

# Improving Generalization in Intent Detection: GRPO with Reward-Based Curriculum Sampling

Zihao Feng<sup>1,2\*†</sup>, Xiaoxue Wang<sup>1\*</sup>, Ziwei Bai<sup>1\*</sup>, Donghang Su<sup>1\*</sup>,  
Bowen Wu<sup>1</sup>, Qun Yu<sup>1</sup>, Baoxun Wang<sup>1</sup>

<sup>1</sup>Platform and Content Group, Tencent

<sup>2</sup>Faculty of Computing, Harbin Institute of Technology

21b903052@stu.hit.edu.cn

{yukixxwang, ziweibai, ashersu, jasonbwu, sparkyu, asulewang}@tencent.com

## Abstract

Intent detection, a critical component in task-oriented dialogue (TOD) systems, faces significant challenges in adapting to the rapid influx of integrable tools with complex interrelationships. Existing approaches, such as zero-shot reformulations and LLM-based dynamic recognition, struggle with performance degradation when encountering unseen intents, leading to erroneous task routing. To enhance the model’s generalization performance on unseen tasks, we employ Reinforcement Learning (RL) combined with a Reward-based Curriculum Sampling (RCS) during Group Relative Policy Optimization (GRPO) training in intent detection tasks. Experiments demonstrate that RL-trained models substantially outperform supervised fine-tuning (SFT) baselines in generalization. Besides, the introduction of the RCS, significantly bolsters the effectiveness of RL in intent detection by focusing the model on challenging cases during training. Moreover, incorporating Chain-of-Thought (COT) processes in RL notably improves generalization in complex intent detection tasks, underscoring the importance of thought in challenging scenarios. This work advances the generalization of intent detection tasks, offering practical insights for deploying adaptable dialogue systems.

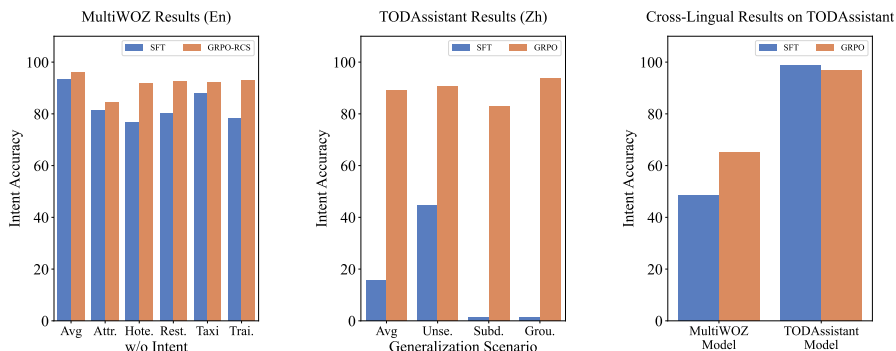


Figure 1: Comparative performance of RL-Trained and SFT-Trained models in intent detection across various generalization scenarios

\*Equal contribution

†Zihao Feng was an intern at Tencent during the preparation of this work

# 1 Introduction

As a crucial component of Task-oriented Dialogue (TOD) systems [1, 2], the intent detection module aims to identify the underlying requirements of users’ queries [3, 4]. Consequently, the intent detection models are expected to efficiently adapt to evolving task priorities and requirements, so as to conduct dynamic task allocation among multiple agents in complicated application scenarios.

The recent development of LLMs has accelerated the evolution of TOD systems, and with the rapid iteration of integrable artificial API tools [5, 6], the number of AI tools that can be incorporated into TOD systems is increasing. This situation leads to a great challenge that, actually, intent detection models need to flexibly adapt to newly introduced tools for unseen tasks, with no timely incremental training processes. In many cases, tools within the management of intent detection modules maintain complex interrelationships, such as functional similarity, overlapping, inclusion, etc. Thus, the generalization of intent detection models is the essence for TOD systems to adjust to complicated practical scenarios, in which a number of tools, with complex relationships and interactions, may be frequently involved.

Previous studies have made much efforts to improve the accuracy of intent detection models by adopting new tools to handle unseen tasks. For example, the model proposed by Siddique et al. introduces external common sense knowledge to address this problem [7]. Comi et al.[8] reformatted the tasks in an NLI format to achieve zero-shot capability. Moreover, LLM-based models [9, 1] dynamically recognized unknown tasks by capitalizing on their inherent zero-shot capabilities. However, these models often experienced significant performance degradation in intent detection models when confronted with unseen or new intent, resulting in the system incorrectly routing user intent to the unmatched agent. This situation indicates that enhancing the generalization of intent detection models is particularly critical.

