PAUSE: PROACTIVE APPROPRIATENESS AND UTILITY FOR SELECTIVE ELICITATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Information-seeking agents must decide not only what information to acquire, but also whether a candidate intervention is appropriate for the current dialogue state. Existing clarification and value-based approaches primarily optimize expected information gain or task utility, implicitly assuming that a valuable intervention is suitable whenever it is selected. We show that this assumption can fail: the same question may be informative but premature at one point in a dialogue and appropriate at another. We introduce PAUSE, a selective elicitation framework that decomposes intervention selection into two decisions: contextual appropriateness and information utility. PAUSE learns an appropriateness score $T(s, R)$ and a utility score $V(s, R)$ from complementary supervision, and performs appropriateness-constrained utility optimization by selecting the highest-value intervention among candidates that satisfy a contextual fit constraint. On NewsInterview, PAUSE improves eight-turn information coverage from 0.5042 with utility-only selection to 0.5676, while leading to more selective clarification behavior on ClariQ. These results suggest that modeling contextual fit as a separate decision factor can improve proactive information seeking, particularly in extended multi-turn interactions.

1 INTRODUCTION

Information-seeking agents must decide not only what information to acquire, but also when and how to intervene as a conversation unfolds. From clarifying an ambiguous request to conducting a long-form interview, an effective agent needs to select questions that advance the interaction while respecting the current conversational state (Aliannejadi et al., 2021; De Lima & Yang, 2026). Consider an interviewer speaking with a guest who is gradually describing a personal experience. A question about a later consequence may eventually reveal valuable information, yet asking it too early may presuppose an event or context that has not been established. Once the relevant context appears, the same question becomes a natural follow-up. Conversely, repeatedly asking questions that are safe but uninformative may leave important aspects undiscovered. Therefore, proactive information seeking requires reasoning about both the potential value of an intervention and its compatibility with the current dialogue state.

Existing approaches largely formulate information-seeking decisions around utility: given that additional information is needed, an agent selects a question expected to maximize information gain or downstream task performance. Turn-taking models study when an agent can enter a conversation by predicting dialogue opportunities (Ekstedt & Skantze, 2020); clarification methods estimate whether additional information is required and select questions that improve retrieval or answerability (Aliannejadi et al., 2021; Chi et al., 2024); and expected-value or value-of-information approaches explicitly model the potential benefit of asking a question, sometimes considering interaction costs (Rao & Daumé III, 2018; Dong et al., 2026). These methods address important aspects of proactive interaction, but they often leave implicit a critical assumption: a high-value question is appropriate whenever it is selected. In practice, the same intervention may have different conversational effects depending on when it is introduced. An opportunity to ask does not imply that every candidate question is suitable for that position, and expected information gain alone may not capture whether a question is premature, redundant, or contextually misaligned.

We argue that question selection requires separating two distinct decisions: whether an intervention fits the current state and how much useful information it may reveal. Contextual appropriateness

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

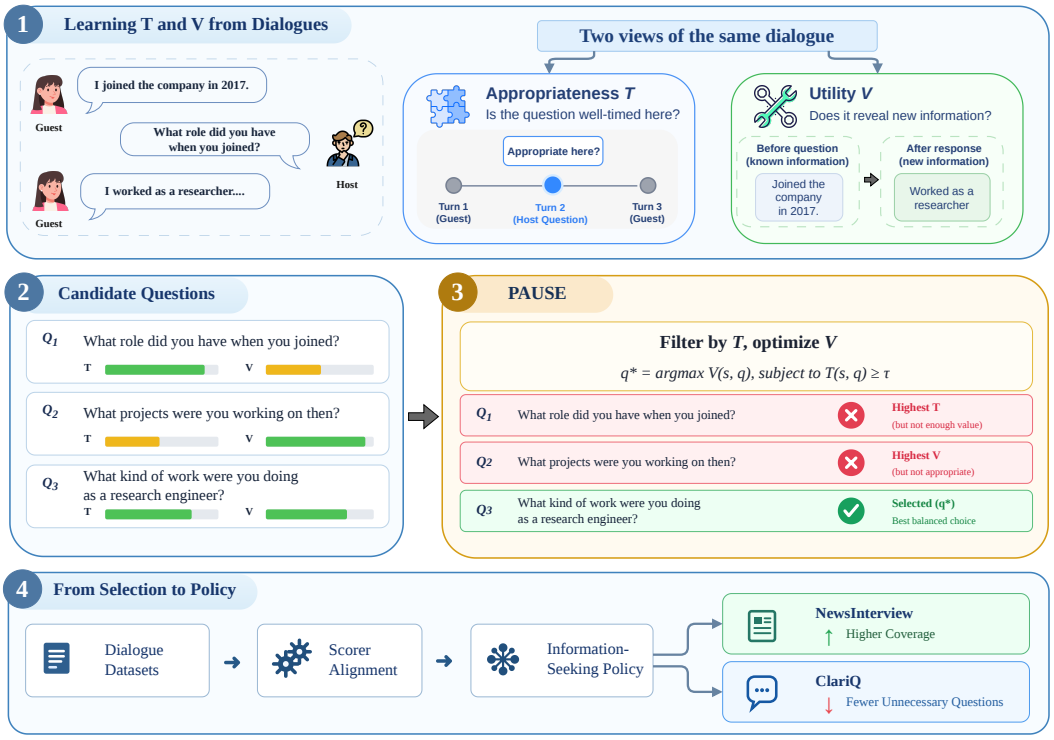


Figure 1: Overview of PAUSE. Contextual appropriateness T evaluates whether a candidate question fits the current dialogue state, while utility V estimates the information it may reveal. PAUSE selects the highest-utility question among appropriate candidates.

