StanceAgent-R1: Agentic Reinforcement Learning for Zero-Shot Conversational Stance Detection

Xing Li, Jinzhong Ning*, Yijia Zhang
School of Information Science and Technology, Dalian Maritime University, Dalian, China
{lllxxx,ningjinzhong,zhangyijia}@dlmu.edu.cn

Abstract

Conversational stance detection requires identifying the stance of a current utterance toward a target within its dialogue context. Existing methods often rely on sequential encoding, fixed knowledge-integration pipelines, or uniform reasoning strategies, which limits their ability to capture cross-turn dependencies and adapt to instance-specific needs. We propose **StanceAgent-R1**, an agentic reinforcement learning framework that adaptively routes each instance to evidence-path selection and/or knowledge elicitation. The model is first warm-started with rejection-sampled expert trajectories and then optimized with a progressive trajectory-faithful reward that encourages correct, well-formed, and route-consistent reasoning. Experiments on ZS-CSD, MT-CSD, and C-MTCSD show that StanceAgent-R1 achieves state-of-the-art performance, with further analyses confirming the effectiveness of its agentic design and reward strategy. Our code is publicly available at https://github.com/lllxxxxx/StanceAgent-R1.

1 Introduction

Conversational stance detection (CSD) (Li et al., 2023b; Niu et al., 2024b) aims to identify the stance (e.g., Favor, Against, or Neutral) of the current utterance toward a specific target, given the conversation history. CSD has been widely applied to downstream tasks such as public opinion analysis (Stefanov et al., 2020), rumor detection (Wei et al., 2019), and argument mining (Bar-Haim et al., 2017), serving as a key research direction for analyzing online discourse.

Early research on stance detection mainly focused on individual sentences and is generally divided into in-target (targets at inference stage are seen during training) (Mohammad et al., 2016), cross-target (transferring across related targets) (Xu et al., 2018; Wei and Mao, 2019), and zero-shot (generalizing to targets unseen) (Allaway and McKeown, 2020). However, on social media, users often express their opinions through conversational exchanges rather than isolated statements, where individual sentence analysis fails to capture inter-speaker dependencies and reply structures. To address this, CSD has been introduced, with research extending from early short-turn replies (Villa-Cox et al., 2020) to multi-turn conversations (Niu et al., 2024b,a). More recently, Ding et al. (2025) propose zero-shot conversational stance detection (ZS-CSD), extending CSD to unseen targets, which better aligns with real-world scenarios where new targets continually emerge.

Despite recent progress, existing methods still

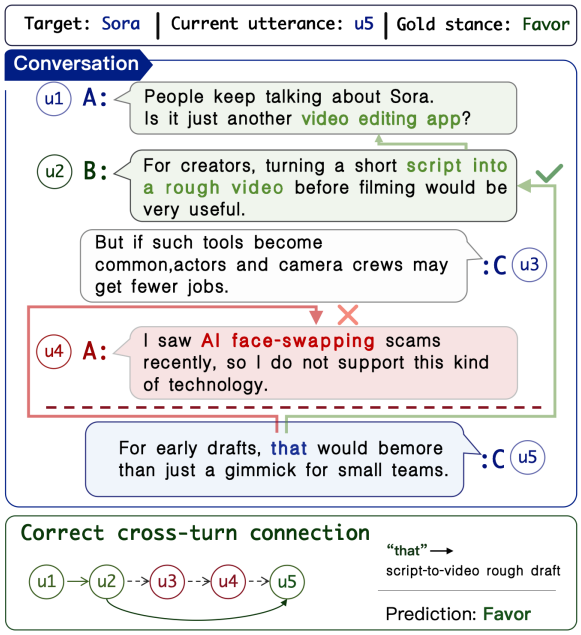


Figure 1: An example of conversational stance detection toward the target *Sora*. The pronoun "that" in u_5 actually refers to the non-adjacent turn u_2, while sequential encoding tends to link it to the adjacent u_4 and produces a wrong stance. This motivates modeling cross-turn dependencies explicitly.

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 23895–23919
October 24-29, 2026 ©2026 Association for Computational Linguistics

exhibit several issues. **Issue 1: Sequential encoding ignores cross-turn dependencies.** Most existing methods (Niu et al., 2024b; Ding et al., 2025; Wang et al., 2025; Niu et al., 2025a) encode conversation history by concatenating utterances in temporal order. This leads models to rely on linear temporal order and ignore cross-turn dependencies that commonly exist in conversations. As illustrated in Figure 1, the opinion in the 5th turn toward the target is actually grounded in a key attribute mentioned in the 2nd turn, not in the unrelated chat in the adjacent 4th turn. Such sequential input construction makes models easily distracted by irrelevant temporal information and limits their ability to build effective semantic connections by skipping irrelevant turns, thereby potentially reducing stance detection performance.

Issue 2: Pipeline knowledge integration prevents joint optimization. To improve generalization to unseen targets in zero-shot settings, existing methods typically incorporate external knowledge in a pipeline manner: they first acquire knowledge from a separate knowledge acquisition module (e.g., by prompting LLMs (Zhang et al., 2024; Taranukhin et al., 2024; Ma et al., 2024) or querying search engines (Li et al., 2023a)), and then use the acquired knowledge to assist stance prediction. However, in such a pipeline the knowledge acquisition stage receives no feedback from the downstream prediction stage, so the acquired knowledge may not align with what the task actually needs, ultimately degrading stance detection performance.

Issue 3: One-size-fits-all ignores instance-specific needs. Most existing methods either leverage fine-grained stance cues (Niu et al., 2024b; Ding et al., 2025) or introduce external knowledge (Wang et al., 2025; Niu et al., 2025a) to better handle the complexity of multi-turn stance reasoning. However, both lines of work adopt a one-size-fits-all processing strategy: uniformly mining fine-grained stance cues or injecting external knowledge across all instances. In practice, different instances may have different needs: utterances with context-dependent stance (e.g., pronouns, agreement, or irony) require fine-grained stance cues mined from the conversation history; utterances mentioning unfamiliar entities or concepts require external knowledge; and utterances with explicit stance words can be judged from the current turn alone. This one-size-fits-all processing strategy may introduce mismatched information to each instance, interfering with stance judgment and degrading detection performance.

We propose **StanceAgent-R1** to address these issues: an agentic RL framework organized around three role-specialized agents, namely the *Routing Agent*, *Evidence Path Agent*, and *Knowledge Elicitation Agent*. **For Issue 1**, the Evidence Path Agent explicitly selects stance-relevant turns from the conversation history to capture cross-turn dependencies. **For Issue 2**, the Knowledge Elicitation Agent elicits utterance related background knowledge under downstream task feedback. **For Issue 3**, the Routing Agent decides per instance which downstream agents to invoke, replacing one-size-fits-all reasoning with an interpretable decision trace. We train StanceAgent-R1 in two stages: behavior cloning on rejection-sampled expert trajectories, followed by RL under a progressive trajectory-faithful reward over answer correctness, format, and trace legality, where the legality term is activated only after base accuracy stabilizes.

The main contributions of this paper are as follows:

- We propose **StanceAgent-R1**, an agentic RL framework that adaptively routes each instance to evidence selection and/or knowledge elicitation for conversational stance detection.
- We introduce a progressive trajectory-faithful reward that optimizes answer correctness, format validity, and trace-route consistency to encourage faithful agentic reasoning.
- Experiments on ZS-CSD, MT-CSD, and C-MTCSD show that StanceAgent-R1 achieves state-of-the-art performance.

2 Related Work

2.1 Conversational Stance Detection

Early stance detection operates at the sentence level, covering in-target (Mohammad et al., 2016; Augenstein et al., 2016; Li et al., 2023b), cross-target (Xu et al., 2018; Wei and Mao, 2019), and zero-shot (Allaway and McKeown, 2020; Liang et al., 2022; Taranukhin et al., 2024; Weinzierl and Harabagiu, 2024; Zhang et al., 2025; Ji et al., 2025) settings.

As social media users increasingly express opinions through multi-turn exchanges, recent research extends the task to conversational scenarios. Prior CSD methods primarily design specialized neural architectures to capture multi-turn conversational

dependencies. GLAN (Niu et al., 2024b) combines global attention with CNN and GCN branches to model both long-range and local conversational patterns. SITPCL (Ding et al., 2025) formalizes the ZS-CSD task and couples a speaker interaction network with target-aware prototypical contrastive learning. More recent work injects LLM-derived signals to enrich conversation reasoning: MRGKD (Wang et al., 2025) distills a multi-agent LLM reasoning graph into a smaller student model, and LLM-CRAN (Niu et al., 2025a) fuses LLM-elicited logical relations and conversational acts with the conversation history via a relational attention network. **Unlike these methods, which apply specialized architectures or prompted LLMs under a fixed reasoning pipeline, we fine-tune an LLM via reinforcement learning into an agentic policy that adapts its reasoning to each instance.**

2.2 Large Language Model Agents

LLM agents tackle complex tasks through multi-step reasoning and action. Existing approaches fall broadly into two categories.

Prompting-based LLM Agents. Building on chain-of-thought (Wei et al., 2022) and the ReAct (Yao et al., 2022) paradigm, some studies prompt a single LLM to acquire external knowledge (Li et al., 2023a; Ma et al., 2024), while others coordinate multiple LLM agents through debate or role assignment (Du et al., 2024; Lan et al., 2024; Wang et al., 2024; Zhang et al., 2026b). Current LLM-based methods for conversational stance detection also fall into this category, where each stage runs in isolation without end-to-end feedback from the final prediction.

Agentic Reinforcement Learning. Inspired by DeepSeek-R1's RL recipe (Guo et al., 2025), recent studies train LLM agents end-to-end to interleave reasoning with search engine calls (Jin et al., 2025), code interpreters (Feng et al., 2025), and general tool use (Qian et al., 2026). Subsequent works refine the RL recipe with step-level credit assignment (Feng et al., 2026), decompose the agent into specialized modules and optimize the planner via in-the-flow RL (Li et al., 2025), or learn multi-round routing across several LLMs via rule-based RL (Zhang et al., 2026a). To our knowledge, our work is the first to bring this paradigm to conversational stance detection.

