When and How Should an Agent Clarify?
CIGAsk: Teaching LLMs to Clarify via Counterfactual Information Gain

Anonymous ACL submission

Abstract

Instruction-tuned LLMs faced with under-specified queries commit to one reading rather than ask, producing confidently wrong answers. Prompting alone does not fix this: the model either asks on everything or asks vague questions that fail to recover the missing fact. The root cause is a missing training signal: the policy must learn two coupled skills — *when* to ask rather than answer and *how* to ask a question that recovers the disambiguating fact — yet existing recipes address at most one axis or require a separately learned critic. We propose **CIGAsk**, an RL recipe that teaches both skills through two complementary reward signals inside a multi-turn GRPO loop. **Counterfactual Information Gain (CIG)** ablates the user response and credits each clarification by how much it raises the gold-answer log-likelihood under a frozen reference, supplying per-turn credit that steers *how* to ask. The **Asymmetric Ambiguity Bonus** pays a signed reward at the terminal token, gating *when* to ask by the gold ambiguity label. On three clarification benchmarks spanning table, passage, and open-domain QA, CIGAsk 7B leads the strongest external recipe on a smaller backbone and transfers across datasets without per-dataset tuning, while preserving single-turn QA on out-of-distribution benchmarks. Code and checkpoints will be released.

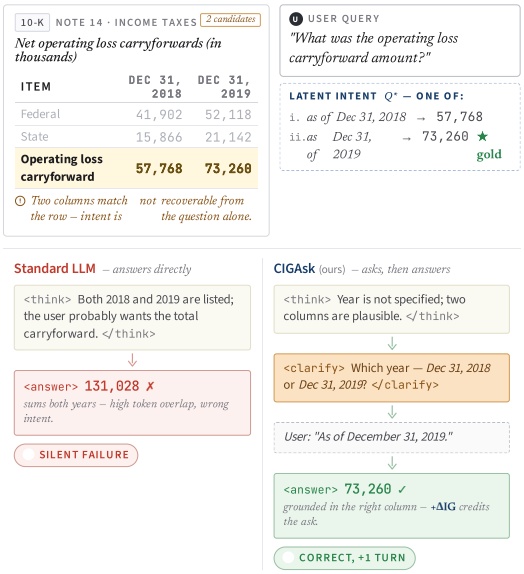

Figure 1: An ambiguous query that omits the target year. The standard model commits to one reading and answers incorrectly; CIGAsk identifies the missing slot, clarifies, and answers correctly. Two reward signals drive this behavior: CIG credits each clarifying turn by how much it raises the gold answer's log-probability under a frozen reference; the asymmetric bonus pays $+\rho$ only when the policy deliberately asks on an ambiguous query.

1 Introduction

Instruction-tuned LLMs (Yang et al., 2024; DeepSeek-AI, 2025) achieve strong single-turn performance on QA benchmarks. Yet when a request is under-specified, say *"What was the operating loss carryforward amount?"* against a table that lists both 2018 and 2019 figures (Deng et al., 2022), the same models commit to one reading rather than ask, producing answers that are confidently wrong rather than usefully partial. The failure is easy to demonstrate, costly in deployment, and hard to remove with prompting alone.

Instruction tuning rewards committing to *an* answer and penalises hedging; the resulting policy answers confidently rather than clarifies. The policy must learn two coupled skills in one training loop: *when* to ask rather than answer, and *how* to ask a question that recovers the disambiguating fact. Each prior line addresses one axis at a cost (§2). SFT-based self-training (Andukuri et al., 2024) filters trajectories by outcome reward and lacks a per-turn signal that grades *which* clarification surfaced the missing fact. Offline action-preference learning (Chen et al., 2025) cannot adapt to the policy's evolving trajectories.

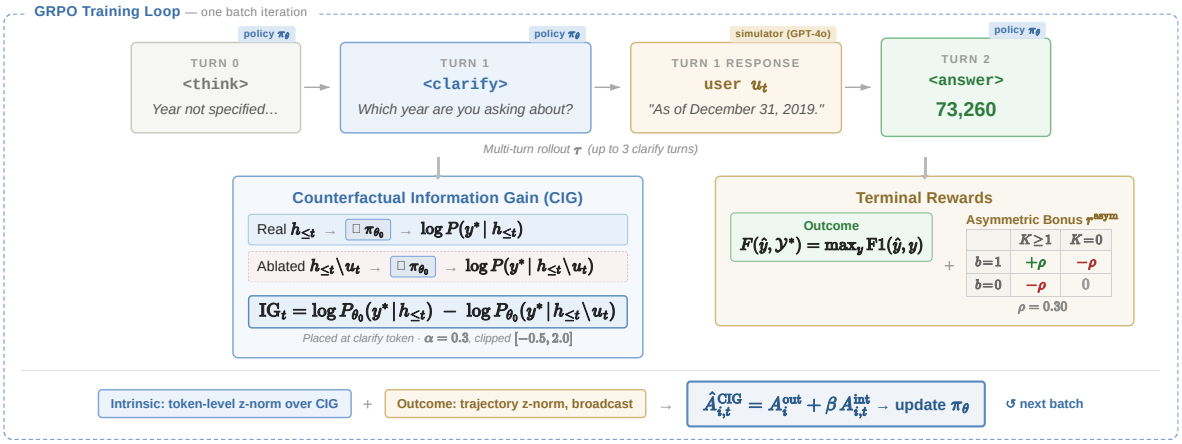

Figure 2: **CIGAsk overview.** *Left:* policy π_θ interacts with a fixed simulator over multi-turn clarification rollouts. *Top right:* rollout structure with clarify and answer turns. *Bottom right:* per-turn **CIG** (illustrated for Turn 2) compares the real history $h_{\leq 2}$ against the ablated $h_{\leq 2} \setminus u_2$ under the same frozen reference π_{θ_0} and credits the log-ratio to that clarify token; the terminal **Asymmetric Ambiguity Bonus** pays $\pm\rho$ along the gold-ambiguity $\times$ clarified axes and gates *when* to ask.

