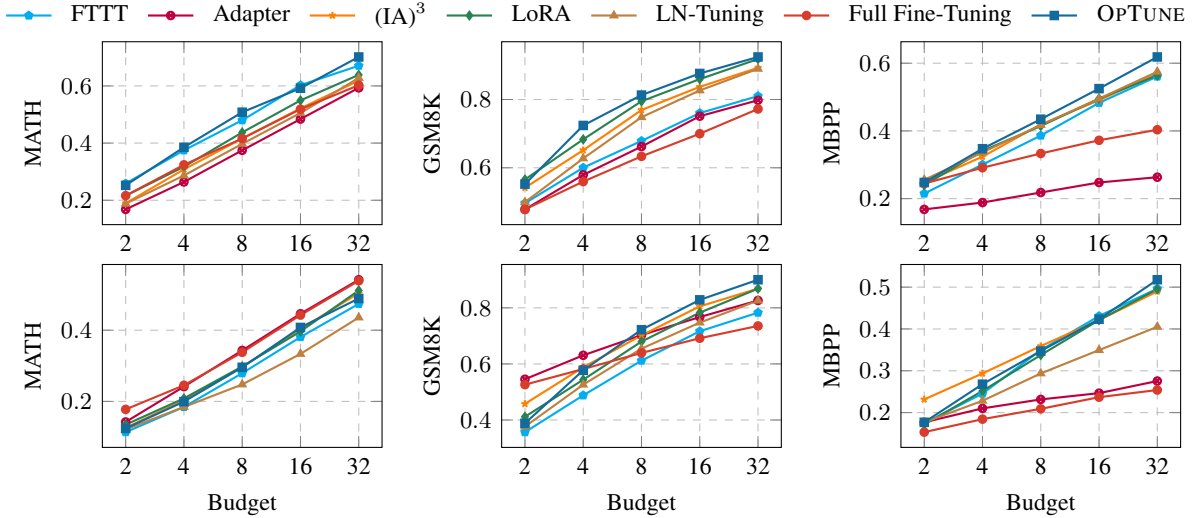


Figure 4: The scaling trends of different fine-tuning methods under varying budgets. We report the mean results of three random trials. The first row is the results of Llama-3.1-8B-Instruct and the second row is Mistral-7B-Instruct-v0.3.

with other PEFT methods with few trainable parameters (e.g., (IA)<sup>3</sup>, LoRA, LN-Tuning). This is likely due to Mistral-7B-Instruct-v0.3’s limited mathematical reasoning capabilities, requiring significant parameter updates to improve performance in this domain.

OPTUNE incurs negligible inference overhead. For example, on GSM8K with Llama-3.1-8B-Instruct and a budget of 32, the best test-time scaling baseline (FTTT) requires 4 GPU hours, whereas OPTUNE uses only 1.5 GPU hours, benefiting from shorter yet accurate predictions.

Finally, Figure 4 examines the scaling behav-

ior of PEFT baselines and OPTUNE. Initially, OPTUNE underperforms compared to other PEFT methods and FTTT, but it mostly achieves superior results when the budget exceeds 2. OPTUNE’s weaker performance with smaller budgets arises from its reliance on an initial attempt sampled from the raw LLM to initiate the process. This initial attempt often fails but is still counted as a valid attempt, making OPTUNE less competitive in low-budget settings.

#### 4.4 Analysis

**Ablation Study.** Table 5 presents an ablation study on the architecture design of OPTUNE. The

**Question:** Blake and Kelly are having a contest to see who can run the most in 15 minutes. They decide to do it on a football field that is 100 yards long. Blake runs back and forth 15 times. Kelly runs back and forth once, and then decides that she doesn't want to run next to Blake, so she starts to run to the 40-yard line and back. She does this 34 times. How much farther does the winner run than the loser?

**LoRA:** Blake runs  $15 * 100 * 2 = 3000$  yards.  
Kelly runs  $100 * 2 = 200$  yards for each run to the 40-yard line and back.

Kelly runs  $200 * 34 = 6800$  yards.  
Kelly runs  $6800 - 3000 = 3800$  yards more than Blake.  
The final answer is **3800 ✗**.

