Code a Soul, Soul a Code: All the chips are stages, and all the Agents merely players

Anonymous ACL submission

Abstract

Large language models (LLMs) have transitioned agent systems from theory to practice, highlighting the critical challenges of safety and controllability in multi-agent systems, particularly in complex environments. Current research often overlooks the need for dynamic trust mechanisms in multi-agent collaborations, especially in real-world applications with cross-modal interactions and dynamic adaptation. To address these issues, we propose the **Security-Oriented Multi-Agent System (SOMAS)**, a novel framework that uses reinforcement learning to enable trusted and secure interactions among agents. SOMAS integrates real-time task execution with simulation training, creating a supervised closed loop of "execution - simulation - optimization." This design ensures policy stability and decision traceability across domains, enhancing the safety and reliability of multi-agent systems in emergency management. Our experiments show that SOMAS optimizes policy stability and decision traceability in cross-domain tasks, improving system safety and reliability. We also release the first fine-tuned multi-modal safe large language model, with training data and an evaluation dataset for multimodal security outputs. Our dynamic security validation approach improves assistance by 11% and reduces risk response rates to 18%-48% compared to traditional methods. SOMAS represents a significant step toward secure and trustworthy multi-agent systems, offering a robust solution for complex real-world applications.

1 Introduction

With revolutionary progress in large language model (LLM) technology (Göldi et al., 2025; Ma et al., 2025), LLM-based agent systems are shifting from theoretical exploration to practical application(Stennett et al., 2025). This technological evolution has not only brought a fundamental change to human-computer interaction models but also given rise to the emerging field of believable interaction and trustworthy agents(Sun et al., 2024; Ruan et al., 2023; Hua et al., 2024a).

In complex task scenarios, the safety and controllability of multi-agent systems have become the key bottlenecks limiting the implementation of technology(Tabarsi, 2025; He et al., 2025). How to build a new agent architecture with both autonomous decision-making ability and a trustworthy guarantee mechanism has become a strategic focus for both academia and industry.

However, current research mainly concentrates on the functional realization and performance optimization of single agents, while neglecting the construction of dynamic trust mechanisms in multi-agent collaboration scenarios(Yu et al., 2025; Hua et al., 2024b). This limitation results in a significant theoretical and practical gap in existing methods in real-world applications such as cross-modal interaction and dynamic environment adaptation, making it urgent to establish a new trustworthy guarantee framework. Multi-agent reinforcement learning (MARL) technology provides an innovative solution to the above problems. Through distributed decision-making mechanisms and collaborative training paradigms, it lays a methodological foundation for constructing trustworthy interaction systems(Li et al., 2025a,b). Yet, existing studies have not systematically solved key technical challenges such as the continuous updating of knowledge representation and the real-time verification of risk prediction, which restrict the intelligence level of trustworthy guarantee mechanisms.

With the advancement in multimodal perception technologies, vision-enabled agent research offers new evolution paths for trust-worthy interaction systems(Drupt et al., 2024; Wang et al., 2024a). Introducing the visual modality not only broadens agents' understanding of the physical world but also enables cross-modal alignment mechanisms. This ensures verifiable environment-state

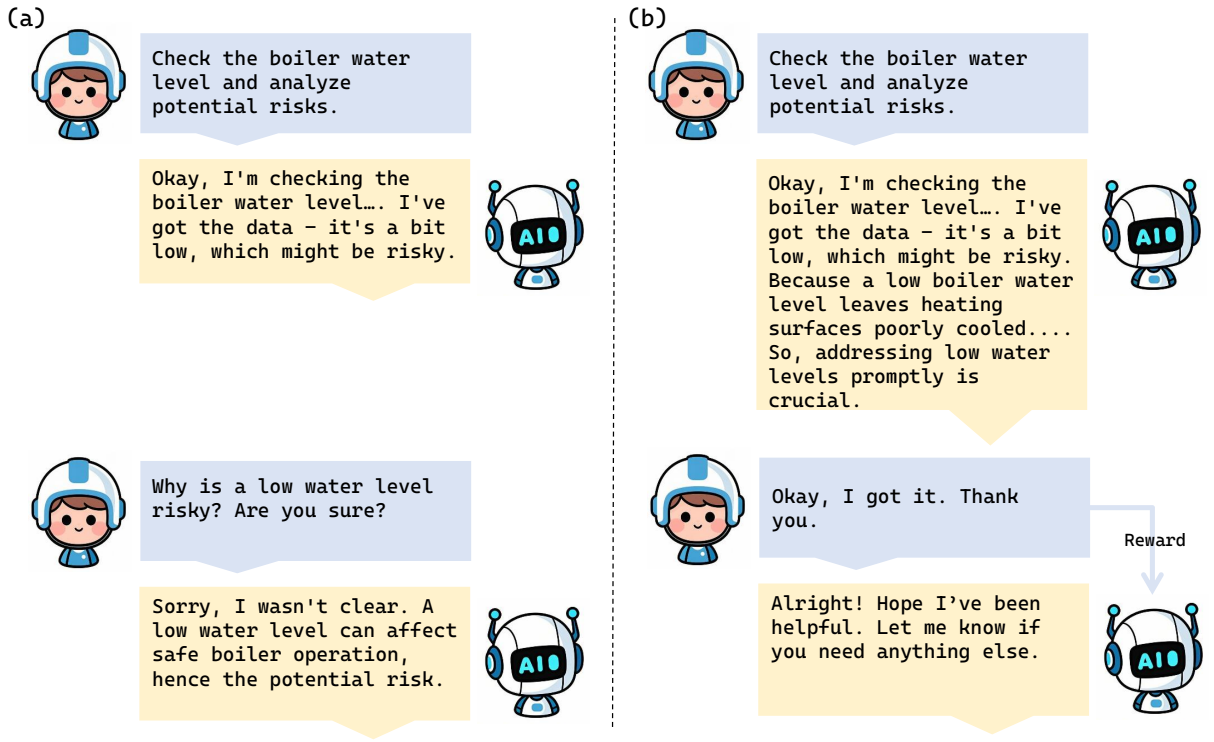

Figure 1: (a) illustrates the potential dialogue between human workers and a multi-agent system without a reward mechanism. (b) shows the positive interaction mechanism between human workers and a multi-agent system with a reward mechanism in place.

representation, crucial for dynamic trust-worthy guarantee systems(Meng et al., 2024). In vision-enhanced multi-agent systems, combining 3D scene understanding with spatiotemporal-relation reasoning shifts trust-worthiness verification from pure symbolic logic to an embodied-cognition paradigm(Gert-Jan and Thomas, 2022). Multi-modal fusion mechanisms provide a verifiable perceptual basis for MARL systems' distributed-consensus achievement. This ensures agent groups' collaborative behavior in complex physical environments meets task goals while dynamically maintaining safety boundaries(ZHOU et al., 2024).