3 Method

3.1 Problem Formulation

Zero-shot conversational stance detection (ZS-CSD) aims to determine the stance of the current utterance toward a given target, leveraging both the conversation history and speaker information. Formally, given a target t and an n-turn conversation $\mathcal{C} = \{(s_i, u_i)\}_{i=1}^{n}$, where u_i denotes the i-th utterance and s_i its corresponding speaker, the task is to predict the stance polarity $y \in \mathcal{Y} = \{\text{FAVOR}, \text{AGAINST}, \text{NEUTRAL}\}$ that the current utterance u_n expresses toward t, based on the conversation history $\{(s_i, u_i)\}_{i=1}^{n-1}$. The zero-shot setting further requires that targets in the test set do not appear in the training set.

3.2 StanceAgent-R1

This section presents **StanceAgent-R1**, an agentic reinforcement learning framework comprising three cooperative agents: a Routing Agent, an Evidence Path Agent, and a Knowledge Elicitation Agent. As illustrated in Figure 2, we describe the agentic trajectory design (§3.2.1) followed by two training stages, Expert-Guided Behavior Cloning (§3.2.2) and Trajectory-Faithful Reinforcement Learning (§3.2.3).

3.2.1 Agentic Trajectory Design

Role-Specialized Agents. The three agents introduced in §1 cooperate within a single rollout: the Routing Agent decides which downstream agents to invoke, the Evidence Path Agent selects stance-relevant history turns, and the Knowledge Elicitation Agent supplies background for entities not explained by the dialogue. Emitting all three in one rollout yields a single-rollout agentic RL framework, avoiding the long-horizon credit assignment of multi-turn paradigms (Jin et al., 2025; Song et al., 2025).

Expert Trajectory Curation. The cold-start corpus consists of expert trajectories that follow a six-field schema. Each trajectory takes the form $\tau = \langle$<plan>, <route>, <select_path>?, <knowledge>?, <think>, <answer>$\rangle$, where ? marks fields populated only when the route requires them. The Routing Agent produces <plan>, which analyzes the current utterance to identify which downstream agents are needed, and <route>, which specifies the subset of the two downstream agents to invoke: DIRECT (neither), RETRIEVE (Evidence Path only), KNOWLEDGE

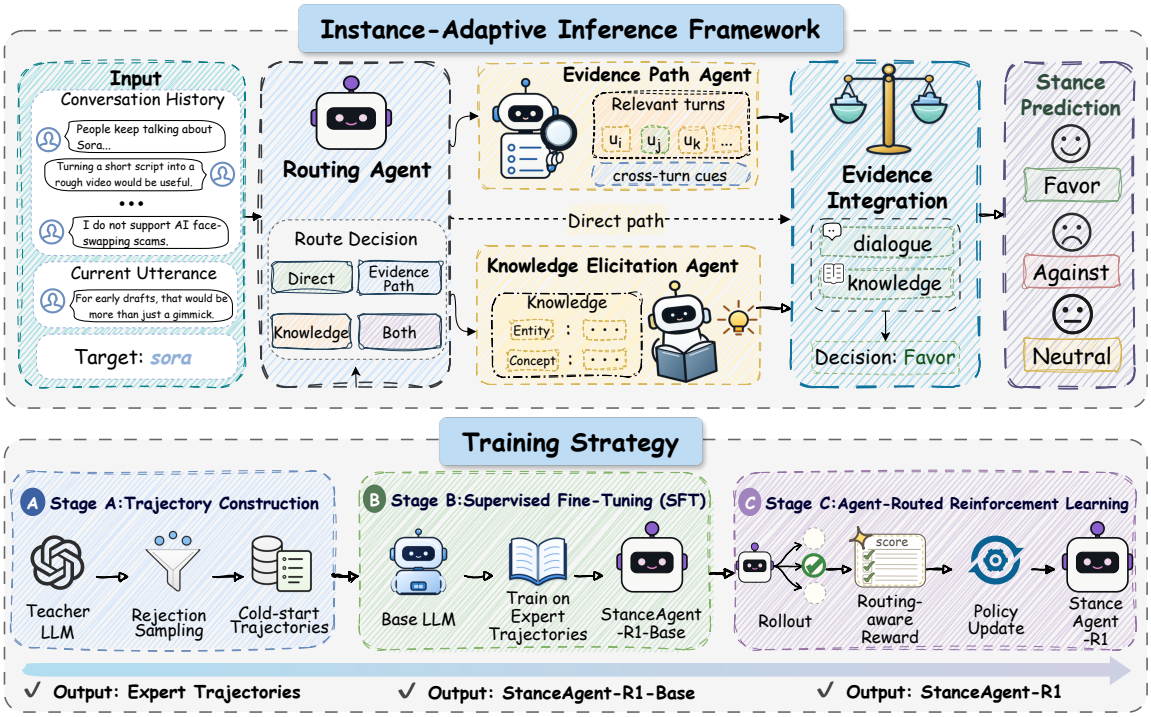


Figure 2: Overview of StanceAgent-R1. Top: the instance-adaptive inference framework, where the Routing Agent decides which downstream agents to invoke per instance. Bottom: the three training stages, including **(A)** cold-start trajectory construction via rejection sampling on a teacher LLM, **(B)** supervised fine-tuning on $\mathcal{D}_{\text{cold}}$ to obtain **StanceAgent-R1-Base**, and **(C)** reinforcement learning with GRPO.

(Knowledge Elicitation only), or BOTH. Conditioned on <route>, <select_path> lists indices of stance-relevant history turns and <knowledge> provides short factual background; <think> integrates the gathered evidence and <answer> outputs the predicted stance $\hat{y}$. We populate this schema by prompting a frozen teacher LLM under a label-blind protocol and keep only candidates that conform to the schema and produce the gold label, yielding $\mathcal{D}_{\text{cold}}$ (Figure 2A; the teacher prompt template is in Appendix J, and a complete example is in Appendix K).

3.2.2 Expert-Guided Behavior Cloning

RL fine-tuning on the base LLM is unstable: most rollouts receive zero reward and the advantage estimator collapses. We therefore warm-start the policy with behavior cloning on $\mathcal{D}_{\text{cold}}$ before RL. The policy π_θ is trained to imitate the expert trajectories via standard token-level cross-entropy:

$$\mathcal{L}_{\text{BC}}(\theta) = -\,\mathbb{E}_{(x_i,\tau_i^{\text{exp}})\sim\mathcal{D}_{\text{cold}}}\left[\log \pi_\theta(\tau_i^{\text{exp}} \mid x_i)\right].$$

The resulting policy, denoted **StanceAgent-R1-Base**, serves as both the initial policy and the KL reference π_{ref} for the RL stage (§3.2.3).

3.2.3 Trajectory-Faithful Reinforcement Learning

GRPO Objective. We optimize π_θ with Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which replaces the value critic with a within-group baseline. For each input x, we sample G trajectories $\{\tau_i\}_{i=1}^{G}$ from $\pi_{\theta_{\text{old}}}$ and compute the advantage $\hat{A}_i$ by standardizing $R(\tau_i)$ against the within-group mean and standard deviation. The policy is updated by maximizing

$$\mathcal{J}(\theta) = \mathbb{E}\Big[\min\big(\rho_i \hat{A}_i,\ \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)\,\hat{A}_i\big)\Big],$$

with $\rho_i = \pi_\theta(\tau_i)/\pi_{\theta_{\text{old}}}(\tau_i)$. Following recent practice (Yu et al., 2026; Liu et al., 2025), we drop the KL penalty since clipping already constrains step size.

Trajectory-Faithful Reward. A reward that checks only the final stance label and the surface output format is insufficient for agentic trajectories. Such a reward can be satisfied by trajectories whose declared <route> disagrees with the evidence or knowledge actually used in <think>, as long as the final <answer> is correct and the tags are well formed. To discourage this shortcut, we define a trajectory-faithful reward with progressive activation. It treats format validity as a prerequisite and

gradually introduces a trace-legality constraint:

$$R_t(\tau) = R_{\text{ans}}(\tau) + \alpha_{\text{fmt}} R_{\text{fmt}}(\tau) + g_t \alpha_{\text{legal}} R_{\text{legal}}(\tau),$$

where $R_{\text{fmt}}, R_{\text{ans}}, R_{\text{legal}} \in \{0, 1\}$. $R_{\text{fmt}} = 1$ iff τ follows the required schema and tag order; if parsing fails we set $R_{\text{ans}} = R_{\text{fmt}} = R_{\text{legal}} = 0$, so the whole reward is zero and format validity acts as a prerequisite. $R_{\text{ans}} = 1$ iff the parsed <answer> equals the gold label y. R_{legal} measures whether the reasoning trace is consistent with the declared route. The gate $g_t \in \{0, 1\}$ implements progressive activation: it is set to 0 at the beginning of RL and switched to 1 once the validation macro-F1 first reaches η_{gate}.

Legality Reward. $R_{\text{legal}} = 1$ when τ satisfies all four trace-route consistency conditions:

- *Route honesty*: <select_path> and <knowledge> appear in τ iff <route> requires them.
- *Path validity*: <select_path> contains only indices of existing history turns.
- *Conclusion alignment*: the terminal stance phrase in <think> agrees with <answer>.
- *Knowledge grounding*: when <route> invokes <knowledge>, <think> reuses the elicited content in at least three distinct character 5-grams.

These conditions are evaluated only when the progressive gate g_t is active.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate StanceAgent-R1 on three conversational stance detection benchmarks covering both zero-shot and in-target settings.

ZS-CSD (Ding et al., 2025) is a large-scale Chinese zero-shot CSD benchmark from Sina Weibo, with 280 targets across noun-phrase and claim types.

MT-CSD (Niu et al., 2024b) is the largest English multi-turn CSD dataset to date, with 15,876 Reddit instances spanning five targets.

C-MTCSD (Niu et al., 2025b) is the largest Chinese multi-turn CSD dataset to date, with 24,264 Sina Weibo instances.

Following Ding et al. (2025); Niu et al. (2024b), we adopt Macro-F1 averaged over the three stance classes as the evaluation metric.

