

Learning Diagnostic Clarifying Questions for Conversational Question Answering via Data-Efficient Dynamic Distillation

Jihyeong Jeon¹ Cheol Ryu² U Kang¹

¹Seoul National University ²Samsung Electronics

{jeonjihyeong, ukang}@snu.ac.kr, cheol.ryu@samsung.com

Abstract

Given a query, how can a QA system ask an informative clarifying question to accurately answer it? Open-domain conversational question answering requires not only detecting when a query is ambiguous but also asking clarifying questions that efficiently uncover the latent intent of the user. However, existing approaches either rely on prompt-only clarify-or-answer decisions with limited robustness or distill clarification trajectories without explicitly modeling intent convergence. In this paper, we propose D4CQ (**D**ata-Efficient **D**iagnosticity-based **D**ynamic **D**istillation for **C**larifying **Q**uestion), a data-efficient framework for learning diagnostic clarifying questions through dynamic distillation. Our approach distills a structured refinement protocol, prioritizes informative clarifications via diagnosticity-based reweighting, and improves sample efficiency through hindsight relabeling. Empirical results demonstrate that D4CQ consistently outperforms strong baselines, while maintaining comparable interaction budgets.

1 Introduction

Given a query, how can a QA system ask an informative clarifying question to accurately answer? Open-domain conversational question answering (Anantha et al., 2021; Kim et al., 2017) resolves missing information through interaction rather than one-shot prediction from an incomplete query (Deng et al., 2023; Andukuri et al., 2024). Its core challenge is to clarify only when needed and answer once the intent is clear. This is particularly difficult in open-domain settings, where a query may admit multiple plausible intents with different valid answers (Min et al., 2020). Nevertheless, addressing this challenge is crucial for building practical QA systems that are both accurate and efficient in real-world use (Yoon et al., 2018; Jung et al., 2017; Lee and Kang, 2025).

Existing approaches fall into two broad categories. Prompt-based ambiguity-aware methods decide whether to clarify or answer, but rely mainly on inference-time prompting, limiting robustness and generalization across ambiguity types (Press et al., 2023; Cole et al., 2023). Trajectory-distillation methods instead learn a clarify-or-answer policy from teacher-generated candidates (Zhang et al., 2025).

However, two limitations remain. First, supervision is solely defined over clarifying question candidates, without teaching the model to *reason* about unresolved ambiguity axes before asking. Thus, the model may learn *which* candidate to prefer, but not *how* to progressively disambiguate the query. Second, preferred clarifying questions are used as positive supervision without explicitly accounting for their diagnostic quality. A question may be preferred because it produces a better downstream answer in one rollout, even if a rejected one is more informative about the user’s latent intent.

To address these, we propose D4CQ (**D**ata-Efficient **D**iagnosticity-based **D**ynamic **D**istillation for **C**larifying **Q**uestion), a data-efficient framework for learning *diagnostic* clarifying questions through dynamic distillation. Our core idea is twofold. *Procedural scaffolding* distills an explicit refinement protocol for rewriting the query and identifying unresolved ambiguity axes. *Dynamic distillation* estimates the diagnosticity of each clarification step and reweights training toward more informative questions. We further improve training with *contrastive intent-convergence alignment*, which moves each refined query toward the ground-truth latent intent and *hindsight relabeling*, which safely recovers supervision from discarded rollouts.

Extensive experiments show that D4CQ consistently improves end-to-end OD-CQA performance. On AmbigQA, it achieves the best performance across multiple backbones under a comparable clarification budget. It also generalizes well to Situat-

edQA, improving ambiguity detection and direct answer precision under distribution shift.

Our contributions are summarized as follows:

- We propose D4CQ, a data-efficient framework for learning to ask diagnostic clarifying questions in open-domain conversational question answering.
- We introduce three key techniques—procedural scaffolding with contrastive intent-convergence alignment, dynamic distillation, and hindsight relabeling—that jointly enable structured ambiguity resolution, quality-aware distillation, and efficient reuse of discarded rollouts.
- Experiments on AmbigQA and zero-shot evaluation on SituatedQA show that D4CQ achieves the state-of-the-art performance in ambiguity resolution and out-of-domain generalization.

2 Background

Modeling conversational question answering systems. Conversational QA (CQA) systems (Deng et al., 2023; Andukuri et al., 2024; Zhang et al., 2025) are designed to interactively elicit missing information through dialogue rather than answering an open-domain question in one shot (Karpukhin et al., 2020; Izacard and Grave, 2021; Lee et al., 2025). In particular, given a user query q , let the dialogue history before turn k be $h_k = (q, c_1, r_1, \dots, c_{k-1}, r_{k-1})$, where c_k and r_k are the clarifying question and user response at turn k , respectively. Let $K_{\max}$ denote the maximum number of clarifying turns. Given a user query q , a CQA pipeline consists of two language models: (i) an *interrogator* that proposes a clarifying question or a direct answer and (ii) a *responzor* that generates a response to the clarifying question. Notably, while the responzor would ordinarily correspond to a human user, we model it in this work as a pretrained language model kept fixed without additional training. Formally, for a trajectory τ with $K(\tau) \leq K_{\max}$ clarifying turns, the interrogator produces $y_k \sim p_\theta(\cdot | h_k)$ with $y_k = c_k$ for $1 \leq k \leq K(\tau)$ and $y_{K(\tau)+1} = \hat{a}$ as the final answer, and the responzor generates $r_k \sim p_\phi(\cdot | q, c_k, x, A(x))$, where x is the latent intent of the question and $A(x)$ is a set of plausible answers—a single answer can be referred to in multiple ways (e.g., *Pope Francis* and *Jorge Mario Bergoglio*) (Min et al., 2020). Here, the latent intent x is equivalent to q if q is unambiguous; otherwise, x is an unambiguous version of q that would di-

rectly lead to some $a \in A(x)$ without further clarification under a perfectly informed interrogator.

In practice, we train the interrogator via sequence-level distillation (Kim and Rush, 2016) from trajectories. Concretely, given an open-domain QA dataset $\{(q, A(x))\}$, a teacher interrogator and a responzor are rolled out to produce trajectories $\tau = (q, \bar{c}_1, r_1, \dots, \bar{c}_{K(\tau)}, r_{K(\tau)}, \bar{a})$, where $\bar{a}$ is the final answer of the teacher interrogator in the rollout. Only trajectories whose final answer is sufficiently close to the gold answer set are retained, e.g., $\psi(\bar{a}, A(x)) \geq \eta$, where ψ can be F1-score and η is a threshold. Then, a student interrogator is fine-tuned by maximizing the likelihood of the observed clarifying questions and the final answer under teacher forcing. Specifically, let the supervised targets be $y_1 = \bar{c}_1, \dots, y_{K(\tau)} = \bar{c}_{K(\tau)}, y_{K(\tau)+1} = \bar{a}$, and write each output as a token sequence $y_k = (y_{k,1}, \dots, y_{k,T_k})$ under the chosen tokenizer, where T_k is the token length of y_k . Under an autoregressive factorization, we have

$$p_\theta(y_k | h_k) = \prod_{t=1}^{T_k} p_\theta(y_{k,t} | h_k, y_{k,<t}),$$

where $y_{k,<t} = (y_{k,1}, \dots, y_{k,t-1})$.