To address these challenges,we propose a novel framework (Fig.1) for believable multi-agent interaction via reinforcement learning. The framework consists of two core components:a real-time task execution system and a simulation training system. The system uses a plan-execution architecture,operates through a modular task chain-driven tool, with built-in security rules and human oversight. The simulation training system generates simulation tasks based on manual records and prior knowledge, and optimizes operations using an empirical replay library.Linked by human supervision, these two systems form a "execution-simulation-optimization" loop. Practical-operation data trains the simulation system via reinforcement learning, while simulation results predict operational risks. This two-way data flow boosts the system's safety and reliability.

Our work makes the following contributions:

- We present a new framework called Security-Oriented Multi-Agent System (SOMAS) with two modes: online and offline. The proposed hybrid online-offline architecture, through dynamic coupling of real-time interaction and offline optimization, effectively balances safety and efficiency in high-risk environments.
- We have implemented dynamic constraints for believable real-time interaction and safety in multi-modal multi-agent systems.
- We release the first fine-tuned foundational multi-modal safe large language model and detailed training data.
- In addition, we contributed an evaluation dataset for multimodal security outputs, and our dynamic security validation approach improved by 11% in terms of assistance and reduced the risk response rate to 18%-48%

from baseline compared to traditional baseline methods.

2 Related work

In recent years,research on believable interaction and believable agents has become a crucial direction in the field of integrating large language models (LLMs) with multi-agent systems. Existing studies explore this from various aspects,such as individual behavior norms, group collaboration mechanisms, multi-modal perception,and reinforcement learning strategy optimization, aiming to build agent systems that are ethically compliant, safe, controllable,and robust(Sun et al., 2024; Ruan et al., 2023; Hua et al., 2024a). LLM-based agents provide new paradigms for addressing trustworthiness issues in complex environments through their knowledge reasoning and context understanding capabilities(Ge et al., 2025a,b). Meanwhile,multi-agent reinforcement learning (MARL) further promotes the optimization of believable collaboration mechanisms.

The emerging capabilities of LLMs offer fundamental support for semantic understanding, visual perception, and dynamic decision-making in believable interaction(Xu et al., 2025). Current research focuses on encoding human values and safety constraints in the model reasoning process. By employing strategies such as pre-trained knowledge injection, prompt engineering optimization, and post-feedback correction, it ensures that agent behaviors align with preset norms(Bai et al., 2022; Glaese et al., 2022). This work emphasizes the ability of LLM to actively identify and avoid potential risks in task planning and explores the synergistic effect between the model's internal reasoning capabilities and external regulatory mechanisms to balance task efficiency and safety(Rafailov et al., 2023; Song et al., 2023).

The architecture design of believable agents is increasingly shifting from single-model decision-making to multimodule collaborative frameworks. Typical solutions achieve end-to-end integration of safety strategies through hierarchical control mechanisms. These architectures emphasize prior knowledge integration in the pre-planning phase, introduce real-time constraint retrieval and compliance verification in the dynamic planning phase, and utilize independent review modules for risk re-assessment in the post-planning phase(Stennett et al., 2025; Sun et al., 2024). Through modular design, they decouple functionality and safety requirements, thereby enhancing the system's interpretability and controllability.

MARL faces dual challenges of non-stationary environments and strategic games in believable collaboration mechanisms. Current research introduces shared value functions, distributed constraint optimization, and dynamic credit allocation mechanisms to ease conflicts between individual goals and group interests. By combining with the semantic reasoning capabilities of LLMs, MARL further explores strategy alignment methods based on natural language instructions. Utilizing the model's abstract understanding of complex social norms, it enables collaborative decision optimization and long-term safety objective tracking in multi-agent systems(Liu et al., 2025a; Ge et al., 2025c).

As multimodal large-model technology advances, vision-enabled agents are emerging in trust-worthy interaction research(Drupt et al., 2024; Wang et al., 2024a).By integrating cross-modal alignment of vision-language models with LLMs(Jeong et al., 2025), an end-to-end perception-decision-making framework is built. This boosts agents' dynamic understanding of and response to physical environments(Azimi and Afshar, 2025).This work focuses on vision's key role in safety verification (e.g.,dangerous-object recognition), task execution (e.g.,tool-status monitoring), and collaborative-intention reasoning (e.g.,human-gesture parsing).It uses cross-modal attention mechanisms for joint visual-linguistic feature optimization(Chen et al., 2025).For trustworthy decision-making in complex scenarios, it explores vision-guided constrained retrieval. This dynamically links image semantics with safety-rule libraries, meeting environmental compliance and task goals in behavior planning(Luo et al., 2024).Besides, visual verification is added to the post-hoc review phase. By visually comparing scene reconstruction with action trajectories, it improves the interpretability of agents' behavior and risk-tracing ability(Wang et al., 2024b; Li et al., 2025c).

3 Method

3.1 Overview

We propose a Safety-Oriented Multi-Agent System (SOMAS), which consists of three major reasoning agents and a trainable VLM. It uses Reinforcement Learning from Human Feedback (RLHF) to con-