4.2 Experimental Settings

We initialize StanceAgent-R1 from Qwen2.5-7B-Instruct, first cold-start-tuned for one epoch on $\mathcal{D}_{\text{cold}}$ generated by GPT-4o. RL post-training uses a learning rate of 5×10^{-6}, $G = 16$ sampled rollouts per prompt, and KL coefficient $\beta = 0$ following Yu et al. (2026); Liu et al. (2025). The trajectory-faithful reward adds answer correctness, format, and legality with format and legality weights $\alpha_{\text{fmt}} = 0.4$ and $\alpha_{\text{legal}} = 0.2$, where the legality term activates once the validation Macro-F1 first reaches $\eta_{\text{gate}} = 0.6$. Full implementation details are provided in Appendix A.

4.3 Baselines

We compare StanceAgent-R1 with several state-of-the-art baselines categorized into three groups.

PLM-based fine-tuning methods. We include RoBERTa (Liu et al., 2019), Branch-BERT (Li et al., 2023b), and GLAN (Niu et al., 2024b). Additionally, we compare with the current state-of-the-art model SITPCL (Ding et al., 2025).

LLM-based prompting methods. We compare against representative large language models in a zero-shot setting to evaluate their inherent CSD reasoning capability without parameter updates. These include closed-source models like GPT-3.5 and GPT-4o-mini, and open-source Llama-3-8B, Llama-3-70B (Grattafiori et al., 2024), and Qwen2.5-14B.

LLM-based SFT methods fine-tune Qwen2.5 at 1.5B/3B/7B and Llama-2-7B under label-only SFT, and additionally fine-tune Qwen2.5 at 3B/7B under CoT-SFT distilled from GPT-4o.

LLM-based RL methods. Qwen2.5-3B-CoT-SFT+GRPO and Qwen2.5-7B-CoT-SFT+GRPO warm-start with CoT-SFT and then apply GRPO under an answer-only reward.

4.4 Experimental Results

4.4.1 Main Results

Table 1 reports the main results on the ZS-CSD benchmark. StanceAgent-R1 consistently achieves the best performance across mixed, noun-phrase, and claim targets. We draw the following observations.

(1) LLM prompting methods substantially trail LLM-based SFT methods, and scaling up the prompted backbone brings little gain, indicating that task-specific fine-tuning, not scale alone, is

Method	Mixed targets				Noun-phrase targets				Claim targets			
	Favor	Against	Neutral	All	Favor	Against	Neutral	All	Favor	Against	Neutral	All
Llama3-8B	35.09	39.34	33.29	35.91	38.31	42.68	37.13	39.37	33.26	37.44	30.68	33.79
GPT4o-mini	24.75	38.63	51.10	38.16	20.78	46.86	52.99	40.21	26.87	32.77	50.07	36.57
GPT3.5	28.04	38.89	53.84	40.25	38.41	48.79	54.84	47.35	20.99	32.29	53.32	35.53
Llama3-70B	40.81	38.04	44.37	41.07	48.77	50.40	47.00	48.72	35.96	30.98	42.75	36.56
Qwen2.5-14B	33.84	44.01	50.48	42.78	27.67	50.56	55.58	44.60	36.86	40.23	46.93	41.34
Roberta	37.37	38.94	44.50	40.27	37.50	42.57	48.31	42.79	37.30	36.64	42.20	38.71
Branch-BERT	43.11	38.19	41.97	41.09	46.29	43.73	43.41	44.48	41.43	34.34	41.13	38.96
GLAN	43.68	34.91	46.79	41.79	45.91	37.65	51.58	45.04	42.43	33.18	43.88	39.83
SITPCL	41.03	41.34	49.07	43.81	40.99	47.14	54.49	47.54	41.06	37.34	46.02	41.47
Qwen2.5-1.5B-SFT	45.01	38.76	53.88	45.88	40.93	39.62	54.98	45.17	49.19	44.39	52.88	48.82
Qwen2.5-3B-SFT	49.22	48.71	55.42	51.11	50.69	49.12	56.46	52.09	48.44	47.72	51.89	49.35
Qwen2.5-7B-SFT	51.36	52.00	55.73	53.03	47.79	53.29	55.50	52.19	54.92	50.71	55.96	53.86
Llama2-7B-SFT	47.36	49.00	50.83	49.06	45.89	51.49	48.23	48.54	48.54	47.79	56.45	50.93
Qwen2.5-3B-CoT-SFT	48.34	47.86	54.21	50.14	49.51	48.04	55.32	50.96	47.32	46.58	50.85	48.25
Qwen2.5-7B-CoT-SFT	50.18	50.92	54.85	51.98	46.68	52.13	54.42	51.08	53.82	49.62	54.83	52.76
Qwen2.5-3B-CoT-SFT+GRPO	51.25	50.84	55.18	52.42	50.83	49.91	54.75	51.83	50.62	49.21	53.42	51.08
Qwen2.5-7B-CoT-SFT+GRPO	53.78	54.42	57.36	55.19	52.62	54.83	56.74	54.73	54.32	53.85	56.16	54.78
StanceAgent-R1-Base	46.32	46.94	51.16	48.14	45.84	45.32	50.84	47.33	47.21	45.46	52.20	48.29
StanceAgent-R1	**58.21**	**59.46**	**60.35**	**59.34**	**58.93**	**60.31**	**57.16**	**58.80**	**60.86**	**58.21**	**59.52**	**59.53**

Table 1: Main results on the ZS-CSD test set. We report per-class F1 (Favor / Against / Neutral) and macro-F1 (All) over three target types: mixed, noun-phrase, and claim. Best score in each column is in **bold**.

what teaches LLMs the reasoning patterns of conversational stance detection.

(2) Compared with PLM-based fine-tuning methods, LLM-based SFT methods achieve clearly better performance. We further observe that PLM-based methods score noticeably lower on claim than on noun-phrase targets, while fine-tuned LLM methods close this gap. This indicates that LLM backbones provide stronger semantic understanding, which is crucial for stance inference toward unseen and complex claim-style targets.

(3) Compared with the corresponding label-only SFT model, CoT-SFT slightly underperforms under the same backbone. This suggests that CoT-SFT only learns to imitate the teacher's reasoning behavior without acquiring genuine reasoning capability, and the CoT tokens act as noise that distracts the label signal during fine-tuning.

(4) Compared with CoT-SFT, CoT-SFT+GRPO consistently improves macro-F1 on both 3B and 7B backbones. This indicates that GRPO with an answer-correctness reward still provides a useful learning signal on top of CoT supervision, although the gain becomes more limited once the SFT baseline is already strong, suggesting that RL alone is not sufficient to fully activate the reasoning capacity carried in the CoT trajectories.

(5) StanceAgent-R1 still outperforms Qwen2.5-7B-CoT-SFT+GRPO by a clear margin despite sharing the same backbone, RL algorithm, and answer reward. This indicates that the gain comes from the agentic schema and the trajectory-faithful reward rather than from RL alone. Within our method, the cold-start variant StanceAgent-R1-Base trails the strongest standard SFT baseline, since the agentic schema is harder to imitate from limited expert trajectories; the subsequent RL stage then turns this richer schema into the main gain, lifting macro-F1 by more than ten points across every split. **It also removes the Neutral bias still present in the cold-start variant and most CSD baselines (Ding et al., 2025), bringing the three per-class scores close to each other.**

Method	Mixed targets			
	Favor	Against	Neutral	All
StanceAgent-R1	**58.21**	**59.46**	**60.35**	**59.34**
Agent components				
w/o Evidence Path	55.34	57.91	57.32	56.86
w/o Knowledge Elicit.	57.13	58.94	58.55	58.21
w/o Routing	55.42	57.41	58.45	57.09
Training pipeline				
w/o SFT	30.06	33.51	34.78	32.78
w/o RL	46.32	46.94	51.16	48.14
Reward design				
w/o R_{legal}	57.49	59.33	59.65	58.82
w/o R_{format}	55.78	58.21	57.85	57.28
w/o $R_{\text{progressive}}$	53.61	56.13	56.67	55.47

Table 2: Ablation on the ZS-CSD mixed-target split, with per-class F1 and macro-F1 (All). *w/o Routing* forces both downstream agents on every instance; *w/o SFT* skips cold-start and runs RL directly on the base LLM. Best score per column in **bold**.

4.4.2 Ablation Studies

Table 2 reports the ablation results on the mixed-target split of ZS-CSD. Among agent components, removing the Evidence Path Agent causes the largest drop, confirming the importance of explicitly selecting stance-relevant turns for cross-turn dependency modeling; removing the Knowledge Elicitation Agent also hurts performance, indicating that utterance-related knowledge aids generalization to unseen targets. The w/o Routing variant underperforms despite invoking both downstream agents, verifying that a fixed reasoning pipeline is suboptimal and may introduce mismatched information. For training, skipping cold-start SFT causes the largest overall drop and removing RL also substantially degrades performance, showing that expert trajectories provide a stable initial policy and RL further optimizes routing and trajectory-faithful reasoning. Finally, removing any reward component hurts performance, with the largest drop coming from the progressive mechanism, indicating that gradually activating the legality constraint is important for stable RL training.

4.4.3 Trajectory Quality Evaluation

To check whether StanceAgent-R1's accuracy is backed by an informative reasoning trace rather than just a correct final label, we evaluate the trajectory in two complementary ways, covering external multi-judge ratings and causal faithfulness under targeted edits. We further test whether the trace alone, with `<answer>` removed, lets other models recover the correct stance; see Appendix D.

Multi-Dimensional Judgment. We grade each trajectory along six axes (planning, routing, evidence-path, knowledge factuality, reasoning coherence, and conclusion support) with three LLM judges (GPT-4o, Claude-Opus-4.1, and Gemini-2.5-Pro) on a 5-point Likert scale. Each cell in Table 3 is averaged across the three judges, and the bottom row reports the per-axis Fleiss' κ, all above 0.66, indicating substantial inter-judge agreement and supporting the averaging. StanceAgent-R1 leads on every axis and outperforms the GPT-4o teacher on reasoning coherence and conclusion support, the two dimensions most directly shaped by R_{legal}. To rule out LLM-judge bias, three human annotators rate 100 sampled R1 trajectories on the same axes (👤 row of Table 3); their scores closely track the LLM judges and preserve the same per-axis pattern. This suggests the legality reward goes beyond enforcing output format and teaches the model to ground its conclusion in the elicited evidence.