and information utility can vary independently. A question may have high prospective value but be unsuitable before the relevant topic has been established; conversely, multiple questions may all fit the current state while differing substantially in the information they are likely to elicit. We formalize this problem as *selective elicitation*: given an evolving dialogue state and a set of candidate interventions, an agent should first identify interventions that are feasible for the current state and then select the most informative among them.

To study selective elicitation, we introduce PAUSE, a framework that explicitly models contextual appropriateness and information utility as two complementary signals. Given a dialogue state s and candidate intervention R, PAUSE learns an appropriateness score $T(s, R)$ measuring whether the intervention fits the current interaction, and a utility score $V(s, R)$ estimating its task-relevant information value. Instead of directly combining these objectives, PAUSE treats appropriateness as an eligibility constraint and optimizes utility within the appropriate candidate set:

$$R^* = \arg \max_{R \in \mathcal{R}(s): T(s,R) \geq \tau} V(s, R).$$

This formulation reflects the observation that an intervention with high information value should not be selected if it violates the current conversational context. Figure 1 illustrates how PAUSE separates contextual fit from information value and applies the resulting decision rule.

We evaluate PAUSE in two information-seeking settings with different decision horizons. In simulations of 348 NewsInterview conversations (Spangher et al., 2025), incorporating appropriateness improves eight-turn information coverage from 0.5042 with utility-only selection to 0.5676. On 50 ClariQ test topics (Aliannejadi et al., 2021), the reported training-mix policy asks on 42 topics and obtains nDCG@20 of 0.1504. These results suggest that contextual appropriateness provides stronger benefits in extended multi-turn elicitation, while its impact on single-turn clarification remains an open question.

Our contributions are as follows:

- We formulate selective elicitation as a candidate-level decision problem that separates intervention feasibility from information utility, highlighting contextual appropriateness as a distinct factor in proactive information seeking.
- We propose PAUSE, which learns appropriateness and utility from complementary supervision, selects interventions through appropriateness-constrained utility optimization, and distills the resulting decisions into a question-generating policy.
- We evaluate PAUSE on NewsInterview and ClariQ, demonstrating improved information recovery in multi-turn elicitation and more selective clarification behavior in single-turn search.

2 Related Work

Turn-taking and contextual fit. Turn-taking work predicts when a participant can take the floor from dialogue cues (Ekstedt & Skantze, 2020; Gervits et al., 2020). Response selection also considers content: Hedayatnia et al. (2022) evaluate candidate replies that are manually labeled as appropriate or inappropriate for the same dialogue context. PAUSE studies this candidate-level fit for information-seeking interventions across dialogue positions and uses it as an eligibility condition before ranking questions by their expected value.

Clarification and question selection. Qulac and ClariQ provide information-seeking conversations and clarification benchmarks (Aliannejadi et al., 2019; 2021); InSCIt includes mixed-initiative turns in which an agent may answer, clarify, or provide relevant information (Wu et al., 2023). Question-selection methods rank candidates by expected answer value (Rao & Daumé III, 2018), conversation context and retrieved passages (Mass et al., 2022), or retrieval state (Chi et al., 2024). Other work generates questions from search facets or explicitly models what and how to ask in conversation (Sekulic et al., 2020; Do et al., 2023). Zhang & Choi (2025) study when clarification is worthwhile under ambiguous user intents. PAUSE uses candidate-level positional fit as an eligibility gate before comparing information value.

Information elicitation and policy learning. NewsInterview exposes strategic weaknesses of LLM interviewers, while YIELD frames information elicitation as a sequential decision problem (Spangher et al., 2025; De Lima & Yang, 2026). RaPSIL trains an interview agent using utterances judged for strategic effectiveness and social sensitivity (Hama et al., 2025). Value-of-information methods weigh the benefit of asking against interaction costs in preference elicitation and human–agent communication (Baarslag & Kaisers, 2017; Dong et al., 2026). PAUSE separates contextual fit from information utility when selecting an intervention, then uses eligible candidates to construct preference pairs for standard DPO training (Rafailov et al., 2023).

3 Method

PAUSE separates two decisions in proactive information seeking: whether an intervention fits the current dialogue state and which fitting intervention is most useful. We treat the former as a feasibility condition and the latter as an optimization objective. The same formulation covers both selecting a question from a candidate pool and deciding that no question is needed.

