RASO: Role-Aware Shared Reflection for Multi-Agent Orchestration in E-Commerce Long-Horizon Planning

Kangjia Niu[1], Yanning Zhang[2], Xiuchong Wang[2], Chennan Ma[2], Siqi Hong[2], Hankz Hankui Zhuo[3,4*], Junxiong Zhu[2], Bo Zheng[2]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Alibaba Group
[3] State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210023, China
[4] School of Artificial Intelligence, Nanjing University, Nanjing 210023, China
niukj3@mail2.sysu.edu.cn, {zhangyanning.zyn, xiuchong.wxc, machennan.mcn, hongsiqi.hsq}@taobao.com, hankz@nju.edu.cn, xike.zjx@taobao.com, bozheng@alibaba-inc.com

Abstract

E-commerce platforms increasingly deploy automated product management systems to help sellers maximize long-term profitability through multi-day planning and scheduling of operational actions. Despite promising results in long-horizon e-commerce planning, existing LLM-based multi-agent systems with self-reflection suffer from ambiguous credit assignment across agents and shallow reflection that fails to diagnose root causes of poor operational outcomes in practice. We propose RASO, a novel multi-agent framework that enhances long-term planning via role-aware and rule-based shared reflection. RASO addresses these challenges through a hybrid reward mechanism that combines global and role-specific counterfactual rewards to enable precise credit attribution across functionally distinct agents, as well as a rule-decision paradigm that requires agents to formalize their reasoning into auditable, structured rules prior to action execution to support logic-level error diagnosis during reflection. Evaluated on a real-world e-commerce platform over extended planning horizons, RASO significantly outperforms baselines in cumulative profit with transparent and interpretable decision processes. Our results demonstrate that integrating role-aware collaboration with structured reflection empowers LLM agents to effectively manage complex, long-term business objectives.

Code — https://github.com/follow-wind-heart/RASO

1 Introduction

E-commerce platforms are increasingly deploying automated pricing systems to help sellers optimize long-term profitability through strategic price adjustments over multi-day horizons. This constitutes a challenging long-horizon sequential decision-making problem: pricing decisions have delayed and compounding effects on future demand and revenue, while optimal policies must continuously adapt to a non-stationary environment characterized by market dynamics, competitor strategies, and inventory constraints. Critically, in real-world applications, such systems must not only maximize cumulative profit but also produce transparent and interpretable pricing decisions to support seller trust, regulatory compliance, and actionable business insights.

Recent work has applied Large Language Model (LLM) to multi-agent systems with self-reflection for strategic decision-making(Shinn et al. 2023; Madaan et al. 2024; Du et al. 2023; Hong et al. 2023; Qian et al. 2024; Gao et al. 2024; Lu et al. 2025). However, these approaches suffer from two limitations. First, they exhibit ambiguous credit assignment(Foerster et al. 2017; Sunehag et al. 2018; Rashid et al. 2018): when functionally specialized agents collaborate on pricing decisions, a shared reward signal fails to disentangle individual contributions without role-aware credit decomposition. Second, their reflection mechanisms remain shallow, relying primarily on outcome-level feedback and failing to diagnose root causes of suboptimal pricing outcomes at the logic level—a critical shortcoming that impedes effective learning over extended planning horizons.

We propose RASO (Role-Aware Shared Reflection for Multi-Agent Orchestration), a novel multi-agent framework for long-horizon automated pricing that overcomes the limitations of existing LLM-based approaches through role-aware collaboration and structured reflection.

Our main contributions are as follows:

- RASO, the first multi-agent pricing framework integrating role-aware credit assignment with rule-based shared reflection for transparent, long-horizon profit growth.
- We design a hybrid counterfactual reward mechanism combining global and role-specific signals to disentangle the contributions of distinct pricing agents, resolving ambiguous credit assignment in collaborative planning.
- We introduce a rule-decision paradigm that requires agents to formalize reasoning into auditable, structured rules prior to action execution, enabling logic-level error diagnosis and ensuring interpretability by construction.
- We validate RASO through extensive real-world experiments on a large-scale e-commerce platform, demonstrating significant improvements in cumulative profit over strong baselines while maintaining decision transparency—confirming the critical role of role-awareness and structured reflection in practical pricing automation.