Reinforcement learning has been proved to be valuable in improving the generalization of LLMs [10], which has also been supported by the exceptional cross-task generalization of the recent model DeepSeek-R1 [11]. Inspired by the principle of DeepSeek-R1, we propose to apply the Group Relative Policy Optimization (GRPO) methodology to enhance the generalization of the intent detection model. In particular, to ensure that the R1-style RL process achieves expected performances on the intent detection problem, a sampling strategy is presented in this work. As depicted in Figure 1, the experimental results demonstrate that in varying generalization scenarios, the reinforcement learning (RL) model successfully predicts user query intents, significantly outperforming the supervised fine-tuned (SFT) model. This superiority is particularly evident in terms of generalization across unseen intents, subdivided intents, grouped intents, and cross-language. In conclusion, our work offers the following findings:

- We demonstrate that models trained with RL significantly outperform those trained with SFT on the intent detection problem, in terms of generalization across unseen intents, subdivided intents, grouped intents, and cross-language.
- To stimulate the capability of GRPO training, we introduce the Rewards-based Curriculum Sampling Strategy, which is found to be valuable for enabling models to focus more on challenging cases during the training process.
- Incorporating COT [12] processes during reinforcement learning significantly enhances model generalization on complex intent detection tasks, highlighting the importance of thought processes for improving generalization in challenging scenarios.
- Furthermore, our experiments also show that even a base model without instruction training can achieve performance comparable to the instruction model on the intent detection task. This finding suggests that the Function Call capability of the base model may not be a necessary prerequisite for intent detection models trained with RL.

## 2 Method

### 2.1 Task Formulation

In task-oriented dialogue systems, accurate detection of user intents is essential for dialogue state tracking and subsequent API execution. We formulate the intent detection task as follows: Given a

dialogue history  $H = \{(u_1, a_1, y_1), (u_2, a_2, y_2), \dots, (u_{t-1}, a_{t-1}, y_{t-1})\}$ , where  $u_i$ ,  $a_i$ , and  $y_i \in \mathcal{Y}$  represent the user’s utterance, the assistant’s response, and the ground truth intent label at turn  $i$ , respectively.  $\mathcal{Y} = \{c_1, c_2, \dots, c_K\}$  denotes a predefined set of  $K$  actionable intents related to domain-specific operations, with each intent  $c_i$  associated with a natural language description  $d_i$  in the prompt. The objective of an intent detection model  $M$  is to accurately predict the intent  $y_t \in \mathcal{Y}$  of the current user’s utterance  $u_t$ . Formulated as:

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{n=1}^N \log P_{\theta}(y_t^n | H^n, u_t^n, d_1, d_2, \dots, d_K) \quad (1)$$

where  $\theta$  represents the parameters of model  $M$ ,  $N$  is the number of training examples,  $P_{\theta}$  denotes the probability assigned by model  $M$ .

Apparently, the model  $M$  demonstrates non-trivial generalization potential for evolving dialogue systems, as its architecture theoretically supports the discovery of novel intent categories through dynamic prompt engineering. Formally, for  $y_t = c_{K+1} \notin \mathcal{Y}$ , model  $M$  can add the description  $d_{K+1}$  of  $c_{K+1}$  to the prompt to predict  $y_t$ . In particular, this  $y_t$  may represent not only a completely new category distinct from  $\mathcal{Y}$ , but also potentially a division or recombination of previous categories.

## 2.2 Intent Detection via Reinforcement Learning

Directly applying supervised fine-tuning (SFT) to learn the prediction of  $y_t$  has been a conventional approach, however, this method often suffers from poor generalization capabilities. In this paper, inspired by DeepSeek-R1-Zero [11], which demonstrated the significant potential of reinforcement learning combined with model reasoning, we design rule-based rewards and exclusively employ GRPO [13] to guide model training.

Specifically, building upon an arbitrary LLM, we construct a complete prompt using the ReAct Prompting [14] method, where the system prompt is "You are a helpful assistant.". In the final turn of the dialogue, we insert an instruction composed of the user query and other relevant information, such as descriptions of available tools. The specific instruction template is as follows.

**Instruction Template of ReAct Prompting**

You are an agent that helps users choose the right tool or tools from the list of given tools to solve their problems.

For each tool, you are first given its description and required parameters. Then, a logic module specifically explains the logical information needed for this tool to handle multi-turn conversation issues.

## Tool APIs

{tools text}

## Task Logic

{logic text}

## Output Format

Use the following format:

Last Tool: the tool used in last query  
 Question: the input question you must answer  
 Thought: you should always think about what to do  
 Action: the action to take  
 Finish!

Begin!  
 Last Tool: {tool}  
 Question: {query}

Regarding the training objectives, we design two rule-based reward functions to guide reinforcement learning training. Specifically, these include a Format Reward to constrain the model’s output structure and an Answer Reward to evaluate the correctness of intent detection.

$$R = \lambda_{\text{format}} \cdot R_{\text{format}} + \lambda_{\text{answer}} \cdot R_{\text{answer}} \quad (2)$$

where  $\lambda_{\text{format}}$  and  $\lambda_{\text{answer}}$  are weighting coefficients for each respective reward component.