3.1 Decision Setting

At turn t, the agent observes a state s_t, comprising the available task context and dialogue history, and considers a finite candidate set $\mathcal{R}(s_t) = \{R_1, \ldots, R_K\}$. Each R is a proposed information-seeking intervention. The interlocutor's answer A is not available until after the intervention is selected. Let $u(s, R, A)$ denote the downstream payoff after that answer, such as newly covered information or retrieval quality, the ideal one-step value would be

$$V^\star(s, R) = \mathbb{E}_{A \sim p(\cdot|s,R)}[u(s, R, A)].$$

This expectation explains why utility must be estimated before observing an answer. PAUSE learns a task-specific value $V(s, R) \in [0, 1]$ from recorded elicitation outcomes and a separate appropriateness score $T(s, R) \in [0, 1]$ from evidence about whether an intervention fits its context.

3.2 Appropriateness and Information Utility

Appropriateness measures whether a proposed intervention belongs at the current dialogue position. A question can have high prospective information value while presupposing an event or preference not yet established. We denote by $y_T \in \{0, 1\}$ a fit label and by $x_T = \phi_T(s, R)$ the information given to its scorer. For response-conditioned scoring, positive supervision pairs an observed intervention with its observed context, while mismatched responses or premature positions provide contrasting supervision. When the decision concerns whether to ask at all, x_T depends on the state alone and the label indicates whether clarification is needed. The resulting score $T(s)$ is shared across candidate questions. Thus, T describes response–position compatibility or clarification need, according to the action space, rather than generic turn availability.

Utility supervision uses a label $y_V \in \{0, 1\}$ derived from an observed downstream outcome, such as the appearance of new information in a subsequent response or a thresholded retrieval-quality target. The scorer uses an input representation $x_V = \phi_V(s, R)$ available before the answer is observed. The subsequent answer determines the training label but does not enter the inference input. Both scorers are trained as label-token predictors. Let p_{θ_j} be the corresponding model, their scores and training losses take the form

$$S_j(x_j) = p_{\theta_j}(\texttt{1} \mid x_j),$$
$$\mathcal{L}_j = -\sum_i \log p_{\theta_j}(\texttt{y}_{j,i} \mid x_{j,i}), \qquad j \in \{T, V\}.$$

Here $\texttt{y}_{j,i}$ is the token `0` or `1` for the corresponding training label, and $T = S_T(x_T)$, $V = S_V(x_V)$. The two losses are optimized separately, so high utility does not make an intervention appropriate by construction. Their outputs are ranking signals. The decision rule does not require them to be calibrated probabilities of the same event.

3.3 Constrained Selection

PAUSE selects an intervention in two steps. It first keeps candidates whose appropriateness score reaches a threshold τ, then chooses the highest-utility candidate among those that remain:

$$\mathcal{F}_\tau(s) = \{R \in \mathcal{R}(s) : T(s, R) \geq \tau\}, \qquad R^*(s) = \arg\max_{R \in \mathcal{F}_\tau(s)} V(s, R).$$

Here $\mathcal{F}_\tau(s)$ is the set of eligible interventions. A candidate below the threshold cannot be selected, no matter how high its utility score is. This differs from combining the scores as $\alpha T(s, R) + (1 - \alpha)V(s, R)$, where high utility can offset low appropriateness. If no candidate remains eligible, the agent either proceeds without asking or, when an intervention is required, selects the candidate with the highest T. With a state-level gate $T(s)$, all candidate interventions are eligible when $T(s) \geq \tau$, and none are eligible otherwise. Section 5 gives the task-specific implementations.

3.4 Distilling the Decision into a Generator

When interventions are generated, PAUSE uses the scorers to train a generator $\pi_\theta(R \mid s)$ as well as to select its outputs. For each training state s_i, we sample a candidate pool $\mathcal{R}_i$ from the initial generator. We retain the top $\lceil \rho|\mathcal{R}_i| \rceil$ candidates by T, then remove any candidate below the training threshold τ_{tr} or failing a validity check. Let $\mathcal{F}_i^{\text{tr}}$ denote the remaining candidates. The supervised fine-tuning (SFT) target is the one with the highest utility:

$$R_i^+ = \arg\max_{R \in \mathcal{F}_i^{\text{tr}}} V(s_i, R).$$

We fine-tune the generator on (s_i, R_i^+), computing loss on response tokens rather than the state prompt. States with no eligible candidate are skipped.

We then construct preference pairs from the same eligible set. The preferred response is R_i^+, and the rejected response R_i^- has the lowest V score in $\mathcal{F}_i^{\text{tr}}$. We keep a pair only when the responses are distinct, their utility difference is at least δ, and the preferred response meets a minimum utility requirement. Standard DPO (Rafailov et al., 2023) trains on these pairs using the frozen SFT generator as its reference, with sequence likelihoods computed on response tokens only. The training

contribution is how the pairs are selected, not a change to the DPO objective. Because both responses have passed the appropriateness gate, the preference primarily distinguishes their modeled utility. At inference, PAUSE samples candidates from the trained generator and applies the same T-then-V selection rule.