Rubric-guided RLVR (Zhao et al., 2026) needs per-instance rubric annotation that does not scale. Privileged-critic RL (Zhou et al., 2025) relies on train-time-only side information unavailable at deployment. None of these recipes offers a per-turn signal that handles both *when* and *how* using only the supervision the gold answer already provides.

We close this gap with **CIGAsk**, a recipe that pairs two reward signals inside a multi-turn GRPO loop. **Counterfactual Information Gain (CIG)** grades each clarification by ablating the user response and re-scoring the gold answer under a *frozen* reference; the resulting log-ratio is the per-turn reward at that clarify token. Anchoring the score to a frozen model decouples credit from the actor that is trying to improve, avoiding the drift that policy-coupled information-gain variants inherit (§2). The **Asymmetric Ambiguity Bonus** pays $\pm\rho$ along the (ambiguity label) $\times$ (clarified) axes at the terminal token, withholding reward for passive correctness on clear queries. A token-level 2-stream advantage carries both signals through GRPO and reduces to vanilla GRPO when no rollout clarifies (§3.2).

Contributions.

- We introduce **CIGAsk**, pairing a per-turn frozen-reference reward (CIG) that grades *which* clarification recovers the missing fact with a terminal asymmetric bonus that gates *when* to ask.
- On PACIFIC, AbgCoQA, and AmbigNQ, CIGAsk 7B reaches F1 0.795, ahead of the privileged-label SGP-Oracle (0.787) on a smaller backbone, and transfers across datasets and to out-of-distribution benchmarks without per-dataset tuning.

2 Related Work

Multi-turn clarification and ambiguous QA. Two threads frame the problem. Benchmarks establish the under-specification regime: AmbigQA (Min et al., 2020), AbgCoQA (Guo et al., 2021), PACIFIC (Deng et al., 2022), ClariQ (Aliannejadi et al., 2020), QuestBench (Li et al., 2025); adjacent multi-answer disambiguation (Stelmakh et al., 2022; Lee et al., 2024; Kim et al., 2023) treats the same ambiguity from the answer side rather than the question side. Question-generation methods predate LLM agents and span EVPI ranking (Rao and III, 2018), conversational search (Aliannejadi et al., 2019, 2020), and expected-information-gain criteria (Mazzaccara et al., 2024); they target *how* to ask while *when* is assumed externally.

Training agents to clarify. Several lines of work train agents to clarify, differing in supervision regime. ACT (Chen et al., 2025) pairs DPO-style action-level contrastive preferences with a Zephyr-7B base but does not score per-turn information value during training. AskBench (Zhao et al., 2026) uses rubric-guided RLVR; Zhang et al. (2025) relabel RLHF preferences via simulated future turns but yield a binary signal; Suri et al. (2025) apply EVPI inside a fixed tool-

call schema. Privileged-critic methods such as SWEET-RL (Zhou et al., 2025) train a separately learned critic with ground-truth access at training time, but the critic is unavailable at deployment. Most similar in supervision is STaR-GATE (Andukuri et al., 2024), which self-trains on clarification trajectories filtered by outcome reward without per-turn credit; we differ precisely by the per-turn signal that decides *which* clarification surfaced the disambiguating fact. Outcome-only multi-turn RL recipes such as UserRL (Qian et al., 2025) treat clarification as an opaque sub-trajectory.

Per-turn credit assignment. Process reward models supply per-step credit but typically require step-labelled data: Math-Shepherd (Wang et al., 2024) synthesises labels via Monte Carlo, VinePPO (Kazemnejad et al., 2025) via rollout trees, ImplicitPRM (Yuan et al., 2025) via outcome-derived logits. Sullivan (2025) show vanilla GRPO with outcome reward equals a Monte-Carlo PRM only when group rollouts share identical prefixes, a condition multi-turn clarification rarely satisfies. Concurrent agentic variants assign turn-level credit via outcome propagation (Wei et al., 2025) or trajectory-DPO-derived implicit step rewards (Liu et al., 2025), but compute the signal under the training policy and inherit the same collapse risk. Closest to our signal are two concurrent per-turn IG recipes. IGPO (Wang et al., 2025) computes turn-level rewards as the marginal increase in the *training policy*'s probability of the correct answer for search agents; InfoPO (Kong et al., 2026) computes the per-turn reward as the change in the *agent*'s action distribution after user feedback for general user-centric agents. Neither uses a frozen reference: both score IG under the training policy itself, so the per-turn signal drifts as the actor updates, inheriting the degeneracy diagnosed by Setlur et al. (2025). Neither pairs the per-turn signal with a label-driven gate on *when* to ask. Our frozen-reference formulation (§3.2) decouples credit from the actor and targets proactive clarification specifically rather than search or generic collaboration. Information-theoretic and counterfactual rewards originate in active learning (MacKay, 1992) and curiosity-driven exploration (Pathak et al., 2017); our self-supervised process reward sits in this family but consumes only the outcome-level gold answer.

3 Method

3.1 Preliminaries

Multi-turn clarification setup. We formulate clarification as a POMDP with hidden state q^* (the user's true intent). At turn t the agent observes history $h_{<t}=\{(a_1,u_1),\ldots,(a_{t-1},u_{t-1})\}$, context c, and ambiguous question q, then picks an action from $\mathcal{A}=\{\textsc{Think},\textsc{Clarify},\textsc{Answer}\}$: THINK runs internal reasoning; CLARIFY emits a question that triggers a user response $u_t \sim \mathrm{Sim}(\cdot \mid q^*, c, h_{<t})$ from a fixed simulator; ANSWER emits $\hat{y}$ and terminates, after which $\hat{y}$ is scored against gold $y^* \in \mathcal{Y}^*$. The policy conditions on observable history rather than maintaining an explicit belief over q^*; the per-turn reward defined in §3.2 supplies the supervision in its place.