Accordingly, the teacher-forced negative log-likelihood over retained trajectories $\mathcal{D}$ is:

$$\mathcal{L}(\theta) = - \sum_{\tau \in \mathcal{D}} \sum_{k=1}^{K(\tau)+1} \sum_{t=1}^{T_k(\tau)} \log p_\theta(y_{k,t} | h_k, y_{k,<t}), \quad (1)$$

where $T_k(\tau)$ is the token length of y_k in τ .

Open-domain conversational question answering. We give a formal definition of the Open-Domain Conversational Question Answering (OD-CQA) problem as follows:

PROBLEM 1 (OPEN-DOMAIN CONVERSATIONAL QUESTION ANSWERING).

Given:

- A user question q associated with latent intent x and its set of acceptable answer strings $A(x)$.
- A dialogue budget $K_{\max}$.

Find: An answer $\hat{a}$ after interacting with the responzor for $K(\tau) \leq K_{\max}$ turns.

Objective: Maximize the correctness score $\psi(\hat{a}, A(x)) \in [0, 1]$, defined as follows. Let $\lambda(\cdot)$ map a string to a multiset of normalized tokens

(e.g., lowercasing and punctuation removal). Then

$$\text{Prec}(\hat{a}, a) = \frac{|\lambda(\hat{a}) \cap \lambda(a)|}{|\lambda(\hat{a})|}, \quad \text{Rec}(\hat{a}, a) = \frac{|\lambda(\hat{a}) \cap \lambda(a)|}{|\lambda(a)|},$$

$$\psi(\hat{a}, A(x)) = \max_{a \in A(x)} \frac{2 \text{Prec}(\hat{a}, a) \text{Rec}(\hat{a}, a)}{\text{Prec}(\hat{a}, a) + \text{Rec}(\hat{a}, a)}, \quad (2)$$

with the convention that $\psi(\hat{a}, A(x)) = 0$ if the denominator is zero.

3 Proposed Method

In this section, we present D4CQ and summarize its main design motivations. Our approach addresses the following challenges:

- C1 Structured Multi-turn Disambiguation with Intent Convergence.** Standard SFT on unstructured traces provides neither a clear multi-turn protocol nor an explicit signal for convergence to the true latent intent. *How can we induce a structured multi-turn refinement process that explicitly converges to the correct latent intent?*
- C2 Prioritizing Diagnostic Clarifying Questions.** Clarifying questions in retained trajectories may support the final answer, but many are weakly informative and do little to distinguish among plausible latent intents. Training on such low-diagnostic questions can encourage unnecessary interactions and hurt answer quality. *How can we quantify clarifying question diagnosticity and prioritize questions that better discriminate among plausible latent intents?*
- C3 Mitigating Inefficient Data Sampling.** Strict trajectory filtering retains only a small subset of successful rollouts, reducing supervision and often favoring easier cases over challenging or long-tail ambiguity patterns. *How can we improve sample efficiency while preserving coverage of difficult and diverse ambiguity patterns?*

To address these challenges, our D4CQ combines the following three core ideas:

- I1 Procedural Scaffolding with Contrastive Intent-Convergence Alignment (§3.1).** We distill a structured multi-turn refinement protocol from a teacher with a few-shot CoT scaffold and use a within- q contrastive objective to align refined queries with the ground-truth intent against alternatives.
- I2 Diagnosticity-based Dynamic Distillation (§3.2).** We score each retained clarifying question according to how effectively its paired response differentiates the ground-truth intent from competing intents, and use this score to reweight the sequence-level SFT loss.

I3 Hindsight Relabeling with Compatibility

Gating (§3.3). We recycle failed rollouts that match alternative gold answers under other plausible intents, while filtering unreliable relabels with a likelihood-based compatibility gate.

Figure 1 summarizes how D4CQ integrates these components into a unified framework.

3.1 Procedural Scaffolding with Contrastive Intent-Convergence Alignment

Ambiguous queries often require resolving multiple ambiguity axes over several turns (Jeon et al., 2025). In such cases, naïve SFT is insufficient, as token-level imitation provides neither a structured disambiguation procedure nor an explicit signal for convergence to the true latent intent. To address this, we introduce *procedural scaffolding*, a structured protocol for multi-turn query refinement, together with *contrastive intent-convergence alignment*, which pulls refined query representations toward the ground-truth intent and away from alternative intents for the same query. Together, these components train an interrogator to progressively refine the query through interaction and produce the correct final answer. A summary of the scaffold prompt is provided in Appendix A.2.

Procedural scaffolding. We elicit structured clarification trajectories by prompting a large teacher interrogator with a step-wise instruction template that decomposes multi-turn disambiguation into the following stages:

- I. Recap and query refinement.** Rewrite the current question by incorporating information from prior dialogue turns.
- II. Parametric background recall.** Retrieve relevant background knowledge from the model’s internal knowledge related to the question.
- III. Singleton check.** Determine whether the question is intended to have a single atomic answer or inherently allows multiple answers.
- IV. Ambiguity axis discovery by replacement.** Identify answer-changing term substitutions that yield different valid answers.
- V. Hidden ambiguity discovery by addition and deletion.** Identify key qualifiers that can be added or removed in the question.

A key component is the *query refinement* step. As dialogue histories grow, models may fail to reliably use early context, exhibiting positional sensitivity that attenuates the influence of the initial query in long inputs (Liu et al., 2024). By explicitly rewriting the question into a context-consistent form after