**OPTUNE:** Blake runs back and forth 15 times, and each time, he runs 100 yards twice because he runs to one end of the field and back. So, Blake's total running distance is  $15 * 200 = 3000$  yards. Kelly runs back and forth once, which is 200 yards, and then runs to the 40-yard line and back, which is  $40 * 2 = 80$  yards, and does this 34 times. Kelly's total running distance is  $200 + (34 * 80) = 200 + 2720 = 2920$  yards.  
Therefore, Blake runs  $3000 - 2920 = 80$  yards farther than Kelly.  
The final answer is **80 ✓**.

**Question:** Josh decides to take up juggling to perform at the school talent show a month in the future. He starts off practicing juggling 3 balls, and slowly gets better adding 1 ball to his juggling act each week. After the end of the fourth week the talent show begins, but when Josh walks on stage he slips and drops three of his balls. 2 of them are caught by people in the crowd as they roll off the stage, but one gets lost completely since the auditorium is dark. With a sigh, Josh starts to juggle on stage with how many balls?

**LoRA:** Josh started with 3 balls and added 1 ball each week for 4 weeks, so he had  $3 + 4 = 7$  balls.  
He dropped 3 balls, but 2 were caught, so he lost  $3 - 2 = 1$  ball.  
He was left with  $7 - 1 = 6$  balls.  
The final answer is **6 ✗**.

**OPTUNE:** Josh started with juggling 3 balls. After the end of the fourth week, Josh would have added 4 balls, making it a total of  $3+4=7$  balls.  
Josh had 3 balls left after 2 of the balls were caught by people in the crowd, so he was left with  $7-3=4$  balls.  
The final answer is **4 ✓**.

Table 4: Two GSM8K examples with model predictions. ✓ means correct and ✗ means incorrect. Sentences marked in orange are incorrect reasoning steps and those marked in blue are the correct counterparts.

Method	#Params.	Pass@1
OPTUNE	439K	0.2600
+ w/o Dropout	439K	0.2243
+ w/o Norm	439K	0.2150
+ Not Share $\theta_1$ & $\theta_2$	878K	0.2243

Table 5: The ablation study of OPTUNE. We report results of Llama-3.1-8B-Instruct on MBPP with a budget of 2.

results demonstrate that all components are essential, as removing any of them significantly degrades performance. Notably, normalization is the most critical component, as it makes the input gradient scales more stable.

**Case Study.** Table 4 provides two examples on GSM8K where the leading PEFT method, LoRA, fails, but OPTUNE succeeds using Llama-3.1-8B-Instruct. These examples highlight OPTUNE's superior ability to correctly interpret and reason through questions, unlike LoRA.

## 5 Related Work

**Learning from Feedback.** Other than the heuristic binary feedback studied in this work, prior research has explored feedback from various sources, such as humans (Ouyang et al., 2022), other models (Yang et al., 2022), tools (Schick et al., 2023), and knowledge bases (Gao et al., 2023). This pa-

per focuses on demonstrating the effectiveness of the proposed method and other feedback types are beyond the scope of this paper.

**Test-Time Training.** Test-Time Training (TTT) has shown success in the image modality by addressing distribution shifts and enhancing model capacity through self-supervised fine-tuning on each test case (Sun et al., 2020; Liu et al., 2021; Sun et al., 2023). Recent studies have extended TTT to the text modality (Hardt and Sun, 2024; Wang et al., 2024). The most relevant work, by Akyürek et al. (2024), uses TTT to enhance the reasoning ability of LLMs. However, their method relies heavily on human scaffolding for self-supervision and does not generalize beyond ARC-AGI (Chollet, 2019). In contrast, FTTT is generally applicable.

**Learning to Optimize.** Learning to Optimize (L2O) trains a network to act as an optimizer for another network (Chen et al., 2022). Early approaches used reinforcement learning to train such optimizers (Li and Malik, 2017; Chen et al., 2017), while recent work focuses on discovering analytical white-box optimizers (Bello et al., 2017; Chen et al., 2023). The most relevant work, MEND (Mitchell et al., 2022), trains a network to predict weight updates from training gradients. OPTUNE builds on this idea, extending it to learn from test-time feedback with a distinct architecture.