GRPO. We train with Group Relative Policy Optimisation (GRPO, Shao et al., 2024), a critic-free PPO variant. For each prompt q we sample a group of G output trajectories $\{o_i\}_{i=1}^{G}$ from the old policy $\pi_{\theta_{\text{old}}}$ with scalar returns $\{r_i\}_{i=1}^{G}$, compute the group-relative advantage $\hat{A}_{i,t}=\big(r_i-\mathrm{mean}(\{r_j\}_{j=1}^{G})\big)/\mathrm{std}(\{r_j\}_{j=1}^{G})$ broadcast to every token of o_i, and maximise the clipped PPO objective with a token-level KL penalty against a frozen reference π_{ref}:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\{o_i\}}\Big[\tfrac{1}{G}\sum_{i=1}^{G}\tfrac{1}{|o_i|}\sum_{t=1}^{|o_i|}\big\{L_{i,t}^{\text{PPO}}(\theta) - \beta\,\mathbb{D}_{\text{KL}}[\pi_\theta\|\pi_{\text{ref}}]\big\}\Big],$$

$$L_{i,t}^{\text{PPO}}(\theta) = \min\big(\rho_{i,t}\hat{A}_{i,t},\ \mathrm{clip}_\epsilon(\rho_{i,t})\,\hat{A}_{i,t}\big), \qquad (1)$$

where $\rho_{i,t}=\pi_\theta(o_{i,t} \mid q, o_{i,<t})/\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})$ is the per-token importance ratio and $\mathrm{clip}_\epsilon(\rho)=\mathrm{clip}(\rho, 1-\epsilon, 1+\epsilon)$.

3.2 Reward design

Method overview. CIGAsk trains an LLM policy through multi-turn rollouts against a fixed user simulator with oracle access to the latent user intent (Figure 2). On top of the standard outcome reward (token-overlap F1 of the final answer), CIGAsk adds two reward terms placed at specific

tokens:

$$r_{i,t} = \underbrace{F(\hat{y}_i, \mathcal{Y}^*)\mathbb{1}[t{=}t_i^*]}_{\text{outcome}} + \underbrace{\alpha\,\mathrm{IG}_t\mathbb{1}[a_t{=}\textsc{Clarify}]}_{\text{CIG (per-turn)}} + \underbrace{r^{\text{asym}}(b_i, K(\tau_i))\mathbb{1}[t{=}t_i^*]}_{\text{Asym (terminal)}}, \quad (2)$$

$F(\hat{y}, \mathcal{Y}^*){=}\max_{y\in\mathcal{Y}^*}\mathrm{F1}_{\text{token}}(\hat{y}, y)$ is the token-overlap F1 of the final answer; IG_t is the per-turn **Counterfactual Information Gain** scoring *how* a clarification helps; r^{asym} is the trajectory-level **Asymmetric Ambiguity Bonus** scoring *when* the policy chose to clarify, keyed on the gold ambiguity label b_i. Below we define each component, then plug the resulting per-token rewards into GRPO via a minimal 2-stream advantage that preserves per-turn locality.

Counterfactual Information Gain (CIG). Standard process-reward models require step-labelled supervision that clarification lacks (§2). We instead measure value counterfactually: compare the real history $h_{\le t}$ (containing u_t) to the same history with u_t ablated, and score the change in a frozen reference's prediction of the gold answer.

Definition 1 (Self-supervised process reward)**.** *Let θ_0 be a fixed reference policy, y^* the gold target, and $h_{\le t}{=}h_{<t}\cup\{(a_t, u_t)\}$ a prefix containing u_t; write $h_{\le t}\setminus u_t{=}h_{<t}\cup\{(a_t, \varnothing)\}$ for the same prefix with u_t ablated, with context c and question q implicit in $h_{<t}$. The* self-supervised process reward *is the log-likelihood ratio*

$$r^{\text{ssp}}_{\theta_0}(u_t; y^* \mid h_{<t}) = \log \frac{P_{\theta_0}(y^* \mid h_{\le t})}{P_{\theta_0}(y^* \mid h_{\le t}\setminus u_t)}.$$

CIG is the empirical instance: for each CLARIFY action a_t,

$$\mathrm{IG}_t = \log P_{\theta_0}(y^* \mid h_{\le t}) - \log P_{\theta_0}(y^* \mid h_{\le t}\setminus u_t), \quad (3)$$

which we inject as $r_t^{\text{IG}}{=}\alpha\cdot\mathrm{IG}_t$ at the clarify token ($\alpha{=}0.3$, clipped to $[-0.5, 2.0]$ nats). The construction has an information-theoretic reading: under the singleton predictor family $\mathcal{V}{=}\{P_{\theta_0}(\cdot \mid h) : h \in \mathcal{H}\}$, CIG is the pointwise V-information (Xu et al., 2020) gain of u_t for y^* given $h_{<t}$, with expected value $\mathbb{E}[\mathrm{IG}_t \mid h_{<t}] \ge 0$ by the standard V-info theorem (equality iff u_t is uninformative about y^* under P_{θ_0}). It requires no separately learned critic and no per-step supervision beyond the dataset-level gold answer y^* that the outcome reward already consumes.

Why frozen. Scoring IG under the training policy ties the per-turn signal to the actor it tries to improve, risking drift toward outcome-only optimisation: Setlur et al. (2025) report a similar degeneracy for process-reward verifiers, and Sullivan (2025) prove that vanilla GRPO with outcome reward equals a Monte-Carlo PRM only on shared-prefix rollouts, a condition multi-turn clarification rarely satisfies. A single frozen predictor removes the policy dependence and yields a stable per-turn gradient regardless of group structure.

Asymmetric Ambiguity Bonus. CIG grades clarifications the policy emits but cannot push the policy to emit them; a policy that ignores clarification entirely still inherits a useful outcome reward. We pair CIG with a trajectory-level bonus on *when* to clarify. Let $K(\tau_i){=}\sum_t\mathbb{1}[a_t{=}\textsc{Clarify}]$ count clarify actions in rollout τ_i and $b \in \{0, 1\}$ be the gold ambiguity label. The bonus, paid at the terminal token t_i^*, is

$$r^{\text{asym}}(b, K) = \begin{cases} +\rho & b{=}1,\ K{\ge}1 \\ -\rho & b{=}1,\ K{=}0 \\ -\rho & b{=}0,\ K{\ge}1 \\ \ \ 0 & b{=}0,\ K{=}0, \end{cases} \quad (4)$$

with $\rho{=}0.30$. The asymmetry lives in the last cell: passive correctness on a clear query earns zero, not $+\rho$, so the only path to positive reward is clarifying a genuinely ambiguous query and earning positive CIG on that clarification; both decision (asym) and informativeness (CIG) signals must fire. We sweep $\rho{\in}\{0.1, 0.3, 0.5\}$.