*Corresponding author. Email: hankz@nju.edu.cn
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

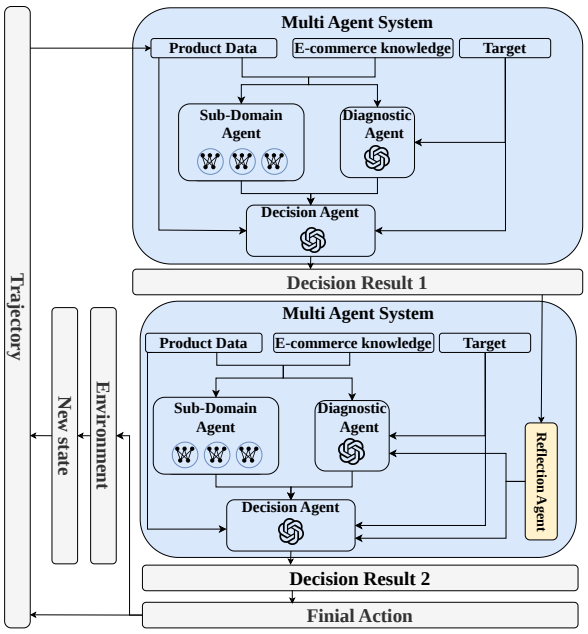

Figure 1: Overview of Our Raso Approach.

2 Related Work

While the remarkable capabilities of LLMs are well-established (Achiam et al. 2023; Yang et al. 2024; Dubey et al. 2024), these models are not without inherent limitations. They remain prone to significant issues such as producing fallacious reasoning (Turpin et al. 2024), fabricating information (hallucination) (Rawte, Sheth, and Das 2023), and generating toxic content (Zhang et al. 2024). Reflection techniques (Pan et al. 2023; Shinn et al. 2023; Madaan et al. 2024) address these issues by utilizing feedback to guide LLMs in refining their outputs. Early representative works like Self-Refine (Madaan et al. 2024) and Reflexion (Shinn et al. 2023) primarily relied on verbal feedback and memory mechanisms to avoid repeating mistakes.

More recent works have started to explore more principled optimization frameworks. For instance, RETROFORMER (Yao et al. 2023) introduced a retrospective model that learns to automatically tune an agent's prompts by learning from environmental feedback via policy gradient. This pushed reflection from simple verbal correction towards reward-based gradient optimization. Building on this, COPPER (Bo et al. 2024) extends this idea to the multi-agent collaboration domain.

3 Method

3.1 Problem Definition

Our research goal is to build a multi-agent system capable of self-optimization through interaction with its environment. The problem can be characterized by a tuple $(\mathcal{S}, \mathcal{A}, R)$, whose components are defined as follows:

At each time step t, the product state $s_t \in \mathcal{S}$ is a high-dimensional vector capturing the product's intrinsic properties, historical performance, operational settings, and environmental context. The agent takes an action $a_t \in \mathcal{A}$ to adjust controllable operational levers such as discount and ad budget. The reward $r_t = f(R_{\text{global}}, R_{\text{role}}) \in \mathcal{R}$ combines outcome-based reward with role-aware reward to maximize long-term cumulative returns.

We consider optimizing a policy $\pi(a|s)$ for a specific product to be sold on a certain e-commerce platform. In our framework, the policy π is not realized by a single model but is executed through a multi agent system. This system involves Diagnostic Agent, Sub-Domain Agent, and Decision Agent.

