

# MindFlow+: A Self-Evolving Agent for E-Commerce Customer Service

Ming Gong<sup>1,2</sup>, Xucheng Huang<sup>1</sup>, Ziheng Xu<sup>1</sup>, Vijayan K. Asari<sup>2</sup>

<sup>1</sup>Xiaoduo AI Lab, Shanghai, China

<sup>2</sup>University of Dayton, Dayton, Ohio, United States

Corresponding author: Ming Gong (e-mail: gongm1@udayton.edu).

arXiv:2507.18884v1 [cs.CL] 25 Jul 2025

**ABSTRACT** High-quality dialogue is crucial for e-commerce customer service, yet traditional intent-based systems struggle with dynamic, multi-turn interactions. We present MindFlow+, a self-evolving dialogue agent that learns domain-specific behavior by combining large language models (LLMs) with imitation learning and offline reinforcement learning (RL). MindFlow+ introduces two data-centric mechanisms to guide learning: tool-augmented demonstration construction, which exposes the model to knowledge-enhanced and agentic (ReAct-style) interactions for effective tool use; and reward-conditioned data modeling, which aligns responses with task-specific goals using reward signals. To evaluate the model's role in response generation, we introduce the AI Contribution Ratio, a novel metric quantifying AI involvement in dialogue. Experiments on real-world e-commerce conversations show that MindFlow+ outperforms strong baselines in contextual relevance, flexibility, and task accuracy. These results demonstrate the potential of combining LLMs, tool reasoning, and reward-guided learning to build domain-specialized, context-aware dialogue systems.

**INDEX TERMS** AI Contribution Ratio, E-Commerce Customer Service, Imitation Learning, Offline Reinforcement Learning

## I. INTRODUCTION

The rapid growth of e-commerce has made it an essential part of modern life. As the industry evolves, customer service emerges as a critical competitive factor, directly influencing user satisfaction and business outcomes [1]. To scale support efficiently, many businesses are turning to automation [2], [3]. However, traditional NLP-based systems, centered on intent recognition and template-based responses, struggle to handle the diverse, dynamic, and often ambiguous nature of real-world e-commerce queries.

In parallel, Large Language Models (LLMs) have demonstrated impressive capabilities across a wide range of tasks [4]. However, their direct application in e-commerce remains limited. Unlike static knowledge retrieval or open-domain chat, e-commerce scenarios demand domain-specific understanding, real-time adaptability, factual grounding, and the ability to make multi-step decisions based on evolving user intent. These requirements exceed the capabilities of general-purpose dialogue systems [5].

To bridge this gap, we introduce MindFlow+, a self-evolving agent specifically designed for e-commerce dialogue automation. Unlike traditional intent-based systems, MindFlow+ is capable of adapting to diverse customer needs through dynamic behavior learning.

MindFlow+ learns domain-specific behavior through supervised fine-tuning (SFT) on interaction data enriched with tool calls, external knowledge, and reward-guided supervision. This approach allows the agent to internalize task-specific patterns, reason over external tools, and generate responses that are both contextually grounded and aligned with human preferences. By unifying tool-use reasoning with preference-aligned generation in a single training framework, MindFlow+ achieves adaptive behavior without requiring architectural modifications to the base LLM.

The primary contributions of this research are summarized as follows:

- 1) **MindFlow+**, a self-evolving agent framework for e-commerce dialogue automation that performs SFT on integrated interaction data enriched with tool calls, external knowledge, and reward-guided supervision, unifying tool-augmented reasoning and preference-aligned response generation into a single training process, enabling adaptive and context-aware responses without modifying the underlying large language model architecture.
- 2) **AI Contribution Ratio**, a novel metric to quantify the agent's autonomy in multi-step workflows. It captures both decision-making efficiency and the model's

practical contribution to real-world tasks, offering fine-grained insight into agent capability.

- 3) **Expansion to Broader Applications in LLM-based Dialogue Systems**, a versatile approach that extends MindFlow+'s adaptability and contextual understanding beyond e-commerce, enabling improved user interactions in diverse domains such as virtual assistants, help desks, and other task-oriented conversational AI systems.

## II. RELATED WORK

### A. NLP IN E-COMMERCE CUSTOMER SERVICE

In the early stages of e-commerce customer service, question-answering (QA) systems were based on rule-based methods, like predefined FAQs and keyword matching [6], [7]. Although these methods offered basic automation, they struggled with complex issues and multi-turn dialogues, limiting their adaptability.

With advancements in deep learning and NLP, rule-based systems have been largely replaced by retrieval- and generation-based models, improving dialogue quality and accuracy [8]–[10]. Modern systems use diverse data sources such as product reviews and past QA pairs to improve answer generation [11], [12], with knowledge graphs further enhancing retrieval accuracy [13], [14]. Scalable solutions that utilize user-generated content also help address repetitive queries [15], [16]. Despite these advancements, challenges remain in addressing complex, context-dependent queries and managing multi-turn interactions.

### B. LLMS IN E-COMMERCE CUSTOMER SERVICE

LLMs are increasingly central to e-commerce customer interactions. Trained on vast datasets, they excel in generating personalized responses in multi-turn conversations [17]–[19]. Abundant studies have been conducted such that interaction optimization by categorizing queries [20], memory architectures enhancement for better conversational continuity [21], and utilizing customer profiles for tailored responses through services like CHOPS [22]. Open-source frameworks like LangChain are transforming customer service from static FAQs to dynamic, context-aware systems [23]. LLMs' ability to adapt via few-shot learning makes them highly effective in fast-evolving e-commerce environments. However, issues like hallucinations and domain adaptation challenges remain, requiring ongoing optimization.