Advantage estimator. Vanilla GRPO broadcasts a single trajectory advantage across all tokens, washing out per-turn rewards. We use a minimal 2-stream variant that keeps each per-turn reward local. Split $r_{i,t}$ at the terminal-token index t_i^* of o_i:

$$r^{\text{out}}_{i,t}{=}r_{i,t}\mathbb{1}[t{=}t_i^*], \quad r^{\text{int}}_{i,t}{=}r_{i,t}\mathbb{1}[t{\neq}t_i^*]. \quad (5)$$

The outcome stream is trajectory-normalised as in GRPO and broadcast across o_i, with $\mu_G^{\text{out}}, \sigma_G^{\text{out}}$ the group mean and std of $\{R_j^{\text{out}}\}_{j=1}^G$:

$$A_i^{\text{out}} = \frac{R_i^{\text{out}} - \mu_G^{\text{out}}}{\sigma_G^{\text{out}} + \epsilon}, \qquad R_i^{\text{out}}{=}\sum_t r^{\text{out}}_{i,t}. \quad (6)$$

The intrinsic stream is z-normalised at the token level over the pool of non-zero r^{int} entries in the prompt group; this token-level pooling keeps CIG at clarify tokens comparable across rollouts with different turn counts:

$$A^{\text{int}}_{i,t} = \frac{r^{\text{int}}_{i,t} - \mu^{\text{int}}_G}{\sigma^{\text{int}}_G + \epsilon} \mathbb{1}[r^{\text{int}}_{i,t} \neq 0]. \tag{7}$$

The combined token-level advantage

$$\hat{A}^{\text{CIG}}_{i,t} = A^{\text{out}}_i + \beta\, A^{\text{int}}_{i,t}, \quad \beta{=}1.0, \tag{8}$$

replaces $\hat{A}_{i,t}$ in Eq. 1. When $r^{\text{int}} \equiv 0$ across the prompt group, the estimator reduces to vanilla GRPO.

4 Experiments

4.1 Setup

Models and datasets. We train Qwen2.5 (Yang et al., 2024) at 3B and 7B as the policy backbone, with GPT-4o (OpenAI, 2024) as the user simulator with oracle access to the latent user intent (prompts in App. A). We evaluate on three clarification benchmarks spanning table-, passage-, and open-domain QA (Table 1). **PACIFIC** (Deng et al., 2022) is the primary in-domain benchmark (table-grounded financial QA); **AbgCoQA** (Guo et al., 2021) tests cross-dataset recipe portability (same recipe re-trained, not zero-shot); **AmbigNQ** (Min et al., 2020) adds an ambiguous-heavy open-domain regime. Single-turn retention checks appear in §5.4.

Table 1: **Training datasets.** Three clarification benchmarks spanning table-, passage-, and open-domain QA.

Dataset	Train	Val/Test	%Ambig
PACIFIC	15,087	1,952	16%
AbgCoQA	7,269	1,184	22%
AmbigNQ	19,244	4,377	79%

Baselines. We compare three categories on the same Qwen2.5 backbone: (i) **Prompting-based**, *Direct* (single-turn answer), *FATA* (first-ask-then-answer with a fixed clarify template), and *ReAct* (Yao et al., 2023) (interleaved think–act prompt where the model chooses Clarify or Answer at each turn); (ii) **SFT-only**, our SFT warmstart without RL; and (iii) **External RL clarification recipes**, reported as context anchors rather than head-to-head: ACT (Chen et al., 2025) (offline action-preference DPO on Zephyr-7B) and the concurrent SGP (Berant et al., 2025) (cost-coefficient steerable self-play on Gemma-2-9b). Neither ACT nor SGP has fully released training code, so we cite their paper-reported numbers and leave cells we cannot fill under our protocol blank (Table 2).

Training protocol. All CIGAsk variants initialise from a supervised warmstart that fine-tunes Qwen2.5 on multi-turn clarification trajectories filtered from STaR-GATE-style self-play (details in App. A). RL training runs for up to 300 GRPO steps with 5 rollouts per prompt, sampling up to 3 clarify turns per rollout. The user simulator is a fixed GPT-4o instance prompted with the gold user intent q^*; it is never updated during training. RL on the 7B backbone takes approximately 22 hours on 8×H100.

Metrics. On PACIFIC and AbgCoQA we report token-overlap F1 (overall) and post-clarify F1; on AmbigNQ we report exact match (EM) following Min et al. (2020). Across all three benchmarks we report clarify rates split by gold ambiguity label: Clr-amb (clarify rate on ambiguous samples, recall) and Clr-clr (clarify rate on clear samples, FPR).

4.2 Main results

CIGAsk produces a single Qwen2.5-7B policy that simultaneously leads PACIFIC fullval on F1, post-clarify F1, and ambiguous-clarify recall. F1 on PACIFIC fullval reaches 0.795, ahead of SGP-Oracle (0.787) despite a smaller backbone (Qwen2.5-7B vs. Gemma-2-9b) and without access to gold ambiguity labels at evaluation time. The clear-query clarify rate stays at 26%, well below FATA-style always-asks baselines and comparable to SGP's 21%. No prompted, SFT-only, or external recipe in the table jointly achieves high F1, productive clarification, and selectivity.