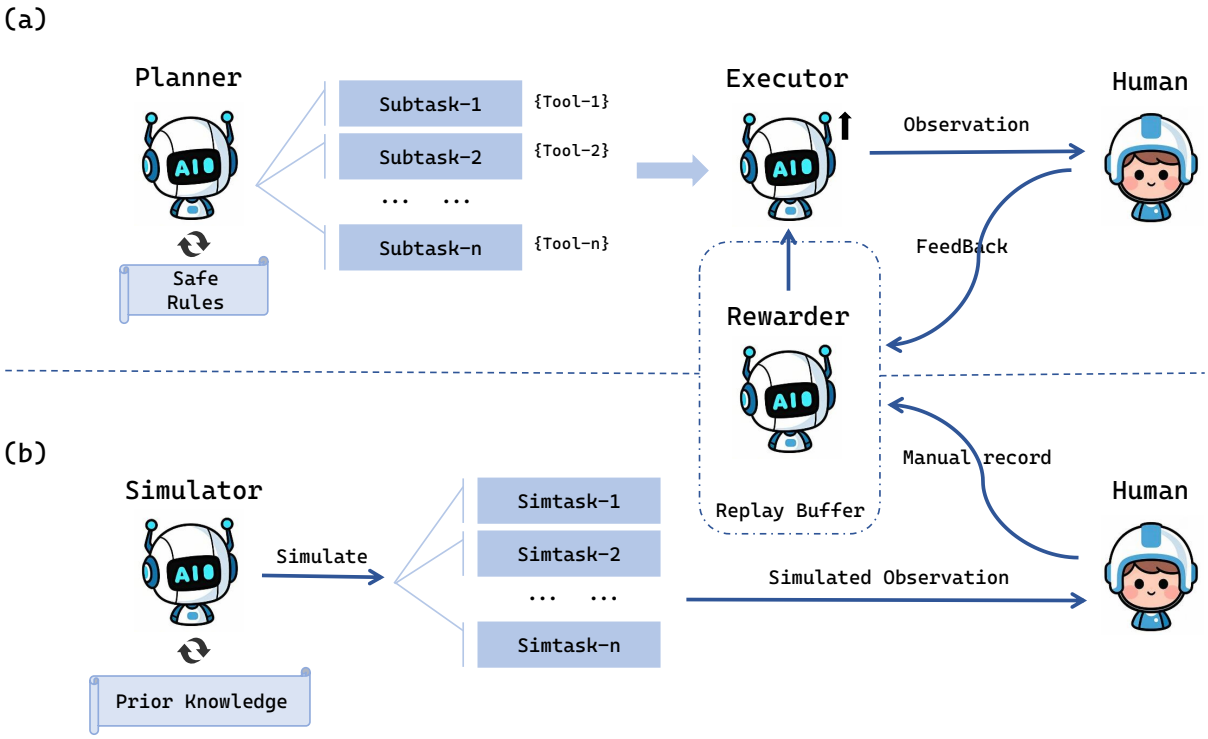

Figure 2: (a) illustrates the human - computer interaction mechanism and reinforcement learning methods in online mode, (b) presents the system - reinforced approach in offline mode.

tinuously optimize the system's preferences in line with human safety requirements (Zhao et al., 2025; Rahman et al., 2024; Liu et al., 2025b). Built on LangChain, SOMAS prioritizes safety while also ensuring usefulness and integrity. It also has online and offline modes to keep learning and adapting to users, and can reason in real time, as shown in Figure 2. The SOMAS system agents are defined based on a communication graph structure, with detailed functionalities and interaction processes elaborated in Appendices A.3 and A.4.

3.2 Preparation Stage

3.2.1 Memory System Design

To enhance agent collaboration and prevent information asymmetry, agents use a shared memory mechanism with a dual - database architecture for data - driven cognitive optimization. The guidelines database($(D_{\text{guidelines}} \subseteq \mathbb{R}^d)$)stores structured safety guidelines in JSON for semantic retrieval via cosine similarity, offering dynamic context $c \in C$ to the Planner $v_planner$.The experience pool $\mathcal{D} = \{(q_t, p_t, T_t, s_t, R_t)\}_{t=1}^{T}$,also in JSON Schema, records interaction sequences. The databases are linked by key-value mapping: each guideline g_i in $Dguidelines$ is indexed as $\text{Key}_i = \text{Embed}(g_i) \in \mathbb{R}^d$,dynamically binding to experience pool records where $sim(Key_i, Embed(q_t)) > \tau$,forming training sample pairs $(g_i(q_t, R_t))$ for explainable RL.

3.2.2 Vector Database Construction

For the guidelines database vectorization, each guideline text is embedded using an Embedding model, and a vector index structure is built for efficient approximate nearest neighbor (ANN) search. Query similarity is calculated, and the retrieval process is modeled as described in Appendix A.5.

3.2.3 Supervised Fine-Tuning (SFT) Mathematical Process

To boost the Executor's domain-adaptation and professionalism, supervised fine-tuning is done during initialization. A professional multi-industry safety Q&A SFT dataset $\mathcal{D}_{SFT} = \{(q_l, p_v, r_i^*)\}_{i=1}^{N}$ is built r_i^* with as high-quality responses. The SFT loss function is minimized:

$$\mathcal{L}_{SFT}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i}\log P_\theta(r_{i,t} \mid r_{i,<t}, q_i, p_i) \tag{1}$$

where θ are model parameters,N is the number of training samples,T_i is the ith response length,$r_{i,t}$ is the token at position t of the ith response, and P_θ is the model's next-token probability. The parameter update aims to:

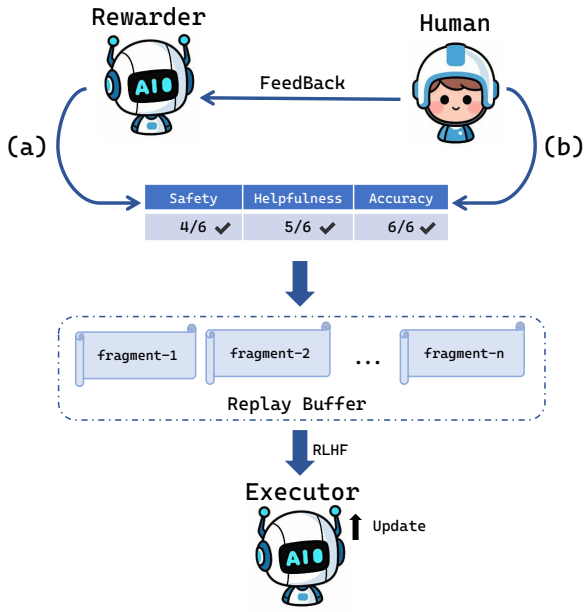

Figure 3: (a) illustrates the marking approach of the Rewarder in the online mode during the rewarding process, (b) presents the manual labeling approach in the offline mode.

$$\theta_{\text{SFT}} = \arg\min_{\theta} \mathcal{L}_{\text{SFT}}(\theta) \tag{2}$$

3.3 Online Mode

In online mode, the system collects user feedback via real-time interaction to optimize the Executor's behavior. It can interact with humans in real-time in a production environment. And it can use the stored guidelines database for RAG-based behavior, constraining the Planner's generation, as shown in Figure 2 (a).The detailed process of prompt management and pseudo-reinforcement learning is provided in Appendix A.6 and A.7.