4 EXPERIMENTS

4.1 DATASETS

We use four datasets with different roles. SPoRC (Litterer et al., 2025) contains over 1.1 million podcast transcripts, with speaker turns available for a subset of roughly 370,000 episodes. Recorded host replies and sampled alternatives provide supervision for contextual appropriateness. YIELD (De Lima & Yang, 2026) contains 2,281 human-to-human information-elicitation dialogues. Question–answer pairs from these dialogues support the factual-novelty target for information utility. NewsInterview (Spangher et al., 2025) contains recorded informational interviews with reference transcripts and information items. We use 152 conversations for development and a disjoint set of 348 conversations for the main interactive evaluation. ClariQ (Aliannejadi et al., 2021) provides open-domain requests, candidate clarification questions, and retrieval outcomes over possible request facets. We select checkpoints and calibrate the clarification threshold on the development set, then evaluate the single-turn policy on a separate test set of 50 topics.

4.2 MODELS

We access GLM-5.1 (Z.ai, 2026), DeepSeek-V4-Pro (DeepSeek-AI, 2026), Kimi-K2.6 (Moonshot AI, 2026), and Qwen3.5-397B-A17B (Qwen Team, 2026) through provider APIs. We locally deploy the open-weight Qwen3-8B and Qwen3-14B models (Yang et al., 2025) and Llama-3.1-8B (Grattafiori et al., 2024) on a server with eight NVIDIA H20 GPUs, using low-rank adaptation (LoRA) (Hu et al., 2021) for learned components. Unless a task-specific value is stated below, generative calls use a sampling temperature of 0.8. Training configurations and implementation details appear in Appendix B.

For ClariQ question reranking, we use two pretrained text embedding models: m3e-base, a Chinese–English sentence embedding model (Wang et al., 2023), and BGE-M3, a multilingual model that supports dense, sparse, and multi-vector retrieval (Chen et al., 2024). In these baselines, we use only their dense embeddings.

4.3 TASKS

Multi-turn information elicitation. On NewsInterview, two agents simulate the interviewer and interviewee. The interviewee is grounded in the reference interviewee transcript, whereas the interviewer sees only the introductory context and the dialogue history so far. The interviewer asks one question per turn for eight turns, aiming to elicit as many of the reference interview's information items as possible. A separate judge checks which reference items are established in the resulting dialogue. Coverage@k averages the fraction of recovered items across conversations after k questions.

Our learned interviewer uses Qwen3-8B as the question generator and as the backbone for the main T and V scorers. We also test Llama-3.1-8B and Qwen3-14B as alternative scorer backbones. The generator samples at temperature 0.8. DeepSeek-V4-Pro plays the interviewee at temperature 0, answering from the reference transcript, and Qwen3.5-397B-A17B serves as the information-item judge at temperature 0. GLM-5.1, DeepSeek-V4-Pro, and Kimi-K2.6 serve as prompt-only interviewer baselines. The main comparison fixes the generator and eight-question budget while contrasting appropriateness-constrained selection with utility-only selection.

Single-turn clarification. On ClariQ, an agent receives an initial information request and decides whether to ask one clarifying question before retrieval. If it asks, it selects a question from the topic-specific candidate pool. Otherwise, retrieval proceeds directly from the original request. The fixed retrieval lookup evaluates the selected action using one facet per topic across 50 test topics. The objective is to improve retrieval while avoiding unnecessary questions. We measure retrieval with

nDCG@20 and report ask rate, unwanted-ask rate, and missed-clarification rate. The latter two use all evaluated topics as their denominator.

The PAUSE result in Table 3 pairs the ClariQ need-to-ask scorer with the YIELD finetuned utility scorer, $T_{\text{ClariQ}} \times V^{\text{ZS}}_{\text{YIELD}}$. Its threshold, $\tau = 0.8964$, is calibrated on dev-25 for a target UAR of 10%. We also report task-specific Qwen3-8B scorer runs and Llama-3.1-8B and Qwen3-14B alternative-backbone results. All policies use the same retrieval lookup; controls include no-ask, always-ask, single-signal, prompt-only, and lexical or embedding-based selection.

5 RESULTS

The clearest decision-rule gain is on NewsInterview: with the generator and question budget fixed, the joint selection rule improves Coverage@8 from 0.5042 to 0.5676 (Table 1). On 50 ClariQ test topics, the training-mix policy asks on 42 topics and obtains nDCG@20 of 0.1504 [0.104, 0.201] (Table 3). In the matched grid, its paired gain over the same-column utility-only policy is 0.008 [−0.005, 0.026], which is not significant (Table 4). We first examine the decision-rule results in both datasets and then report scorer-backbone sensitivity. Further scorer and generator diagnostics appear in the appendix.

5.1 APPROPRIATENESS IMPROVES MULTI-TURN INFORMATION RECOVERY

The main comparison asks whether T improves information recovery when question generation and utility scoring are held fixed. It evaluates 348 NewsInterview conversations disjoint from the development cohort. Each selection-rule arm uses the same frozen DPO generator, with four candidates ($K = 4$) per turn and a total budget of eight questions. We use DeepSeek-V4-Pro as the interviewee and Qwen3.5-397B-A17B as the information-item judge. Each policy rolls out along its own dialogue history; the two main policies therefore share the sampling procedure but not identical candidate texts. We report Coverage@8, the average fraction of reference information items recovered after eight questions in Table 1.