The baselines illuminate why outcome reward alone is insufficient. Direct achieves PACIFIC F1 0.312 by committing to one interpretation of every query. FATA forces every query through a clarification turn, degrading F1 to 0.177: mandatory clarification is counterproductive on clear queries. ReAct (0.481) lets the model choose whether to clarify but its clarify rate on ambiguous and clear queries is near-identical (32.9% vs. 41.0%), indicating zero selectivity. SFT raises F1 to 0.581 through better answer quality but never clarifies,

Table 2: **Main results. Bold** = our best model (CIGAsk 7B). PACIFIC fullval (n=1,952, GPT-4o sim); AbgCoQA / AmbigNQ n=500 seed-42 (GPT-4o sim). — = inapplicable or not reported. ‡SGP / SGP-Oracle: Gemma-2-9b on 50/50 balanced dev; we report ambig/unambig F1 mean and place AmbigQA F1 in the AmbigNQ column.

Method	Backbone	PACIFIC fullval (1,952)				AbgCoQA	AmbigNQ
		F1	$F1_{post}$	Clr-amb	Clr-clr	F1	EM
Prompting-based (no training)							
Direct	Qwen2.5-7B	0.312	—	0.000	0.000	0.320	0.058
FATA	Qwen2.5-7B	0.177	0.177	1.000	1.000	0.228	0.076
ReAct (Yao et al., 2023)	Qwen2.5-7B	0.481	0.418	0.329	0.410	0.554	0.076
SFT only (no RL)							
SFT	Qwen2.5-3B	0.543	—	0.000	0.000	0.589	0.056
SFT	Qwen2.5-7B	0.581	—	0.000	0.000	0.638	0.109
External RL clarification recipes (different base & protocol; anchor only)							
ACT (Chen et al., 2025)	Zephyr-7B	0.681	0.620	—	—	—	—
SGP (Berant et al., 2025)‡	Gemma-2-9b	0.726	—	0.429	0.206	—	0.306
SGP-Oracle‡	Gemma-2-9b	0.787	—	0.435	0.156	—	0.359
Ours (CIGAsk)							
CIGAsk (3B)	Qwen2.5-3B	0.673	0.633	0.539	0.179	0.635	0.271
CIGAsk (7B)	**Qwen2.5-7B**	**0.795**	**0.768**	**0.873**	**0.260**	**0.724**	**0.511**

confirming that supervised fine-tuning does not spontaneously produce clarification behavior.

CIGAsk 3B already clears every prompting and SFT baseline and nearly matches ACT (0.673 vs. 0.681), a DPO-trained Zephyr-7B, with half the parameters. The gain from 3B to 7B is not uniform across metrics. F1 rises 12 points (0.673→0.795) while ambiguous-clarify recall jumps 33 points (0.539 → 0.873); the clear-query false-positive rate increases only modestly (0.179→0.260). The selectivity gap (recall minus FPR) widens from 36 to 61 percentage points, indicating that the larger backbone's improvement comes primarily from asking on more ambiguous queries rather than asking on everything indiscriminately. A component ablation at 7B isolates each reward term (Table 3; §5.1): removing CIG while keeping the asymmetric bonus preserves clarify rates but lowers post-clarify F1 by 7 points; removing the asymmetric bonus instead preserves per-clarify quality but drops recall below 20%. The selectivity gap is driven by the asymmetric bonus and is acquired only during RL; it is absent at the SFT warmstart (Fig. 3).

The recipe transfers across datasets without per-dataset hyperparameter search. AbgCoQA differs from PACIFIC in grounding modality (passage vs. table) and dialogue structure (multi-turn conversational QA with history); the same recipe re-trained from scratch reaches 0.635 at 3B and 0.724 at 7B. AmbigNQ pushes the regime further: 79% of items are ambiguous, no grounding context is provided, and evaluation uses exact match rather than token-overlap F1. CIGAsk 7B reaches EM 0.511, more than double every prompting baseline and well above the 3B variant (0.271), suggesting that the CIG signal scales with backbone capacity on harder open-domain regimes. Both CIG and the asymmetric bonus consume only the gold answer y^* and the ambiguity label b that the training data already provides.

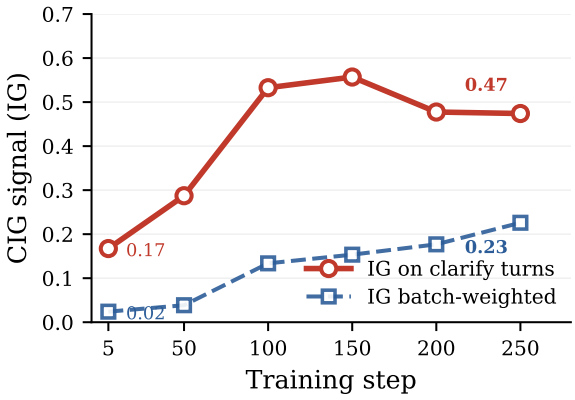

Figure 3: **CIG signal grows with training** (CIGAsk 7B). Six ±25-step anchors. Clarify-only IG: 0.17→0.47. Batch-weighted IG (non-clarify rollouts = 0): 0.02→0.23, crossing 0.2 at step ∼149 where the recall–FPR gap opens.

5 Analysis

5.1 Does the Model Learn *How* to Ask?

CIG is the per-turn term that grades *which* clarification recovers the disambiguating fact. We

test whether it changes the clarifications a policy learns to produce, beyond what an outcome reward already shapes. At 3B, CIGAsk and an outcome-only GRPO variant share the SFT base, seed, asymmetric bonus, KL, and rollout protocol, differing only in whether CIG is active. Both variants learn to clarify at similar rates. Only the CIG-equipped policy turns those clarifications into post-clarify F1; the outcome-only sibling collapses to near-zero (Table 2, 3B row vs. Table 3, −both). The training trajectories overlap early and diverge once CIG begins to separate informative from uninformative clarify turns (Fig. 3). At 7B the same swap shaves only a small slice of post-clarify F1: a stronger base is stable under outcome-only RL, and CIG's role shifts from rescuing stability to tightening selectivity. The magnitude of the CIG contribution thus depends on backbone capacity: at 3B, where the base policy lacks the capacity to self-discover productive clarification from outcome reward alone, CIG is load-bearing; at 7B it refines an already functional policy. This scale dependence is consistent with the broader observation that weaker models benefit more from denser reward signals (Sullivan, 2025). Across late-training checkpoints, items the policy chose to clarify score higher F1 than items it answered directly, with bootstrap 95% CIs above zero at every checkpoint, a self-consistency check on the partition the policy itself induced.