**Format Reward** We restrict the model’s output to strictly follow a fixed format, as specified in the Instruction Template of ReAct Prompting. Specifically, the model’s output must strictly conform to a

three-line structure where each line begins with “Thought:”, “Action:”, and “Finish!” respectively. Each of these keywords must appear exactly once in the entire output and the content of the third line is limited to solely “Finish!”.

$$R_{\text{format}} = \begin{cases} 1, & \text{if format is correct} \\ 0, & \text{otherwise} \end{cases}$$

**Accuracy Reward** The accuracy-based reward is a binary metric that evaluates the exact match between the predicted intent  $\hat{y}_t$  and the ground truth label  $y_t$ . We employ a regular expression-based method to extract the predicted intent from the model’s output.

$$R_{\text{answer}} = \begin{cases} 1, & \text{if the answer } \hat{y}_t \text{ fully matches the ground truth } y_t \\ 0, & \text{otherwise} \end{cases}$$

### 2.3 Reward-Based Curriculum Sampling

Research indicates that low reward variance leads to a flat landscape in the RLHF objective, resulting in suboptimal convergence [15]. Our observations on intent detection tasks reveal that GRPO-trained models converge remarkably quickly, reaching accuracy comparable to SFT models within dozens of training steps. Consequently, in subsequent training phases, the reward variance becomes extremely small, and the model’s focus on challenging examples diminishes. To address this issue, we employ an offline Reward-based Curriculum Sampling Strategy to enhance both the efficiency and effectiveness of the training process.

**Offline Reward Collection** To select the most challenging sample for RL, we first apply the GRPO method to the entire training dataset, recording the rewards for each data across all samples throughout the GRPO training process. Just as shown in Eq 3, the  $G$  represents the sampling number of each data,  $R^{i,j}$  represents the reward of  $j$ -th sampling of the  $i$ -th data, and the  $Score_i$  represents the score of  $i$ -th data.

$$Score_i = \sum_{j=1}^G (\lambda_{\text{format}} \cdot R_{\text{format}}^{i,j} + \lambda_{\text{answer}} \cdot R_{\text{answer}}^{i,j}) \quad (3)$$

**Curriculum Sampling** After obtaining the training rewards for each sample, we employ a two-stage training method. In the first stage, we train the model for dozens of steps on the entire dataset until the accuracy on the validation set changes less than a certain threshold. We intentionally avoid using easier data during this initial stage because the model demonstrated significantly low rewards across all examples at the beginning of the training process. In addition, this approach facilitates the transition of our proposed method to an online format in subsequent work. In the second stage, we define the  $i$ -th data is challenging when the  $Score_i < (\lambda_{\text{format}} + \lambda_{\text{answer}}) * G$ . We select the challenging data to continue training the model trained in the first stage. This approach allows the model to concentrate on these difficult examples during the second stage.

## 3 Experimental Setup

### 3.1 Dataset

We conduct experiments on two task-oriented dialogue datasets.

The first dataset is the widely used MultiWOZ benchmark, specifically a subset of **MultiWOZ 2.2** [16]. This large-scale multi-domain task-oriented dialogue dataset contains 10,437 conversations spanning 7 domains. These domains encompass tasks that require multiple interaction turns to complete, such as flight booking and hotel reservations. We extract the intent subtask from this dataset for training and evaluation of our proposed method.

Additionally, considering the rapid iteration of integrable artificial API tools, we construct a dataset that simulates interactions with a general AI assistant that integrates various task capabilities, named

**TODAssistant.** This dataset encompasses 10 task categories, including traditional task-oriented functions such as signature settings, friend recommendations, and chatbot recommendations, as well as AI-driven task types, including text-to-image generation, image style transformation, and text-based conversation. All dialogue data for these tasks were simulated using GPT-4o [17] to generate conversations representing possible interaction scenarios for each tool, with specific details omitted here. In summary, this is a task-oriented dialogue dataset containing 10 tasks, covering both traditional task-oriented dialogue-focused areas and emerging AI-driven tasks. The data is entirely generated by LLMs and comprises 9,500 training samples and 500 test samples.

To better evaluate model adaptability to situations involving new domains, subdivided, or grouped tasks, we further develop three generalization test sets with new intents that are not included in the known 10 categories:

- **TODAssistant-Unseen5:** Introduces 5 completely novel tasks not encountered in the dataset, including singing children’s songs and storytelling, which are oriented toward children’s scenarios.
- **TODAssistant-Subdivided:** For the text chat task already included in the 10 categories, we divide it into three more granular intents to simulate real-world scenarios where finer-grained capabilities might better address specific user needs. Specifically, we split the text chat task into:
  - Various text processing intents: Covering purpose-specific text generation tasks including translation, text classification, text generation, mathematical calculation, and code generation.
  - Safety topics: Involving content related to pornography, violence, etc.
  - Free topic conversation: Chit-chat or intents not belonging to the other two categories.
- **TODAssistant-Grouped:** This set simulates situations where, due to agent upgrades, multiple previously separate tasks may be completed by a single agent. Specifically, we regroup two relatively similar intents — "friend recommendations" and "chatbot recommendations" into a single intent.