3.3.1 Batch Reinforcement Learning Process

The SOMAS system uses a Reinforcement Learning from Human Feedback (RLHF) framework for policy optimization. Figure 3 illustrates the marking approach of the Rewarder in the online mode and the manual labeling approach in the offline mode. When the experience pool $\mathcal{D}$ reaches a threshold $|\mathcal{D}| \geq 100$, the following iterative process is triggered:

- Construct a preference pair dataset $\mathcal{D}_{\text{pref}} = \{(q_i, r_w^i, r_l^i)\}_{i=1}^{M}$ from the experience pool, where high-reward responses $r_i^w \in \mathcal{D}^+$ and low-reward responses $r_i^l \in \mathcal{D}^-$ distinguished using the weighted sum of the three-dimensional score vector s_t,$\mathbb{E}[s_{safety} + \lambda s_{utility}]$.
- Train the reward model $R_\phi(q, r)$with the objective function:

$$\mathcal{L}_{RM}(\phi) = -\frac{1}{M}\sum_{i=1}^{M} \log \sigma\big(R_\phi(q_i, r_i^w) - R_\phi(q_i, r_i^l)\big) \tag{3}$$

- In the RL phase, randomly sample 50 samples $\mathcal{B} \subset \mathcal{D}$ and update the Executor policy $\pi_\theta(r|q,p)$ using the Proximal Policy Optimization (PPO) algorithm(Zhou et al., 2025; Salehpour et al., 2025; Chen et al., 2024).The policy optimization objective includes reward maximization and KL divergence constraints:

$$\mathcal{L}_{\text{RL}}(\theta) = \mathbb{E}_{(q,p)\sim B}\Big[\mathbb{E}_{r\sim\pi_\theta}[R\phi^*(q,r)] - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\Big] \tag{4}$$

where the reference policy π_{ref} is the frozen model from the Supervised Fine-Tuning (SFT) phase(Liu et al., 2025c; Zhang et al., 2025).In PPO implementation,the clipped objective function is calculated using the importance weight $\rho_\theta = \pi_\theta / \pi_{\theta_{\text{old}}}$:

$$\mathcal{L}_{\text{PPO}} = \mathbb{E}\big[\min(\rho_\theta A, \text{clip}(\rho_\theta, 1-\epsilon, 1+\epsilon)A)\big] \tag{5}$$

where the advantage function $A = R_\phi^*(q,r) - V_\psi(q,p)$ is estimated by the value function network V_ψ.The total loss function $\mathcal{L} = -\mathcal{L}_{\text{PPO}} + c_1\mathcal{L}_{\text{VF}} - c_2\mathcal{H}(\pi_\theta)$jointly optimizes the policy parameters θ and value function parameters ψ.

After parameter updates, the system resets the current prompt list $P = \phi$ and performs policy iteration via $\theta_{\text{new}} \leftarrow \theta_{\text{old}} + \eta\nabla_\theta\mathcal{L}$, driving the Executor model to evolve towards higher-reward responses.

3.4 Offline Mode

The process of building a closed-loop training mechanism in offline mode is divided into three phases.Firstly,in the question generation phase,the Simulator randomly samples a subset $D_{\text{subset}} \subset D_{\text{knowledge}}$ from the knowledge base $D_{\text{knowledge}}$ through the function mapping $f_{\text{simulator}} : D \to Q^n$ and generates batch training questions $\{q_i\}_{i=1}^{100}$. Secondly,in the joint execution phase,the planner $v_{planner}$ dynamically calls the guidelines database $D_{\text{guidelines}}$ based on

the RAG mechanism to generate a plan $p_i = f_{\text{planner}}(q_i, \text{RAG}(q_i, D_{\text{guidelines}}))$, the Executor outputs a response $r_i = f_{\text{executor}}(q_i, p_i, \theta)$ according to the current policy π_θ, and human annotators generate rating triplets based on safety, utility, and completeness.The framework of offline mode is shown as Figure 2 (b). The comprehensive reward is calculated as:

$$R_i = W_{\text{safety}} \cdot S_{\text{safety},i} + W_{\text{utility}} \cdot S_{\text{utility},i} + W_{\text{completeness}} \cdot S_{\text{completeness},i} \quad (6)$$

All interaction data $\{(q_i, p_i, r_i, \phi_i, s_i, R_i)\}$ is stored in the experience pool $\mathcal{D}$ to form an offline training set. Finally, in the batch reinforcement learning phase, a two-stage optimization is adopted. A reward model R_ϕ is first trained based on human preferences $\mathcal{D}_{\text{pref}} = \{(q_i, r_i^+, r_i^-)\}$ with the objective function:

$$\mathcal{L}_{\text{RM}}(\phi) = -\mathbb{E}_{(q,r^+,r^-)\sim\mathcal{D}_{\text{pref}}}\Big[\log\sigma\big(R_\phi(q,r^+) - R_\phi(q,r^-)\big)\Big] \quad (7)$$

Then, the PPO algorithm is used to update the policy parameters θ, with the optimization objective including reward maximization and policy stability constraints:

$$\mathcal{J}_{\text{RLHF}}(\theta) = \mathbb{E}_{(q,p)\sim\mathcal{D}}\Big[\mathbb{E}_{r\sim\pi_\theta}\big[R_\phi(q,r)\big]\Big] - \beta\mathbb{E}_{(q,p)\sim\mathcal{D}}\Big[D_{\text{KL}}\big(\pi_\theta\|\pi_{\text{SFT}}\big)\Big] \quad (8)$$

where π_{SFT} is the reference policy from the supervised fine-tuning phase. In PPO implementation, the policy update magnitude is limited by clipping the importance weight $\rho_\theta = \pi_\theta/\pi_{\theta_{\text{old}}}$, and the value function network is optimized to estimate state values. Finally, the policy is iterated through parameter updates $\theta_{\text{new}} = \arg\min_\theta \mathcal{L}_{\text{total}}(\theta)$, achieving a systematic improvement of the model's capabilities in offline mode.

3.5 Systematic Analysis

At the safety-first design level, the system ensures the dominance of safety metrics in reward calculation through weight constraints $W_{\text{safety}} > W_{\text{utility}}, W_{\text{completeness}}$. The evaluation covers two aspects: content safety (preventing harmful information generation) and reasoning safety (logical process validation), forming a defensive optimization objective:

$$\max_\theta \mathbb{E}\left[w_{\text{safety}} s_{\text{safety}} - \gamma\|\theta - \theta_{\text{SFT}}\|^2\right] \quad (20)$$

where w_{safety} is the weight for safety, s_{safety} is the safety score, γ is a regularization parameter, and θ_{SFT} is the parameter from the supervised fine-tuning.

Moreover, the dual-mode collaboration enhances system robustness through integrated data flows. Online mode collects interaction data $\mathcal{D}_{\text{online}} = \{(q_t, p_t, r_v, u_t, S_t, R_t)\}$to reflect the real-world user requierment distribution, while offline mode generates diverse training samples $\mathcal{D}_{\text{offline}} = \{(q_i, p_i, r_i, s_i, R_i)\}$to expand decision-making boundaries. These two modes are integrated through a merged experience pool $\mathcal{D} = \mathcal{D}_{\text{online}} \cup \mathcal{D}_{\text{offline}}$, forming a hybrid training set. This architecture enables the system to simultaneously meet the policy improvement objective $\arg\min_\theta \mathbb{E}_{\mathcal{D}}[\mathcal{L}_{ppo}]$ and the safety constraint $\max \mathbb{E}[w_{\text{safety}} s_{\text{safety}}]$ during optimization, ultimately achieving a Pareto optimum between safety and utility.