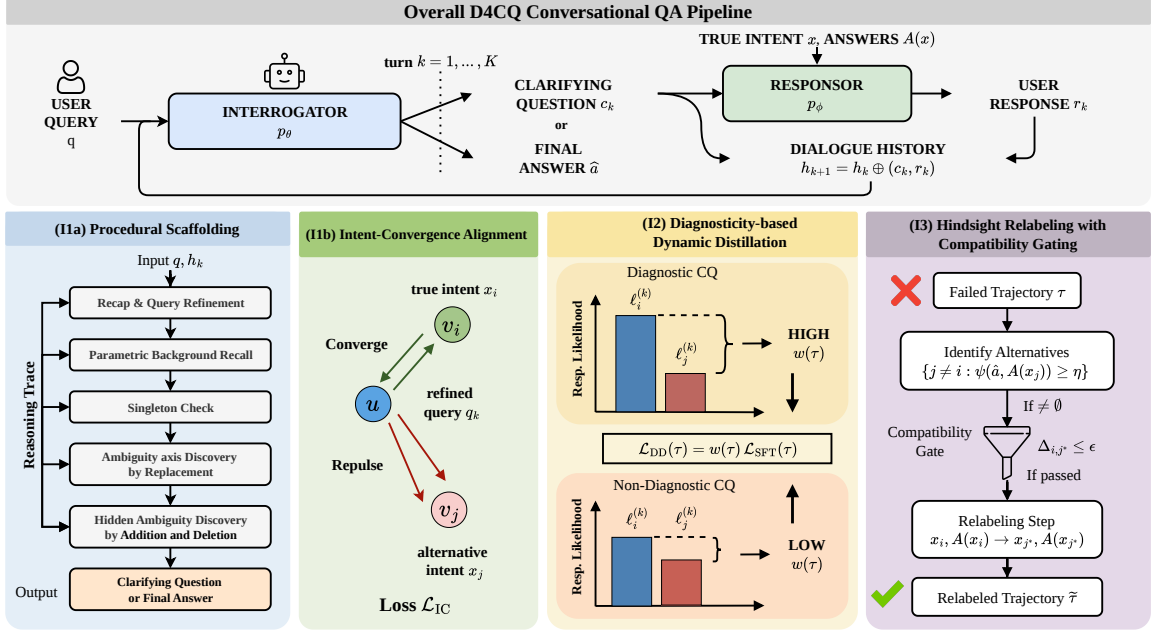


Figure 1: Overview of D4CQ, a data-efficient framework for learning diagnostic clarifying questions via dynamic distillation. The top panel depicts the multi-turn conversational QA pipeline and the bottom panels present its four main components: (I1a) procedural scaffolding for structured clarification, (I1b) intent-convergence alignment for refining intent representations, (I2) diagnosticity-based dynamic distillation for prioritizing informative clarifying questions, and (I3) hindsight relabeling with compatibility gating for recycling failed trajectories.

each user response, the interrogator re-centers subsequent reasoning on the most up-to-date intent hypothesis, improving focus on the remaining unresolved ambiguity axes.

Contrastive intent-convergence alignment.

Beyond token-level imitation, we introduce a representation-level objective on the final refined query produced in the answer-generation turn, encouraging it to move toward the ground-truth intent and away from within-query alternatives. Consider a retained trajectory $\tau = (q, \bar{c}_1, r_1, \dots, \bar{c}_{K(\tau)}, r_{K(\tau)}, \bar{a})$ for an ambiguous query q with ground-truth latent intent x_i . Let $h_{K(\tau)+1}$ denote the dialogue history before the final interrogator turn and let $\pi(\tau)$ denote the teacher-forced reasoning prefix in that turn up to, but excluding the final refined query $\bar{q}$. We then define the conditional representation of any token sequence s as

$$z_\theta(\tau, s) = f_\theta(h_{K(\tau)+1} \oplus \pi(\tau) \oplus s), \quad (3)$$

where $\oplus$ denotes sequence concatenation, and $f_\theta(\cdot)$ is the final-layer hidden state at the last token of s obtained under teacher forcing.

Let $u = z_\theta(\tau, \bar{q})$ and $v_j = z_\theta(\tau, x_j)$, where $\{x_j\}_{j=1}^{m_q}$ denotes the set of plausible latent intents for q , and m_q is the total number of such intents. We align the final refined query representation with the ground-truth intent x_i while repelling alterna-

tives by minimizing the following InfoNCE-like (Oord et al., 2018) loss:

$$\mathcal{L}_{IC}(\tau) = -\log \frac{\exp(\text{sim}(u, v_i)/\gamma)}{\sum_{j=1}^{m_q} \exp(\text{sim}(u, v_j)/\gamma)}, \quad (4)$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and $\gamma > 0$ is a temperature parameter. This objective provides a direct optimization signal that encourages the final refined query under the same answer-generation context to converge toward the true latent intent.

3.2 Diagnosticity-based Dynamic Distillation

Not every successful teacher trajectory is equally useful for distillation. Some clarifying questions are verbose or confirmatory, and their paired responses remain compatible with many candidate intents. Uniformly imitating such trajectories can teach the student to ask unnecessary questions without meaningfully reducing intent uncertainty. Thereafter, we score the diagnosticity of each clarifying question-response pair and reweight distillation toward trajectories that more sharply identify the true latent intent.

Diagnosticity scoring with response likelihood. Consider a retained trajectory $\tau = (q, \bar{c}_1, r_1, \dots, \bar{c}_{K(\tau)}, r_{K(\tau)}, \bar{a})$ for query q with ground-truth intent x_i . For the k th clarifying question $\bar{c}_k$ and its paired response r_k , we evaluate how selectively the response supports each can-

didate intent x_j by computing the teacher-forced log-likelihood under the responsor:

$$\ell_j^{(k)} = \frac{1}{|r_k|} \log p_\phi(r_k \mid q, \bar{c}_k, x_j, A(x_j)), \quad (5)$$

for $j \in \{1, \dots, m_q\}$, where ϕ is the fixed parameters of the responsor and length normalization makes scores comparable across responses of different lengths. A clarifying question is diagnostic if the observed response is substantially more likely under the true intent than under competing intents. We measure this with the log-mean-exp margin:

$$d^{(k)}(\tau) = \ell_i^{(k)} - \log \left(\frac{1}{m_q - 1} \sum_{j \neq i} \exp(\ell_j^{(k)}) \right). \quad (6)$$

Large $d^{(k)}(\tau)$ indicates that r_k provides intent-specific evidence elicited by $\bar{c}_k$, rather than a generic response that fits many alternatives.

Dynamic distillation weighting. We aggregate turn-level diagnosticity into a trajectory-level score $\bar{d}(\tau) = \frac{1}{K(\tau)} \sum_{k=1}^{K(\tau)} d^{(k)}(\tau)$, and convert it into a nonnegative weight:

$$w(\tau) = \text{softplus}(\beta \bar{d}(\tau)).$$

For trajectories with $K(\tau) = 0$, we set $w(\tau) = 1$. We then scale the token-level distillation loss over the interrogator outputs by

$$\mathcal{L}_{\text{DD}}(\tau) = w(\tau) \sum_{t \in \mathcal{T}(\tau)} \text{CE}_t, \quad (7)$$

where $\mathcal{T}(\tau)$ denotes the interrogator-token positions in the trajectory and CE_t denotes the token-level cross-entropy loss at position t under teacher forcing. The weight is computed from the fixed teacher rollout and detached from student gradients. Consequently, highly diagnostic interactions contribute more strongly to learning, whereas low-diagnostic supervision is downweighted.

3.3 Hindsight Relabeling with Compatibility Gating

We next address the sample inefficiency caused by strict trajectory filtering. Requiring exact or near-exact terminal correctness under the original intent discards many rollouts that are locally sensible but land on a different plausible answer for the same ambiguous query. Rather than treating all such failures as useless, we recycle trajectories whose terminal prediction is compatible with an alternative

intent, while explicitly checking that the observed responses are not grossly inconsistent with the re-labeled intent.

Hindsight relabeling. Consider a failed rollout $\tau = (q, \bar{c}_1, r_1, \dots, \bar{c}_{K(\tau)}, r_{K(\tau)}, \hat{a})$ generated under the ground-truth intent x_i , i.e., $\psi(\hat{a}, A(x_i)) < \eta$. Let $\{x_j\}_{j=1}^{m_q}$ be the candidate intents associated with q . We identify alternative intents whose answer sets are sufficiently compatible with the terminal prediction:

$$\mathcal{I}(\tau) = \{j \neq i : \psi(\hat{a}, A(x_j)) \geq \eta\}.$$

If $\mathcal{I}(\tau) \neq \emptyset$, we have the best-matching alternative:

$$j^* = \arg \max_{j \in \mathcal{I}(\tau)} \psi(\hat{a}, A(x_j)).$$

To avoid oversampling near-duplicate dialogues for the same target answer set, we relabel only if no retained trajectory already exists for the pair $(q, A(x_{j^*}))$. The resulting relabeled trajectory $\tilde{\tau}$ preserves the original dialogue tokens but replaces the intent label with x_{j^*} .