To clarify, TODAssistant-Unseen5 introduces 5 entirely new task categories, TODAssistant-Subdivided uses a portion of the test samples originally belonging to the text chat task and divides them into three new intents, and TODAssistant-Grouped modifies the intent of two test set categories into one new intent. It is important to emphasize that none of these categories were encountered during the training process.

### 3.2 Setup

We selected Qwen2.5-7B-Instruct<sup>3</sup> [18] as our foundation model, which represents a widely adopted open-source large language model.

For the MultiWOZ2.2 dataset, we utilize the 10k conversations to conduct reinforcement learning. We conduct 60 steps for the first stage of the curriculum learning, and 1 epoch (153 steps) for the second stage. For both of the two stages, we train our model with a learning rate of  $9.0 \times 10^{-6}$ , incorporating a sampling strategy that generated 7 responses per prompt at a temperature parameter of 0.7. In the case of the TODAssistant dataset, we employ the entire training set for our experiments. We train the model with a learning rate of  $3.0 \times 10^{-6}$ , incorporating a sampling strategy that generated 7 responses per prompt at a temperature parameter of 0.9. For all the datasets, we utilize a global batch size of 448 for our training.

Regarding the supervised fine-tuning approach, we fully fine-tune the model with the same epoch of the corresponding GRPO-based method. On both datasets, we employ **Accuracy** as the metric to measure the effectiveness of intent detection.

## 4 Experiments

### 4.1 Comparison of Reinforcement Learning and Supervised Fine-Tuning Effects

We conduct intent detection training on two datasets using both GRPO and SFT approaches. Our evaluation strategy involves testing in-domain intent categories (those present in the training data)

<sup>3</sup><https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

Table 1: Results of the in-domain evaluation on two datasets.

Model	TODAssistant	MultiWOZ 2.2	Avg
Qwen2.5-7B-Instruct	22.4	23.2	22.8
Qwen2.5-7B-Instruct + SFT	98.8	93.3	96.1
Qwen2.5-7B-Instruct + GRPO	96.8	93.3	95.1

Table 2: Results of the out-of-domain evaluation on MultiWOZ 2.2 dataset. The symbol “†” denotes the performance on the excluded intent category that was unseen in the training data.

Model	Attraction	Hotel	Restaurant	Taxi	Train	Avg
<u>Baseline</u>						
Qwen2.5-7B-Instruct + SFT	92.6	93.5	93.4	95.6	92.3	93.3
Qwen2.5-7B-Instruct + GRPO	94.5	91.6	91.9	93.9	92.6	93.3
<u>w/o Attraction</u>						
Qwen2.5-7B-Instruct + SFT	<b>43.8<sup>†</sup></b>	<u>94.3</u>	<u>93.7</u>	96.4	92.9	81.3
Qwen2.5-7B-Instruct + GRPO	43.1 <sup>†</sup>	92.7	93.0	<u>97.5</u>	<u>93.3</u>	<u>84.4</u>
<u>w/o Hotel</u>						
Qwen2.5-7B-Instruct + SFT	93.5	37.1 <sup>†</sup>	<u>92.3</u>	95.0	91.3	76.9
Qwen2.5-7B-Instruct + GRPO	<u>95.3</u>	<b>87.1<sup>†</sup></b>	92.3	<u>96.1</u>	<u>92.6</u>	<u>91.8</u>
<u>w/o Restaurant</u>						
Qwen2.5-7B-Instruct + SFT	92.6	89.7	57.1 <sup>†</sup>	93.6	<u>92.1</u>	80.3
Qwen2.5-7B-Instruct + GRPO	<u>95.1</u>	<u>93.0</u>	<b>91.2<sup>†</sup></b>	<u>95.3</u>	91.9	<u>92.8</u>
<u>w/o Taxi</u>						
Qwen2.5-7B-Instruct + SFT	87.0	90.0	92.7	53.4 <sup>†</sup>	89.6	88.0
Qwen2.5-7B-Instruct + GRPO	<u>95.9</u>	<u>92.5</u>	<u>92.6</u>	<b>74.2<sup>†</sup></b>	<u>92.9</u>	<u>92.3</u>
<u>w/o Train</u>						
Qwen2.5-7B-Instruct + SFT	92.1	91.1	<u>94.1</u>	91.8	47.9 <sup>†</sup>	78.4
Qwen2.5-7B-Instruct + GRPO	<u>95.9</u>	<u>93.1</u>	92.6	<u>96.8</u>	<b>90.6<sup>†</sup></b>	<u>93.0</u>

and out-of-domain intent categories (those unseen during training). It is important to note that the GRPO training discussed in this subsection corresponds to the methodology described in Section 2.1, which does not incorporate curriculum learning. Our primary objective is to analyze the performance differences between models trained using GRPO versus those trained through standard SFT.