Method	Pl.	Rt.	Pa.	Kn.*	Rs.	Cn.	Avg
🤖 GPT-4o	4.4	4.1	4.2	4.5	4.0	4.3	4.3
🤖 R1-Base	4.0	3.9	3.8	4.0	3.7	3.9	3.9
🤖 **StanceAgent-R1**	**4.5**	**4.5**	**4.6**	**4.4**	**4.6**	**4.6**	**4.5**
Fleiss' κ	*.71*	*.78*	*.69*	*.82*	*.66*	*.74*	–
👤 **StanceAgent-R1**	**4.4**	**4.3**	**4.6**	**4.3**	**4.4**	**4.4**	**4.3**

Table 3: Trajectory quality along Plan (Pl.), Route (Rt.), Path (Pa.), Knowledge (Kn.), Reasoning (Rs.), and Conclusion (Cn.). 🤖: avg of three LLM judges; 👤: avg of three human annotators on 100 sampled R1 trajectories. The *Fleiss' κ* row reports per-axis Fleiss' κ across the three LLM judges. *Only on traces invoking `<knowledge>`.

Faithfulness via Causal Intervention. A high judge score still leaves open whether the model truly uses what its trace declared. We test this with three interventions (Figure 3). We remove the turn cited by `<select_path>` against a random-turn control; we corrupt `<knowledge>` into a plausible falsehood; and we flip the `<route>` tag between RETRIEVE and DIRECT while checking whether `<think>` still references turn indices. Removing the selected turn drops accuracy by an order of magnitude more than removing a random one for our model, with a much narrower gap for the SFT-only variant, showing that the declared selection actually drives the prediction. Knowledge corruption flips the answer in most cases for our model but only sporadically for the baselines, so the model conditions on the elicited fact rather than on the dialogue alone. After a route flip, our model rarely keeps referencing turn indices while the baselines continue to do so. Together these results show that the legality reward yields a causal, not just surface-level, faithfulness.

4.4.4 In-Target Evaluation

We further evaluate StanceAgent-R1 on the in-target splits of MT-CSD and C-MTCSD. As shown in Table 4, StanceAgent-R1 attains the best score on every target of both datasets, and the gap to prior methods widens on the targets they find hardest, such as Joe Biden on MT-CSD and Non-marriage Doctrine on C-MTCSD. Stance cues for these targets are sparse or spread across distant turns; instance-adaptive routing recovers them by selectively invoking evidence path selection or knowledge elicitation, while fixed pipelines tend to miss

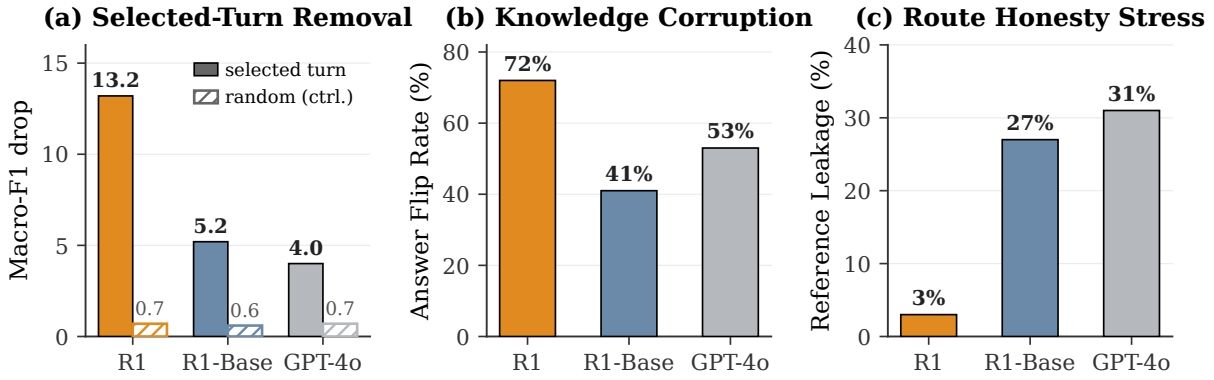


Figure 3: Causal interventions on the trajectory components of StanceAgent-R1 (R1), StanceAgent-R1-Base (R1-Base), and the GPT-4o teacher. (a) Macro-F1 drop after removing the turn cited by <select_path>, compared against a random unselected turn. (b) Answer flip rate after corrupting <knowledge> with a plausible falsehood; higher means more faithful. (c) Reference leakage after swapping the <route> tag; lower means more honest.

MT-CSD						
Method	BC	TS	SX	JB	DT	**Avg**
TTS	53.60	46.08	52.41	31.23	44.41	45.55
KPT	53.22	46.67	52.65	32.22	43.97	45.75
GPT-4	49.39	50.71	55.34	45.09	40.33	48.17
GLAN	56.95	52.38	55.98	38.15	48.91	50.47
Qwen2.5-7B-SFT	67.41	64.83	63.27	56.74	61.91	62.83
StanceAgent-R1	**72.83**	**73.14**	**70.52**	**68.78**	**69.96**	**71.05**
C-MTCSD						
Method	i15	AG	PM	NR	ND	**Avg**
KEPrompt	52.85	64.63	64.17	70.13	51.34	60.62
RoBERTa	50.93	66.06	65.25	72.40	48.67	60.66
XLNet	61.92	71.78	64.38	75.39	53.19	65.33
GLAN	63.20	73.05	66.52	71.45	56.27	66.10
Qwen2.5-7B-SFT	77.18	82.46	81.83	83.04	74.27	79.76
StanceAgent-R1	**86.41**	**89.72**	**90.16**	**88.39**	**87.07**	**88.35**

Table 4: In-target results on MT-CSD (BC, TS, SX, JB, DT) and C-MTCSD (i15, AG, PM, NR, ND) with macro-F1 (%). Baseline numbers are taken from the corresponding source papers. Best score in each column is in **bold**.

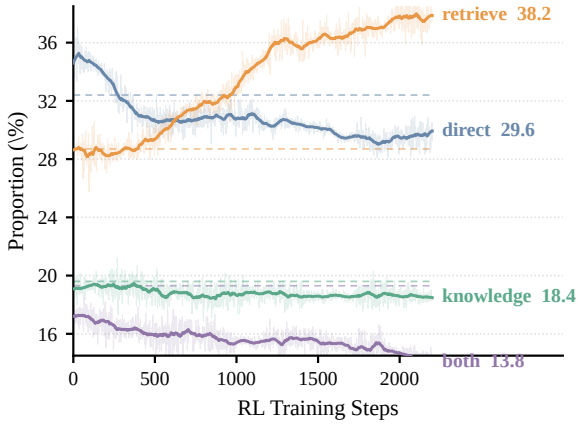


Figure 4: Routing distribution across RL training steps on the ZS-CSD mixed-target test set. Step 0 corresponds to StanceAgent-R1-Base; dashed lines mark the $\mathcal{D}_{\text{cold}}$ routing prior.

or dilute these signals.

4.4.5 Routing Behavior Analysis

As shown in Figure 4, the cold-start policy over-uses DIRECT and under-uses RETRIEVE relative to the teacher prior. RL then makes RETRIEVE the dominant route, pulls DIRECT and BOTH down, and leaves KNOWLEDGE essentially unchanged. The converged distribution matches neither the cold-start nor the teacher prior, showing that RL reshapes the routing policy rather than preserving imitation. Appendix F shows that no fixed routing policy matches adaptive routing on any split, and Appendix G reports the reward and response-length curves.

5 Conclusion

We presented **StanceAgent-R1**, an agentic reinforcement learning framework for conversational stance detection. By adaptively routing each instance to evidence-path selection and/or knowledge elicitation, StanceAgent-R1 addresses cross-turn dependencies, utterance related knowledge needs, and instance-specific reasoning within a unified policy. We further introduced a progressive trajectory-faithful reward to encourage correct, well-formed, and route-consistent reasoning. Experiments on ZS-CSD, MT-CSD, and C-MTCSD show that StanceAgent-R1 achieves state-of-the-art performance, while ablation and trajectory analyses confirm the effectiveness of its agentic design and reward strategy. Future work will scale the agentic training paradigm to broader domains.

Limitations

Our study has several limitations. First, our model is trained and evaluated on conversational stance detection benchmarks (ZS-CSD, MT-CSD, C-MTCSD) whose targets and discourse styles are skewed toward social-media platforms; generalization to other domains such as parliamentary debate or product reviews remains to be verified. Second, the cold-start trajectories are distilled from a single frozen teacher LLM, which inherits the teacher's biases in routing decisions and knowledge content; using a more diverse pool of teachers or self-distillation may further improve trajectory quality. Third, we restrict the backbone to a 7B-parameter open-weight model and use LoRA for both SFT and RL; scaling to larger or alternative architectures may change the trade-off between routing accuracy and inference cost. Finally, the routing-aware reward couples answer correctness, format validity, and trajectory legality with fixed weights; an adaptive curriculum over these components could yield more stable RL training but is left to future work.

Acknowledgments

This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (NSFC) [Grant No. 62506058] and the Young Teachers Science and Technology Innovation Project of the Fundamental Research Funds for the Central Universities [Grant No. 3132026362].

References

Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 876–885.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261.

Yuzhe Ding, Kang He, Bobo Li, Li Zheng, Haijun He, Fei Li, Chong Teng, and Donghong Ji. 2025. Zero-shot conversational stance detection: Dataset and approaches. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3221–3235.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. 2025. ReTool: Reinforcement learning for strategic tool use in LLMs. *arXiv preprint arXiv:2504.11536*.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. 2026. Group-in-Group policy optimization for LLM agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yanxu Ji, Jinzhong Ning, Yijia Zhang, Zhi Liu, and Hongfei Lin. 2025. LLM-driven implicit target augmentation and fine-grained contextual modeling for zero-shot and few-shot stance detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5872–5884.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused LLM-based agents. In *Proceedings of the international AAAI conference on web and social media*, volume 18, pages 891–903.

Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023a. Stance detection on social media with background knowledge. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 15703–15717.