Compatibility gating. A relabeled trajectory is useful only if the observed responsor turns remain reasonably plausible under the relabeled intent. Using the per-turn log-likelihoods from Eq. 5, we define the compatibility gap between the original intent x_i and the relabeled intent x_{j^*} as

$$\Delta_{i,j^*}(\tau) = \sum_{k=1}^{K(\tau)} (\ell_i^{(k)} - \ell_{j^*}^{(k)}). \quad (8)$$

We accept $\tilde{\tau}$ only if $\Delta_{i,j^*}(\tau) \leq \epsilon$, which filters out cases where the relabeled intent explains the observed responses substantially worse than the original one. For accepted relabeled samples, we supervise only the interrogator reasoning and clarifying question tokens, excluding the final answer span. This allows the student to inherit useful disambiguation behavior from failed rollouts without forcing answer tokens that were produced under a different intent-conditioning context.

3.4 Overall Training Objective

Let $\mathcal{D}_o$ and $\mathcal{D}_r$ denote the originally retained trajectories and the accepted hindsight-relabeled trajectories, respectively. For any training sample τ , the token-level term is $\mathcal{L}_{\text{DD}}(\tau)$ in Eq. (7); for $\tau \in \mathcal{D}_r$, the supervised token set $\mathcal{T}(\tau)$ excludes the final answer span. Our overall objective is

$$\mathcal{L} = \frac{1}{|\mathcal{D}_o \cup \mathcal{D}_r|} \sum_{\tau \in \mathcal{D}_o \cup \mathcal{D}_r} (\mathcal{L}_{\text{DD}}(\tau) + \lambda \mathcal{L}_{\text{IC}}(\tau)), \quad (9)$$

where $\lambda > 0$ balances token-level distillation and representation-level intent convergence. For hindsight-reabeled samples, we use the relabeled intent when computing both the diagnosticity scores and the positive target in $\mathcal{L}_{IC}$. The inference-time procedure is described in Appendix D.3.

4 Experiments

We aim to answer the following research questions:

- RQ1 **Overall effectiveness (§4.2)**. Does D4CQ improve end-to-end OD-CQA performance on both ambiguous and unambiguous questions under a comparable interaction budget?
- RQ2 **Component ablation and hyperparameter sensitivity (§4.3)**. How much does each component contribute, and how sensitive is performance to key hyperparameters?
- RQ3 **Intent-convergence analysis (§4.4)**. Does the learned reasoning process move the refined query toward the ground-truth latent intent over turns?
- RQ4 **Ambiguity handling under distribution shift (§4.5)**. Does D4CQ ask for clarification on ambiguous inputs and answer directly on unambiguous inputs, and does this behavior transfer to an out-of-domain benchmark?

4.1 Experimental Setup

Dataset. Our primary benchmark is **AmbigQA** (Min et al., 2020), which provides open-domain questions with multiple plausible latent intents and answer aliases. For zero-shot evaluation in Section 4.5, we use **SituatedQA-Geo** and **SituatedQA-Temp** (Zhang and Choi, 2021), which focus on geographically and temporally underspecified questions, respectively. Detailed dataset statistics are summarized in Appendix C.

Evaluation metrics. For AmbigQA, we report alias-max F1 as defined in Eq. (2) and separately present $F1_u$, $F1_a$, and $F1_{all}$ for the unambiguous, ambiguous, and full subsets, respectively. We also report the average number of generated clarification turns, denoted by “#”. For SituatedQA, we report ambiguity detection F1 ($F1_{detect}$) and unambiguous prediction F1 ($F1_{unp}$), which respectively evaluate whether a method asks for clarification when needed, and answers directly when the input is already sufficiently specified.

Baselines. We compare D4CQ against several representative ambiguity-handling baselines, including single-turn direct answering, prompt-based

Method	LLAMA3 8B				GEMMA 7B			
	#	$F1_u$	$F1_a$	$F1_{all}$	#	$F1_u$	$F1_a$	$F1_{all}$
<i>Vanilla baselines</i>								
DIRECT ANSWER	0	25.0	16.8	21.1	0	31.2	19.2	24.8
AMBIG-AWARE	0.12	27.4	20.9	23.0	0.17	34.8	23.6	28.7
<i>Clarify-or-Answer baselines</i>								
APAGEN	0.37	31.6	24.1	27.9	0.39	29.8	23.9	25.3
CDA _{DPO}	1.00	42.9	31.7	36.5	1.00	40.7	28.6	33.9
D4CQ (proposed)	1.08	47.2	34.3	40.1	1.06	42.6	31.1	36.0

Table 1: Main results on **AmbigQA**. “#” denotes the average number of clarification turns. All methods are evaluated under the same simulated interaction protocol; whenever clarification is requested, the same frozen responder are used. The best results are in bold.

ambiguity-aware methods, and trained clarification-or-answer models based on alignment or preference optimization. For zero-shot evaluation on SituatedQA (Section 4.5), we further consider baselines that explicitly decide whether to ask a follow-up question or answer directly based on self-asking or answer disagreement. Detailed baseline descriptions and implementation details are provided in Appendix D.1 and Appendix D.2, respectively. Whenever a method requires a user response, we use the same frozen Llama3-70B-Instruct responder for fair comparison; within each result table, all compared methods also share the same student backbone and decoding constraints.

4.2 Overall Effectiveness (RQ1)

Table 1 reports the main results on **AmbigQA**. We evaluate the correctness on the unambiguous ($F1_u$), ambiguous ($F1_a$), and full test subsets ($F1_{all}$).

Across both student backbones, D4CQ achieves the strongest overall performance with the clearest gains on the ambiguous subset where explicit intent resolution is the most critical. Under Llama3-8B, D4CQ improves over the strongest baseline CDA_{DPO} from 31.7 to 34.3 on $F1_a$ (+8.2%) and from 36.5 to 40.1 on $F1_{all}$ (+9.9%). We observe that these gains do not come from indiscriminately asking more questions. The average number of clarification turns remains close to 1 overall and is nearly identical to that of CDA_{DPO}, indicating that D4CQ improves the quality of clarification rather than merely increasing interaction frequency. Moreover, D4CQ improves $F1_u$ on both backbones, suggesting that the learned reasoning process does not overfit to ambiguity-heavy cases and remains effective when the question is already sufficiently specified.

4.3 Component Ablation and Hyperparameter Sensitivity (RQ2)

To better understand the contribution of each design choice, we conduct two analyses: component ablation and hyperparameter sensitivity.