#### 4.1.1 Performance on In-Domain Test Set

As shown in Table 1, both SFT and GRPO-trained models significantly improve intent recognition performance on in-domain categories. However, using only RL (GRPO) on the same training data as SFT does not surpass SFT’s performance on in-domain testing. While both approaches achieve comparable convergence results on the more complex MultiWOZ 2.2 dataset, GRPO performs slightly worse on the machine-generated TODAssistant dataset.

#### 4.1.2 Performance in generalization scenarios

To assess the performance of RL methodologies across various generalization scenarios, we conduct a comparative analysis of the GRPO model and the SFT model, focusing on their adaptability as the intent label set progressively evolves and deviates from the training dataset.

Table 3 shows performance on the three generalization test sets of TODAssistant. Compared to the untuned Qwen2.5-7B-Instruct model, the performance of the SFT model shows a notable decline across all three test sets. This deterioration is especially evident on the Subdivided and Grouped test sets, where the SFT-trained model limits its predictions to the 10 categories seen during training,

Table 3: Results of the out-of-domain evaluation on TODAssistant dataset

Model	TODAssistant	Unseen5	Subdivided	Grouped	Avg
Qwen2.5-7B-Instruct	-	63.0	40.2	21.6	41.6
+ SFT	-	44.5	0.0	0.0	14.8
+ GRPO	-	<b>90.6</b>	<b>83.1</b>	<b>93.6</b>	<b>89.1</b>
+ GRPO (MultiWOZ)	65.2	-	-	-	-

rather than producing new labels as instructed by the input prompts. It suggested that the SFT model primarily learned a straightforward mapping from user queries to intent labels. In contrast, models trained with GRPO demonstrate significant improvements across all three test sets, maintaining over 90% accuracy on both the Unseen5 and Grouped tests. These results indicate that the GRPO model effectively learns instruction understanding and reasoning, leading to superior generalization capabilities.

In order to further validate the above findings, we conduct additional generalization testing on the MultiWoz 2.2 dataset. Specifically, we entirely exclude the data corresponding to a particular intent from the training set and then evaluate the model on the official test set, which includes both the unseen category and other categories. As illustrated in Table 2, models trained with GRPO surpass those trained with SFT by over 20% in most categories, except on the "Attraction" category where both methods yield subpar performance. These findings underscore that GRPO training improves the generalization capability for intent detection tasks.

Interestingly, when excluding an intent categories, models trained with GRPO demonstrated stronger in-domain performance than those fine-tuned through SFT - a finding that contrasts with the primary results shown in Table 1. This divergence suggests that SFT models exhibit greater sensitivity to reductions in training data diversity and sample size, while GRPO-trained models maintain more consistent robustness. Specifically, category removal leads to performance declines of 5%-17% in SFT models, whereas GRPO models maintain stable performance, with accuracy reductions remaining consistently below 2% in most cases.

To further validate the generalization capabilities of the GRPO method, we design and execute a rigorous cross-domain experiment, as summarized in Table 3. Specifically, we train a model exclusively on the MultiWOZ dataset and subsequently evaluate its zero-shot performance on the TODAssistant corpus. Notably, TODAssistant presents a distinct challenge as an artificially generated Chinese dialogue dataset, differing fundamentally from MultiWOZ in both linguistic structure (Chinese vs. English) and data provenance (machine-generated vs. human-curated). The results demonstrate that the GRPO approach maintains robustness even in such challenging cross-lingual and cross-task scenarios, thereby highlighting its superiority over models trained by SFT method.

In conclusion, our comprehensive comparative analysis across diverse test sets demonstrates that the GRPO approach (similar to R1) consistently maintains robust generalization capabilities. While SFT achieves competitive performance on in-domain evaluations, this method exhibits significant limitations in practical task-oriented dialogue scenarios, particularly when faced with dynamic adjustments to the intent schema or novel domain adaptations.

## 4.2 Results of Reward-based Curriculum Sampling

### 4.2.1 Results of Curriculum Method

To better understand the impact of our proposed Reward-based Curriculum Sampling (RCS) method, we conduct a comparative analysis against both the SFT method and the original GRPO approach, with results presented in Table 4. The first stage of our RCS method requires only 60 training steps—significantly fewer than the 150 steps needed for the original GRPO method—yet achieves comparable performance outcomes. We therefore deliberately terminate the first stage of training at 60 steps to transition to the subsequent curriculum-based phase. Notably, our proposed methodology enables the original GRPO to exceed SFT performance during the second training stage. What is particularly significant is that throughout all training phases, RCS utilizes merely 60% of the complete training dataset compared to the full dataset employed by both SFT and GRPO methods, while still

Table 4: Results of our proposed RCS method on the MultiWOZ dataset.