Yupeng Li, Haorui He, Shaonan Wang, Francis CM Lau, and Yunya Song. 2023b. Improved target-specific stance detection on social media platforms by delving into conversation threads. *IEEE Transactions on Computational Social Systems*, 10(6):3031–3042.

Zhuofeng Li, Haoxiang Zhang, Seungju Han, Sheng Liu, Jianwen Xie, Yu Zhang, Yejin Choi, James Zou, and Pan Lu. 2025. In-the-Flow agentic system optimization for effective planning and tool use. *arXiv preprint arXiv:2510.05592*.

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. JointCL: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 81–91.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Junxia Ma, Changjiang Wang, Hanwen Xing, Dongming Zhao, and Yazhou Zhang. 2024. Chain of stance: Stance detection with large language models. In *CCF International conference on natural language processing and chinese computing*, pages 82–94. Springer.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Fuqiang Niu, Zebang Cheng, Xianghua Fu, Xiaojiang Peng, Genan Dai, Yin Chen, Hu Huang, and Bowen Zhang. 2024a. Multimodal multi-turn conversation stance detection: A challenge dataset and effective model. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 3867–3876.

Fuqiang Niu, Genan Dai, Yisha Lu, Jiayu Liao, Xiang Li, Hu Huang, and Bowen Zhang. 2025a. MT2-CSD: A new dataset and multi-semantic knowledge fusion method for conversational stance detection. *arXiv preprint arXiv:2506.21053*.

Fuqiang Niu, Min Yang, Ang Li, Baoquan Zhang, Xiaojiang Peng, and Bowen Zhang. 2024b. A challenge dataset and effective models for conversational stance detection. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 122–132.

Fuqiang Niu, Yi Yang, Xianghua Fu, Genan Dai, and Bowen Zhang. 2025b. C-MTCSD: A Chinese multi-turn conversational stance detection dataset. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 769–772.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tur, Gokhan Tur, and Heng Ji. 2026. ToolRL: Reward is all tool learning needs. *Advances in Neural Information Processing Systems*, 38:105523–105553.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-Searcher: Incentivizing the search capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2503.05592*.

Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 527–537.

Maksym Taranukhin, Vered Shwartz, and Evangelos Milios. 2024. Stance reasoner: Zero-shot stance detection on social media with explicit reasoning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15257–15272.

Ramon Villa-Cox, Sumeet Kumar, Matthew Babcock, and Kathleen M Carley. 2020. Stance in replies and quotes (SRQ): A new dataset for learning stance in Twitter conversations. *arXiv preprint arXiv:2006.00691*.

Bingbing Wang, Zhixin Bai, Qianlong Wang, Jingjie Lin, Min Yang, Xi Zeng, and Ruifeng Xu. 2025. More than just a conversation: A multi-agent reasoning graph knowledge distillation for conversational stance detection. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2895–2899.

Xiaolong Wang, Yile Wang, Sijie Cheng, Peng Li, and Yang Liu. 2024. DEEM: Dynamic experienced expert modeling for stance detection. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)*, pages 4530–4541.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Penghui Wei and Wenji Mao. 2019. Modeling transferable topics for cross-target stance detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1173–1176.

Penghui Wei, Nan Xu, and Wenji Mao. 2019. Modeling conversation structure and temporal dynamics for jointly predicting rumor stance and veracity. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 4787–4798.

Maxwell Weinzierl and Sanda Harabagiu. 2024. Tree-of-counterfactual prompting for zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–880.

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–783. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. 2026. DAPO: An open-source LLM reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244.

Haozhen Zhang, Tao Feng, and Jiaxuan You. 2026a. Router-R1: Teaching LLMs multi-round routing and aggregation via reinforcement learning. *Advances in Neural Information Processing Systems*, 38:141233–141265.

Yuanshuo Zhang, Aohua Li, Bo Chen, Jingbo Sun, and Xiaobing Zhao. 2026b. MSME: A multi-stage multi-expert framework for zero-shot stance detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 34879–34887.

Zhao Zhang, Yiming Li, Jin Zhang, and Hui Xu. 2024. LLM-driven knowledge injection advances zero-shot and cross-target stance detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 371–378.

Zhaodan Zhang, Jin Zhang, Hui Xu, Jiafeng Guo, and Xueqi Cheng. 2025. MPRF: Interpretable stance detection through multi-path reasoning framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 454–470.

A Implementation Details

Table 5 reports the full hyperparameter configuration for cold-start SFT and RL post-training of StanceAgent-R1.

Hyperparameter	**Value**
Cold-start SFT	
LoRA rank	64
LoRA scaling	rsLoRA
LoRA target modules	attention & MLP projections
Effective batch size	16
Learning rate	5×10^{-5}
LR schedule	cosine, 10% warmup
Weight decay	0.01
Gradient clipping	1.0
Epochs	1
Reinforcement learning	
Rollouts per prompt G	16
Per-device batch size	16
Gradient accumulation	8
Effective batch size	128
Learning rate	5×10^{-6}
Sampling temperature	0.9
Top-p	0.99
Max completion length	812 tokens
Clipping ratio ϵ	0.2
KL coefficient β	0
α_{fmt}	0.4
α_{legal}	0.2
η_{gate}	0.6
Epochs	1
Hardware	2× NVIDIA A800-80GB

Table 5: Hyperparameter configuration for cold-start SFT and RL post-training of StanceAgent-R1.

B Per-Judge Score Breakdown

Table 6 reports the score from each individual LLM judge. The relative ordering of the three methods is preserved across all three judges, which justifies the averaging used in Table 3. We do observe that GPT-4o tends to score the GPT-4o teacher slightly above what Claude-Opus-4.1 and Gemini-2.5-Pro give, consistent with self-preference biases reported in the LLM-as-judge literature; the gap is small and does not change the ranking of methods.

C Human Evaluation

To verify that StanceAgent-R1's scores in Table 3 are not an artifact of LLM-judge bias, we recruit three NLP-background annotators and ask them to rate 100 randomly sampled StanceAgent-R1 trajectories from the ZS-CSD mixed-target test set, using the same six axes and 5-point Likert scale as the LLM judges and following the rubric of Appendix L. Each trajectory is rated independently by all three annotators, and the Knowledge axis is rated only when the trace invokes `<knowledge>`.

The aggregated human scores are reported in the 👤 row of Table 3. They are slightly lower than

Method	Judge	Plan	Route	Path	Know.	Reas.	Concl.	Avg
GPT-4o (Teacher)	GPT-4o	4.5	4.3	4.3	4.6	4.2	4.4	4.4
	Claude-Opus-4.1	4.3	4.0	4.2	4.4	3.9	4.2	4.2
	Gemini-2.5-Pro	4.4	4.0	4.1	4.5	3.9	4.3	4.2
StanceAgent-R1-Base	GPT-4o	4.1	4.0	3.9	4.1	3.8	4.0	4.0
	Claude-Opus-4.1	3.9	3.8	3.7	3.9	3.6	3.8	3.8
	Gemini-2.5-Pro	4.0	3.9	3.8	4.0	3.7	3.9	3.9
StanceAgent-R1	GPT-4o	4.6	4.6	4.5	4.5	4.7	4.7	4.6
	Claude-Opus-4.1	4.4	4.4	4.3	4.3	4.5	4.5	4.4
	Gemini-2.5-Pro	4.5	4.5	4.4	4.4	4.6	4.6	4.5

Table 6: Per-judge scores from the three LLM judges (GPT-4o, Claude-Opus-4.1, and Gemini-2.5-Pro), averaging to the values in Table 3. The relative ranking of the three methods is preserved across all three judges.

the LLM-judge scores, consistent with the common observation that human raters are more critical than LLM judges, but the gap is within 0.2 on every axis and the relative ordering across axes (e.g., Rs. and Cn. being the strongest) is preserved. The per-axis Fleiss' κ across the three annotators ranges from 0.62 to 0.79, all in the substantial range. These results indicate that the high LLM-judge scores of StanceAgent-R1 reflect genuine trajectory quality rather than LLM-judge preference for our model's style.

D Trajectory Transferability Analysis

A faithful trace should make the reasoning visible in text rather than hide it in policy weights. We test this by feeding each trajectory, with `<answer>` stripped, as a prompt to three frozen small LLMs and a frozen GPT-3.5 (Table 7). StanceAgent-R1 traces lift every target LLM well above its no-trace baseline, and Qwen2.5-1.5B with our trace even exceeds the same backbone LoRA-tuned on labelled data, showing that the trace itself carries more transferable signal than the weights learned from labels. The ranking of the three trace sources is also preserved across all four targets, with our model strictly above both the GPT-4o teacher and the cold-start variant. The RL gain therefore transfers across model families rather than residing in the 7B policy, indicating that StanceAgent-R1 produces a portable reasoning artifact that opens a natural path toward trace-level distillation.

E Static Routing Distribution

As shown in Table 8, all four routes remain in active use at convergence and the policy does not collapse onto any single route. RL substantially rebalances the cold-start distribution: RETRIEVE rises by about 10 percentage points and becomes the dominant route, DIRECT drops correspondingly, BOTH decreases modestly, and KNOWLEDGE stays close to its cold-start level.

F Forced-Routing Comparison

To test whether any fixed routing policy can match adaptive routing, we hold the `<route>` tag constant at inference time across all instances. *Forced both* coincides with the *w/o Routing* row of Table 2.

As shown in Table 9, no fixed policy reaches adaptive macro-F1 on any split. The best fixed policy varies by target type: RETRIEVE is strongest on Mixed and Claim, while KNOWLEDGE is strongest on Noun-phrase. Adaptive routing handles both target types within a single model.

G RL Training Dynamics

As shown in Figure 5(a), the answer reward rises steadily from the cold-start initialization and gradually converges, while the format reward stays near one throughout training, indicating that schema compliance is already learned during cold-start. The legality reward remains at zero in the early phase and activates once the validation macro-F1 first crosses $\eta_{\text{gate}} = 0.6$, after which it quickly rises and stabilizes. As shown in Figure 5(b), the average response length first increases during early exploration and then drops sharply, stabilizing below the cold-start initialization. This pattern indicates that RL initially expands the reasoning trace before learning to produce more concise outputs, consistent with the response-length dynamics reported for RL-trained reasoning agents (Jin et al., 2025; Feng et al., 2025).