Component ablation. Table 2 isolates the effects of the proposed components under the Llama3-8B. PROCEDURAL-SCAFFOLDING denotes an inference-only pipeline using the structured interrogator prompt from Section 3.1 without student training. PROCICA, DYNDIST, and HINDREL correspond to I1, I2, and I3, respectively. We use $F1_a$ on the ambiguous subset of AmbigQA as the primary metric, because all three ideas are designed to improve ambiguity resolution.

As a result, we observe that the procedural scaffolding alone improves over the simpler AMBIG-AWARE prompt baseline by a large margin (27.8 vs. 20.9), confirming that a structured clarification protocol is beneficial even without parameter updates. Among the trained variants, every two-component configuration substantially outperforms the inference-only scaffold, reaching 32.5 to 33.9 in $F1_a$. The full model achieves the best result at 34.3, which shows that the three components are complementary rather than redundant. In particular, the performance gap among the pairwise variants indicates that hindsight relabeling contributes the largest single gain in this setting, while dynamic distillation and intent-convergence alignment further improve the final model when combined.

Hyperparameter sensitivity analysis. Figure 2 studies the four key hyperparameters: the trajectory-retention threshold η , the intent-convergence loss weight λ , the diagnosticity reweight factor β , and the compatibility threshold ϵ . Overall, D4CQ is reasonably robust across a broad range of values, but the trends are still informative. Performance increases as η becomes more selective, suggesting that filtering out weak rollouts is beneficial in our setting. For λ , adding the intent-convergence loss is clearly helpful, with the best performance obtained at a moderate value rather than at either extreme. For β , moderate reweighting works the best: too little underuses diagnosticity information, whereas overly aggressive weighting begins to over-focus on a narrow subset of trajectories. Finally, a small positive compatibility threshold is preferable to both extremes, where overly strict filtering discards useful relabels, while a loose threshold allows response-intent mismatches.

Method Configuration	$F1_a \uparrow$	$\Delta F1_a(\%)$
<i>Inference only</i>		
AMBIG-AWARE	20.9	–
PROCEDURAL-SCAFFOLDING	27.8	–
<i>Training</i>		
PROCICA (I1) + DYNDIST (I2)	32.5	+16.9
PROCICA (I1) + HINDREL (I3)	33.3	+19.8
DYNDIST (I2) + HINDREL (I3)	33.9	+21.9
D4CQ (proposed)	34.3	+23.4

Table 2: Ablation study on **AmbigQA**. $F1_a$ denotes alias-max F1 on the ambiguous subset. $\Delta F1_a(\%)$ is the relative improvement over the inference-only PROCEDURAL-SCAFFOLDING baseline.

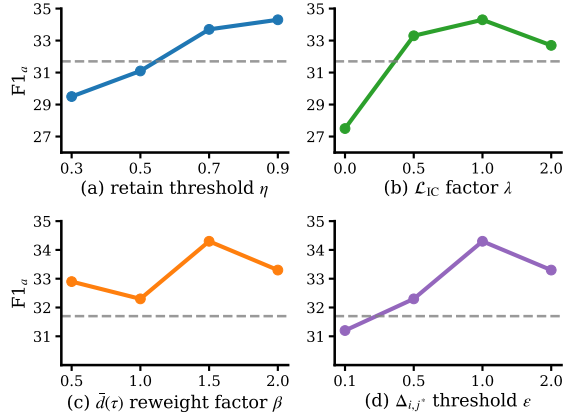


Figure 2: Hyperparameter sensitivity for the trajectory-retention threshold η , the intent-convergence loss weight λ , the diagnosticity scaling factor β , and the compatibility threshold ϵ .

4.4 Intent-Convergence Analysis (RQ3)

To verify that the learned multi-turn reasoning process actually moves the refined query toward the correct latent intent, we analyze how the refined query evolves over turns. Specifically, we encode the rewritten query at each turn with a SentenceTransformer encoder (Reimers and Gurevych, 2019; Kim et al., 2026) and compare it against the ground-truth intent and all other intents associated with the same ambiguous question. For a refined query $\tilde{q}_k$ after turn k , we define

$$\Delta_{IC}^{(k)}(\tau) = \text{sim}(\tilde{q}_k, x_i) - \max_{j \neq i} \text{sim}(\tilde{q}_k, x_j), \quad (10)$$

where $\text{sim}(\cdot, \cdot)$ is the cosine similarity between SentenceTransformer embeddings, x_i is the ground-truth intent, and $\{x_j\}_{j=1}^{m_q}$ are the candidate intents for the same ambiguous query. A larger value means that the refined query is semantically closer to the correct intent than to its nearest competing interpretation. Using the similarity measure, we derive three more metrics: $\text{Hit}@1_{\text{int}}$ is the fraction of trajectories whose final refined query is closest to the gold intent among the candidate intents.

Method	Intent convergence		Turn progress	
	$\Delta_{IC}^{(K(\tau))} \uparrow$	Hit@1 _{int} $\uparrow$	MR $\uparrow$	$\overline{\Delta}_{sim} \uparrow$
LLAMA3 8B				
D4CQ-ICA	0.034	0.481	0.419	0.042
D4CQ-DYNDIST	0.065	0.569	0.505	0.080
D4CQ (proposed)	0.079	0.612	0.608	0.094

Table 3: Intent-convergence analysis on the ambiguous subset of AmbigQA. D4CQ-ICA removes the intent-convergence alignment loss, and D4CQ-DYNDIST removes the dynamic-distillation component.

MonoRate (MR) measures the fraction of trajectories for which $\text{sim}(\tilde{q}_k, x_i)$ increases monotonically across turns. Finally, $\overline{\Delta}_{sim}$ is the average per-turn increase in similarity to the gold intent. We evaluate only trajectories containing at least two clarifying turns so that sequential refinement is observable.

Intent convergence. Table 3 shows that the full model achieves the strongest final-turn alignment to the gold intent. Removing the intent-convergence loss causes the largest drop, reducing $\Delta_{IC}^{(K(\tau))}$ from 0.079 to 0.034 and Hit@1_{int} from 0.612 to 0.481. This directly supports the role of the alignment objective in pulling the refined query representation toward the correct latent intent. Removing dynamic distillation also hurts performance, though less severely, suggesting that prioritizing diagnostic trajectories further sharpens the final refined query.

Turn progress. The full model also performs the best on the turn-wise progression metrics. It attains the highest MR and the largest average similarity gain per turn, indicating that successive clarification turns are more consistently useful rather than noisy or redundant. In other words, the model is not merely ending closer to the correct intent at the final turn; it is also refining the query in a more stable and progressive manner throughout the dialogue.

4.5 Ambiguity Handling under Distribution Shift (RQ4)

To examine whether the ambiguity-handling behavior learned on AmbigQA transfers beyond the training distribution, we conduct zero-shot evaluation on the out-of-domain SituatedQA-Geo and SituatedQA-Temp benchmarks.