5.2 Does the Model Learn *When* to Ask?

A clarification policy that asks on everything is no more useful than one that asks on nothing. The metric that matters is the gap between recall on truly-ambiguous queries and false-positive rate on clear ones. Removing the asymmetric bonus while keeping CIG and the outcome reward active collapses recall to 17% without recovering any post-clarify F1 (Table 3, −asym). Removing CIG instead leaves selectivity nearly intact but cuts post-clarify F1 at 7B (−CIG, −both). The asymmetric bonus governs *when* to ask; CIG governs *how* well the ask is targeted. Removing the asymmetric bonus also doubles the any-failure rate among clarify turns (14.6% → 30.2% on PACIFIC ambiguous fullval). The dominant failure modes are self-completion, where the model generates both a question and a plausible answer within the same turn without ceding control to the simulator, and declarative reformulation, where the model rephrases the query as a statement rather than posing a genuine question. Both modes bypass the interaction channel that CIG rewards: the user response carries negligible information because the model's turn gave the simulator nothing actionable to answer. Without the decision-level gate the policy asks more often and less carefully, degrading both selectivity and surface form. The operating point on this frontier is governed by ρ: higher ρ tightens selectivity at the cost of recall; our chosen value sits at a productive midpoint, and deployments pick along this axis without retuning the recipe.

Table 3: **Component ablation on PACIFIC fullval.**

Variant	CIG	Asym	F1	$F1_{post}$	Clr-amb	Clr-clr
CIGAsk 7B	✓	✓	**.795**	**.768**	**.873**	.260
−asym	✓	—	.736	.733	.171	.036
−CIG	—	✓	.744	.700	.763	.167
−both	—	—	.683	.657	.087	.035

5.3 Qualitative Case Study

A single PACIFIC trajectory illustrates the point (Box 4). The prompted baseline does ask a question, but it is vague (*"Could you provide more context?"*), so the simulated user simply restates the original query and the model falls back to summing both years. Without a per-turn signal that rewards information recovery, the prompted model has no mechanism to learn *which* slot to ask about; its clarifications remain generic or target irrelevant dimensions. CIGAsk asks about the specific missing slot (the year), recovers the disambiguating fact in one turn, and answers correctly.

Box 4: clarify quality dictates correctness.

Context: *[table: deferred-tax-asset components by year, 2019 & 2018]*
Query: *What was the operating loss carryforward amount?*
Intent: … *as of December 31, 2019.* **Gold answer:** `73,260`

Prompted × *generic clarify → wrong answer*	**CIGAsk (ours)** ✓ *targeted clarify → correct answer*
`#CLARIFY: Could you provide more context?` `#RESPOND: I'm asking about the operating loss carryforward.` `#ANSWER: 131,028`	`#CLARIFY: Which period are you asking about?` `#RESPOND: as of December 31, 2019.` `#ANSWER: 73,260`

The contrast isolates CIG's effect: without a per-turn information-gain signal, a model can learn

when to clarify but not *what* to ask. Its questions stay vague or latch onto an irrelevant dimension of the query, so the user response carries no new information and the model defaults to the same wrong answer it would have produced without clarifying. CIG closes the loop by crediting only clarifications whose user response moves the frozen reference closer to the gold answer, steering the policy toward the specific missing slot. The failure is not a lack of world knowledge — the prompted model knows the table has two years — but a lack of training signal connecting question content to downstream answer quality.

5.4 Out-of-Domain Retention and Selectivity

Reward-shaping for clarification might teach the policy to over-ask at the expense of basic single-turn QA, or fail to transfer the *when*-to-ask skill outside PACIFIC's table-grounded regime. Two never-trained open-QA benchmarks (TriviaQA (Joshi et al., 2017) closed-book, NQ Open (Kwiatkowski et al., 2019) closed-book) rule out both (Table 5). When forced to answer directly, CIGAsk 7B matches the Qwen2.5-7B base within bootstrap CI on both benchmarks, so the PACIFIC gains are not bought by regressing on standard QA. When given the multi-action prompt, the policy clarifies on only a small fraction of items (an order of magnitude below its PACIFIC ambiguous-clarify rate), and the clarifications take the form of targeted disambiguations rather than search-like rewrites (App. A.4). Manual inspection confirms the transferred skill: questions target genuine lexical ambiguity or temporal scope (e.g. "Which season are you referring to?") rather than generic requests for more context. This pattern mirrors the PACIFIC training distribution, where CIG rewards only clarifications whose user response raises the gold answer's log-probability, and suggests that the policy has internalised a general disambiguation skill rather than overfitting to financial-table structure. The *when*-to-ask skill transfers out-of-domain at a sensible rate; the *how*-to-ask quality holds.

Table 5: **OOD retention** on TriviaQA + NQ Open closed-book (n=500, seed 42).

Bench	CIGAsk F1	Base F1	Clarify-rate
TriviaQA	56.2	54.0	5.6%
NQ Open	25.9	25.1	9.4%

6 Conclusion

We introduced **CIGAsk**, a reinforcement learning recipe that teaches an instruction-tuned LLM both *when* and *how* to clarify under-specified queries inside a multi-turn GRPO loop. The recipe rests on two signals that divide the problem cleanly: Counterfactual Information Gain supplies a per-turn frozen-reference reward that steers clarification quality, while the Asymmetric Ambiguity Bonus supplies a terminal gate that steers selectivity. The ablation in §5.1–5.2 shows each signal is necessary and insufficient alone: CIG without the gate produces high-quality but rare clarifications; the gate without CIG produces frequent but less targeted ones.

Empirically, CIGAsk 7B leads PACIFIC full-val F1 (0.795) on a smaller backbone than the strongest external recipe while transferring to AbgCoQA and AmbigNQ with no per-dataset tuning and preserving single-turn QA on TriviaQA and NQ Open. The broader implication is that a frozen-reference counterfactual signal, which requires only the gold answer the outcome reward already consumes, can replace both a learned critic and step-labelled supervision for multi-turn credit assignment. Whether this principle extends to other agentic actions beyond clarification, such as tool calls or search queries, is a natural next step.