Source of Trajectory	Qwen-1.5B	Qwen-3B	Llama-3B	GPT-3.5	Avg
No trace (zero-shot)	22.4	33.0	28.1	40.3	31.0
GPT-4o trace	44.5	48.2	45.0	51.8	47.4
StanceAgent-R1-Base trace	41.5	45.7	42.3	49.0	44.6
StanceAgent-R1 trace	**53.4**	**57.2**	**54.8**	**58.1**	**55.9**
Ref: target LoRA-tuned	45.9	51.1	41.2	–	–

Table 7: Trajectory transferability on the ZS-CSD mixed-target test set with macro-F1 (%). Each generated trace, with <answer> stripped, is fed to four frozen LLMs (Qwen-1.5B, Qwen-3B, Llama-3B, GPT-3.5) to predict the stance. The bottom row reports the same backbones LoRA fine-tuned on the labelled data as a reference ceiling.

H Hyperparameter Sensitivity

We assess the sensitivity of three RL-stage hyperparameters introduced in §3.2.3: the legality gate η_{gate}, the legality weight α_{legal}, and the format weight α_{fmt}. Each is varied independently while the others are held at their default values ($\eta_{\text{gate}} = 0.6$, $\alpha_{\text{legal}} = 0.2$, $\alpha_{\text{fmt}} = 0.4$). All runs use the same cold-start checkpoint and the same RL recipe. We report macro-F1 on the ZS-CSD mixed-target test set.

Across all three sweeps, performance forms a concave curve around the default and stays within roughly two macro-F1 points of the peak inside the explored range, indicating that StanceAgent-R1 is not brittle to small changes in these hyperparameters. The η_{gate} curve drops on both sides for distinct reasons: a low gate activates the legality term before the answer reward stabilizes, which destabilizes RL; a high gate is rarely crossed during the available training budget, so legality stays effectively inactive. α_{legal} shows a similar trade-off between trace faithfulness and answer-signal dilution, and α_{fmt} a similar trade-off between schema compliance and answer-signal dilution.

I Inference Efficiency Analysis

A central motivation of the Routing Agent is that not every instance needs the full reasoning pipeline.

Source	DIR.	RETR.	KNOW.	BOTH
$\mathcal{D}_{\text{cold}}$	32.4	28.7	19.6	19.3
StanceAgent-R1-Base	35.2	28.1	19.4	17.3
StanceAgent-R1	29.6	38.2	18.4	13.8

Table 8: Routing distribution (%) on the ZS-CSD mixed-target test set, corresponding to the endpoints of Figure 4. $\mathcal{D}_{\text{cold}}$ denotes the rejection-sampled cold-start corpus and is included as a reference prior. DIR., RETR., KNOW. abbreviate the four routes.

Routing policy	Mixed	NP	Claim
Adaptive (Ours)	**59.34**	**58.80**	**59.53**
Forced DIRECT	53.42	53.85	52.94
Forced RETRIEVE	55.86	54.96	56.74
Forced KNOWLEDGE	54.31	56.12	53.42
Forced BOTH	56.09	55.81	56.22

Table 9: Forced versus adaptive routing on the ZS-CSD test set (macro-F1, %). NP denotes noun-phrase targets. Best score in each column is in **bold**.

Route	Share (%)	Tokens	Latency (s)
DIRECT	29.6	215	2.7
RETRIEVE	38.2	315	3.9
KNOWLEDGE	18.4	415	5.2
BOTH	13.8	515	6.4
weighted avg.	–	331	4.1

Table 10: Per-route inference cost of StanceAgent-R1. Tokens is the average number of generated tokens per instance; Latency is the wall-clock generation time per instance on a single A800-80GB. The bottom row is weighted by the routing distribution from Table 8.

We measure inference cost on the ZS-CSD mixed-target test set under a single A800-80GB GPU with vLLM serving, batch size 1, and the same decoding configuration as the main experiments. Table 10 reports the per-route cost of StanceAgent-R1, and Figure 7 contrasts adaptive routing against the *w/o Routing* variant, where BOTH is forced for every instance, and the strongest LLM-SFT baseline on the cost–accuracy plane.

Two observations follow. First, the per-route cost increases with the number of schema fields invoked by the route. Therefore, the routing distribution directly controls the average generation cost: a policy that always selects BOTH spends the most tokens, whereas a policy that always selects DIRECT spends the fewest. Under the learned

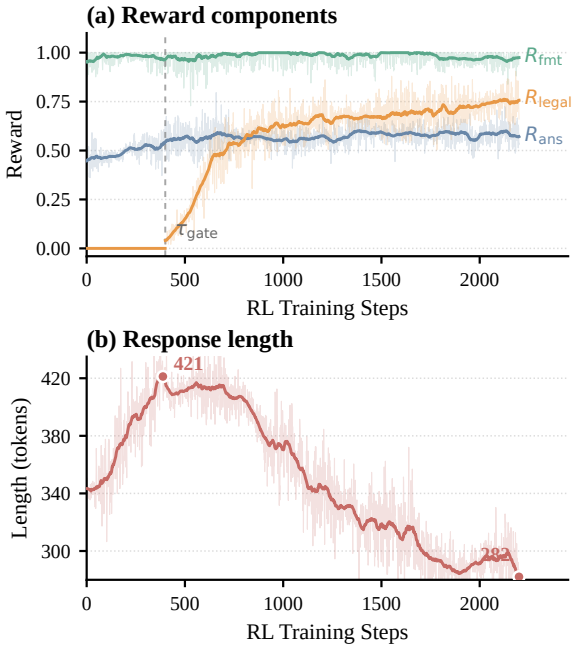


Figure 5: RL training dynamics of StanceAgent-R1 on the ZS-CSD validation set. (a) The three reward components R_{ans}, R_{fmt}, and R_{legal} across training steps, where the vertical dashed line marks the η_{gate} threshold at which R_{legal} is activated. (b) Average response length in tokens.

routing distribution, StanceAgent-R1 generates 331 tokens and takes 4.1 seconds per instance on average. Compared with the forced-BOTH variant, which generates 515 tokens and takes 6.4 seconds per instance, adaptive routing reduces the number of generated tokens by $1 - 331/515 \approx 35.7\%$ and also lowers latency.

Second, StanceAgent-R1 is not designed to be uniformly cheaper than a label-only SFT baseline. Instead, it targets a more favorable cost–accuracy trade-off: compared with the SFT-only model, it generates longer structured trajectories but yields substantially higher macro-F1; compared with the forced-BOTH pipeline, it achieves both higher macro-F1 and lower inference cost. We therefore view adaptive routing as a practical operating point for offline opinion analysis, where several seconds of latency per instance is acceptable.

J Teacher Prompt Template

We use a frozen teacher LLM (`gpt-4o`) to generate each cold-start trajectory via label-blind rejection sampling. The system prompt fixes the output schema and the routing rules; the user prompt formats one training instance with a stable per-turn index so the teacher can reference history turns inside `<select_path>`. The gold stance is never included in the teacher input. Instead, an offline validator checks schema validity, routing consistency, and whether the generated `<answer>` matches the gold label. Rejected trajectories are re-submitted with the validator's failure reason appended, but still without exposing the gold label (Appendix J.3). The original prompts are written in the language of the corresponding dataset; English translations below are functionally equivalent.

J.1 System Prompt

The system prompt has two parts: the output schema (Table 11) defining the six to eight tags that compose a trajectory, and the routing rules plus hard constraints (Table 12) that govern when each tag is required and what content it may contain.

J.2 User Prompt Template

The user prompt is assembled per instance (Table 13). The `[turn N]` index prefix gives the teacher a stable address space for `<select_path>`. The gold stance is never included in the prompt; it is used only by the offline validator for rejection sampling.

J.3 Retry Suffix on Rejection

When a generated trajectory fails the validator (schema, routing consistency, or answer mismatch against the gold label held only by the validator), the failure reason is appended to the same user prompt and the teacher is asked to regenerate, up to a fixed retry budget (Table 14). The gold label is never exposed to the teacher during regeneration.

K Trajectory Example

We show one accepted cold-start trajectory produced by the teacher (translated from the original Chinese). The target is *Sora*; the conversation discusses whether Sora has practical industrial value. The current utterance (u_3) first concedes the prior framing and then dismisses it as unimportant, yielding stance AGAINST. The router selects `retrieve` and points to a non-adjacent ancestor turn u_2, illustrating the cross-turn dependency motivated in §1.

The input fed to the teacher (Table 15) and the accepted output trajectory (Table 16) are shown below.

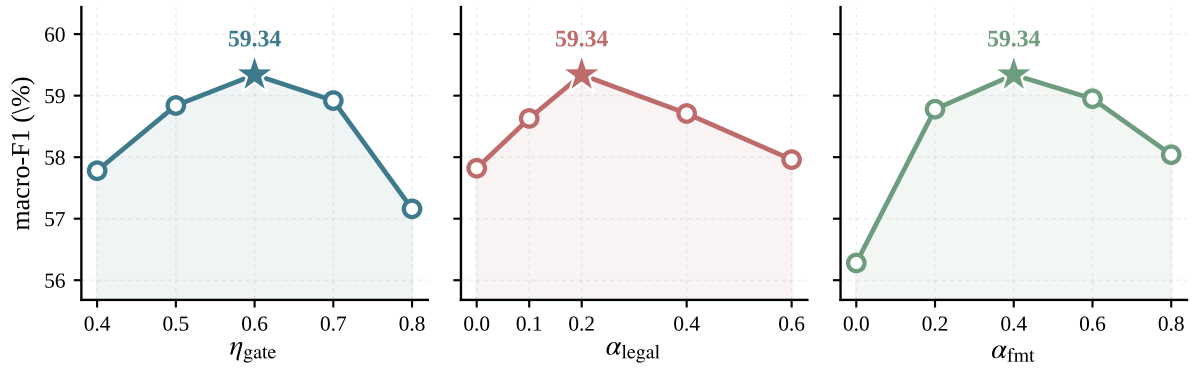


Figure 6: Sensitivity of StanceAgent-R1 to the three RL-stage hyperparameters on the ZS-CSD mixed-target test set. Each panel varies one hyperparameter while holding the others at their default values. Stars mark the default setting; the $\alpha_{\text{legal}} = 0$ and $\alpha_{\text{fmt}} = 0$ points coincide with the *w/o* R_{legal} and *w/o* R_{format} rows of Table 2, respectively.