3.2 Multi Agent System

We formalize the operational trajectory of a product as a sequence $\tau = (s_0, a_0, s_1, a_1, \dots)$. At any given time step t, we employ a multi agent system to generate the action a_t required for the current state s_t under the operational target $Target$. The process is as follows:

- **Diagnostic Agent ($\mathcal{A}_{\text{diag}}$)**: It takes the historical trajectory up to the current state s_t, denoted as $H_t = (s_0, a_0, \dots, s_{t-1}, a_{t-1}, s_t)$, external e-commerce knowledge $\mathcal{K}_{\text{ecom}}$, and operational target $Target$ as input, to output an in-depth diagnostic report $diag_t$.

$$diag_t = f_{\text{diag}}(H_t, \mathcal{K}_{\text{ecom}}, Target)$$

- **Sub-Domain Agent ($\mathcal{A}_{\text{sub}}$)**: It also takes the historical trajectory H_t and e-commerce knowledge $\mathcal{K}_{\text{ecom}}$ as input to independently generate a sub-domain information sub_t.

$$sub_t = f_{\text{sub}}(H_t, \mathcal{K}_{\text{ecom}})$$

- **Decision Agent ($\mathcal{A}_{\text{deci}}$)**: It synthesizes all upstream outputs to craft a final plan aligned with the operational objective $Target$. It takes the diagnostic report and the sub-domain information as context to first generate explicit decision rules $rules_t$, and subsequently, a coherent decision text $deci_t$.

$$(rules_t, deci_t) = f_{\text{deci}}(s_t, diag_t, sub_t, Target)$$

Ultimately, an executable action a'_t is extracted from the decision $deci_t$. This action is referred to as the Initial Action in our framework.

3.3 Reflection

Our framework incorporates a Reflection Agent that operates within each planning cycle. Before final execution, the Decision Agent first generates an initial action along with its underlying reasoning formalized as a structured rule. The Reflection Agent then audits the rule-action pair generated by the Decision Agent and the diagnostic report from the Diagnostic Agent by evaluating their logical consistency, alignment with business objectives, and potential risks under current market conditions.

The resulting reflection $refl_t^i$ is fed back as additional context. Conditioned on this targeted feedback, the relevant agent revises its reasoning and the process produces an improved action a_t, which we refer to as the Final Action. This two-stage process ensures that decisions are not only profit-driven but also interpretable and logically sound.

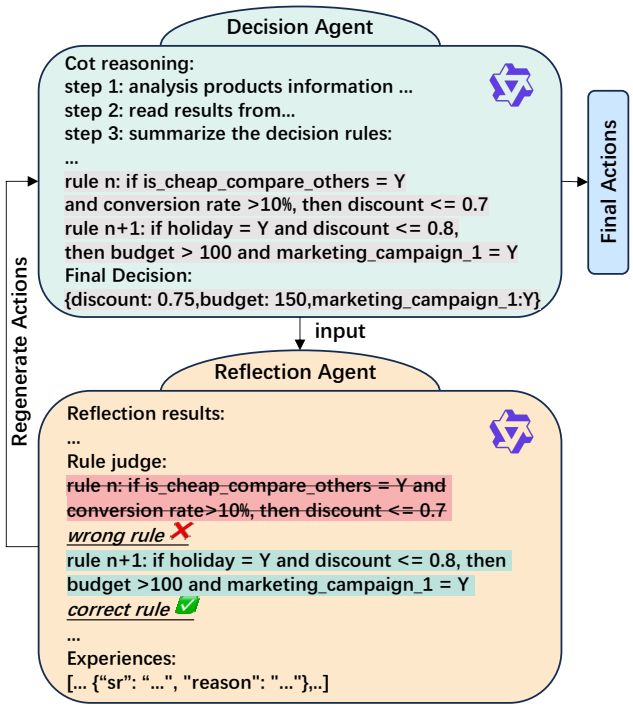

Figure 2: The rule-decision and reflection loop in RASO.

3.4 Optimization of the Reflection

Rule-decision Paradigm in Decision Agent. As shown in Figure 2, before outputting a decision, the Decision Agent externalizes its reasoning into a set of explicit IF-THEN rules. These rules provide a structured and auditable object for the subsequent reflection process, thereby shifting the reflection task from a vague critique of prose to a traceable audit of a concrete decision-making logic chain. Formally, we define a rule r as a tuple $< C, A_{\text{bounds}} >$:

$$r := \text{IF } C(s_t) \text{ THEN } a_t \in A_{\text{bounds}}$$

- $C(s_t)$ is a logical predicate over the state vector.
- A_{bounds} defines the valid action space region.