### C. LLM-BASED RL AGENT

LLM-based agents have become integral to modern intelligent systems, utilizing advanced language generation and understanding to manage complex, multi-turn dialogues and provide personalized responses. In fields like e-commerce, they analyze customer queries to generate accurate answers and optimize interactions. Advanced prompting techniques, such as Chain-of-Thought (CoT) [24] and ReAct [25], enhance LLMs by enabling step-by-step reasoning and task-specific actions. A key advancement is tool use, where models

integrate external tools to access real-time information and perform actions beyond their language capabilities. This allows the agents to tackle more complex tasks, improve their decision-making, and broaden their applicability, helping to address challenges like hallucinations and domain adaptation that require ongoing optimization [26]–[29]. While these advancements have significantly enhanced LLM-based agents, further optimization can be achieved by integrating RL techniques, which allow agents to improve performance through continuous learning and feedback.

RL enhances LLM-based agents by allowing them to optimize decision-making and improve task performance over time through feedback [30]. In online RL, agents continuously adjust their strategies based on real-time interactions with the environment, whereas offline RL leverages pre-collected data for optimization, refining agent's behavior without the need for real-time interaction [31]–[33]. This integration provides LLM-based agents with greater adaptability, enabling them to tackle complex tasks more effectively [34]–[37]. A promising direction in RL research is framing the problem as sequence modeling, which offers new opportunities for improving both the efficiency and scalability of LLM-based agents. One notable approach is Decision Transformer [38], which utilizes the simplicity and scalability of Transformer architecture to treat RL as conditional sequence modeling. Unlike traditional RL methods that rely on value functions or policy gradients, Decision Transformer predicts optimal actions by conditioning on the desired return (reward), past states, and actions using a causally masked Transformer. This autoregressive approach allows the model to generate actions that align with the specified return, providing a novel perspective on RL as a sequence prediction problem. Based on the definition from [39], MindFlow+ is motivated by world simulation and aims to simulate customer representatives' behavior in the e-commerce customer service domain. The system is structured around an LLM-powered agent with the ability to use external tools and interact with the environment, enabling context-aware, adaptive responses. Agent profiling generated is achieved through both pre-defined and data-derived methods, with a focus on simulating the behavior of customer representatives. Capabilities are continuously refined through feedback from both humans and the environment, with memory-based adjustments that allow the agent to evolve and improve over time.

Another key approach to enhance the performance of LLM-based agents is through Human-in-the-loop RL (HIRL), which integrates human feedback during the learning process. When combined with Human-in-the-loop RL (HIRL), LLM-based agents' adaptability is further enhanced [40]–[46]. HIRL incorporates human feedback during the learning process, allowing LLM-based agents to optimize their behavior in real-world applications, where both real-time and offline learning can be utilized to fine-tune (FT) responses [47]. This combination is particularly beneficial for domains that require continuous interaction or large-scale datasets, offering significant improvements in handling dy-

dynamic environments and complex tasks [48].

### III. MINDFLOW+

Pre-trained LLMs, while strong in general-purpose tasks, often fall short in delivering context-aware and domain-aligned responses for complex multi-turn dialogue QA. To address this, we introduce MindFlow+, a self-evolving dialogue agent that continuously improves its domain-specific behavior through SFT on a reward-guided corpus. By learning from curated examples that reflect task-specific quality standards and user preferences, MindFlow+ becomes increasingly aligned with practical needs in e-commerce dialogue systems.

Here, we introduce a framework designed to align LLMs with domain-specific behaviors in multi-turn QA settings. Our approach integrates two complementary, data-centric strategies: imitation learning through expert-like demonstrations augmented with knowledge-enhanced demonstrations and ReAct-based agentic demonstrations, empowering the model with factual knowledge and tool-use capabilities; and reward-based supervision, in which user preferences are explicitly annotated using specialized tokens to highlight desirable model responses.

Rather than treating these strategies independently, we unify them at the data construction level to form a reward-aware, tool-augmented training corpus. This hybrid dataset encodes both expert-level task knowledge and preference-aligned behavior signals, enabling the model to learn interaction patterns that reflect both correct domain behavior and downstream feedback. Figure 1 illustrates the full pipeline of this domain-specific training data construction.

Through this integrated training process, the resulting model acquires three key capabilities: structured domain knowledge, contextualized tool usage, and fine-grained preference alignment. These enhancements enable even relatively compact models to outperform larger general-purpose counterparts in specialized domains, delivering higher accuracy, faster inference, and reduced computational cost in practical applications.

#### A. TOOL-AUGMENTED DEMONSTRATION CONSTRUCTION

To equip the model with expert-level reasoning and task execution capabilities, we construct imitation data through two complementary augmentation strategies: knowledge-enhanced demonstrations and ReAct-based agentic demonstrations. These strategies enrich raw dialogue data by injecting factual context and demonstrating structured tool-use reasoning.

##### 1) Knowledge-Augmented Demonstrations

Rather than performing retrieval at inference time as in traditional retrieval augmented generation (RAG) systems [49], we take a more controlled, training-centric approach. For each multi-turn dialogue instance, relevant background information is heuristically selected from internal knowledge bases and inserted into the system message at the beginning of the

conversation. This static context grounding ensures that the model has access to accurate and task-relevant information during response generation.

By conditioning the model on high-quality, domain-specific knowledge without requiring an active retriever component, this approach mitigates hallucination and improves factual consistency, especially in scenarios with stable or repeatable task patterns.

##### 2) Agentic Demonstrations via ReAct

To further enhance the model's reasoning capabilities and tool-use behaviors, we simulate agent-style decision-making using ReAct prompting [25]. Specifically, for each user query, the assistant's response is structured as a sequence of alternating steps. First, the model formulates a *Thought* by interpreting the user's intent based on the current context and planning how to address the problem. Next, guided by this *Thought*, it takes an *Action* by selecting a specific operation such as issuing a query or invoking an external tool like search, compare, or recommend. Finally, the model receives an *Observation* which is the outcome of the executed action and uses this information for subsequent reasoning and decision-making.