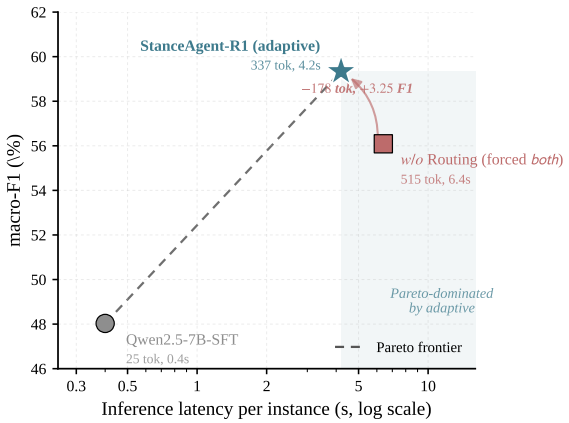


Figure 7: Cost–accuracy trade-off on the ZS-CSD mixed-target test set. The dashed line marks the Pareto frontier formed by the SFT-only baseline and StanceAgent-R1. The forced-BOTH variant falls inside the shaded region, indicating that it is Pareto-dominated by adaptive routing: lower macro-F1 *and* higher latency. Numbers in parentheses are average generated tokens and wall-clock latency per instance on a single A800-80GB.

L LLM-as-Judge Prompt

Each judge in Table 3 grades a trajectory along six axes on a 5-point Likert scale in a single call. The system prompt (Table 17) defines the role and the rubric for each axis; the user prompt (Table 18) supplies the instance and the trajectory to evaluate. The Knowledge axis is graded only when the trajectory's `<route>` invokes `<knowledge>`; in all other cases the judge is instructed to return `null` for that axis and the corresponding cell is excluded from the per-axis average.

M Case Study

We illustrate the four routing decisions with one representative case each, drawn from the ZS-CSD mixed-target test set. Each case shows the dialogue history, the current utterance, the routing decision returned by the Routing Agent, and the resulting reasoning trajectory.

Case 1. Route = Direct

Target	**Pinduoduo**
Gold stance	**Favor**
Depth	3
Dialogue	[turn 1] Speaker 0: Pinduoduo lets a lot of people live with quality. [turn 2] Speaker 9: And the stuff on Pinduoduo is actually decent – I once compared the same item on Taobao and Pinduoduo: Pinduoduo was cheaper, the shipping address was the same as Taobao, and the goods were identical. [turn 3] Speaker 23: You know what? Five years ago I bought a hairdryer on Pinduoduo – the kind with a blue glow inside. I wanted to switch to a smaller one, but it just wouldn't break and the airflow was monstrous. I finally tossed it when I moved this year. [laughing-crying]
Routing rationale	The current utterance names the target directly and unfolds a personal shopping anecdote anchored by a time cue ("five years ago"), a product detail ("hairdryer with a blue glow inside"), and an outcome ("just wouldn't break, airflow was monstrous"). Although the wording is colloquial and exaggerated, it carries no pronoun reference or anaphoric dependency on any prior turn, and the target is a common e-commerce platform that requires no specialised background. The literal content is self-sufficient, so direct routing applies.
Route	→ the routing agent chose **Direct**
Reasoning	The utterance centres on a hairdryer the speaker bought on Pinduoduo. After mentioning the wish to swap it for a smaller one, the speaker delivers the key experience – "it just wouldn't break" and "airflow was monstrous" – indicating the product is durable and performs strongly. The teasing description ("blue glow inside") and the final disposal ("tossed it when I moved") do not negate the durability claim: disposal here is a move-out clean-up, not a quality failure. Read literally, the utterance uses exaggerated phrasing to emphasise that a five-year-old Pinduoduo purchase has held up. Therefore the speaker holds a Favor stance toward the target.
Predicted stance	→ predicted stance is **Favor**

Figure 8: Case 1, DIRECT: the current utterance is self-contained and the target is widely known, so the Routing Agent invokes neither downstream agent and the trajectory omits <select_path> and <knowledge>.

Case 2. Route = Retrieve

Target	**Sora**
Gold stance	**Against**
Depth	4
Dialogue	[turn 1] Speaker 0: With Sora out, China's AI scene is left more than a tier behind. In a sense, a new industrial revolution has begun. [turn 2] Speaker 2: Since you know so much, tell me – how exactly is this thing applied in industry? [turn 3] Speaker 22: The film industry is also an industry. [turn 4] Speaker 2: Sure, that's true. But films are dispensable – totally useless!
Routing rationale	The current utterance opens with "Sure, that's true," a clear anaphoric link to the previous turn. It then introduces a contrast with "but" and delivers a colloquial verdict: "films are dispensable, totally useless." The implicit "that" must refer back to "the film industry is also an industry," so the evaluation target cannot be located from the current sentence alone – history retrieval is required. The target Sora is common knowledge, so no extra background is needed.
Route	→ the routing agent chose **Retrieve**
Selected turns	**[turn 3]**
Reasoning	[turn 3] says "the film industry is also an industry," answering [turn 2]'s challenge about industrial application and pinning Sora's use case to film. The current utterance accepts that framing with "sure, that's true," but uses "but films are dispensable, totally useless" to dismiss the value of this application – thereby undermining Sora's claim as evidence of an industrial revolution. Speaker 2 was previously questioning in [turn 2]; the current utterance is not self-correction but a sharpened judgment after the reply. Belittling Sora's application denies its significance, so the stance is Against.
Predicted stance	→ predicted stance is **Against**

Figure 9: Case 2, RETRIEVE: the utterance refers back to a non-adjacent ancestor turn through a pronoun, so the Routing Agent invokes the Evidence Path Agent to select the referenced turn and the trajectory grounds its reasoning on the selected ancestor.

Case 3. Route = Knowledge

Target	**Medical Bulk-Procurement**
Gold stance	**Against**
Depth	3
Dialogue	[turn 1] Speaker 0: Anti-corruption in healthcare is necessary, but we also need to see the downsides of bulk-procurement. Few procured products are actually effective, and manufacturers have very thin profits – with no profits, how can they invest in R&D? [turn 2] Speaker 7: The price already preserves the profit needed for R&D. You think they leave nothing for the company? [turn 3] Speaker 11: National drug-price negotiations really do squeeze hard – profits are basically very, very low.
Routing rationale	The utterance is a declarative claim – "national drug-price negotiations really do squeeze hard, profits basically very, very low" – with stacked intensifiers ("really", "very", "basically"). "Squeeze" carries a directional judgment, but "national drug-price negotiations" is not literally the target term "medical bulk-procurement", so policy background is needed to decide whether they refer to the same cost-control regime. The sentence has no pronoun reference and is not a mere agreement marker – its content is self-contained while the target is a domain concept, hence the knowledge route.
Route	→ the routing agent chose **Knowledge**
Background knowledge	Medical bulk-procurement refers to centralised volume-based purchasing of drugs or consumables organised by national or local authorities, lowering bid prices through volume-for-price exchanges. National drug-price negotiation is a separate mechanism between insurance regulators and pharma firms over reimbursement standards – but both are price-governance / cost-control policies and are routinely grouped together in public discussion. "Profits being squeezed" is a common criticism extended from one to the other.
Reasoning	The utterance's core meaning is that the policy crushes corporate profit. "Squeeze hard" + "basically very, very low" form a chain of intensifiers, expressing dissatisfaction with this cost-control approach. Per the background, national drug-price negotiations and bulk-procurement are distinct programs but belong to the same price-governance category and are routinely paired in discussion. Framing "low profits" as a negative consequence therefore extends the critique "bulk-procurement squeezes profit and harms the industry." The speaker holds a negative view of the target – Against.
Predicted stance	→ predicted stance is **Against**

Figure 10: Case 3, KNOWLEDGE: the utterance is self-contained at the dialogue level but the target is a niche entity, so the Routing Agent invokes the Knowledge Elicitation Agent and the trajectory grounds its reasoning on the elicited factual background.

Case 4. Route = Both

Target	**Sora**
Gold stance	**Favor**
Depth	5
Dialogue	[turn 1] Speaker 0: With Sora out, China's AI scene is left more than a tier behind. A new industrial revolution has begun, in a sense. [turn 2] Speaker 3: Revolution my foot – this thing barely improves productivity at all. [turn 3] Speaker 8: Don't draw conclusions before you understand it. Sora's videos can replace entire low-end teams – shooting, makeup, props, sets, actors, scriptwriters. Low- to mid-tier teams make up 70-80% of the film industry. That's not an improvement? [turn 4] Speaker 3: At its current level, do you think those videos are actually usable in industrial or commercial production? [laughing-crying] [turn 5] Speaker 0: Algorithms simulate the human brain – if powerful enough, they vastly surpass humans. Ancestors once thought airplanes and cars were absurd fantasies – just like you now. Do you know how fast algorithms generate things today? Have you even looked into it?
Routing rationale	The current utterance uses second-person "you" and consecutive rhetorical questions ("do you know...?", "have you even looked into it?"), clearly addressing the previous turn. It uses an analogy – "ancestors thought airplanes and cars were absurd fantasies, just like you now" – to align the opponent with outdated thinking. It also unfolds technical claims ("algorithms simulate the human brain", "powerful enough to surpass humans") about the target. Because it depends on prior disagreement and carries technical content, both history retrieval and background knowledge are required.
Route	→ the routing agent chose **Both**
Selected turns	**[turn 3] [turn 4]**
Background knowledge	Sora is OpenAI's text-to-video model, capable of generating relatively long, coherent video clips with complex camera motion from text prompts. It shows strong scene consistency, object-motion modelling and visual detail, marking an important step for generative AI from text/image into video. Limits remain – unstable physics, weak long-horizon temporal consistency, limited controllability – so it does not yet fully replace mature film-industry pipelines.
Reasoning	[turn 4] challenges whether Sora is industrially usable at the current level. [turn 3] lists Sora replacing whole low-end production teams, emphasising Sora's pipeline impact. The current utterance lifts the discussion from "current usability" to technical trajectory and potential ("if powerful enough, vastly surpass humans"), then uses the airplane/car analogy to cast the opponent's scepticism as historical underestimation of new technology, and finally challenges its basis with "do you know how fast algorithms generate things today?". The background confirms Sora is a real text-to-video milestone with caveats – but the utterance does not stress the caveats, only speed and potential. Also, in [turn 1] the same speaker said "China's AI scene is left more than a tier behind"; the current turn extends that line under fire, reinforcing rather than softening it. Stance: Favor.
Predicted stance	→ predicted stance is **Favor**

Figure 11: Case 4, BOTH: the utterance refers back to the dialogue history *and* the target is technical, so the Routing Agent invokes both the Evidence Path Agent and the Knowledge Elicitation Agent, and the trajectory combines the selected ancestor turn with the elicited background.