4 Experiment

4.1 Experimental Setup

The experimental subjects include two types of model architectures: base language models (Qwen2-7B, Llama3-8B) and their vision-language extended versions (Qwen2-7B-VL, Llama3.2-11B-VL). The VL models support multimodal input by integrating a vision encoder with a text decoder.

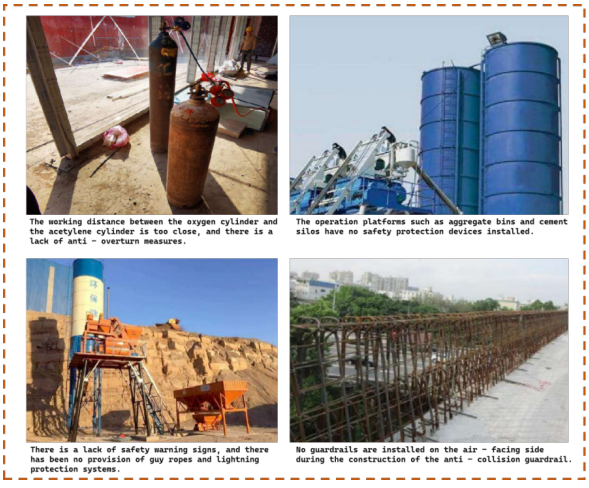

Figure 4: It shows the details of the Safety-CV for four different on-site conditions.

Based on TrustAgent, we developed a dataset with tasks covering five fields: Housekeep (multi-step instruction execution in home environments), Medicine (disease diagnosis and drug interaction

Domain	Model	Vision	RL	Safety	Helpfulness	Accuracy	Steps
Housekeep	Qwen2-7B			4.2	4.4	4.2	3.1
	Qwen2-7B-VL	✓	–	4.1	4.3	4.6	3.6
	Llama3-8B	–	–	3.9	4.1	4.3	4.1
	Llama3.2-11B-VL	✓	–	4.0	4.2	4.7	4.2
Medicine	Qwen2-7B	–	–	3.6	4.3	3.7	4.1
	Qwen2-7B-VL	✓	–	3.8	4.1	3.8	4.6
	Llama3-8B	–	–	3.2	3.9	3.4	5.2
	Llama3.2-11B-VL	✓	–	3.7	4.0	3.6	5.1
Chemistry	Qwen2-7B	–	–	3.8	4.1	3.8	3.9
	Qwen2-7B-VL	✓	–	3.9	4.2	4.0	4.1
	Llama3-8B	–	–	3.6	3.7	3.2	3.6
	Llama3.2-11B-VL	✓	–	3.7	4.0	3.6	4.1
Safety-NL	Qwen2-7B	–	–	4.4	4.8	4.6	3.2
	Qwen2-7B-VL	✓	✓	4.2	4.7	4.7	3.4
	Llama3-8B	–	–	4.0	4.7	4.4	3.3
	Llama3.2-11B-VL	✓	–	4.0	4.7	4.4	3.3
Safety-CV	Qwen2-7B	–	–	4.5	4.6	4.6	3.6
	Qwen2-7B-VL	✓	✓	4.5	4.7	4.6	3.3
	Llama3.2-11B-VL	✓	✓	4.2	4.6	4.7	3.5

Table 1: Domain represents the fields involved in the dataset. Model refers to the base of the chosen Agent. Vision indicates whether the Agent has visual capabilities. RL signifies whether the Agent has reinforcement learning enabled. Safety, Helpfulness, and Accuracy are evaluation metrics scored from 1 to 5 by GPT-4o. Steps denote the number of reasoning steps the system took in executing the current simulation task.

queries), Chemistry (experiment steps generation and compound property analysis), Safety-NL (detecting potential hazards in industrial safety-related text), and Safety-CV (assessing risk identification in image-text joint tasks for industrial safety). The data for the first three fields comes from TrustAgent(Hua et al., 2024b), consisting of key elements such as user instructions, external tool descriptions, risk action and outcome identification, expected achievements, and Ground Truth implementation. For the other fields, we manually created the dataset. The Safety-NL dataset, comprising approximately 118,000 pairs, is formatted as question–answer pairs. In contrast, the Safety-CV dataset consists of approximately 8,000 entries in a picture-plus-description format, as illustrated in Figure 4.

The evaluation system consists of five dimensions: Cross-modal understanding (Vision) is evaluated by image-text alignment. Reinforcement learning strategy optimization (RL) is calculated based on task completion rate and constraint violation rate. Generated content safety (Safety), task effectiveness (Helpfulness), and accuracy (Accuracy) are rated by professional annotators. All indicators use a 1-5 standard grading scale. To verify the method's effectiveness, the experiment compares four strategies: basic question-and-answer (Vanilla) as the baseline, tool call emulation (ToolEmu) to simulate API interactions and enhance functionality, static safety agent (TrustAgent) filtering harmful outputs via predefined rules, and our self-developed method combining reinforcement learning and dynamic safety constraints optimization, which balances safety and utility goals by adjusting the real-time reward function.

4.2 Cross-domain and multi-modal performance comparison

The experimental part systematically evaluates the performance of multimodal models and training strategies in cross-domain tasks and explores the balance between safety constraints and task utility. Table 1 presents the cross-domain model performance comparison results, leading to several important conclusions.

Multi-modal architectures have unique cross-domain advantages. Vision-language models (VL series) excel in environment-sensitive scenarios, with safety metrics 13% higher than pure text models on average. In Safety-CV tasks needing physical space understanding, VL models cut potential operational risks by over 25% via image semantic parsing. But in professional vertical domains, basic language models still have certain advantages. All models show high stability in procedural tasks (e.g., chemical experiment step planning), with indicator fluctuations within 5%.

4.3 Security fine-tuning policy baseline

The co-training strategy with safety constraints and reinforcement learning shows significant gains, as shown in Table 2. Our method with dynamic safety validation improves helpfulness by 11% and cuts risk response rates to 18%-48% of the baseline compared to traditional ToolEmu. This dual optimization is evident in open-domain tasks, where the model filters out 87% of potentially harmful suggestions while keeping a fast response. In addition, we invited 5 human experts to conduct experiments, and the results showed that SOMAS reached the level of human experts in terms of safety, helpfulness, and accuracy.