Algorithm 1 illustrates the overall ReAct prompting workflow, where the assistant iteratively alternates between internal reasoning and interactions with external tools in a closed-loop process.

These sequences are embedded directly into the assistant's turn, allowing the model to observe multiple iterations of decision-making within a single example. Demonstrations are generated via few-shot prompting with open-source LLMs and post-processed for coherence and correctness.

This format equips the model with the ability to emulate real-world workflows: deciding when to use a tool, selecting the appropriate tool for the query, interpreting the returned information, and integrating it into a final response.

Together, these two augmentation strategies form the foundation of our tool-augmented training data. The former exposes the model to grounded factual context, while the latter teaches multi-step decision-making and tool invocation. Importantly, all demonstrations are constructed offline and do not require runtime tool execution, making them suitable for SFT.

#### B. REWARD-CONDITIONED DATA MODELING

To enable preference-aware generation without relying on costly online exploration, we adopt a reward-conditioned modeling approach inspired by offline RL. Unlike traditional RL methods requiring environment interaction to estimate reward signals or update policies, offline RL leverages static, pre-collected datasets annotated with feedback signals. This provides a stable and scalable alternative for aligning LLMs to domain-specific or task-specific goals.

We formally model dialogue generation as a Markov Decision Process (MDP) with the tuple  $(\mathcal{S}, \mathcal{A}, P, \mathcal{R})$ , where states  $s \in \mathcal{S}$  correspond to user inputs, actions  $a \in \mathcal{A}$  are assistant responses, and  $\mathcal{R}(s, a)$  denotes the reward function.

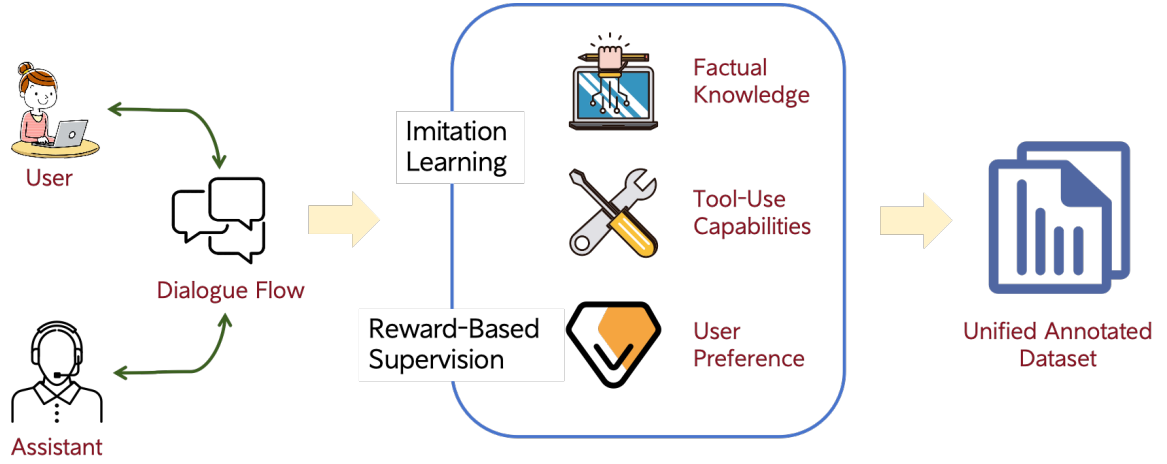


FIGURE 1. Domain-Specific Training Data Construction Pipeline

**Algorithm 1** Tool-Augmented Reasoning Loop**Input:** Query *query*, History Msgs *hist\_msgs*, Toolset *tools*, Model *model***Output:** Final Response *response*

```

1: Initialize observation  $\leftarrow \emptyset$ 
2: Initialize actionprev  $\leftarrow \emptyset$ 
3: while True do
4:   thought  $\leftarrow$  GenerateThought(model, query, hist_msgs, observation, tools)
5:   action  $\leftarrow$  ExtractAction(thought)
6:   if name(action) = finish then
7:     break
8:   end if
9:   if action = actionprev then
10:    break
11:  end if
12:  observation  $\leftarrow$  Execute(action)
13:  actionprev  $\leftarrow$  action
14: end while
15: response  $\leftarrow$  GenerateResponse(model, query, hist_msgs, observation)
16: return response

```

To capture temporal dependencies of multi-turn dialogues, we represent dialogues as interleaved sequences of states, rewards, and actions at each timestep  $t$ , arranged in a "state-reward-action" order that reflects the causal structure of conversations:

$$\tau = (s_1, r_1, a_1, s_2, r_2, a_2, \dots, s_t, r_t, a_t) \quad (1)$$

This design allows the model to learn how its actions influence subsequent responses and feedback over time.

Inspired by the Decision Transformer (DT) framework [38], which formulates RL as an autoregressive sequence modeling problem, we adapt this paradigm for dialogue modeling as shown in Figure 2. Unlike typical DT applications relying on numeric embeddings, we represent both states and actions as natural language token sequences, enabling seamless integration with LLM pretraining objectives.

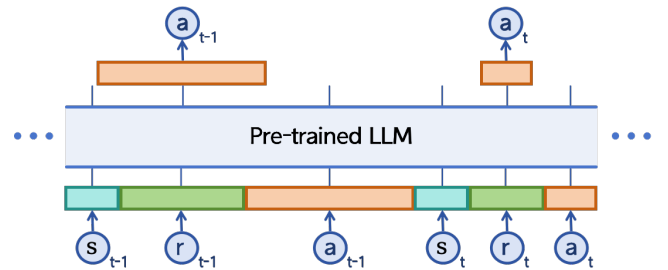


FIGURE 2. Adaptation of Decision Transformer for Reward-Conditioned Dialogue Modeling

To support learning, we construct a reward-annotated dataset where each user query is paired with multiple assistant responses labeled with scalar feedback (e.g., 1 for preferred, 0 for suboptimal). These reward signals are derived from human preferences or automated metrics assessing relevance,



factuality, and coherence.