Model	Attraction	Hotel	Restaurant	Taxi	Train	Avg
Qwen2.5-7B-Instruct						
+ SFT	92.6	93.5	93.4	95.6	92.3	93.3
+ GRPO	94.5	91.6	91.9	93.9	92.6	93.3
+ GRPO + RCS (First Stage)	94.6	91.9	92.3	<b>96.1</b>	91.7	92.6
+ GRPO + RCS (Second Stage)	<b>96.2</b>	<b>94.8</b>	<b>94.7</b>	95.7	<b>94.6</b>	<b>96.0</b>

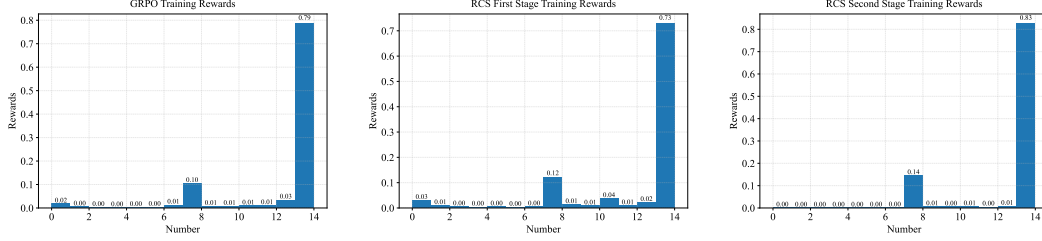


Figure 2: Histogram of rewards during the training process.

delivering superior performance. These findings suggest that easier examples within the GRPO training framework introduce redundancy, potentially hindering the model’s ability to concentrate on error-prone and more challenging cases. Our RCS method effectively addresses this limitation by prioritizing more informative training instances.

To facilitate a clearer analysis of the RCS method, we present the distribution of rewards across all training data for different methods throughout the training process in Figure 2. For each data point, we design two reward metrics and sampled seven instances, yielding a maximum possible score of 14 points per data point in the graph. The results reveal that, compared to the original GRPO method, the RCS-based GRPO training strategy increases the proportion of perfect-score examples during the second stage, even when processing more challenging data. These experimental findings demonstrate that the **Reward-based Curriculum Sampling Strategy enables the model to focus more effectively on challenging examples**, thereby enhancing overall model performance.

#### 4.2.2 Result of Positive Instance Sampling

Due to the fact that the second stage of Curriculum Learning exclusively comprises erroneous data encountered, the initial accuracy of the second stage is considerably low, showing a substantial deviation from the foundation model of the first stage. This deviation may lead to a risk of model degradation. Therefore, we randomly sample data from the whole training set as **positive instance** to combine with the curriculum samples in order to evaluate the performance.

Table 5 presents a comprehensive comparison of different ratios between positive and challenging samples in our curriculum. We first evaluate the effectiveness of training exclusively with challenging examples, addressing concerns about potential catastrophic forgetting. Contrary to conventional beliefs, our results reveal that **exclusive training with challenging data does not lead to the expected performance degradation issues**. Furthermore, we systematically vary the proportion of positive examples in the training mixture, observing a clear inverse relationship between the percentage of positive examples and the overall performance of the model. This finding strongly suggests that **the relative concentration of challenging data in the second training phase critically determines the model’s capacity to address difficult cases**. The curriculum’s effectiveness appears to depend not only on the inclusion of challenging samples but on ensuring that they constitute a substantial proportion of the training distribution, allowing sufficient learning signal for the model to improve on precisely those examples where performance gains are most valuable.



Table 5: Results on different ratios between challenging data and positive data during the sampling process.

Model	Attraction	Hotel	Restaurant	Taxi	Train	Avg
Qwen2.5-7B-Instruct + GRPO + RCS (1:2)	97.0	94.6	94.0	96.1	94.1	94.8
Qwen2.5-7B-Instruct + GRPO + RCS (1:1)	96.2	94.8	94.7	95.7	<b>94.6</b>	95.0
Qwen2.5-7B-Instruct + GRPO + RCS (2:1)	96.7	<b>95.3</b>	95.0	96.8	<b>94.6</b>	95.4
Qwen2.5-7B-Instruct + GRPO + RCS (1:0)	<b>98.2</b>	94.9	<b>96.4</b>	<b>98.6</b>	94.4	<b>96.0</b>

Table 6: Ablation results on the "Thought" during the GRPO training process.

Model	TODAssistant				MultiWOZ2.2	Avg
	in-domain	Unseen5	Subdivided	Grouped		
Qwen2.5-7B-Instruct + GRPO						
w/o think	<b>97.8</b>	86.4	72.7	<b>94.4</b>	76.1	85.5
w/ think	96.8	<b>90.6</b>	<b>83.1</b>	93.6	<b>93.3</b>	<b>91.5</b>

### 4.3 Evaluating the Effect of "Thought"

Considering that intent detection is inherently simpler than tasks like math or coding, we investigate whether incorporating thought processes during reinforcement learning (which we term "Thought"), similar to the R1 training methodology, is truly necessary. To explore this question, we remove the "Thought"-related format loss and instructions from our reinforcement learning process and observe the resulting performance changes. We conduct experiments on both datasets.