For the evaluation, we first partition model outputs into five cases (Kim et al., 2024): ① clarification request or CQ generation for an ambiguous query, ② non-clarification or direct answer to an ambiguous query, ③ correct direct answer to an unambiguous query, ④ incorrect direct answer to an unambiguous query, and ⑤ clarification request or

Method	SituatedQA-Geo		SituatedQA-Temp	
	F1 _{unp}	F1 _{detect}	F1 _{unp}	F1 _{detect}
MISTRAL 7B				
DIRECT ANSWER	11.29	0.00	15.34	0.00
AMBIG-AWARE	3.66	26.01	8.43	22.48
SAMPLE REP	7.64	25.31	7.83	21.13
SELF-ASK	11.29	0.00	15.34	0.00
APAFIXED	38.43	41.84	45.01	43.95
D4CQ (proposed)	41.01	44.75	47.64	48.12

Table 4: The zero-shot performance on SituatedQA-Geo and SituatedQA-Temp benchmarks.

CQ generation for an unambiguous query. Based on these five cases, we define two complementary metrics. **Ambiguity Detection F1** ($F1_{\text{detect}}$) measures whether the model correctly identifies ambiguous queries and requests clarification only when needed. We define this metric using the precision $\frac{①}{①+⑤}$ and the recall on ambiguous queries $\frac{①}{①+②}$. **Unambiguous Prediction F1** ($F1_{\text{unp}}$) evaluates whether the model answers unambiguous queries directly and correctly, while avoiding arbitrary direct answers to still-ambiguous inputs. Accordingly, we compute this metric using the precision of direct-answer predictions $\frac{③}{②+③+④}$ and the recall on unambiguous queries $\frac{③}{③+④+⑤}$.

Table 4 shows that D4CQ generalizes well to both geographic and temporal ambiguity, despite being trained on a different ambiguity distribution. Notably, on **SituatedQA-Geo**, D4CQ improves over APAFIXED from 38.43 to 41.01 on $F1_{\text{unp}}$ and from 41.84 to 44.75 on $F1_{\text{detect}}$. These improvements indicate that D4CQ does not merely fit the annotation style of AmbigQA; rather, it learns a more general clarify-or-answer policy that remains effective under distribution shift. In particular, consistent gains in both $F1_{\text{detect}}$ and $F1_{\text{unp}}$ suggest that the model improves not only ambiguity recognition, but also decision calibration, which is precisely the behavior required for robust open-domain clarification systems.

5 Related Work

Ambiguous and Context-dependent Open-domain QA Benchmarks. Prior work on open-domain and conversational question answering progressively moves beyond single-turn answer generation toward settings in which the intended meaning of a question must be inferred from dialogue history or external context. Early conversational QA benchmarks such as QuAC (Choi et al., 2018)

and CoQA (Reddy et al., 2019) establish multi-turn evaluation settings where follow-up questions are highly dependent on preceding turns and shared context, making interpretation itself a core part of the task. Elgohary et al. (2019) further formalizes this issue through question rewriting, focusing on transforming context-dependent questions into self-contained forms suitable for downstream retrieval and reasoning. More recent work makes ambiguity explicit in open-domain QA. AmbigQA (Min et al., 2020) frames a question as potentially corresponding to multiple plausible latent intents and answer sets, requiring systems to recover all valid interpretations rather than committing prematurely to a single answer. SituatedQA (Zhang and Choi, 2021) extends this perspective by showing that even surface-identical questions may require clarification when the answer depends on extra-linguistic context such as time or location.

Ask-or-answer Decision Making for OD-QA.

A closely related line of work studies how systems should ask clarifying questions when user requests are underspecified. In open-domain conversational search, Qulac (Aliannejadi et al., 2019) and the ClariQ challenge (Aliannejadi et al., 2020) provide benchmark settings in which the system is rewarded for selecting or generating clarification questions that help disambiguate user intent. Earlier work (Rao and Daumé III, 2018) also studies clarification as an information-theoretic decision problem, for example by ranking candidate questions using neural approximations to expected value of perfect information, thereby emphasizing that a good clarification should be diagnostically useful rather than merely fluent or generic. More recent LLM-based studies (Yun et al., 2026) move from question ranking to full ask-or-answer behavior. In particular, APA (Kim et al., 2024) explicitly aligns language models to recognize ambiguous inputs and respond with clarification requests when appropriate, rather than always producing direct answers. This shift highlights that ambiguity handling is not only a generation problem but also a policy-learning problem. Our method builds on this view, but focuses on learning diagnostic clarifications in a data-efficient manner under limited interaction budgets.

Data-efficient Alignment and Preference-based Learning for Conversational QA. Our training strategy is also related to a broader body of work on data-efficient alignment and supervision design for language models. Instruction tun-

ing (Wei et al., 2021) and RLHF (Ouyang et al., 2022) show that model behavior can be aligned with user intent through preference-informed supervision rather than task-specific labels alone. DPO (Rafailov et al., 2023) further simplifies this paradigm by directly optimizing preferences without an explicit reward model, providing an efficient alternative for preference-based policy learning. In parallel, self-improvement and bootstrapping approaches such as STaR (Zelikman et al., 2022) show that models can iteratively improve reasoning quality by generating and filtering their own intermediate supervision, while Constitutional AI (Bai et al., 2022) explores how AI-generated feedback can substitute for expensive human annotation in alignment pipelines. At the representation level, contrastive objectives such as InfoNCE (Oord et al., 2018) have become a standard mechanism for pulling semantically aligned instances together while separating distractors. Our framework draws inspiration from these directions. Contrastive intent-convergence alignment encourages refined queries to move toward the target latent intent, dynamic distillation prioritizes more informative interaction traces, and hindsight relabeling improves sample efficiency by recovering useful supervision from failed trajectories. Together, these connections place our method at the intersection of clarification policy learning and efficient alignment for interactive QA systems.

6 Conclusion

In this paper, we propose D4CQ, a data-efficient framework for learning diagnostic clarifying questions in open-domain conversational question answering. D4CQ combines procedural scaffolding, contrastive intent-convergence alignment, diagnosticity-based dynamic distillation, and hindsight relabeling to learn when to clarify, what to ask, and how to progressively refine an ambiguous query toward its correct latent intent. Experiments on AmbigQA and zero-shot evaluation on SituatedQA show that D4CQ consistently improves both ambiguity resolution and clarify-or-answer calibration under a comparable interaction budget, highlighting its effectiveness as a robust approach to open-domain conversational disambiguation.

Limitations

Our work is developed and evaluated primarily for open-domain conversational QA with short-form

answers and a small clarification budget. Extending the framework to settings that require longer-horizon reasoning, richer multi-turn interactions, or long-form answer generation remains outside the scope of the present study. In addition, our formulation starts from a single underspecified query and does not explicitly model extra context such as persistent user profiles, prior conversation history, or external situational signals that may further shape ambiguity in real deployments. Our experimental study also covers a limited set of student backbones and benchmarks. Broader evaluation across larger models, additional architectures, and more diverse domains could reveal different strengths or failure modes. Finally, although the target setting is interactive, the responder in our experiments is instantiated as a fixed pretrained language model rather than a human user, and the training pipeline is built around supervised distillation from teacher-generated trajectories. Alternative supervision sources or alignment strategies may offer complementary benefits, but investigating them is left for future work.

Acknowledgments

This work was supported by Mobile eXperience(MX) Business, Samsung Electronics Co., Ltd. U Kang is the corresponding author.