## 6 Conclusion

In this paper, we propose a novel paradigm that leverages optimization to address the challenge of exploiting test-time feedback, resulting in improved scaling performance. We further present a learnable test-time optimizer, OPTUNE, which surpasses various PEFT baselines. Both FTTT and OPTUNE are efficient in terms of speed and trainable parameter count.

## Limitations

The current evaluation setting limits FTTT’s potential by providing only binary feedback (i.e., correct or incorrect) for each attempt. However, developing complex reasoning environments with rich feedback is beyond the scope of this work. Additionally, while continuous feedback, such as that from reward models (Yang et al., 2024), has been extensively studied, it is not examined here. Our method can be adapted to continuous feedback with minimal modifications, such as using REINFORCE (Williams, 1992). For coherence, we leave this exploration to future work.

## Acknowledgments

This work is supported by National Key R&D Program of China (Project No. 2022ZD0161200, 2022ZD0161201). This work is also supported by the Research Grants Council of the HKSAR - Early Career Scheme (Grant No. 24200223) and Hong Kong Innovation and Technology Commission Project No. ITS/228/22FP. It is also partially supported by RGC Senior Research Fellow Scheme Ref. No. SRFS2425-4S03.

## References

Ekin Akyürek, Mehul Damani, Linlu Qiu, Han Guo, Yoon Kim, and Jacob Andreas. 2024. [The surprising effectiveness of test-time training for abstract reasoning](#). Preprint, arXiv:2411.07279.

Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. 2021. [Program synthesis with large language models](#). CoRR, abs/2108.07732.

Jimmy Ba, Geoffrey E. Hinton, Volodymyr Mnih, Joel Z. Leibo, and Catalin Ionescu. 2016. [Using fast weights to attend to the recent past](#). In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4331–4339.

Irwan Bello, Barret Zoph, Vijay Vasudevan, and Quoc V. Le. 2017. [Neural optimizer search with reinforcement learning](#). In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 459–468. PMLR.

Bradley C. A. Brown, Jordan Juravsky, Ryan Saul Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. [Large language mon-keys: Scaling inference compute with repeated sampling](#). CoRR, abs/2407.21787.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri 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 Hebggen 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. 2021. [Evaluating large language models trained on code](#). CoRR, abs/2107.03374.

Tianlong Chen, Xiaohan Chen, Wuyang Chen, Howard Heaton, Jialin Liu, Zhangyang Wang, and Wotao Yin. 2022. [Learning to optimize: A primer and A benchmark](#). J. Mach. Learn. Res., 23:189:1–189:59.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. 2023. [Symbolic discovery of optimization algorithms](#). In *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*.

Yutian Chen, Matthew W. Hoffman, Sergio Gomez Colmenarejo, Misha Denil, Timothy P. Lillicrap, Matthew M. Botvinick, and Nando de Freitas. 2017. [Learning to learn without gradient descent by gradient descent](#). In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 748–756. PMLR.

François Chollet. 2019. [On the measure of intelligence](#). Preprint, arXiv:1911.01547.

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. 2021. [Training verifiers to solve math word problems](#). *CoRR*, abs/2110.14168.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Alonso, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelier van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Luyu Gao, Zhu Yun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. [RARR: Researching and revising what language models say, using language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.
- Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. [rstar-math: Small llms can master math reasoning with self-evolved deep thinking](#). Preprint, arXiv:2501.04519.
- Moritz Hardt and Yu Sun. 2024. [Test-time training on nearest neighbors for large language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the MATH dataset](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. [Parameter-efficient transfer learning for NLP](#). In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. [Large language models cannot self-correct reasoning yet](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *CoRR*, abs/2310.06825.
- Yoon Kim and Alexander M. Rush. 2016. [Sequence-level knowledge distillation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Ke Li and Jitendra Malik. 2017. [Learning to optimize](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