The results in Table 6 demonstrate that on the TODAssistant dataset, models without thought processes performed better on in-distribution tests, with results more closely matching those achieved after SFT. However, these models exhibit significantly reduced generalizability. However, compared to pre-trained models and SFT-trained models, their generalization ability still shows substantial improvement, **indicating that the reinforcement learning methodology itself provides inherent benefits to model generalization beyond what SFT can achieve.**

For the MultiWOZ dataset, we observe markedly different results that the performance declining considerably as thought processes are removed. We attribute this difference to the inherent characteristics of the two datasets: TODAssistant contains machine-synthesized data, resulting in statistically similar distributions between the training and testing sets. In contrast, MultiWOZ is a human-constructed dataset specifically designed to evaluate task-oriented dialogue capabilities, demanding a stronger understanding of known intents and better generalization to varied expressions.

Our analysis of model output lengths provides additional evidence for this disparity of difficulty: models trained on TODAssistant data generate responses averaging 37 tokens in length, while MultiWOZ-trained models produce significantly longer outputs, averaging 56 tokens. This quantitative difference further confirms the variation in task complexity between the datasets. Consequently, **the thought process appears more beneficial for MultiWOZ (i.e., more challenging intent detection tasks) as it helps models learn recognition logic under reinforcement learning guidance.**

### 4.4 Base Model or Instruction Model

Since intent detection requires models to have strong task comprehension and classification capabilities, it shares many similarities with function call tasks. Given that instruct models undergo extensive alignment training to better understand and differentiate tools, we are curious whether these models, which demonstrate significant performance improvements on function call tasks compared to base models, will also show superior results on intent detection tasks after RL training. Surprisingly, our findings align with observations from mathematical tasks: **the base model achieved performance comparable to the instruct model on the intent detection task**, as shown in Table 7. We present a comparison of rewards and completion lengths during the training process for both models in Figure 3a and 3b. Notably, while the base model converges more slowly, it ultimately achieves comparably strong performance. This discovery seems to confirm that model capabilities are primarily

Table 7: Results of the base model and the instruct model trained with GRPO on the MultiWOZ dataset.

Model	Attraction	Hotel	Restaurant	Taxi	Train	Avg
Qwen2.5-7B + GRPO	<b>94.98</b>	88.98	<b>92.09</b>	<b>93.91</b>	92.09	91.93
Qwen2.5-7B-Instruct + GRPO	94.46	<b>91.55</b>	91.94	<b>93.91</b>	<b>92.55</b>	<b>93.25</b>

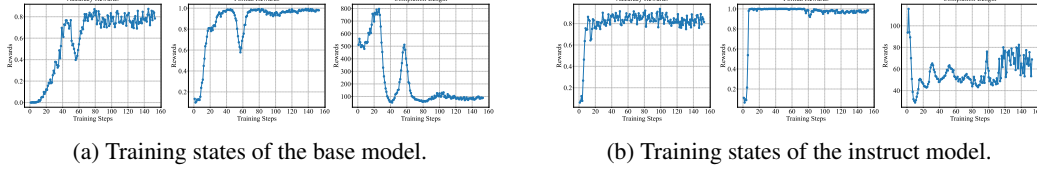


Figure 3: Training curves including the accuracy, format reward, and completion length for various models.

acquired during pre-training, with subsequent training merely helping models better utilize their inherent abilities.

To further investigate the completion lengths of the models and determine whether "aha moments" exist in this task, we reduce the learning rate and increase the training epochs for both models. Additionally, we implement two types of format rewards: 1) A strict format that rigidly restricts the output to the prescribed content, prohibiting any superfluous information; 2) A relaxed format, where the output is deemed correct as long as it encompasses the specified content. As shown in Figure 4a and 4b, the completion length of the instruct model remains constant under both reward functions. However, the base model displays an initial decrease followed by an increase in completion length under the relaxed format reward. This phenomenon is absent under the stricter format reward. Importantly, the increased length does not contribute valuable information but rather introduces task-irrelevant content. **This comparison reveals that R1-like reinforcement learning training indeed attempts to increase the length to achieve higher rewards, but true "aha moments" are less likely to emerge in relatively simple intent detection (single-task setting) tasks, as the contextual logic is limited and does not require deep reasoning from the model.**

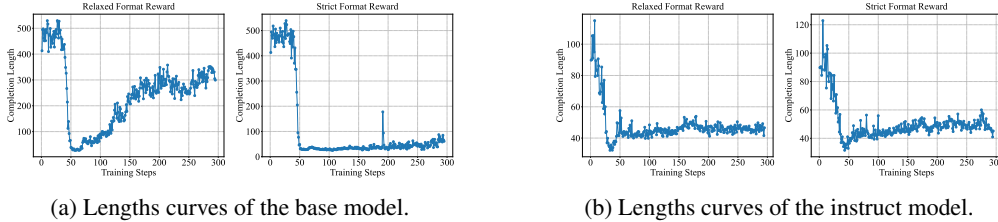


Figure 4: Comparison of completion lengths for various models with different format rewards.