When reflecting on the Decision Agent, the Reflection Agent examines this entire decision process from multiple facets, by auditing the reasonability of the rules, verifying if the action aligns with the rule logic, and assessing the soundness of the causal chain within the textual analysis. Finally, it distills these findings into a generalizable Self-Reflection (sr), structured in JSON format.

Hybrid Counterfactual Reward Mechanism. Our reward function guides the agent to improve both external performance and internal logic. It consists of two components:

Global Reward (R_{global}): This reward measures the objective, long-term value of the final action. We use a pre-trained Q-network to compute it, which takes the state and final action as input and outputs an estimated long-term value.

The Q-network was trained on historical operational tuples, learning to predict the cumulative Gross Merchandise Volume (GMV) over a 7/14 day horizon by minimizing the Mean Squared Error (MSE). We evaluated it on a held-out expert dataset and observed a strong monotonic correlation (Spearman's rank correlation coefficient of 0.935) between the predicted Q-scores and the realized future GMV.

Role-Aware Reward (R_{role}): This reward focuses on assessing the intrinsic quality and logical coherence of the combined diagnostic or decision results, enabling role-aware credit attribution. We develop an LLM-as-Judge rubric for this task. This rubric evaluates the reasoning chain across three dimensions corresponding to our agent roles:

- Diagnostic Accuracy (S_{diag}): Evaluates if the Diagnostic Agent correctly identified the root cause.
- Logical Consistency (S_{logic}): Assesses if the Decision Agent's plan effectively solves the identified problem.
- Rule Alignment (S_{rule}): Verifies if the final action aligns with the generated rules.

The final R_{role} is calculated as the average of these sub-scores. This mechanism operationalizes credit assignment by identifying the bottleneck in the reasoning chain. We posit that the agent role with the lowest sub-score (S_{min}) is the failure point. We employ Direct Preference Optimization (DPO) (Rafailov et al. 2023), treating reflections that target the agent responsible for S_{min} as the chosen responses in preference pairs. Thus, the system is optimized to favor reflections that specifically critique this agent.

Fine-tuning the Reflection Agent. We fine-tune an open-source model, Qwen3-30B-A3B-Instruct, to serve as a professional and efficient Reflection Agent via DPO. This approach requires a preference dataset $\mathcal{D}$, where each sample consists of a prompt and a pair of better and worse reflections. The construction of $\mathcal{D}$ is detailed in Algorithm 1.

By training on the preference dataset $\mathcal{D}$, DPO aligns the Reflection Agent's behavior with our hybrid reward criteria, teaching it to generate reflections that lead to improvements in both global action value and role-aware logical quality.

Algorithm 1: Preference Dataset Construction

Require: Training trajectory H_t, multi-agent system $\mathcal{M}$, initial reflection LLM $f_{\text{init\_refl}}$
Ensure: Preference dataset $\mathcal{D} = \{(p, y_w, y_l), \dots\}$

$\mathcal{D} \leftarrow \emptyset$
$a_0 \leftarrow \mathcal{M}(H_t)$
$(R^0_{\text{global}}, R^0_{\text{role}}) \leftarrow \text{Eval}(a_0)$
$Refls \leftarrow \{refl_1, \dots, refl_N\} \sim f_{\text{init\_refl}}(H_t)$
for $k \in \{1, \dots, N\}$ **do**
  $a_k \leftarrow \mathcal{M}(H_t, refl_k)$
  $(R^k_{\text{global}}, R^k_{\text{role}}) \leftarrow \text{Eval}(a_k)$
  $\Delta R^k_{\text{global}} \leftarrow R^k_{\text{global}} - R^0_{\text{global}}$
  $\Delta R^k_{\text{role}} \leftarrow R^k_{\text{role}} - R^0_{\text{role}}$
end for
for each pair (i, j) where $i \neq j$ **do**
  if $\Delta R^i_{\text{global}} > \Delta R^j_{\text{global}}$ **and** $\Delta R^i_{\text{role}} > \Delta R^j_{\text{role}}$ **then**
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(H_t, refl_i, refl_j)\}$
  end if
end for
return $\mathcal{D}$

4 Experiment

4.1 Experimental Setup

Training Dataset. 1 million transition samples from 6 months of historical logs on a real-world e-commerce platform (used for DPO base model alignment).