- Yanyang Li, Shuo Liang, Michael Lyu, and Liwei Wang. 2024. [Making long-context language models better multi-hop reasoners](#). In [Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics \(Volume 1: Long Papers\)](#), pages 2462–2475, Bangkok, Thailand. Association for Computational Linguistics.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In [The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024](#). OpenReview.net.
- Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohata, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. [Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning](#). In [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](#).
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). [Trans. Assoc. Comput. Linguistics](#), 12:157–173.
- Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. 2025. [Can 1b llm surpass 405b llm? rethinking compute-optimal test-time scaling](#). [Preprint, arXiv:2502.06703](#).
- Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. 2021. [TTT++: when does self-supervised test-time training fail or thrive?](#) In [Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual](#), pages 21808–21820.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-refine: Iterative refinement with self-feedback](#). In [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](#).
- Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022. [Fast model editing at scale](#). In [The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022](#). OpenReview.net.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. [s1: Simple test-time scaling](#). [Preprint, arXiv:2501.19393](#).
- OpenAI, :, Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaev, Daniel Selsam, David Do-han, Francis Song, Hunter Lightman, Ignasi Clavera, Jakub Pachocki, Jerry Tworek, Lorenz Kuhn, Lukasz Kaiser, Mark Chen, Max Schwarzer, Mostafa Rohaninejad, Nat McAleese, o3 contributors, Oleg Mürk, Rhythm Garg, Rui Shu, Szymon Sidor, Vineet Kosaraju, and Wenda Zhou. 2025. [Competitive programming with large reasoning models](#). [Preprint, arXiv:2502.06807](#).
- 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. 2022. [Training language models to follow instructions with human feedback](#). In [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](#).
- Peng Si Ow and Thomas E Morton. 1988. Filtered beam search in scheduling. [The International Journal Of Production Research](#), 26(1):35–62.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. [Toolformer: Language models can teach themselves to use tools](#). In [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](#).
- Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. 2021. [Linear transformers are secretly fast weight programmers](#). In [Proceedings of the 38th International Conference on Machine Learning](#), volume 139 of [Proceedings of Machine Learning Research](#), pages 9355–9366. PMLR.
- Jürgen Schmidhuber. 1992. [Learning to control fast-weight memories: An alternative to dynamic recurrent networks](#). [Neural Comput.](#), 4(1):131–139.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Reflexion: language agents with verbal reinforcement learning](#). In [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](#).

- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. [Scaling LLM test-time compute optimally can be more effective than scaling model parameters](#). *CoRR*, abs/2408.03314.
- Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. [Dropout: a simple way to prevent neural networks from overfitting](#). *J. Mach. Learn. Res.*, 15(1):1929–1958.
- Yu Sun, Xinhao Li, Karan Dalal, Chloe Hsu, Sanmi Koyejo, Carlos Guestrin, Xiaolong Wang, Tatsunori Hashimoto, and Xinlei Chen. 2023. [Learning to \(learn at test time\)](#). *CoRR*, abs/2310.13807.
- Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. 2020. [Test-time training with self-supervision for generalization under distribution shifts](#). In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research*, pages 9229–9248. PMLR.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [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.
- Yan Wang, D. Ma, and Deng Cai. 2024. [With greater text comes greater necessity: Inference-time training helps long text generation](#). *CoRR*, abs/2401.11504.
- Ronald J. Williams. 1992. [Simple statistical gradient-following algorithms for connectionist reinforcement learning](#). *Mach. Learn.*, 8:229–256.
- Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024. [Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models](#). *Preprint*, arXiv:2408.00724.
- Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Qizhe Xie. 2023. [Self-evaluation guided beam search for reasoning](#). In *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*.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. [Qwen2.5-math technical report: Toward mathematical expert model via self-improvement](#). *CoRR*, abs/2409.12122.
- Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. [Re3: Generating longer stories with recursive reprompting and revision](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. [React: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. 2024a. [Physics of language models: Part 2.2, how to learn from mistakes on grade-school math problems](#). *Preprint*, arXiv:2408.16293.
- Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. 2024b. [Differential transformer](#). *Preprint*, arXiv:2410.05258.
- Bingchen Zhao, Haoqin Tu, Chen Wei, Jieru Mei, and Cihang Xie. 2024. [Tuning layernorm in attention: Towards efficient multi-modal LLM finetuning](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

## A Hyperparameter Settings

Below is the detailed configurations of different PEFT methods:

- **Adapter** uses a learning rate of  $1e-4$  and the reduction factor of the bottleneck is 16.
- **(IA)<sup>3</sup>** uses a learning rate of  $5e-5$ .
- **LoRA** uses a learning rate of  $2e-5$ . We only apply LoRA to the query and value projections in the last 8 layers, with a rank of 16 and a dropout ratio of 0.05.
- **LN-Tuning** uses a learning rate of  $4e-4$ .
- **Full Fine-Tuning** uses a learning rate of  $1e-5$ .