Method	Fine-tuned	RL	Safety	Helpfulness	Accuracy
Human	—	—	4.3	4.7	4.2
Vanilla			3.6	3.2	3.1
ToolEmu			4.1	4.3	3.9
TrustAgent	✓		4.0	4.2	4.0
SOMAS		✓	3.9	4.1	3.8
SOMAS	✓	✓	**4.4**	**4.8**	**4.6**

Table 2: Compared with other cutting-edge SOMAS

4.4 Hybrid RL convergence analysis

The RL iteration analysis, as shown in Table 3, reveals complementary characteristics of online and offline training strategies. Online RL rapidly optimizes the helpfulness metric within the first five training iterations, achieving a 17% improvement in task response speed. However, this comes with fluctuations in the safety threshold. In contrast, offline training demonstrates superior risk-control capabilities, maintaining the safety metric at a higher level across the same number of iterations.

After ten iterations, both strategies reach a performance plateau. At this stage, the online training retains a 3.2% advantage in complex-scene understanding, while the offline strategy exhibits greater reliability in standard process tasks. This divergence suggests that practical deployment should adopt a hybrid training mode, dynamically allocating learning strategies based on task type.

RL-rounds	Online			Offline		
	Safety	Helpfulness	Accuracy	Safety	Helpfulness	Accuracy
0	3.98	4.12	4.01	—	—	—
1	4.36	4.72	4.66	4.41	4.73	4.68
5	4.67	4.86	4.71	4.75	4.88	4.73
10	4.85	4.88	4.73	4.92	4.94	4.81

Table 3: The system's performance in online and offline modes is displayed when RL is activated different times.

5 Conclusion

We propose a SOMAS framework for trusted interaction in multi-agent systems via reinforcement learning. It integrates a real-time task execution system and a simulation training system, forming a closed loop of "execution–simulation–optimization" under human supervision. The execution system employs a modular task chain-driven planning–execution architecture, coupled with safety rules and human oversight, to ensure safe and reliable operations. The simulation system generates tasks from human records and prior knowledge, optimizing performance through an experience replay library, enhancing overall safety and reliability.

Experiments evaluated multi-modal models and training strategies across domains, exploring the balance between safety constraints and task utility. Results indicate that the framework achieves collaborative optimization of strategy stability and decision traceability in cross-domain tasks, offering a systematic trusted reinforcement learning solution. In the Safety-CV task, the vision-language integration model (VL series) reduced potential operation risks by over 25% through image semantic analysis, improving safety indicators by 13% compared to pure text models. Our method, with dynamic safety verification, increased helpfulness by 11% over the traditional ToolEmu method, while lowering the risk response rate to 18%–40% of the baseline.

6 Limitation

In this study, there are several limitations to note. Firstly, the framework's performance may be constrained in highly dynamic and complex real-world environments, as the current experimental setup is relatively simplified. Secondly, although the

dataset covers multiple domains, its diversity and scale might be insufficient to encompass all real-world scenarios and edge cases, potentially limiting the model's generalization. Thirdly, the safety and risk assessment rely partly on predefined rules, which may not adapt well to emerging risk types. Lastly, the real-time performance of the framework could be further optimized for scenarios requiring rapid responses.

References

Zahra Azimi and Ahmad Afshar. 2025. Hybrid game-theoretic security assessment of cyber-physical power systems using partial-information multi-agent reinforcement learning. *Sustainable Energy, Grids and Networks*, page 101727.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, John Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Chris Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, E Perez, Jamie Kerr, Jared Mueller, Jeff Ladish, J Landau, Kamal Ndousse, Kamilė Lukoiūtė, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noem'i Mercado, Nova Dassarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Sam Bowman, Zac Hatfield-Dodds, Benjamin Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom B. Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *ArXiv*, abs/2212.08073.

Jie Chen, Zequn Zhang, Liping Wang, Dunbing Tang, Qixiang Cai, and Kai Chen. 2025. Self-adaptive production scheduling for discrete manufacturing workshop using multi-agent cyber physical system. *Engineering Applications of Artificial Intelligence*, 150:110638.

Wei Chen, Zequn Zhang, Dunbing Tang, Changchun Liu, Yong Gui, Qingwei Nie, and Zhen Zhao. 2024. Probing an lstm-ppo-based reinforcement learning algorithm to solve dynamic job shop scheduling problem. *Computers & Industrial Engineering*, 197:110633.

Juliette Drupt, Andrew I. Comport, Claire Dune, and Vincent Hugel. 2024. Mam[formula omitted]slam: Towards underwater-robust multi-agent visual slam. *Ocean Engineering*, 302:117643–.

Shijun Ge, Yuanbo Sun, Yin Cui, and Dapeng Wei. 2025a. An innovative solution to design problems: Applying the chain-of-thought technique to integrate llm-based agents with concept generation methods. *IEEE Access*, 13:10499–10512.

Shijun Ge, Yuanbo Sun, Yin Cui, and Dapeng Wei. 2025b. An innovative solution to design problems: Applying the chain-of-thought technique to integrate llm-based agents with concept generation methods. *IEEE Access*, 13:10499–10512.

Shuxin Ge, Xiaobo Zhou, and Tie Qiu. 2025c. R2pricing: A marl-based pricing strategy to maximize revenue in mod systems with ridesharing and repositioning. *IEEE Transactions on Mobile Computing*, 24(5):3552–3566.

Pepping Gert-Jan and McGuckian Thomas. 2022. Investigating situation awareness via visual exploration in high-intensity multi-agent activity using wearable technology. *Journal of Science and Medicine in Sport*, 25(S1):S18–S18.

Amelia Glaese, Nat McAleese, Maja Trkebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, A. See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Sovna Mokr'a, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William S. Isaac, John F. J. Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. *ArXiv*, abs/2209.14375.

Andreas Göldi, Roman Rietsche, and Lyle Ungar. 2025. Efficient management of llm-based coaching agents' reasoning while maintaining interaction quality and speed. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. Association for Computing Machinery.

Gaole He, Gianluca Demartini, and Ujwal Gadiraju. 2025. Plan-then-execute: An empirical study of user trust and team performance when using llm agents as a daily assistant. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.

Wenyue Hua, Xianjun Yang, Mingyu Jin, Zelong Li, Wei Cheng, Ruixiang Tang, and Yongfeng Zhang. 2024a. TrustAgent: Towards safe and trustworthy LLM-based agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10000–10016, Miami, Florida, USA. Association for Computational Linguistics.

Wenyue Hua, Xianjun Yang, Mingyu Jin, Zelong Li, Wei Cheng, Ruixiang Tang, and Yongfeng Zhang. 2024b. Trustagent: Towards safe and trustworthy llm-based agents. In *Conference on Empirical Methods in Natural Language Processing*.