Limitations. Training and evaluation rely on a single GPT-4o simulator with oracle access to the true intent; we have not tested how sensitive CIG is to simulator quality, and a weaker or noisier simulator may degrade the per-turn signal. The asymmetric bonus requires a per-instance ambiguity label b that may not exist in unannotated deployment streams; relaxing this to a soft ambiguity score or learning it from data is an open problem. All experiments use Qwen2.5 at 3B and 7B; whether the recipe transfers to other model families or to larger scales remains untested. We evaluate on English-only benchmarks; multilingual transfer is unexplored.

Ethical considerations. The asymmetric bonus inherits annotator judgements of what counts as ambiguous; downstream policies may behave unevenly across user populations whose phrasing conventions differ. Deployers should re-tune ρ on their own distribution.

Reproducibility Statement

We will release our code and the remixed dataset upon acceptance.

References

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2020. Convai3: Generating clarifying questions for open-domain dialogue systems (clariq). *CoRR*, abs/2009.11352.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 475–484. ACM.

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2024. Star-gate: Teaching language models to ask clarifying questions. *CoRR*, abs/2403.19154.

Jonathan Berant, Maximillian Chen, Adam Fisch, Reza Aghajani, Fantine Huot, Mirella Lapata, and Jacob Eisenstein. 2025. Learning steerable clarification policies with collaborative self-play. *CoRR*, abs/2512.04068.

Maximillian Chen, Ruoxi Sun, Tomas Pfister, and Sercan Ö. Arik. 2025. Learning to clarify: Multi-turn conversations with action-based contrastive self-training. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. 2022. PACIFIC: towards proactive conversational question answering over tabular and textual data in finance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6970–6984. Association for Computational Linguistics.

Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. Abg-coqa: Clarifying ambiguity in conversational question answering. In *3rd Conference on Automated Knowledge Base Construction, AKBC 2021, Virtual, October 4-8, 2021*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron C. Courville, and Nicolas Le Roux. 2025. Vineppo: Refining credit assignment in RL training of llms. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 996–1009. Association for Computational Linguistics.

Fanqi Kong, Jiayi Zhang, Mingyi Deng, Chenglin Wu, Yuyu Luo, and Bang Liu. 2026. Infopo: Information-driven policy optimization for user-centric agents. *CoRR*, abs/2603.00656.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM.

Yoonsang Lee, Xi Ye, and Eunsol Choi. 2024. Ambigdocs: Reasoning across documents on different entities under the same name. *CoRR*, abs/2404.12447.

Belinda Z. Li, Been Kim, and Zi Wang. 2025. Questbench: Can llms ask the right question to acquire information in reasoning tasks? *CoRR*, abs/2503.22674.

Xiaoqian Liu, Ke Wang, Yuchuan Wu, Fei Huang, Yongbin Li, Junge Zhang, and Jianbin Jiao. 2025. Agentic reinforcement learning with implicit step rewards. *CoRR*, abs/2509.19199.

David J. C. MacKay. 1992. Information-based objective functions for active data selection. *Neural Comput.*, 4(4):590–604.

Davide Mazzaccara, Alberto Testoni, and Raffaella Bernardi. 2024. Learning to ask informative questions: Enhancing llms with preference optimization and expected information gain. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 5064–5074. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5783–5797. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Proceedings of Machine Learning Research, pages 2778–2787. PMLR.

Cheng Qian, Zuxin Liu, Akshara Prabhakar, Jielin Qiu, Zhiwei Liu, Haolin Chen, Shirley Kokane, Heng Ji, Weiran Yao, Shelby Heinecke, Silvio Savarese, Caiming Xiong, and Huan Wang. 2025. Userrl: Training interactive user-centric agent via reinforcement learning. *CoRR*, abs/2509.19736.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2737–2746. Association for Computational Linguistics.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2025. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems, EuroSys 2025, Rotterdam, The Netherlands, 30 March 2025 - 3 April 2025*, pages 1279–1297. ACM.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8273–8288. Association for Computational Linguistics.

Michael Sullivan. 2025. GRPO is secretly a process reward model. *CoRR*, abs/2509.21154.

Manan Suri, Puneet Mathur, Nedim Lipka, Franck Dernoncourt, Ryan A. Rossi, and Dinesh Manocha. 2025. Structured uncertainty guided clarification for LLM agents. *CoRR*, abs/2511.08798.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, and Zhenzhe Ying. 2025. Information gain-based policy optimization: A simple and effective approach for multi-turn llm agents. *arXiv preprint arXiv:2510.14967*.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9426–9439. Association for Computational Linguistics.

Quan Wei, Siliang Zeng, Chenliang Li, William Brown, Oana Frunza, Wei Deng, Anderson Schneider, Yuriy Nevmyvaka, Yang Katie Zhao, Alfredo Garcia, and 1 others. 2025. Reinforcing multi-turn reasoning in llm agents via turn-level reward design. *arXiv preprint arXiv:2505.11821*.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, and 16 others. 2025. DAPO: an open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. 2025. Free process rewards without process labels. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Michael J. Q. Zhang, W. Bradley Knox, and Eunsol Choi. 2025. Modeling future conversation turns to teach llms to ask clarifying questions. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Jiale Zhao, Ke Fang, and Lu Cheng. 2026. When and what to ask: Askbench and rubric-guided RLVR for LLM clarification. *CoRR*, abs/2602.11199.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. Pytorch FSDP: experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16(12):3848–3860.

Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. 2025. SWEET-RL: training multi-turn LLM agents on collaborative reasoning tasks. *CoRR*, abs/2503.15478.

A Implementation Details

This appendix lists training hyperparameters (§A.1), dataset sources and licenses (§A.2), verbatim prompts (§A.3), and qualitative clarify-turn examples on in-domain and out-of-domain probes (§A.4).