Reward tokens are added to assistant responses as part of the input context, enabling the model to learn preference alignment without modifying the loss function. This offline RL-style setup offers a stable training alternative to traditional policy optimization, supporting context-aware generation in complex multi-turn dialogue tasks.

### C. SFT ON TOOL-AUGMENTED AND REWARD-ALIGNED DEMONSTRATIONS

SFT provides a stable and scalable framework for adapting pre-trained LLMs to downstream tasks [50]. In our approach, SFT is performed over a unified training corpus that integrates both tool-augmented reasoning traces and reward-conditioned responses. This hybrid data structure enables the model to simultaneously learn structured decision-making and preference-aligned generation.

Building upon the previously introduced "state–reward–action" sequence representation, the model is trained to predict the next action given prior states, rewards, and actions by minimizing the standard cross-entropy loss:

$$\mathcal{L}_{\text{SFT}} = - \sum_t \log P(a_t \mid s_{\leq t}, r_{\leq t}, a_{<t}) \quad (2)$$

This formulation treats the reward signal as part of the input context, injected via special tokens such as `<reward>`, rather than modifying the loss function itself. As a result, the model learns to condition its generation on reward levels while preserving the simplicity and generalization benefits of supervised learning.

The inclusion of both positive and negative responses within the same dialogue context encourages contrastive behavior learning, enabling the model to distinguish high-quality outputs from suboptimal ones. When this training is combined with tool-augmented demonstrations that incorporate structured reasoning and action-observation feedback, the unified SFT process empowers the model to ground its responses in external knowledge through static context or tool observations, perform multi-step, agentic reasoning aligned with user intent, and adapt output quality based on explicit reward feedback.

Overall, this approach bridges long-horizon planning and preference alignment within a single training stage, allowing even moderately sized LLMs to exhibit expert-like behavior in complex, domain-specific QA settings.

## IV. EXPERIMENTS AND EVALUATION

### A. SFT

We fine-tune models from the Qwen2.5 series, focusing on those with fewer than 10 billion parameters, to adapt them for complex, multi-turn dialogues in the professional e-commerce domain. These models provide a strong foundation with established language understanding and generation capabilities, enabling efficient SFT tailored to domain-specific tasks.

The SFT process is conducted on 8 NVIDIA A800 GPUs (each with 80GB memory). The maximum input sequence length is set to 4096 tokens. We use a learning rate of  $2 \times 10^{-5}$ , train for 4 epochs, and adopt an effective batch size of 256.

To better equip the model for practical e-commerce scenarios, we incorporate four domain-specific tools into the dialogue system during training and evaluation:

- **Product Information Retrieval Tool:** retrieves detailed specifications and descriptions for queried products.
- **Product Recommendation Tool:** suggests relevant products based on user preferences and contextual information.
- **Product Comparison Tool:** provides side-by-side comparisons of multiple products to assist user decision-making.
- **Image Descriptor Tool:** analyzes dialogue-shared images to extract key visual features, enabling enhanced multimodal understanding.

These tools simulate realistic functionalities commonly used in e-commerce customer service and help the model learn to generate tool-augmented responses, improving its practical utility in multi-turn dialogues.

### B. EVALUATION

#### 1) Evaluation Metric

We define the framework for intelligent customer service and highlight the characteristics and application scenarios of each level outlined in the Table. 1. X1, representing the traditional AI level, features AI that primarily assists human agents, requiring frequent intervention and significant configuration expertise.

To quantify AI's contribution to the customer service process, we introduce the AI Contribution Ratio, calculated as follows:

$$\text{AI Contribution Ratio} = \frac{V_{\text{AI}}}{T_{\text{AI}} + T_{\text{CR}}} \quad (3)$$

where  $V_{\text{AI}}$  denotes the number of AI-generated messages judged by customer representatives to be contextually appropriate and relevant within the dialogue.  $T_{\text{AI}}$  denotes the total number of AI-generated responses, and  $T_{\text{CR}}$  denotes the total number of responses from customer representatives.

The calculation is performed within a defined time window during which messages are monitored. This metric provides a clear indication of AI's participation and automation level across various collaboration models, offering insights into how AI's role evolves from X1 to higher levels. Unlike traditional evaluation metrics such as accuracy or F1 score, the AI Contribution Ratio evaluates the model's specific contribution to user decision-making and task completion in real-world dialogues. It considers not only the accuracy of the generated content but also its practical value in the e-commerce context. By assessing performance across multi-turn dialogues, this metric measures adaptability and collaborative effectiveness, providing deeper insights into the model's performance in e-commerce question-answering tasks.

Rating	Automation Level	Collaboration Models	AI Proficiency Levels
X0	No Automation	Human-Exclusive	-
X1	Robot Assistance	AI-Assisted, Human-Centric	Entry-Level
X2	Partial Automation	AI-Centric, Human-Assisted	Intermediate; Advanced in Some Scenarios
X3	Conditional Automation	AI Primarily Independent, with Human Supervision for Complex Scenarios	Advanced; Human-Machine Integration at Expert Level
X4	High Automation	AI Fully Independent, with Human Involvement in Training	Expert-Level
X5	Full Automation	AI Autonomous Training and Operation	Beyond Expert-Level

**TABLE 1. Taxonomy of Intelligent Customer Service Standards**

To efficiently evaluate model performance, we automate the scoring process using a preset prompt that incorporates customer inquiries, customer representatives' responses, and MindFlow+ outputs (which are the AI-generated messages). We invoke an LLM API to score these outputs (denoted as 0 for Invalid AI Messages and 1 for Valid AI Messages). To ensure alignment with manual scoring, we label a small test set and calculate the Spearman's correlation between manual and API scores. This correlation is used to iteratively refine the prompt, ensuring statistical significance between API and manual scores [51].