Table 1: NewsInterview Coverage@8 on 348 test conversations. Paired contrasts are each selection rule minus utility-only, constrained (T+V) minus the named prompt-only system, or constrained minus utility-only within an alternative scorer backbone. Brackets give paired 95% bootstrap confidence intervals.

System / selection rule	Coverage@8 ↑	Paired difference [95% CI]
Various selection rules		
Utility-only (*baseline*)	0.5042	n/a
$0.25T + 0.75V$	0.5529	+0.0486 [+0.0178, +0.0802]
$0.50T + 0.50V$	0.5478	+0.0436 [+0.0095, +0.0768]
Utility threshold, $c = 0.25$	0.5104	+0.0062 [−0.0279, +0.0412]
Utility threshold, $c = 0.40$	0.5149	+0.0106 [−0.0219, +0.0441]
Constrained T+V (*PAUSE*)	**0.5676**	+0.0634 [+0.0302, +0.0968]
Prompt-only baselines		
GLM-5.1	0.5227	+0.0449 [+0.0135, +0.0754]
DeepSeek-V4-Pro	0.4311	+0.1365 [+0.1039, +0.1687]
Kimi-K2.6	0.4360	+0.1316 [+0.0984, +0.1644]
Alternative scorer backbones		
Llama-3.1-8B, utility-only	0.4654	n/a
Llama-3.1-8B, constrained T+V	0.5625	+0.0971 [+0.0629, +0.1306]
Qwen3-14B, utility-only	0.4936	n/a
Qwen3-14B, constrained T+V	0.5266	+0.0330 [−0.0001, +0.0650]

PAUSE reaches Coverage@8 of 0.5676, improving on the matched utility-only baseline by 0.0634 [0.0302, 0.0968]. The weighted combinations reach 0.5529 and 0.5478. Both exceed utility-only selection, while PAUSE has the highest observed coverage among the decision rules. The two utility-threshold controls reach 0.5104 and 0.5149. These controls select the highest-V candidate when its score exceeds c and otherwise use a generic follow-up. The comparisons support incorpo-

rating appropriateness, although Table 1 does not establish that hard filtering outperforms weighted scoring.

Prompt-only interviewers are the system-level baselines in Table 1. They use the same cohort, respondent, judge, and question budget but generate questions directly rather than selecting from the frozen Qwen3-8B generator's candidate pool. PAUSE has higher coverage than the strongest prompt-only baseline, GLM-5.1 (0.5676 versus 0.5227; difference +0.0449 [0.0135, 0.0754]). Figure 2 traces coverage through all eight questions for the two selection rules and the human interviewer reference, which reaches 0.7225 at question eight. We treat human coverage as the ceiling for this evaluation.

The NewsInterview rule advantage is not confined to Qwen3-8B scorers. We also replace the T/V scorer pair with Llama-3.1-8B or Qwen3-14B while fixing other settings. Llama yields a Coverage@8 difference of $+0.0971$ [0.0629, 0.1306], compared to its utility-only counterpart. Qwen3-14B yields $+0.0330$ [-0.0001, 0.0650], whose interval narrowly includes zero.

We next evaluate three scorer-training states across three backbone sizes. Zero-shot (ZS) uses a scorer adapter trained on the source task directly on the target task, without target-task training. Fine-tuning (FT) continues training that source adapter on target-task data, whereas scratch trains a new adapter on target-task data from the base model. These definitions apply separately to T and V: their source tasks are SPoRC and YIELD, respectively.

Table 2 summarizes the NewsInterview 3-backbone by 3-training-state grid. For T, scratch training lowers MRR relative to ZS on both 8B backbones (Qwen3-8B: -0.064 [-0.116, -0.015]; Llama-3.1-8B: -0.049 [-0.121, -0.001]), while the difference is not significant for Qwen3-14B. Continued fine-tuning raises MRR numerically for all three backbones; the paired gain is significant for Qwen3-14B (+0.076 [0.012, 0.147]) and not significant for the two 8B models. V AUROC also rises numerically after in-domain adaptation in these scorer-level evaluations. These metric changes are distinct from downstream policy effects; the separate rollout comparisons are reported in Appendix A.4.

Table 2: NewsInterview scorer results across three backbones and training states. Each cell reports T MRR / V AUROC. Scratch is trained on NewsInterview; fine-tuning continues from the source-domain adapter.

Backbone	Zero-shot	Scratch	Fine-tuned
Qwen3-8B	0.8581 / 0.7217	0.7943 / 0.7935	0.8841 / 0.8217
Llama-3.1-8B	0.8346 / 0.7233	0.7852 / 0.7854	0.8906 / 0.8160
Qwen3-14B	0.8516 / 0.7425	0.8451 / 0.7779	0.9271 / 0.8195

In a separate judge-sensitivity analysis of previously recorded rollouts, the policy advantage has a similar direction and magnitude under Qwen3.5-397B-A17B and DeepSeek-V4-Pro (Appendix A.5).