References

- Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2020. Convai3: Generating clarifying questions for open-domain dialogue systems (clariq). *arXiv preprint arXiv: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*, pages 475–484.
- Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chapidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534.
- Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. 2024. Star-gate: Teaching language models to ask clarifying questions. *arXiv preprint arXiv:2403.19154*.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wenta Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. [Quac: Question answering in context](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184. Association for Computational Linguistics.
- Jeremy Cole, Michael Zhang, Dan Gillick, Julian Eisen-schlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543.
- Michael Han Daniel Han and Unsloth team. 2023. [Unsloth](#).
- Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.
- Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. [Can you unpack that? learning to rewrite questions-in-context](#). 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 5918–5924. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open

- domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 874–880.
- Jihyeong Jeon, Jiwon Lee, Cheol Ryu, and U Kang. 2025. Entity-aware generative retrieval for personalized contexts. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 1078–1090.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). Preprint, arXiv:2310.06825.
- Jinhong Jung, Namyong Park, Sael Lee, and U Kang. 2017. Bepi: Fast and memory-efficient method for billion-scale random walk with restart. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 789–804.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781.
- Hyuhng Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. 2024. Aligning language models to explicitly handle ambiguity. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1989–2007.
- Kyung-Min Kim, Jinhong Jung, Jihee Ryu, Ha-Myung Park, Joseph P Joohee, Seokwoo Jeong, U Kang, and Sung-Hyon Myaeng. 2017. A new question answering approach with conceptual graphs. In *CORIA*, pages 218–234.
- Minjun Kim, Jaeri Lee, Jongjin Kim, Jeongin Yun, Yongmo Kwon, and U Kang. 2026. Lampq: Towards accurate layer-wise mixed precision quantization for vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 5691–5699.
- Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1317–1327.
- 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*, pages 611–626.
- Jaeri Lee and U Kang. 2025. Context-aware sequential bundle recommendation via user-specific representations. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 1468–1476.
- Seungjoo Lee, Yong-chan Park, and U Kang. 2025. Swagner: Leveraging span-aware grid transformers for accurate nested named entity recognition. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 1487–1496.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paran-jape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- 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)*, pages 5783–5797.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.
- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Sudha Rao and Hal Daum   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 (Volume 1: Long Papers)*, pages 2737–2746. Association for Computational Linguistics.
- Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. [Coqa: A conversational question answering challenge](#). *Transactions of the Association for Computational Linguistics*, 7:249–266.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Minji Yoon, Jinhong Jung, and U Kang. 2018. Tpa: Fast, scalable, and accurate method for approximate random walk with restart on billion scale graphs. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 1132–1143. IEEE.

Jeongin Yun, Jaeri Lee, Jongjin Kim, Minjun Kim, Jinho Song, and U Kang. 2026. Sharvet: Similarity-aware parameter sharing with vector-based tuning for efficient llm compression. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 47921–47937.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *arXiv preprint arXiv:2203.14465*.

Michael Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387.

Michael JQ 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*.

Appendix

A Interrogator Prompt for Procedural Scaffold

This section summarizes the prompt used to elicit procedurally scaffolded clarification trajectories from the interrogator. The prompt is designed to

(i) enforce a consistent multi-turn disambiguation protocol and (ii) explicitly enumerate ambiguity axes. Both teacher and student interrogators are instructed to produce structured chain-of-thought in a dedicated reasoning block, followed by a single terminal action indicating either a clarifying question or a direct answer.

A.1 Output Schema and Parsing.

The interrogator must output *exactly one* terminal line in the form of: <FINAL> <Q> [clarifying question] or <FINAL> <ANS> [direct answer].

All reasoning must be placed inside <thinking> with structured <step> and <reflection> tags. This structural constraint allows reliable extraction of the final response (e.g., CQ or answer) and the CQ string for downstream training and analysis.

A.2 Step-wise Disambiguation Protocol.

The scaffold decomposes ambiguity resolution into ordered stages. Specifically, an interrogator (I) rewrites the question given prior turns, (II) recalls relevant internal knowledge to ground the ambiguity search, (III) checks whether the question is a singleton, and (IV-V) enumerates ambiguity axes via controlled edits (replacement, addition, and deletion). Expressions of the form $\{ \cdot \}$ denote instantiated input variables, whereas italicized symbols (e.g., q and h) not written in this form are shown verbatim as literal prompt tokens.

Summary of Interrogator p_θ Prompt

System. You are a careful interrogator for clarifying question answering. Given a potentially ambiguous question q and the dialogue history h , you must either **ask clarifying question(s)** that resolve *all* important ambiguities, **or** produce **one direct answer**. Never do both.

Formatting constraints.

- Put all reasoning in <thinking> ... </thinking>.
- Inside <thinking>, write analysis as <step> ... </step> and periodic summaries as <reflection> ... </reflection>.
- After </thinking>, output *exactly one* terminal line: <FINAL> <Q> ... or <FINAL> <ANS>

Procedure (inside <thinking>).

- I. Recap and question refinement.** If prior dialogue exists, recap the key constraints

the user already confirmed and rewrite a refined question q ; otherwise mark N.A. and keep the original q .

II. Parametric background recall. Recall relevant background knowledge from internal knowledge (entities, roles, timelines, etc.) that is plausibly connected to q .

III. Singleton check. Decide whether q is inherently multi-answer by design (e.g., “list all”) or structurally single-answer once specified. If multi-answer, propose clarification to narrow to a single atomic target.

IV. Ambiguity axis discovery by replacement. For terms already present in q , consider whether replacing a term with another value of the same type would yield a different valid answer. Each such term defines an ambiguity axis (e.g., time-rank, nationality/team, institution, competition level).

V. Hidden ambiguity discovery by addition and deletion. Consider whether adding or deleting plausible qualifiers/modifiers (e.g., gender, age group, format, time window) would yield different answers. Each such edit defines a hidden ambiguity axis.

Critical decision policy.

- Be conservative: any plausible ambiguity axis discovered via replacement or, addition and deletion is treated as unresolved unless the user explicitly resolved it in prior turns.
- Do not rely on world knowledge to assume an axis is fixed; only explicit user confirmation counts.
- When asking, include at least one explicit sub-question per unresolved axis.
- Keep <FINAL> free of reasoning or meta-commentary; the answer (if any) must be a short atomic span.

Few-shot exemplars. A small set of worked examples illustrating (i) ambiguity-axis identification through replace, add, and delete, and (ii) query refinement after a user response:

- **Example k .** User: query
→ Assistant: <thinking>...</thinking>
<FINAL> <Q> ...
→ User: response
→ Assistant: follow-up output.

Dialogue history. ($\{q\}, \{c_1\}, \{r_1\}, \dots$)

A.3 Notes on Intent-Convergence Training.

The scaffold produces trajectories with (a) explicitly enumerated ambiguity axes and (b) refined questions that summarize the current intent hypothesis after each interaction. These refined-question slots are subsequently used for contrastive intent-convergence alignment (§3.1), where the refined query representation is encouraged to move toward the ground-truth intent representation and away from within-query alternative intents.

A.4 Practical Considerations.