## 2) Experimental Setting

We conduct our evaluation on 150 real-world user queries collected from an e-commerce customer service platform. The conversations originate from a store selling products in a representative category (e.g., consumer electronics), covering diverse scenarios such as parameter inquiries, setup assistance, greetings, product recommendation, product comparison, multimodal understanding, and multi-turn dialogues that often require deeper user intent interpretation. The main challenge lies in the complexity and diversity of the queries, many of which require reasoning over background knowledge, identifying user needs over multiple turns, and generating helpful, context-aware responses.

## 3) Experimental Results

Table 2 summarizes the evaluation results. Baseline model is trained on dialogues augmented with relevant background information, such as store descriptions and product names. This information encodes key product attributes and enhances the informativeness of each turn. As a result, the model achieves reasonable performance even without task-specific alignment.

The Tool-Augmented model is trained to follow chain-of-thoughts and invoke tools appropriately. However, the original dataset's answer annotations often mimic realistic service responses containing delays or vague phrases (e.g., "please wait", "I can't help with that"), leading to poor response quality despite learned tool-calling behaviors.

The Reward-Guided model is trained with both positive and negative response samples. Although it does not invoke tools, it has access to the same background information as the baseline, and learns to generate responses aligned with expert-level behavior, achieving notable improvements.

MindFlow+ model combines both reward-guided and tool-augmented signals. It learns to both make appropriate tool calls and generate effective, preference-aligned responses. This dual ability leads to superior performance.

## C. ABLATION STUDY AND DISCUSSION

In this section, we conduct a series of ablation experiments to examine the effects of different modeling choices on the AI contribution ratio. Specifically, we analyze the impact of the reward token position and compare multiple pretraining strategies.

### 1) Effect of Reward Token Position

We adopt a standardized reward token format denoted as `<reward>score</reward>` for all reward-aligned demonstrations. To investigate the impact of reward token placement, we compare two sequence configurations:  $\{s, r, a\}$  (state  $\rightarrow$  reward  $\rightarrow$  action) and  $\{r, s, a\}$  (reward  $\rightarrow$  state  $\rightarrow$  action). The results are summarized in Table 3.

Unlike Decision Transformers that operate over numeric trajectories, our LLM-based model is sensitive to sequence order. The above results suggest that the  $\{s, r, a\}$  configuration yields better alignment, likely due to the proximity of state and action tokens reinforcing reward semantics via the attention mechanism.

### 2) Comparison of Pretraining Strategies

We compare three pretraining setups: (1) Baseline, which includes background information such as product descriptions and store attributes; (2) Baseline w. Tool, which introduces explicit tool usage prompts in the input to test prompt-based tool reasoning capabilities; and (3) Background-Free, which excludes background knowledge but retains user and assistant roles.

Pretrained Models	Baseline	Tool-Augmented	Reward-Guided	MindFlow+
Qwen2.5-0.5B-Instruct	35.33%	4.00%	66.67%	<b>72.67%</b>
Qwen2.5-1.5B-Instruct	44.00%	4.67%	61.33%	<b>85.33%</b>
Qwen2.5-3B-Instruct	56.00%	6.67%	57.33%	<b>87.33%</b>
Qwen2.5-7B-Instruct	58.00%	14.00%	72.67%	<b>94.00%</b>

**TABLE 2.** Evaluation Results on AI Contribution Ratio

Pretrained Models	$\{s, r, a\}$ order	$\{r, s, a\}$ order
Qwen2.5-0.5B-Instruct	<b>66.67%</b>	33.33%
Qwen2.5-1.5B-Instruct	<b>61.33%</b>	15.33%
Qwen2.5-3B-Instruct	<b>57.33%</b>	52.00%
Qwen2.5-7B-Instruct	<b>72.67%</b>	48.00%

**TABLE 3.** Comparison of Reward Token Position

As shown in Table 4, the Baseline model performs reasonably well, supported by accessible background features. The Background-Free setting highlights the performance drop when factual context is unavailable, revealing the importance of knowledge-augmented demonstrations. Meanwhile, the Tool-Prompt baseline shows limited improvement, indicating that pretraining alone is insufficient for tool learning without supervision or reward guidance.

#### D. CROSS-DOMAIN GENERALIZATION

To evaluate the generalizability of our proposed model trained on the unified-annotated dataset, we apply it directly to a new domain, a store that sells household essentials, without any additional training, prompt engineering, or domain-specific adaptation.

We randomly select 30 real-world conversations from this new store for evaluation. As shown in Table 5, the AI contribution ratio remains consistently high across different model scales, demonstrating the robustness and reusability of our method in unseen scenarios.

Overall, these results demonstrate that our method is both robust and reusable, with the model exhibiting strong generalization capabilities in unseen domains. This indicates that its behavior can effectively transfer to structurally similar customer service environments with minimal adaptation effort.

#### V. CONCLUSION

We present MindFlow+, a self-evolving agent framework tailored for e-commerce customer service. MindFlow+ acquires domain-specific capabilities through SFT on interaction traces enriched with tool calls, external knowledge, and reward-guided, preference-aligned supervision at the token level. These enriched signals support alignment with task-specific goals and enable the agent to dynamically interact

with external tools and incorporate domain-specific knowledge in a controllable and interpretable manner.

To assess real-world performance, we introduce the AI Contribution Ratio metric to quantify agent autonomy in tool-assisted workflows. Extensive experiments on real-world customer service data from a large-scale e-commerce platform validate the effectiveness of MindFlow+.