The number of training epochs is 100, 10, and 3 for MBPP, GSM8K, and MATH, respectively. We use the Adam optimizer with a batch size of 20 for all methods in all datasets, including OPTUNE. We use the checkpoint of the last epoch for inference.

For OPTUNE, we sample 10 attempts for each training example together with the raw question to construct the model input. We employ nucleus sampling (Holtzman et al., 2020) with a temperature of 0.6 and  $p = 0.95$  to generate attempts. The number of training epochs for MBPP, GSM8K, and MATH is set to 10, 3, and 3, respectively. The learning rate is  $1e-5$ . In inference, we sample an attempt using the same hyperparameters as in data generation before applying OPTUNE to mitigate

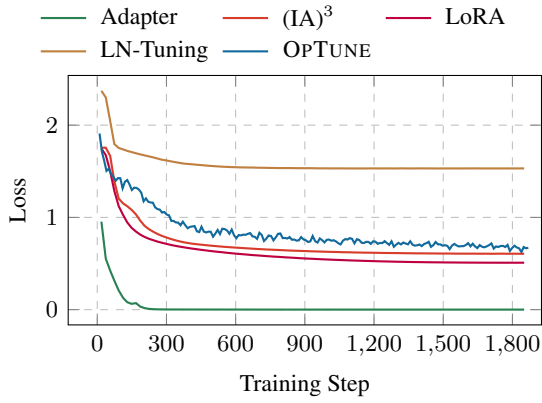


Figure 5: The training curves of PEFT methods when fine-tuning Llama-3.1-8B-Instruct on MBPP.

the train-test discrepancy. We alternate between sampling attempts from the raw LLM and predicting refined attempts from sampled attempts when scaling OPTUNE with more budgets.

## B Prompts

Below is the reflection generation prompts  $P$  for Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3:

Llama-3.1-8B-Instruct
<p><b>User:</b> Solve the following math problem . . .</p> <p><b>Assistant:</b> . . . the final answer is: . . .</p> <p><b>User:</b> Your answer is incorrect. Please carefully check the solution and summarize all mistakes in short. Do NOT provide the corrected solution. Do NOT say “my solution”.</p> <p><b>Assistant:</b> Here is the summary of the mistakes in the previous solution . . .</p>
Mistral-7B-Instruct-v0.3
<p><b>User:</b> Solve the following math problem . . .</p> <p><b>Assistant:</b> . . . the final answer is: . . .</p> <p><b>User:</b> Your answer is incorrect. Carefully check the solution step-by-step and list all mistakes in short. MUST NOT provide the correct answer. Your response MUST be in the third person tone.</p> <p><b>Assistant:</b> Here is the summary of the mistakes in the previous solution . . .</p>

Sentences in gray are the prompt  $P$  and the one in blue is the generated reflection.

## C Training Curves

Figure 5 is the training curves of various PEFT methods, including OPTUNE. We observe that Adapter shows a clear signal of overfitting, where it has a training loss close to 0, while its performance

Model	GSM8K (trained on MATH)	HumanEval (trained on MBPP)
Llama-3.1-8B-Instruct		
Raw Model	0.8069	0.6728
+ OPTUNE	<b>0.8914</b>	<b>0.7037</b>
Mistral-7B-Instruct-v0.3		
Raw Model	0.7807	0.6500
+ OPTUNE	<b>0.7921</b>	<b>0.7100</b>

Table 6: The cross-task transferability of OPTUNE.

on the test set is the worst. All PEFT methods seem to converge smoothly. For OPTUNE, its training is not as stable as baselines, suggesting the difficulty of learning to optimize problem.

## D Transferability

While generalization and transfer are not the primary focus of this work, we demonstrate OPTUNE’s transferability by training it on one dataset (MBPP/MATH) and testing it on another (HumanEval/GSM8K). Table 6 indicates that OPTUNE consistently improves baseline performance, even when tested on unseen datasets. This highlights OPTUNE’s potential for generalization and transferability.