5.2 Clarification Reduces Question Use while Improving Retrieval Gains

On ClariQ, the reported PAUSE policy asks on 42 of 50 test topics, as shown in Table 3. It pairs the ClariQ need-to-ask scorer with the YIELD zero-shot utility scorer, $T_{\text{ClariQ}} \times V^{\text{ZS}}_{\text{YIELD}}$, using $\tau = 0.8964$ calibrated on dev-25 at a target UAR of 10

The first group of Table 3 compares question-selection rules on the same candidate pools. The always-ask lexical baseline selects the question with the highest word-set Jaccard overlap with the initial request; m3e-base and BGE-M3 instead rank questions by cosine similarity between dense embeddings. Their nDCG@20 values are 0.1254, 0.1334, and 0.1408, respectively, and each asks on every topic with UA=0.08. The PAUSE training-mix cell reaches 0.1504 [0.104, 0.201], asks on 42 topics, and has UA=0.02. Table 4 uses the same ZS/FT/scratch definitions for its T and V variants; its best cell pairs a ClariQ-trained scratch T with the YIELD-trained ZS V. For this cell, the paired gate difference is $+0.008$ [-0.005, $+0.026$] against the same-column V-argmax; this interval includes zero.

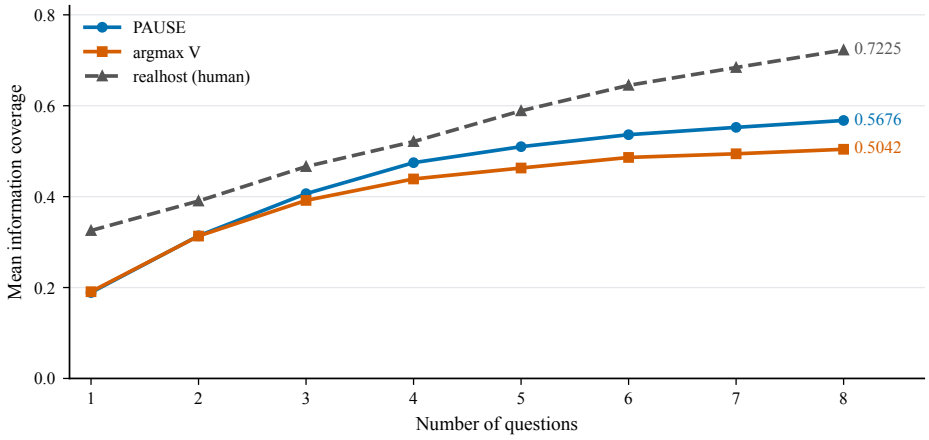


Figure 2: NewsInterview mean information coverage across eight questions (348 conversations). The curves show the direct PAUSE policy (keep), utility-only selection ($\arg\max V$), and the human interviewer reference. Endpoint labels give the values for this turn-wise rollout series.

Table 3: Single-turn ClariQ results on 50 test topics. nDCG@20 intervals are topic-clustered bootstrap 95% CIs (5,000 resamples; seed 2026). PAUSE uses the held-out best cell, $T_{\text{ClariQ}} \times V^{\text{ZS}}_{\text{YIELD}}$, with $\tau = 0.8964$ calibrated on dev-25 at target UAR=10%. UA is the unwanted-ask rate over all topics. Gold-need and answer-informed rows are diagnostics for the original task-specific scorer.

System / policy	nDCG@20 ↑	Ask rate	UA
Decision rules and diagnostics			
No ask	0.1112 [0.072, 0.154]	0.00	0.00
Always ask + lexical	0.1254 [0.085, 0.169]	1.00	0.08
m3e cosine ranking	0.1334 [0.085, 0.187]	1.00	0.08
bge-m3 cosine ranking	0.1408 [0.090, 0.198]	1.00	0.08
Always ask + V	0.1390 [0.088, 0.194]	1.00	0.08
T gate + lexical	0.1264 [0.084, 0.171]	0.78	0.00
T **gate +** V **(*PAUSE*)**	**0.1504 [0.104, 0.201]**	0.84	**0.02**
Alternative scorer backbones			
Llama-3.1-8B	0.1330 [0.087, 0.182]	0.64	0.00
Qwen3-14B	0.1392 [0.092, 0.188]	0.88	0.00
Prompt-only baselines			
GLM-5.1 (basic)	0.0970 [0.062, 0.136]	0.98	0.06
DeepSeek-V4-Pro (basic)	0.1087 [0.070, 0.152]	0.98	0.06
Kimi-K2.6 (basic)	0.1123 [0.071, 0.156]	0.90	0.06
GLM-5.1 (rubric)	0.1153 [0.071, 0.163]	0.74	0.04
DeepSeek-V4-Pro (rubric)	0.1219 [0.079, 0.168]	0.54	0.06
Kimi-K2.6 (rubric)	0.1312 [0.086, 0.179]	0.66	0.04
Oracle experiments			
Gold-need gate + V	0.1472 [0.094, 0.204]	0.92	0.00
T gate + answer-informed	0.1914 [0.134, 0.253]	0.78	0.00
Always ask + random	0.1197 [0.074, 0.171]	1.00	0.08

The gold-need and answer-informed rows in Table 3 retain diagnostics for the earlier task-specific scorer configuration; we discuss them in Appendix A.6. The reported PAUSE cell uses $\tau_{\text{C}} = 0.8964$, calibrated on dev-25 at target UAR=10%. Appendix A.6 retains the threshold sweep for the earlier task-specific $T_{\text{C}} + V_{\text{C}}$ scorer configuration.