While MindFlow+ demonstrates strong alignment and robustness, limitations remain. The current pipeline is tailored to specific toolsets and domains, and future work will explore generalization across verticals, including low-resource or zero-shot settings. We also plan to extend MindFlow+ to high-stakes applications such as healthcare and finance, where safety, reliability, and alignment are even more critical.

#### APPENDIX A DATA FORMATTING

##### 1) Original Paired QA

Standard user–assistant QA pairs

```
{ "role": "user", "content": "query" }
{ "role": "assistant", "content": "
  ↪ original answer" }
```

##### 2) Knowledge-Augmented Demonstrations (Baseline)

Dialogues augmented with relevant background context

```
{ "role": "system", "content": "
  ↪ conversation background" }
{ "role": "user", "content": "query" }
{ "role": "assistant", "content": "
  ↪ original answer" }
```

##### 3) Agentic Demonstrations via ReAct

Baseline with reasoning and action steps in assistant responses

```
{ "role": "system", "content": "
  ↪ conversation background" }
{ "role": "user", "content": "query" }
```

Pretrained Models	Baseline	Baseline w. Tool	Background-Free
Qwen2.5-0.5B-Instruct	<b>35.33%</b>	26.00%	28.67%
Qwen2.5-1.5B-Instruct	<b>44.00%</b>	38.00%	25.33%
Qwen2.5-3B-Instruct	<b>56.00%</b>	36.67%	28.00%
Qwen2.5-7B-Instruct	<b>58.00%</b>	28.00%	37.77%

**TABLE 4.** Effect of Pretraining Strategies

Pretrained Models	AI Contribution Ratio
Qwen2.5-0.5B-Instruct	73.33%
Qwen2.5-1.5B-Instruct	80.00%
Qwen2.5-3B-Instruct	86.67%
Qwen2.5-7B-Instruct	90.00%

**TABLE 5.** AI Contribution Ratio On A New Domain Without Any Further Tuning

```

{"role": "assistant", "content": "<
  ↳ thought>...</thought><action
  ↳ >...</action><observation
  ↳ >...</observation><answer>
  ↳ original answer</answer>"}

```

**4) Reward-Conditioned Data Modeling**

Baseline with multiple responses per query annotated with scalar rewards

```

{"role": "system", "content": "
  ↳ conversation background"}
{"role": "user", "content": "query"}
{"role": "assistant", "content": "<
  ↳ reward>1</reward>preferred
  ↳ answer"}
{"role": "assistant", "content": "<
  ↳ reward>0</reward>suboptimal
  ↳ answer"}

```

**5) Integrated Reasoning-Reward Demonstrations**

Baseline combining reasoning traces and reward annotations

```

{"role": "system", "content": "
  ↳ conversation background"}
{"role": "user", "content": "query"}
{"role": "assistant", "content": "<
  ↳ thought>...</thought><action
  ↳ >...</action><observation
  ↳ >...</observation><reward>1</
  ↳ reward>preferred answer"}
{"role": "assistant", "content": "<
  ↳ thought>...</thought><action

```

```

↳ >...</action><observation
↳ >...</observation><reward>0</
↳ reward>suboptimal answer"}

```

**APPENDIX B  
CHAT TEMPLATE**

To ensure compatibility with the pretrained tokenizer and conversation prompt design, we adopt the ChatML format. Each message is wrapped with `<|im_start|>` and `<|im_end|>` tags, which clearly delineate conversational turns. This format preserves role-based structure and supports variable-length reasoning blocks.

The following example illustrates how a tool-augmented, reward-scored assistant response is formatted within the ChatML framework. It presents a hypothetical academic research assistant scenario where the model helps users find relevant scientific literature, explains complex concepts, and provides thoughtful, context-aware answers. The dialogue showcases the integration of external knowledge from the system prompt and the use of the ReAct paradigm for reasoning and tool-based actions during multi-turn interactions. Additionally, the model generates responses aligned with predefined reward scores, ensuring high-quality and usable outputs.

```
<|im_start|>system
```

You are an expert academic research assistant, skilled in helping users find relevant scientific papers, explaining complex concepts, and providing thoughtful answers. Below is the background information for this session: ...

```
<|im_end|>
```

```
<|im_start|>user
```

Can you help me find papers about graph neural networks (GNNs) applied in chemistry?

```
<|im_end|>
```

```
<|im_start|>assistant
```

<thought> The user is requesting scientific literature on GNNs in chemistry. I should search relevant databases. </thought>

```
<action><tool_call>"name":"Search",
"arguments": "query": "graph neural networks"
```



```
chemistry"</tool_call></action>
<observation> Found several papers, including
"Molecular Property Prediction with Graph Neural Net-
works". </observation>
<reward>1</reward> Sure! One relevant paper
is "Molecular Property Prediction with Graph Neu-
ral Networks". It demonstrates how GNNs can predict
molecular properties effectively.
<|im_end|>
```

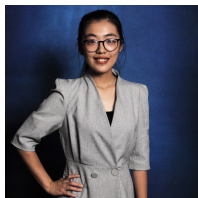
The integration of ChatML and structured special tokens equips the model with the ability to perform multi-step reasoning, invoke tools, and align responses with reward signals, all within a unified fine-tuning framework.