System Prompt (part 1 of 2): output schema.

```
You are a stance-detection annotator for multi-turn conversations. Given a target,
a dialogue history, and a current utterance, generate one reasoning trajectory that
predicts the stance of the current utterance toward the target. You are not given
the gold stance. Infer the stance only from the target, the dialogue context, and
the current utterance.

Output format (strict, in order, no extras):

<plan>
Analyze the linguistic structure of the current utterance (pronoun reference,
rhetorical question, irony, enumeration, technical terms, etc.) and choose a route
accordingly. Do not give a stance conclusion and do not use any evaluative verbs.
</plan>

<route>direct | retrieve | knowledge | both</route>

<select_path>[index list]</select_path>
  Output only when route = retrieve or both. Each index is the N in [turn N] from
  the dialogue context. Must be a subset of the ancestor chain, size 1..min(5,
  depth), written as a Python list, e.g. [0,2].

<knowledge>
Relevant factual background about the target, 100-250 characters.
</knowledge>
  Output only when route = knowledge or both.

<think>
Integrate the literal meaning of the current utterance, the selected ancestor
turns (if any), and the target background knowledge (if any), and reason step by
step about the speaker's stance. End with an explicit conclusion such as
"therefore the speaker holds a Favor/Against/Neutral stance".
</think>

<answer>Favor | Against | Neutral</answer>
```

Table 11: Teacher system prompt (1/2): the output schema describing each tag and its required content.

System Prompt (part 2 of 2): routing rules and hard constraints.

```
Routing rules:
- direct:    utterance self-contained AND target widely known.
- retrieve:  utterance refers back to history (pronoun, agreement, ironic
             callback) AND target common.
- knowledge: target niche or technical AND utterance self-contained at the
             dialogue level.
- both:      when both above hold.

Hard constraints:
(1) The gold stance is not provided. Infer <answer> only from the target, the
    dialogue context, and the current utterance.
(2) <plan> describes only linguistic structure and routing rationale. No evaluative
    verbs (agree/support/endorse/defend; deny/oppose/criticize/belittle/mock;
    neutral/non-committal) and no conclusive phrases.
(3) <plan> must base its routing choice on observable linguistic cues.
(4) If retrieve or both is chosen, <think> must actually quote content from the
    selected ancestor turns.
(5) If knowledge or both is chosen, <knowledge> must be factually correct and
    <think> must use it.
(6) If the ancestor chain has earlier turns by the current speaker and the current
    utterance defends/revises that stance, <think> must explicitly state
    strengthens / weakens / shifts to neutral / preserves -- do not default to
    neutral.
(7) When depth = 1, route must be direct or knowledge.
(8) Do not fabricate facts in <knowledge>; if unsure, switch route to retrieve or
    direct.
(9) Output exactly the 6-8 tags above. First character must be <, last character
    must be >. No Markdown code fences, no explanatory text.
```

Table 12: Teacher system prompt (2/2): the routing rules and hard constraints enforced by the validator.

User Prompt Template (per-instance: target, dialogue context, and current utterance).

```
The following is a multi-party conversation on some topic. Output, in the
required format, the reasoning trajectory for the current speaker's stance
toward the target.

[Target]
{target}

[Dialogue context]
[turn 0] speaker{s_0}: {u_0}
[turn 1] speaker{s_1}: {u_1}
...
[turn n-2] speaker{s_{n-2}}: {u_{n-2}}
(or "No dialogue context" when depth = 1)

[Current utterance]
[turn n-1] speaker{s_{n-1}}: {u_{n-1}}
```

Table 13: Teacher user prompt template, instantiated per training instance with the target, dialogue context, and current utterance. The gold stance is not included in the prompt and is used only by the offline validator for rejection sampling.

Retry Suffix (appended to the user prompt on validator rejection).

```
The previous generation failed. Reason: {failure_reason}. Please regenerate and
fix this issue. Do not assume any gold label; infer <answer> only from the target,
the dialogue context, and the current utterance.
```

Table 14: Retry suffix appended to the user prompt when the validator rejects a generated trajectory, with the failure reason inserted. The retry prompt remains label-blind.

Trajectory Example — input: target, dialogue history, and current utterance fed to the teacher.

```
[Target] Sora

[Dialogue context]
[turn 0] speaker A: Sora looks impressive in the demos.
[turn 1] speaker B: But how is this thing going to be applied in industry?
[turn 2] speaker A: The film industry is also an industry -- Sora has obvious
                    applications there.

[Current utterance]
[turn 3] speaker B: That is true, but movies are optional -- so it is basically
                    useless.
```

Table 15: Trajectory example: the input (target, dialogue history, and current utterance) fed to the teacher LLM.

Trajectory Example — accepted reasoning trajectory returned by the teacher.

```
<plan>
The utterance opens with "that is true," picking up the previous turn, then
introduces a contrast with "but" followed by the colloquial verdict "movies are
optional, so it is basically useless." The demonstrative "that" refers back to
"film industry is also an industry," so the current utterance alone does not
fully locate the evaluation target or the argumentative chain; history is needed.
Sora is a common entity, so extra technical background is not required.
</plan>

<route>retrieve</route>

<select_path>[2]</select_path>

<think>
[turn 2] says "the film industry is also an industry," answering [turn 1]'s
question about industrial applications by anchoring Sora to film production. The
current utterance first accepts "film is also an industry," but then uses
"movies are optional, so it is basically useless" to redirect the discussion to
the value of that application, claiming the use case does not matter and thereby
weakening Sora's standing as evidence of an "industrial revolution." Speaker B
was previously in a questioning posture at [turn 1]; the current utterance is
not a self-correction but a sharper judgment after receiving an answer. Given
the target is Sora, the speaker is dismissing its application value, therefore
the speaker holds an Against stance toward the target.
</think>

<answer>Against</answer>
```

Table 16: Trajectory example: the reasoning trajectory returned by the teacher LLM and accepted by the validator for the input in Table 15.

LLM-as-Judge System Prompt: role, six-axis rubric, and output schema.

```
You are an expert evaluator of reasoning trajectories for conversational stance
detection. You will receive (i) a target, (ii) a dialogue history with per-turn
indices, (iii) a current utterance, and (iv) a reasoning trajectory written by a
model. The trajectory follows a fixed schema with six to eight tags:
<plan>, <route>, <select_path>(optional), <knowledge>(optional), <think>, <answer>.

Grade the trajectory along the six axes below on a 5-point Likert scale
(1 = very poor, 5 = excellent). For each axis, anchor your judgment to the
written content of the trajectory; do NOT use any external knowledge that the
trajectory itself does not invoke.

Axes and rubric:

1. Planning (Pl.): Does <plan> describe the linguistic structure of the current
   utterance (pronoun reference, irony, technical terms, etc.) without leaking a
   stance conclusion or using evaluative verbs?
   5: thorough and stance-free; 3: partial; 1: missing or already states a stance.

2. Routing (Rt.): Given <plan>, is the value of <route> consistent with the
   stated routing rules (direct / retrieve / knowledge / both)?
   5: clearly the best route; 3: defensible but not optimal; 1: contradicts <plan>.

3. Evidence Path (Pa.): If <route> requires <select_path>, are the selected turn
   indices valid (within history range, on the ancestor chain) and actually
   relevant to the stance of the current utterance? Penalize over- and
   under-selection equally. Score 5 if <route> does not require <select_path>
   and the field is correctly omitted.

4. Knowledge Factuality (Kn.): If <route> invokes <knowledge>, is the elicited
   content factually correct, on-topic for the target, and free of fabrication?
   Return null if <knowledge> is not invoked.

5. Reasoning Coherence (Rs.): Does <think> integrate the current utterance with
   the selected history turns (if any) and the elicited knowledge (if any) in a
   logically connected chain? Penalize abrupt jumps, repetition, and unsupported
   leaps.

6. Conclusion Support (Cn.): Does the terminal stance phrase in <think> follow
   from the preceding reasoning, and does it agree with <answer>?
   5: clearly entailed and aligned; 1: contradicts the reasoning or <answer>.

Output strictly the following JSON, no Markdown, no extra text:
{
  "Pl": <int 1-5>,
  "Rt": <int 1-5>,
  "Pa": <int 1-5>,
  "Kn": <int 1-5 or null>,
  "Rs": <int 1-5>,
  "Cn": <int 1-5>,
  "justification": "<one short sentence, <= 40 words>"
}
```

Table 17: LLM-as-judge system prompt used for all three judges (GPT-4o, Claude-Opus-4.1, Gemini-2.5-Pro). The same rubric is applied to every method, ensuring comparability of the scores in Tables 3 and 6.

LLM-as-Judge User Prompt: the per-instance evaluation request.

```
Grade the following trajectory along the six axes defined in the system prompt.

[Target]
{target}

[Dialogue history]
[turn 0] speaker{s_0}: {u_0}
[turn 1] speaker{s_1}: {u_1}
...
[turn n-2] speaker{s_{n-2}}: {u_{n-2}}
(or "No dialogue history" when depth = 1)

[Current utterance]
[turn n-1] speaker{s_{n-1}}: {u_{n-1}}

[Trajectory to grade]
{trajectory}

Return one JSON object as specified.
```

Table 18: LLM-as-judge user prompt, instantiated per evaluation instance with the target, dialogue history, current utterance, and the trajectory to grade.