A.1 Training hyperparameters

Table 6 lists every hyperparameter. The reward configuration is identical across PACIFIC and AbgCoQA; the AbgCoQA warmstart adds a small per-turn cost (-0.03) for long-horizon stability. All reported numbers are single-seed (PyTorch / vLLM seed 42); the 3B paired comparison in Table 3 is matched-seed across configurations.

Table 6: **CIGAsk training hyperparameters.** Identical across PACIFIC and AbgCoQA.

Group	Parameter	Value
Reward	α (CIG weight)	0.3
	CIG clip range (nats)	$[-0.5, 2.0]$
	ρ (asym bonus)	0.30
	malformed-clarify penalty	-0.30
GRPO	framework	verl (Sheng et al., 2025)
	rollouts per prompt	5
	KL coefficient	0.08
	learning rate	5×10^{-7}
	steps (max)	300
	batch size	32
	$\epsilon_{\text{high}}/\epsilon_{\text{low}}$	1.4 (Yu et al., 2025)
Rollout	engine	vLLM (Kwon et al., 2023)
	temperature	1.0
	prompt tokens (max)	6,144
	response tokens (max)	2,048
	max clarify turns	3
SFT warmstart	engine	FSDP (Zhao et al., 2023)
	steps	84
	learning rate	1×10^{-5}
	corpus size	$\sim$1,300 multi-turn
Compute	7B RL wall-clock	$\sim$22 h
	3B RL wall-clock	$\sim$10 h
	SFT wall-clock	$\sim$1.5 h
	hardware	8×H100

A.2 Dataset details

Table 7 lists size, split, prompting regime, source citation, and license for every dataset. PACIFIC, AbgCoQA, and AmbigNQ supply training and in-domain evaluation; TriviaQA and NQ Open are held-out probes for out-of-domain retention and the multi-action clarify probe (§5.4). All evaluations decode greedily (T=0 for direct extraction, T=1.0 for rollouts as in Table 6); scoring uses token-overlap F1 on PACIFIC and AbgCoQA and exact match on AmbigNQ following Min et al. (2020), with regex extraction from the ⟨answer⟩...⟨/answer⟩ span.

A.3 Prompts

Verbatim system and user prompts used during training and evaluation.

POLICY SYSTEM PROMPT (training & eval)

```
You are a helpful agent.  Available actions:
THINK, CLARIFY, ANSWER.
- THINK: reason internally.  Wrap in
<think></think>.
- CLARIFY: ask ONE question.  Wrap in
<clarify></clarify>.
- ANSWER: give final answer.  Wrap in
<answer></answer>.  ONLY the value, no
explanation.
First THINK, then CLARIFY if ambiguous, then
ANSWER.
```

Table 7: **Dataset details.** Source, used size, prompting regime, and license for every dataset in the paper.

Dataset	Description	Used size & protocol	Source	License
Training and in-domain evaluation				
PACIFIC	Proactive conversational QA over financial tables and text; each example carries a gold ambiguity label and clarification reference.	Train 15,087; fullval 1,952 (n_{amb}=316, n_{cl}=1,636). Multi-turn with GPT-4o simulator; up to 3 clarify turns.	Deng et al. (2022)	MIT
AbgCoQA	Conversational QA derived from CoQA with annotated ambiguous turns and clarification references.	Train 7,269; fullval 1,184; eval subset n=500 seed 42 with GPT-4o simulator.	Guo et al. (2021)	MIT
AmbigNQ	Ambiguous open-domain questions derived from Natural Questions, with multiple disambiguated targets per query.	Train 19,244; test 4,377; eval subset n=500 seed 42 with GPT-4o simulator.	Min et al. (2020)	CC BY-SA 3.0
Out-of-distribution probes (never trained on)				
TriviaQA (closed-book)	Trivia QA over short factoid questions; closed-book setting used as a single-turn retention check.	n=500 seed 42; forced-direct and multi-action prompts (§5.4).	Joshi et al. (2017)	Apache 2.0
NQ Open (closed-book)	Open-domain factoid questions from Natural Questions; closed-book setting.	n=500 seed 42; same protocol as TriviaQA.	Kwiatkowski et al. (2019)	Apache 2.0

SIMULATOR SYSTEM PROMPT (GPT-4o; PACIFIC)

```
You simulate a user who asked a question to a
QA assistant.
ABSOLUTE RULES (violating any = failure):
1. You do NOT know the answer. NEVER include
numbers, dollar amounts, percentages, or
computed values.
2. Answer ONLY the ONE dimension asked. If
asked ``which year?'' → say ONLY the year.
NEVER combine multiple pieces of info.
3. If the assistant's question is vague (e.g.
``more details?''), push back: ask THEM to be
specific. Do NOT dump your info.
4. If the question is irrelevant, say so.
5. Keep response under 20 words.
GOOD examples. Q: ``What is the revenue?''
| Meant: 2019. Assistant ``Which year?''
→ ``Fiscal year 2019.'' Assistant ``Can
you provide more details?'' → ``What
specifically do you need?''
BAD examples. ``From Q1 to Q2 in 2019, the
gross profit increased.'' (combined period
+ metric + direction); ``The increase was
$13,407.'' (included a number); ``Sure, here
are all the details ...'' (dumped everything).
```

PER-TURN USER MESSAGE (each clarify turn)

```
Your original question: ``[ambiguous q]''
What you actually meant: ``[gold intent q*]''
The assistant asks: ``[clarification
question]''
How do you respond? (Brief, natural, NO
numbers or data values)
```

A.4 OOD clarify-turn examples

On the small subset where CIGAsk 7B chooses to clarify rather than answer (5.6% on TriviaQA, 9.4% on NQ Open; §5.4), the clarifications are sensible disambiguations: *"Do you mean which US gangster was scheduled to be released from Alcatraz prison in November 1939?"* (year-specific Trivia query); *"Which year are you asking about?"* on *"who won oscar for best director this month"* (NQ query with a relative time reference); *"Which episode are you asking about?"* on a series-by-name query that names the season but not the episode. The selectivity learned on PACIFIC table-grounded ambiguity transfers to text-grounded open QA without retraining.