## REFERENCES

- [1] T. Gajewska, D. Zimon, G. Kaczor, and P. Madziński, "The impact of the level of customer satisfaction on the quality of e-commerce services," *International Journal of Productivity and Performance Management*, vol. 69, no. 4, pp. 666–684, 2020.
- [2] R. Chaturvedi and S. Verma, "Opportunities and challenges of ai-driven customer service," *Artificial Intelligence in customer service: The next frontier for personalized engagement*, pp. 33–71, 2023.
- [3] Q. Ren, Z. Jiang, J. Cao, S. Li, C. Li, Y. Liu, S. Huo, T. He, and Y. Chen, "A survey on fairness of large language models in e-commerce: progress, application, and challenge," *arXiv preprint arXiv:2405.13025*, 2024.
- [4] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, "A survey of large language models," 2024. [Online]. Available: <https://arxiv.org/abs/2303.18223>
- [5] S. Bamberger, N. Clark, S. Ramachandran, and V. Sokolova, "How generative ai is already transforming customer service," *Boston Consulting Group*, 2023.
- [6] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.
- [7] N. Malik and M. Bilal, "Natural language processing for analyzing online customer reviews: A survey, taxonomy, and open research challenges," *PeerJ Computer Science*, vol. 10, p. e2203, 2024.
- [8] M. Mashaabi, A. Alotaibi, H. Qudaih, R. Alnashwan, and H. Al-Khalifa, "Natural language processing in customer service: a systematic review," *arXiv preprint arXiv:2212.09523*, 2022.
- [9] P. A. Olujimi and A. Ade-Ibija, "Nlp techniques for automating responses to customer queries: a systematic review," *Discover Artificial Intelligence*, vol. 3, no. 1, p. 20, 2023.
- [10] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia tools and applications*, vol. 82, no. 3, pp. 3713–3744, 2023.
- [11] S. Gao, X. Chen, Z. Ren, D. Zhao, and R. Yan, "Meaningful answer generation of e-commerce question-answering," *ACM Transactions on Information Systems (TOIS)*, vol. 39, no. 2, pp. 1–26, 2021.
- [12] Q. Yu, W. Lam, and Z. Wang, "Responding e-commerce product questions via exploiting qa collections and reviews," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 2192–2203.
- [13] Z. Xu, M. J. Cruz, M. Guevara, T. Wang, M. Deshpande, X. Wang, and Z. Li, "Retrieval-augmented generation with knowledge graphs for customer service question answering," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 2905–2909.
- [14] A. G. Tapeh and M. Rahgozar, "A knowledge-based question answering system for b2c e-commerce," *Knowledge-Based Systems*, vol. 21, no. 8, pp. 946–950, 2008.
- [15] L. Cui, S. Huang, F. Wei, C. Tan, C. Duan, and M. Zhou, "Superagent: A customer service chatbot for e-commerce websites," in *Proceedings of ACL 2017, system demonstrations*, 2017, pp. 97–102.
- [16] H. Mittal, A. Chakrabarti, B. Bayar, A. A. Sharma, and N. Rasiwasia, "Distantly supervised transformers for e-commerce product qa," *arXiv preprint arXiv:2104.02947*, 2021.
- [17] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong et al., "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.
- [18] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," *arXiv preprint arXiv:2307.06435*, 2023.
- [19] Z. Yi, J. Ouyang, Y. Liu, T. Liao, Z. Xu, and Y. Shen, "A survey on recent advances in llm-based multi-turn dialogue systems," *arXiv preprint arXiv:2402.18013*, 2024.
- [20] C. Nandkumar and L. Peternel, "Enhancing supermarket robot interaction: A multi-level llm conversational interface for handling diverse customer intents," *arXiv preprint arXiv:2406.11047*, 2024.
- [21] N. Liu, L. Chen, X. Tian, W. Zou, K. Chen, and M. Cui, "From llm to conversational agent: A memory enhanced architecture with fine-tuning of large language models," *arXiv preprint arXiv:2401.02777*, 2024.
- [22] J. Shi, J. Li, Q. Ma, Z. Yang, H. Ma, and L. Li, "Chops: Chat with customer profile systems for customer service with llms," *arXiv preprint arXiv:2404.01343*, 2024.
- [23] K. Pandya and M. Holia, "Automating customer service using langchain: Building custom open-source gpt chatbot for organizations," *arXiv preprint arXiv:2310.05421*, 2023.
- [24] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.
- [25] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafraan, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," *arXiv preprint arXiv:2210.03629*, 2022.
- [26] E. Karpas, O. Abend, Y. Belinkov, B. Lenz, O. Lieber, N. Ratner, Y. Shoham, H. Bata, Y. Levine, K. Leyton-Brown, D. Muhlgay, N. Rozen, E. Schwartz, G. Shachaf, S. Shalev-Shwartz, A. Shashua, and M. Tenenholz, "Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning," 2022. [Online]. Available: <https://arxiv.org/abs/2205.00445>
- [27] A. Parisi, Y. Zhao, and N. Fiedel, "Talm: Tool augmented language models," 2022. [Online]. Available: <https://arxiv.org/abs/2205.12255>
- [28] T. Schick, J. Dwivedi-Yu, R. Dessi, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," 2023. [Online]. Available: <https://arxiv.org/abs/2302.04761>
- [29] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face," 2023. [Online]. Available: <https://arxiv.org/abs/2303.17580>
- [30] Y. Lee, S. Shin, W.-J. Park, and H. Woo, "Llm-based offline learning for embodied agents via consistency-guided reward ensemble," *arXiv preprint arXiv:2411.17135*, 2024.
- [31] R. Agarwal, D. Schuurmans, and M. Norouzi, "An optimistic perspective on offline reinforcement learning," in *International conference on machine learning*. PMLR, 2020, pp. 104–114.
- [32] R. F. Prudencio, M. R. Maximo, and E. L. Colombini, "A survey on offline reinforcement learning: Taxonomy, review, and open problems," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [33] D. Ghosh, A. Ajay, P. Agrawal, and S. Levine, "Offline rl policies should be trained to be adaptive," in *International Conference on Machine Learning*. PMLR, 2022, pp. 7513–7530.
- [34] Z. Qi, X. Liu, I. L. Iong, H. Lai, X. Sun, X. Yang, J. Sun, Y. Yang, S. Yao, T. Zhang et al., "Webri: Training llm web agents via self-evolving online curriculum reinforcement learning," *arXiv preprint arXiv:2411.02337*, 2024.
- [35] S. Yun, "Pretrained llm adapted with lora as a decision transformer for offline rl in quantitative trading," *arXiv preprint arXiv:2411.17900*, 2024.
- [36] S. Morad, A. Shankar, J. Blumenkamp, and A. Prorok, "Language-conditioned offline rl for multi-robot navigation," *arXiv preprint arXiv:2407.20164*, 2024.
- [37] J.-C. Pang, S.-H. Yang, K. Li, J. Zhang, X.-H. Chen, N. Tang, and Y. Yu, "Knowledgeable agents by offline reinforcement learning from large language model rollouts," *arXiv preprint arXiv:2404.09248*, 2024.
- [38] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learn-