Testing Dataset. a dataset of 20,000 decision instances from the operational logs of a real e-commerce platform.

Evaluation Metrics. For offline evaluation, we employ the Global Reward and Role-Aware Reward, as detailed in Section 3.4, to assess the final action's value and the system's reasoning quality. For the online test, we measure key business indicators: Item Transition Rate (ITR), Gross Merchandise Volume (GMV), and Return on Investment (ROI).

Baselines and Ablation Models. We compare against two main baselines CoT and Relfexion, and conduct an ablation study with several variants of our model.

- **Main Baselines**:
  - CoT (Wei et al. 2022): A standard Chain-of-Thought prompting approach where the agent reasons step-by-step but lacks a reflective feedback loop.
  - Reflexion (Shinn et al. 2023): A reflection baseline that generates unstructured, verbal self-critiques to guide subsequent decisions.
- **Ablation Variants**:
  - RASO (w/o Rule-Decision): Simulates verbal reflection within our framework by removing the structured rule paradigm.
  - RASO (w/o Finetuning): Uses the full RASO structure but the Reflection Agent is not fine-tuned.
  - RASO (Global Reward FT): Fine-tuned using only the global counterfactual reward (ΔR_{global}).
  - RASO (Hybrid Reward FT): Our full proposed model.

Implementation Details. All agents are built on the Qwen3-30B-A3B-Instruct (Yang et al. 2025). The Reflection Agent is fine-tuned from this base model using DPO.

4.2 Main Results

Offline Evaluation. We compare our RASO model against the CoT and Reflexion baselines.

Method	R_{global}	R_{role}
CoT	0.85	9.4183
Reflexion	0.98	9.4237
RASO	**1.16**	**9.5084**

Table 1: Main comparison against baseline methods.

The results show that while CoT provides some reasoning, its lack of a feedback loop yields poor performance. Reflexion outperforms CoT by incorporating self-correction, but its unstructured nature limits effectiveness. Our RASO framework, with its structured reflection and targeted fine-tuning, achieves the best scores on both metrics, demonstrating its significant advantage over existing methods.

Ablation Study. To dissect the sources of RASO's effectiveness, we conducted a detailed ablation study.

Method	R_{global}	R_{role}
RASO (w/o Rule-Decision)	1.09	9.4317
RASO (w/o Finetuning)	1.11	9.4342
RASO (Global Reward FT)	1.13	9.4659
RASO (Hybrid Reward FT)	**1.16**	**9.5084**

Table 2: Ablation study of the RASO framework.

The results lead to several key observations:

1. Structured rule is Essential: Removing the Rule-Decision paradigm(`w/o Rule-Decision`) causes a sharp drop in performance. This proves a structured object for reflection is critical to avoid the ambiguity of verbal critiques.
2. Fine-tuning is Necessary: The `w/o Finetuning` model, while better than unstructured methods, is still suboptimal. This indicates that a generic LLM struggles to produce high-quality reflections without being explicitly aligned with a reward signal.
3. Hybrid Reward is Crucial: Our full model (`Hybrid Reward FT`) outperforms the one tuned only on global rewards (`Global Reward FT`). The pronounced gain in R_{role} shows that the role-aware reward is vital for solving credit assignment and improving reasoning quality.

Online A/B Test Results. To bridge the gap between offline metrics and real-world value, we deployed reflection model in an online A/B test against the `CoT` baseline over one week period. The results are presented in Table 3.

Method	ITR	GMV (Rel.)	ROI (Rel.)
CoT	10.82%	Baseline	Baseline
RASO	**11.52%**	**+1.34%**	**+5.81%**

Table 3: Online A/B test results.

The online results show that the reflection led to a 0.7 percentage point increase in ITR (the percentage of items that achieve the target), a 1.34% uplift in GMV, and a 5.81% improvement in ROI. This confirms that the enhanced reasoning and decision quality observed in offline evaluations successfully translate into tangible, significant business value.