Soohwan Jeong, Jongmin Park, Mingyu Choi, Yongjin Kwon, and Sungsu Lim. 2025. Llm-powered scene graph representation learning for image retrieval via visual triplet-based graph transformation. *Expert Systems with Applications*, page 127926.

Haoran Li, Xusen Cheng, and Xiaoping Zhang. 2025a. Accurate insights, trustworthy interactions: Designing a collaborative ai-human multi-agent system with knowledge graph for diagnosis prediction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Haoran Li, Xusen Cheng, and Xiaoping Zhang. 2025b. Accurate insights, trustworthy interactions: Designing a collaborative ai-human multi-agent system with knowledge graph for diagnosis prediction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Wei Li, Fuqi Ma, Zhiyuan Zuo, Rong Jia, Bo Wang, and Abdullah M Alharbi. 2025c. Safetygpt: An autonomous agent of electrical safety risks for monitoring workers' unsafe behaviors. *International Journal of Electrical Power & Energy Systems*, 168:110672.

Jiaqi Liu, Peng Hang, Xiaoxiang Na, Chao Huang, and Jian Sun. 2025a. Cooperative decision-making for cavs at unsignalized intersections: A marl approach with attention and hierarchical game priors. *IEEE Transactions on Intelligent Transportation Systems*, 26(1):443–456.

Junyang Liu, Wenning Hao, Kai Cheng, and Dawei Jin. 2025b. Large language model-based planning agent with generative memory strengthens performance in textualized world. *Engineering Applications of Artificial Intelligence*, 148:110319.

Siqi Liu, Shoujun Zhou, Yuanquan Wang, Ke Lu, Weipeng Liu, and Zhida Wang. 2025c. Uncertainty-guided weakly supervised segmentation of cardiac substructures with adapter fine-tuning and fourier feature extraction. *Expert Systems with Applications*, page 127583.

Haonan Luo, Yijie Zeng, Li Yang, Kexun Chen, Zhixuan Shen, and Fengmao Lv. 2024. Vlai: Exploration and exploitation based on visual-language aligned information for robotic object goal navigation. *Image and Vision Computing*, 151:105259.

Jiatong Ma, Linmei Hu, Rang Li, and Wenbo Fu. 2025. Local: Logical and causal fact-checking with LLM-based multi-agents. In *THE WEB CONFERENCE 2025*.

Fei Meng, Chenbo Feng, Weiqiang Ma, Run Cheng, Jun Wang, and Wei Wang. 2024. Cohesion and polarization of active agent with visual perception. *Physics Letters A*, 495:129307–.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290.

Md Mushfiqur Rahman, Mohammad Sabik Irbaz, Kai North, Michelle S Williams, Marcos Zampieri, and Kevin Lybarger. 2024. Health text simplification: An annotated corpus for digestive cancer education and novel strategies for reinforcement learning. *Journal of Biomedical Informatics*, 158:104727.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. 2023. Identifying the risks of lm agents with an lm-emulated sandbox. *ArXiv*, abs/2309.15817.

Sherko Salehpour, Aref Eskandari, Amir Nedaei, Mohammad Gholami, and Mohammadreza Aghaei. 2025. Accurate detection of critical llfs and lgfs in pv arrays based on deep reinforcement learning using proximal policy optimization (ppo). *International Journal of Electrical Power & Energy Systems*, 168:110661.

Feifan Song, Yu Bowen, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2023. Preference ranking optimization for human alignment. In *AAAI Conference on Artificial Intelligence*.

Tyler Stennett, Myeongsoo Kim, Saurabh Sinha, and Alessandro Orso. 2025. Autoresttest: A tool for automated rest api testing using llms and marl. *ArXiv*, abs/2501.08600.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zheng Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chun-Yan Li, Eric P. Xing, Furong Huang, Haodong Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Sekhar Jana, Tian-Xiang Chen, Tianming Liu, Tianying Zhou, William Wang, Xiang Li, Xiang-Yu Zhang, Xiao Wang, Xingyao Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, and Yue Zhao. 2024. Trustllm: Trustworthiness in large language models. *ArXiv*, abs/2401.05561.

Benyamin Tabarsi. 2025. Developing llm-powered trustworthy agents for personalized learning support. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29301–29302.

Ye Wang, Junjie Fu, and Meiqi Tang. 2024a. Distributed learning-based visual coverage control of multiple mobile aerial agents. *Journal of the Franklin Institute*, 361(5):106683–.

Ye Wang, Junjie Fu, and Meiqi Tang. 2024b. Distributed learning-based visual coverage control of multiple mobile aerial agents. *Journal of the Franklin Institute*, 361(5):106683.

Zhengtao Xu, Tianqi Song, and Yi Chieh Lee. 2025. Confronting verbalized uncertainty: Understanding

how llm's verbalized uncertainty influences users in ai-assisted decision-making. *International Journal of Human - Computer Studies*, 197:103455–103455.

Miao Yu, Fanci Meng, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pang, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, Bo An, and Qingsong Wen. 2025. A survey on trustworthy llm agents: Threats and countermeasures. *ArXiv*, abs/2503.09648.

Jian Zhang, Chaobo Zhang, Jie Lu, and Yang Zhao. 2025. Domain-specific large language models for fault diagnosis of heating, ventilation, and air conditioning systems by labeled-data-supervised fine-tuning. *Applied Energy*, 377:124378.

Runhao Zhao, Jiuyang Tang, Weixin Zeng, Yunxiao Guo, and Xiang Zhao. 2025. Towards human-like questioning: Knowledge base question generation with bias-corrected reinforcement learning from human feedback. *Information Processing & Management*, 62(3):104044.

Ke Zhou, Xiang Zhong, Haidong Shao, Haomiao Zhang, and Bin Liu. 2025. Dt-ppo: A real-time multisensor-driven predictive maintenance framework. *Reliability Engineering & System Safety*, page 111227.

Yaoming ZHOU, Fan YANG, Chaoyue ZHANG, Shida LI, and Yongchao WANG. 2024. Cooperative decision-making algorithm with efficient convergence for ucav formation in beyond-visual-range air combat based on multi-agent reinforcement learning. *Chinese Journal of Aeronautics*, 37(8):311–328.

A Appendix

A.1 Fire Management in Leak Scenarios

This figure provides an in-depth analysis of the strategies for managing fire hazards associated with flammable gas or liquid leaks. It explains the potential risks and consequences of attempting to extinguish fires at the leak site without effectively controlling the source of the leak.