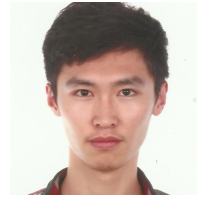
ing via sequence modeling,” *Advances in neural information processing systems*, vol. 34, pp. 15 084–15 097, 2021.

- [39] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, “Large language model based multi-agents: A survey of progress and challenges,” *arXiv preprint arXiv:2402.01680*, 2024.
- [40] C. Chai and G. Li, “Human-in-the-loop techniques in machine learning,” *IEEE Data Eng. Bull.*, vol. 43, no. 3, pp. 37–52, 2020.
- [41] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, “A survey of human-in-the-loop for machine learning,” *Future Generation Computer Systems*, vol. 135, pp. 364–381, 2022.
- [42] D. J. Hejna III and D. Sadigh, “Few-shot preference learning for human-in-the-loop rl,” in *Conference on Robot Learning*. PMLR, 2023, pp. 2014–2025.
- [43] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, “Deep reinforcement learning from human preferences,” *Advances in neural information processing systems*, vol. 30, 2017.
- [44] D. Chen, Q. Zhang, and Y. Zhu, “Efficient sequential decision making with large language models,” *arXiv preprint arXiv:2406.12125*, 2024.
- [45] H. Sun, “Reinforcement learning in the era of llms: What is essential? what is needed? an rl perspective on rlhf, prompting, and beyond,” *arXiv preprint arXiv:2310.06147*, 2023.
- [46] J. Hu, L. Tao, J. Yang, and C. Zhou, “Aligning language models with offline reinforcement learning from human feedback,” *arXiv preprint arXiv:2308.12050*, 2023.
- [47] S. Emmons, B. Eysenbach, I. Kostrikov, and S. Levine, “Rvs: What is essential for offline rl via supervised learning?” *arXiv preprint arXiv:2112.10751*, 2021.
- [48] A. K. Kalusivalingam, A. Sharma, N. Patel, and V. Singh, “Enhancing customer service automation with natural language processing and reinforcement learning algorithms,” *International Journal of AI and ML*, vol. 1, no. 2, 2020.
- [49] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” 2021. [Online]. Available: <https://arxiv.org/abs/2005.11401>
- [50] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu *et al.*, “Instruction tuning for large language models: A survey,” *arXiv preprint arXiv:2308.10792*, 2023.
- [51] Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto, “Length-controlled alpaca-eval: A simple way to debias automatic evaluators,” *arXiv preprint arXiv:2404.04475*, 2024.



**MING GONG** received the M.S. and Ph.D. degrees in Electrical Engineering from the University of Dayton. She was previously affiliated with the Vision Lab at the University of Dayton, where she conducted research in computer vision, including object recognition, line segment detection, and semantic segmentation, and published multiple papers in these areas. She is currently an intern at the Xiaoduo AI Lab, focusing on natural language processing, large language models (LLMs), agent

frameworks, and multimodal LLMs, with recent publications in these domains.



**XUCHENG HUANG** served as a Research Assistant at the Xiaoduo & Aiiit Union AI Laboratory from 2023 to 2025, he. Since 2019, he has been working as a student assistant in the Fifth Information Department at Friedrich-Alexander University Erlangen-Nürnberg (FAU). He has participated in two notable projects: DIBCO handwritten document image denoising, and medical acoustic and linguistic analysis for Alzheimer’s disease. His research interests include image instance segmentation, image denoising, text classification, NER, spell correction, LLM pre-training, and multimodal LLM information comprehension.



**ZIHENG XU** received the B.S. degree in Geographic Information Science from East China Normal University in 2022 and the M.S. degree in Geomatics Engineering from the same university in 2025. During 2023-2024, he served as an NLP Algorithm Intern at the AI Department of Banna. From 2024 to 2025, he was an Algorithm Intern at the AI Department of Xiaoduotech. He has published three papers, with primary research interests including: Human-Computer Interaction intelligent cockpits, E-commerce AI customer service.



**DR. VIJAYAN K. ASARI** is a Professor in Electrical and Computer Engineering and Ohio Research Scholars Endowed Chair in Wide Area Surveillance at University of Dayton. He is the director of the Center of Excellence for Computational Intelligence and Machine Vision at UD. Dr. Asari received his PhD degree in Electrical Engineering from Indian Institute of Technology, Madras. Prior to joining UD in February 2010, Dr. Asari worked as Professor in ECE at Old Dominion University, Norfolk, Virginia. Dr. Asari holds five patents and has published more than 800 research articles, including 142 peer-reviewed journal papers in the areas of image processing, pattern recognition, machine learning and artificial neural networks. He has so far mentored 33 PhD dissertations and 50 MS theses in electrical and computer engineering. Dr. Asari received several awards for teaching, research, advising, and technical leadership. He is an elected Fellow of SPIE and a Senior Member of IEEE.

...