5 Conclusion

We presented RASO, a framework for continuous self-improvement in LLM-based multi-agent systems. By combining a Rule-Decision paradigm with a hybrid-reward DPO objective, RASO addresses both credit assignment and shallow reflection in long-horizon decision-making. Experiments on a real-world e-commerce platform show improved offline metrics and online business performance, indicating that structured, role-aware reflection can make multi-agent systems more reliable and interpretable. Future work will incorporate action model learning into RASO (Zhuo, Nguyen, and Kambhampati 2013), reduce reflection overhead, and extend the framework to more complex domains.

Acknowledgments

This work was supported by Alibaba Group through Alibaba Innovative Research Program and Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bo, X.; Zhang, Z.; Dai, Q.; Feng, X.; Wang, L.; Li, R.; Chen, X.; and Wen, J.-R. 2024. Reflective Multi-Agent Collaboration based on Large Language Models. In *NeurIPS*.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Foerster, J. N.; Farquhar, G.; Afouras, T.; Nardelli, N.; and Whiteson, S. 2017. Counterfactual Multi-Agent Policy Gradients. In *AAAI*.

Gao, S.; Wen, Y.; Zhu, M.; Wei, J.; Cheng, Y.; Zhang, Q.; and Shang, S. 2024. Simulating Financial Market via Large Language Model based Agents. *ArXiv*, abs/2406.19966.

Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z.; et al. 2023. MetaGPT: Meta programming for a multi-agent collaborative framework. In *ICLR*.

Lu, X.; Qiu, J.; Yang, Y.; Zhang, C.; Lin, J.; and An, S. 2025. Large Language Model-Based Bidding Behavior Agent and Market Sentiment Agent-Assisted Electricity Price Prediction. *IEEE Transactions on Energy Markets, Policy and Regulation*, 3: 223–235.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. 2024. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*, volume 36.

Pan, L.; Saxon, M.; Xu, W.; Nathani, D.; Wang, X.; and Wang, W. Y. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.

Qian, C.; Liu, W.; Liu, H.; Chen, N.; Dang, Y.; Li, J.; Yang, C.; Chen, W.; Su, Y.; Cong, X.; et al. 2024. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15174–15186.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Rashid, T.; Samvelyan, M.; Schroeder, C.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *ICML*, 4295–4304.

Rawte, V.; Sheth, A.; and Das, A. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Shinn, N.; Cassano, F.; Berman, E.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv:2303.11366.

Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W. M.; Zambaldi, V. F.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J. Z.; Tuyls, K.; and Graepel, T. 2018. Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward. In *Adaptive Agents and Multi-Agent Systems*.

Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *NeurIPS*, volume 36.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, volume 35, 24824–24837.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Zhou, J.; Lin, J.; Dang, K.; Bao, K.; Yang, K.; Yu, L.; Deng, L.; Li, M.; Xue, M.; Li, M.; Zhang, P.; Wang, P.; Zhu, Q.; Men, R.; Gao, R.; Liu, S.; Luo, S.; Li, T.; Tang, T.; Yin, W.; Ren, X.; Wang, X.; Zhang, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Wang, Z.; Cui, Z.; Zhang, Z.; Zhou, Z.; and Qiu, Z. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yao, W.; Heinecke, S.; Niebles, J. C.; Liu, Z.; Feng, Y.; Xue, L.; Murthy, R.; Chen, Z.; Zhang, J.; Arpit, D.; et al. 2023. Retroformer: Retrospective large language agents with policy gradient optimization. *arXiv preprint arXiv:2308.02151*.

Zhang, J.; Wu, Q.; Xu, Y.; Cao, C.; Du, Z.; and Psounis, K. 2024. Efficient toxic content detection by bootstrapping and distilling large language models. In *AAAI*, volume 38, 21779–21787.

Zhuo, H. H.; Nguyen, T. A.; and Kambhampati, S. 2013. Model-Lite Case-Based Planning. In desJardins, M.; and Littman, M. L., eds., *AAAI*, 1077–1083.