## 5 Parameter Tuning Tricks

In this section, we will discuss our experimental trials with various parameters in the MultiWOZ dataset. As illustrated in Figure 5, we conduct experiments with different learning rates. The results indicate that the performance of the model first increases and then decreases as the learning rate increases, achieving optimal performance at a learning rate of  $9 \times 10^{-6}$ . To investigate whether the low learning rates contributed to the non-convergence of the model, we extend the training for an additional epoch. We observe that increasing the epochs does not improve performance, which demonstrates that one epoch is sufficient for convergence on the intent detection task.

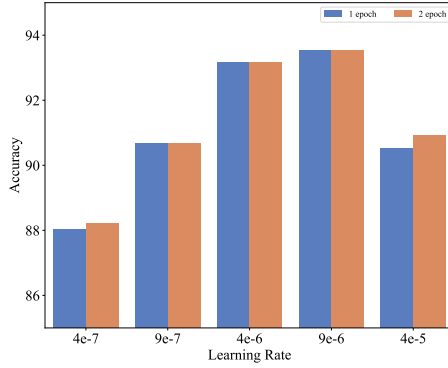


Figure 5: Figure of the accuracy variation with different parameters.

## 6 Conclusion

In this work, to enhance the model’s ability to adapt to complex and dynamic scenarios, we apply reinforcement learning to the intent detection task using the GRPO method. We introduce a Reward-based Curriculum Sampling (RCS) method, which leverages the reward function of the GRPO method during the training process to select data of varying difficulty levels. We conduct the curriculum learning approach and sample more challenging data in the second phase. In this way, the model is able to continuously focus on data it does not yet understand, thereby improving its performance and outperforming the SFT method. Furthermore, we empirically demonstrate that the RL-based model exhibits superior generalization capabilities on both in-domain and out-of-domain data. Moreover, we also disclose some interesting findings and share insights regarding parameter tuning encountered during our experimental process.

## 7 Next Step

Moving forward, we intend to channel our research efforts into the following areas:

- 1) At present, the Reward-based Curriculum Sampling (RCS) we employ is offline. In the future, we plan to transition to an online RCS, which will allow for more efficient selection of superior samples.
- 2) We aspire to shift our focus from single-intent detection tasks to addressing multi-intent detection tasks, which will significantly improve our capacity to deal with the intricacies of dialogue tasks found in real-world situations.
- 3) In addition to intent detection tasks, we are set to explore the utilization of reinforcement learning within other facets of Task-Oriented Dialogue (TOD) systems, including but not limited to Dialogue Policy and Response Generation.
- 4) We are committed to further investigating the deep-seated reasons behind the "aha moment" phenomenon, to augment the task-oriented dialogue model’s abilities in self-reflection, self-correction, and self-direction.

## References

- [1] Aman Gupta, Anirudh Ravichandran, Ziji Zhang, Swair Shah, Anurag Beniwal, and Narayanan Sadagopan. Dard: A multi-agent approach for task-oriented dialog systems. *arXiv preprint arXiv:2411.00427*, 2024.
- [2] Heng-Da Xu, Xian-Ling Mao, Puhai Yang, Fanshu Sun, and He-Yan Huang. Rethinking task-oriented dialogue systems: From complex modularity to zero-shot autonomous agent. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2748–2763, 2024.

- [3] Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. A survey of joint intent detection and slot filling models in natural language understanding. *ACM Computing Surveys*, 55(8):1–38, 2022.
- [4] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*, 2020.
- [5] Yu Du, Fangyun Wei, and Hongyang Zhang. Anytool: Self-reflective, hierarchical agents for large-scale api calls. *arXiv preprint arXiv:2402.04253*, 2024.
- [6] Kunyang Qu and Xuande Wu. Chatgpt as a call tool in language education: A study of hedonic motivation adoption models in english learning environments. *Education and Information Technologies*, pages 1–33, 2024.
- [7] A. B. Siddique, Fuad T. Jamour, Luxun Xu, and Vagelis Hristidis. Generalized zero-shot intent detection via commonsense knowledge. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1925–1929. ACM, 2021.
- [8] Daniele Comi, Dimitrios Christofidellis, Pier Francesco Piazza, and Matteo Manica. Zero-shot-bert-adapters: a zero-shot pipeline for unknown intent detection. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 650–663. Association for Computational Linguistics, 2023.
- [9] Soham Parikh, Quaizar Vohra, Prashil Tumbade, and Mitul Tiwari. Exploring zero and few-shot techniques for intent classification. *arXiv preprint arXiv:2305.07157*, 2023.
- [10] Gokul Swamy, Sanjiban Choudhury, Wen Sun, Zhiwei Steven Wu, and J. Andrew Bagnell. All roads lead to likelihood: The value of reinforcement learning in fine-tuning, 2025.
- [11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [12] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [13] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [14] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *11th International Conference on Learning Representations, ICLR 2023*, 2023.
- [15] Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason D Lee, and Sanjeev Arora. What makes a reward model a good teacher? an optimization perspective. *arXiv preprint arXiv:2503.15477*, 2025.
- [16] Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. *arXiv preprint arXiv:2007.12720*, 2020.
- [17] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [18] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.