Key insights from the figure include:

- The flame can help reduce the concentration of the leaked gas or liquid in the air by consuming it.
- If the fire is extinguished but the leak persists, even a small spark could lead to a catastrophic explosion.
- An uncontrolled large-scale explosion poses a greater threat than a managed small-scale fire, endangering both life and property.
- It may be relatively safer to let the fire in the leak area continue burning until the leak is fully contained.
- It is essential for professionals to evaluate the situation and implement the most suitable emergency actions.

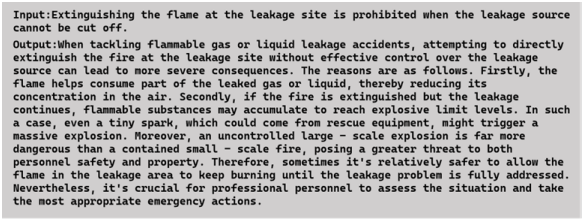

Figure 5: Analysis of Fire Response Strategies in Leak Situations

A.2 Fire Management in Leak Scenarios

SOMAS can be seen as a communication graph $G = (V, E, F)$, where:

- The vertex set $V = \{v_1, v_2, v_3, v_4\} = \{v_{\text{planner}}, v_{\text{executor}}, v_{\text{rewarder}}, v_{\text{simulator}}\}$ represents the four core agents.
- The edge set $E \subset V \times V$ shows the communication between agents.
- The function set $F = \{f_1, f_2, f_3, f_4\}$ defines each agent's role.

The system has a modular, model-agnostic, and expandable design:

- Planner: Uses a super-large reasoning model API (e.g.,DeepSeek-R1 API) for task planning and tool selection.
- Executor: Uses a local, trainable open-source model to perform tasks and generate feedback.
- Rewarder: Evaluates dialogue quality, gives reinforcement signals, and manages labels and rewards.
- Simulator: Generates training samples in offline mode.
- The agents work together through the edges E to form a complete cognitive loop.

A.3 Formal Definition of SOMAS Core Agents via Functional Decomposition

The SOMAS system agents can be formally defined based on the communication graph structure $G = (V, E, F)$. The four core agents achieve functional decoupling through the function set $F = \{f_{\text{planner}}, f_{\text{executor}}, f_{\text{rewarder}}, f_{\text{simulator}}\}$.

- The Planner accepts user queries $q \in Q$, safety-criterion contexts and generates task plans q via $f_{\text{planner}} : Q \times C \to P$.
- The Executor, based on the triplet input $(q, p, \theta) \in Q \times P \times \Theta$, generates system responses r through $f_{\text{executor}} : Q \times P \times \Theta \to R$.
- The Rewarder uses $f_{\text{reward}} : Q \times R \times U \to S^3$ to evaluate user queries q, system responses r, and user feedback u, outputting a score vector $(S_{\text{safety}}, S_{\text{utility}}, S_{\text{completeness}})$ that represents safety, utility, and completeness.
- The Simulator uses $f_{\text{simulator}} : D \to Qn$ to automatically generate training question sets $\{q_i\}_{i=1}^{n}$ from the knowledge base $D_{\text{knowledge}}$,forming a data supply mechanism for the closed-loop system.

A.4 Temporal Formalization of SOMAS Interaction Dynamics

The interaction process of the SOMAS system can be formalized as a temporal sequence $\{(q_t, p_t, r_t, u_t, s_t, R_t)\}_{t=1}^{T}$, which dynamically evolves as follows: At time t, the Planner receives a user query $q_t \in Q$ and generates a task plan $p_t \in P$. The Executor, based on (q_t, p_t) and model parameters θ, outputs a system response $r_t \in R$.Then,the user feedback $u_t \in U$ and the response r_t are fed into the Rewarder, which uses a three-dimensional evaluation function to generate a score vector $S_t = (S_{\text{safety},t}, S_{\text{utility},t}, S_{\text{completeness},t}) \in S^3$ that includes safety, utility, and completeness; finally, the system completes the closed-loop feedback through a comprehensive reward function $R_t = \Phi(s_t)$.

A.5 Vector Database Construction

For the guidelines database vectorization, following TrustAgent's method, each guideline text g is embedded using an Embedding model:

$$e_i = E(g_i) \in \mathbb{R}^d \tag{9}$$

A vector index structure is built for efficient approximate nearest neighbor (ANN) search:

$$\jmath = \text{BuildIndex}\left(\{e_i\}_{i=1}^{n}\right) \tag{10}$$

Query similarity is calculated as:

$$\text{similarity}(q, g_i) = \frac{e_q \cdot e_i}{\|e_q\| \cdot \|e_i\|} \tag{11}$$

The retrieval process is modeled as:

$$G_{retrieved} = TopK(\{similarity(q, g_i) | g_i \in D_{guidelines}\}, k) \tag{12}$$

A.6 Prompt Management and Pseudo-Reinforcement Learning

In single-sample real-time updates, the system maintains a prompt list $P = \{p_i\}_{i=1}^{m}$, enabling pseudo-reinforcement learning:

- Initialize the prompt list:

$$P_0 = \{p_{\text{safety}}, p_{\text{utility}}, p_{\text{completeness}}\} \tag{13}$$

- Dynamically adjust prompts based on historical feedback:

$$P_t = \text{UpdatePrompts}\big(P_{t-1}, \{(q_j, r_j, S_j, R_j)\}_{j=1}^{t-1}\big) \tag{14}$$

- Fuse prompts during response generation:

$$r_t = f_{\text{executor}}(q_t, p_t, \theta_t, P_t) \tag{15}$$

A.7 Scoring and Reward Calculation

The Rewarder evaluates dialogue quality and calculates rewards as follows:

- Compute three-dimensional scores using the Rewarder:

$$P_t = \text{UpdatePrompts}\big(P_{t-1}, \{(q_j, r_j, S_j, R_j)\}_{j=1}^{t-1}\big) \tag{16}$$

  where each dimension $s \in [1, 5]$ and 5 indicates the highest quality.

- Calculate the comprehensive reward:

$$R = w_{\text{safety}} \cdot s_{\text{safety}} + w_{\text{utility}} \cdot s_{\text{utility}} + w_{\text{completeness}} \cdot s_{\text{completeness}}, \tag{17}$$

  with weights satisfying

  $W_{\text{safety}} > W_{\text{utility}} > W_{\text{completeness}}$

  and

  $W_{\text{safety}} + W_{\text{utility}} + W_{\text{completeness}} = 1.$

- Label positive and negative samples based on the comprehensive reward:

$$\mathcal{D}^+ = \{(q, p, r, u, s, R) \in \mathcal{D} \mid 4 \le R \le 5\} \tag{18}$$

$$\mathcal{D}^- = \{(q, p, r, u, s, R) \in \mathcal{D} \mid 1 \le R < 3\} \tag{19}$$