We compare three API-accessed LLMs—GLM-5.1, DeepSeek-V4-Pro, and Kimi-K2.6—using prompt-based question selection. In the zero-shot setting, each model directly selects a question from the topic-specific candidate pool; the rubric setting adds explicit decision criteria and three demonstrations from training topics. The zero-shot prompt policies obtain nDCG@20 between

Table 4: ClariQ nDCG@20 across the T and V training variants on 50 held-out topics. Cells show topic-clustered bootstrap 95% CIs; bold denotes the best cell.

	V_{ZS}	V_{ft}	V_{scratch}
T_{ZS}	0.1334 [0.091, 0.179]	0.1165 [0.073, 0.162]	0.1123 [0.068, 0.158]
T_{ft}	0.1467 [0.102, 0.195]	0.1182 [0.079, 0.162]	0.1147 [0.072, 0.161]
T_{scratch}	**0.1504 [0.104, 0.201]**	0.1308 [0.084, 0.180]	0.1239 [0.077, 0.174]

0.0970 and 0.1123; adding the rubric and demonstrations yields 0.1153–0.1312. The PAUSE point estimate is 0.0192 above the strongest rubric baseline, Kimi-K2.6 at 0.1312.

The alternative scorer configurations reported in Table 3 are Llama-3.1-8B at two epochs and Qwen3-14B at six epochs. Their additional training runs and scorer-level metrics appear in Appendix A.6.

6 Discussion

PAUSE reframes intervention selection from a single utility optimization problem into a constrained decision problem: contextual appropriateness determines which interventions are feasible, while utility determines preferences among feasible candidates. The NewsInterview results support the value of this decomposition, as adding appropriateness improves information recovery when the generator and question budget are fixed. At the same time, weighted combinations also outperform utility-only selection, and the observed differences between weighted scoring and PAUSE do not establish that hard filtering is universally preferable. The evidence therefore supports treating appropriateness as a distinct decision variable, while leaving the optimal mechanism for combining appropriateness and utility open.

The role of appropriateness depends on the decision horizon and action space. In NewsInterview, appropriateness captures whether a candidate question fits a particular position in an evolving interaction, whereas in ClariQ it mainly determines whether clarification is warranted before retrieval. On the 50 ClariQ test topics, PAUSE exhibits more selective asking behavior, but the retrieval improvement over utility-only selection remains uncertain because the paired confidence interval includes zero. Since the two benchmarks differ in data, action space, and evaluation objectives, their contrast does not isolate dialogue length as the only factor determining when appropriateness is useful.

7 Conclusion

We introduced PAUSE, a framework for selective elicitation that reframes proactive information seeking as a constrained decision problem. Rather than treating question selection as utility maximization alone, PAUSE separates two complementary judgments: whether an intervention fits the current conversational state and how much useful information it may reveal. By enforcing contextual appropriateness before optimizing information utility, PAUSE provides a simple mechanism for selecting interventions that are both informative and timely. Experiments on NewsInterview show that incorporating appropriateness improves multi-turn information recovery when the question budget and generator are fixed. On ClariQ, PAUSE reduces unnecessary clarification behavior, while the limited retrieval improvement highlights the remaining challenge of estimating question utility before observing the answer. These findings suggest that contextual fit is an important but underexplored factor in proactive dialogue systems, especially for long-horizon interactions where the suitability of an intervention depends on the evolving conversational state. Future work includes developing more adaptive approaches for jointly estimating appropriateness and utility, extending selective elicitation to open-ended generation, and studying how these signals interact under broader conversational objectives.

AI USE STATEMENT

We used generative AI tools to assist with editing and polishing the manuscript and to support the retrieval and discovery of related work; the selected literature is discussed in Section 2. We also used AI to generate synthetic training data for generator distillation: the base Qwen3-8B model sampled candidate questions at the training boundaries, which were filtered to construct the SFT targets and DPO pairs described in Appendix B. The authors are responsible for the final content of this paper and the reported results.

REFERENCES

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 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*, pp. 475–484. ACM, 2019. doi: 10.1145/3331184.3331265. URL https://doi.org/10.1145/3331184.3331265.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4473–4484. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.367. URL https://aclanthology.org/2021.emnlp-main.367/.

Tim Baarslag and Michael Kaisers. The value of information in automated negotiation: A decision model for eliciting user preferences. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems*, pp. 391–400. International Foundation for Autonomous Agents and Multiagent Systems, 2017. URL https://www.ifaamas.org/Proceedings/aamas2017/pdfs/p391.pdf.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024. doi: 10.48550/arXiv.2402.03216. URL https://arxiv.org/abs/2402.03216.

Yizhou Chi, Jessy Lin, Kevin Lin, and Dan Klein. CLARINET: Augmenting language models to ask clarification questions for retrieval. *arXiv preprint arXiv:2405.15784*, 2024. doi: 10.48550/arXiv.2405.15784. URL https://arxiv.org/abs/2405.15784.