We implement strict parsing to discard malformed generations that violate the <FINAL> schema. In addition, we optionally truncate overly long reasoning blocks to control context length during rollout, while preserving the refined question and terminal action fields needed for downstream distillation.

B Responzor Prompt for Simulated User Feedback

This section summarizes the prompt used for the fixed responzor p_ϕ that simulates user feedback during rollout. While the interrogator actively searches for missing intent-defining information, the responzor plays the role of the *original user* who asked the ambiguous question q . Concretely, the responzor is conditioned on the ambiguous question q , the latent intent x , the answer set $A(x)$, and the current clarifying question c , and then generates a short response r .

The prompt is designed to satisfy two requirements. First, the responzor should reveal only the information that is legitimately implied by the latent intent x , rather than directly exposing the final answer. Second, the response should remain brief and easy to parse so that the resulting interaction resembles a natural user reply rather than a long explanation. This design makes the simulated interaction more faithful to the OD-CQA setting, where the interrogator must infer the latent intent through targeted clarification rather than through answer leakage.

B.1 Output Schema and Behavioral Constraints.

The response is constrained to be short and directly address the clarifying question. Crucially, although the prompt includes the answer set $A(x)$ for conditioning consistency, the responzor is explicitly instructed to *pretend not to know the answer* and to

answer only from the information contained in the intended question x whenever the clarifying question is resolvable from x . In the prompt summary below, expressions of the form $\{\cdot\}$ denote instantiated input variables, whereas italicized symbols (e.g., q , x , $A(x)$, and c) not written in this form are shown verbatim as literal prompt tokens.

Summary of Responder p_ϕ Prompt

System. Assume that you are the original asker of an ambiguous question q . Your true intent corresponds to a single intended question x with answer set $A(x)$, but the question q was written too hastily and omitted important details, making it ambiguous.

Another agent now asks you a clarifying question c in order to identify your intended question x and eventually answer it.

Generate a brief response to the clarifying question while *pretending that you do not know the final answer*. You do, however, know your intended question x . If the answer to the clarifying question is contained in x , respond using only that information.

Formatting constraints.

- Output exactly one line in the format: "response":
- Keep the response brief and directly answer the clarifying question.
- Do not produce reasoning, explanations, or meta-commentary.
- Do not directly reveal the final answer unless it is explicitly entailed by the intended question itself.

Prompt variables.

- Ambiguous question q : $\{q\}$
- Originally intended question x : $\{x\}$
- Expected answer aliases $A(x)$: $\{A(x)\}$

User. Clarifying question c : $\{c\}$

B.2 Role in Trajectory Construction.

During rollout, the responder is used after each interrogator-generated clarifying question to produce the next user turn:

$$r_k \sim p_\phi(\cdot \mid q, c_k, x, A(x)). \quad (11)$$

The generated response is then appended to the dialogue history and becomes part of the next interrogator input (Appendix A.2).

Dataset	# Amb.	# Unamb.	Avg. intents
AmbigQA	1,172	788	3.6
SituatedQA-Geo	129	506	4.0
SituatedQA-Temp	876	2,795	3.2

Table 5: Summary statistics of the evaluation datasets. Avg. intents denotes the average number of latent intents per ambiguous question.

C Dataset Statistics

Table 5 summarizes the key statistics of the datasets used in our experiments. AmbigQA is used as the primary benchmark, while SituatedQA-Geo and SituatedQA-Temp are used only for zero-shot evaluation in Section 4.5.

D Experiments Details

D.1 Baselines

Baselines for AmbigQA. For the main comparison on AmbigQA (Section 4.2), we consider the following baselines.

- **DIRECT ANSWER.** A single-turn baseline that directly generates an answer from q without asking clarifying questions.
- **AMBIG-AWARE.** A prompt-only baseline that first determines whether the input is ambiguous and, if so, asks one clarifying question; otherwise, it answers directly. After receiving a simulated user response, it generates the final answer from the accumulated dialogue history.
- **APAGEN/APAFIXED** (Kim et al., 2024). An ambiguity-aware alignment baseline that explicitly distinguishes between directly answering and requesting clarification. We use APAGEN for the end-to-end AmbigQA comparison and APAFIXED for SituatedQA, where the fixed ambiguity-decision template is better suited to detection-oriented evaluation.
- **CDA_{DPO}** (Zhang et al., 2025). A preference optimization baseline that learns whether to clarify or answer directly by modeling downstream conversation outcomes after clarification.

Baselines for SituatedQA. For the zero-shot evaluation in Section 4.5, we reuse DIRECT ANSWER, AMBIG-AWARE, and APAFIXED, and additionally compare against the following baselines.

- **SELF-ASK** (Press et al., 2023). A prompt-based method that first decides whether an additional follow-up question is needed before producing the final answer.

- **SAMPLE REP** (Cole et al., 2023). An uncertainty-based method that asks for clarification when sampled answer candidates disagree sufficiently with the greedy prediction.

Whenever a method requires a user response, we use the same frozen Llama3-70B-Instruct responder for fair comparison. Within each result table, all compared methods also share the same student backbone and the same decoding constraints.

D.2 Implementation Details

We train the student interrogator p_θ with the objective in Eq. (9). For the main AmbigQA experiments, the student backbone is an instruction-tuned Llama3-8B (Grattafiori et al., 2024) or Gemma-7B (Team et al., 2024); for the zero-shot experiment on SituatedQA, we report results with Mistral-7B (Jiang et al., 2023). Teacher rollouts are generated by Llama3-70B-Instruct under the procedural scaffold described in Section 3.1. For the interrogator prompt, we additionally provide a small number of few-shot chain-of-thought exemplars (Wei et al., 2022), enabling the teacher to follow the scaffold and produce high-quality structured trajectories via in-context learning (Brown et al., 2020) without additional training. For efficient fine-tuning, we use Unsloth (Daniel Han and team, 2023) with 4-bit quantization (Dettmers et al., 2023) and LoRA adapters (Hu et al., 2022). When computing Eq. (4), we compute the shared prefix KV-cache (Kwon et al., 2023) for $h_{K(\tau)+1} \oplus \pi(\tau)$ once and reuse it across all intent continuations $\{x_j\}_{j=1}^{m_q}$. We optimize with AdamW (Loshchilov and Hutter, 2017), learning rate $2e-4$, batch size 64, and 4 training epochs on 4 A100-80GB GPUs. We report the mean over three random seeds. The hyperparameters are set to $\eta = 0.8$, $\lambda = 1.0$, $\beta = 1.5$, and $\epsilon = 1.0$. When a benchmark does not provide an official validation split, we randomly split the training set into 80 to 20 for train and validation partitions, and tune hyperparameters on the held-out portion.

D.3 Inference Procedure

At inference time, only the student interrogator is used. Starting from the initial query, the model autoregressively generates either a clarifying question c_k or a final answer $\hat{a}$ at each turn. If a clarifying question is generated, the system obtains a response from the responder p_ϕ , appends it to the dialogue history, and proceeds to the next turn. Interaction terminates when the interrogator produces a final answer or when the turn budget $K_{\max}$ is reached,

in which case the final answer is generated from the last dialogue state.