Victor De Lima and Grace Hui Yang. YIELD: A large-scale dataset and evaluation framework for information elicitation agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14882–14895. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.678. URL https://aclanthology.org/2026.acl-long.678/.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Xuan Long Do, Bowei Zou, Shafiq Joty, Tran Tai, Liangming Pan, Nancy Chen, and Ai Ti Aw. Modeling what-to-ask and how-to-ask for answer-unaware conversational question generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10785–10803. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.603. URL https://aclanthology.org/2023.acl-long.603/.

Yijiang River Dong, Tiancheng Hu, Zheng Hui, Caiqi Zhang, Ivan Vulić, Andreea Bobu, and Nigel Collier. Value of information: A framework for human–agent communication. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42879–42896. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1987. URL https://aclanthology.org/2026.acl-long.1987/.

Erik Ekstedt and Gabriel Skantze. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2981–2990. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.268. URL https://aclanthology.org/2020.findings-emnlp.268/.

Felix Gervits, Ravenna Thielstrom, Antonio Roque, and Matthias Scheutz. It's about time: Turn-entry timing for situated human-robot dialogue. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 86–96. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.sigdial-1.12. URL https://aclanthology.org/2020.sigdial-1.12/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Kenta Hama, Atsushi Otsuka, Masahiro Mizukami, Hiroaki Sugiyama, and Makoto Naka. RaPSIL: A preference-guided interview agent for rapport-aware self-disclosure. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 355–369. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.sigdial-1.29. URL https://aclanthology.org/2025.sigdial-1.29/.

Behnam Hedayatnia, Di Jin, Yang Liu, and Dilek Hakkani-Tur. A systematic evaluation of response selection for open domain dialogue. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 298–311. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.sigdial-1.30. URL https://aclanthology.org/2022.sigdial-1.30/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. doi: 10.48550/arXiv.2106.09685. URL https://arxiv.org/abs/2106.09685.

Benjamin Roger Litterer, David Jurgens, and Dallas Card. Mapping the podcast ecosystem with the structured podcast research corpus. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25132–25154. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1222. URL https://aclanthology.org/2025.acl-long.1222/.

Yosi Mass, Doron Cohen, Asaf Yehudai, and David Konopnicki. Conversational search with mixed-initiative - asking good clarification questions backed-up by passage retrieval. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pp. 65–71. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.dialdoc-1.7. URL https://aclanthology.org/2022.dialdoc-1.7/.

Moonshot AI. Kimi-K2.6. Hugging Face model card, 2026. URL https://huggingface.co/moonshotai/Kimi-K2.6.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Sudha Rao and Hal Daumé III. 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)*, pp. 2737–2746. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1255. URL https://aclanthology.org/P18-1255/.

Ivan Sekulic, Mohammad Aliannejadi, and Fabio Crestani. Towards facet-driven generation of clarifying questions for conversational search. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, 2020.

Alexander Spangher, Michael Lu, Sriya Kalyan, Hyundong Justin Cho, Tenghao Huang, Weiyan Shi, and Jonathan May. NewsInterview: a dataset and a playground to evaluate LLMs' grounding gap via informational interviews. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32895–32925. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1580. URL https://aclanthology.org/2025.acl-long.1580/.

Yuxin Wang, Qingxuan Sun, and Sicheng He. M3E: Moka massive mixed embedding model. Hugging Face model card, 2023. URL https://huggingface.co/moka-ai/m3e-base.

Zeqiu Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf, and Hannaneh Hajishirzi. InSCIt: Information-seeking conversations with mixed-initiative interactions. *Transactions of the Association for Computational Linguistics*, 11:453–468, 2023. doi: 10.1162/tacl_a_00559. URL https://aclanthology.org/2023.tacl-1.27/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Z.ai. GLM-5.1. Hugging Face model card, 2026. URL https://huggingface.co/zai-org/GLM-5.1.

Michael JQ Zhang and Eunsol Choi. Clarify when necessary: Resolving ambiguity through interaction with LMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5541–5558. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.306. URL https://aclanthology.org/2025.findings-naacl.306/.

A Additional Experimental Analyses

A.1 Appropriateness and Utility Prediction

The appropriateness scorer must recognize both the dialogue position and the proposed response. On 884 SPoRC test ranking groups, the Qwen3-8B scorer achieves mean reciprocal rank 0.792 and top-1 accuracy 0.636. The YIELD-trained utility scorer achieves AUROC 0.7429 on 18,022 held-out question–answer pairs.

The response–position construct check makes the meaning of T more specific (Table 5). The observed host response receives a high score at its actual boundary, but its score falls when placed prematurely. A matched decoy is scored low at either position. The difference-in-differences interaction is $+0.558$, with a paired 95% interval of $[0.465, 0.646]$. Thus, the learned signal responds to the fit between a particular response and its position, rather than merely detecting a turn boundary or general topical relevance.

Table 5: Mean appropriateness score in the SPoRC response–position construct check. The observed response is favored at its actual boundary, while matched decoys remain low at both positions.

Position	Observed response	Matched decoy
Observed boundary	**0.664**	0.103
Premature boundary	0.161	0.105

A.2 Calibrating the NI Gate with PBSR

We calibrate the NewsInterview appropriateness threshold using the pre-boundary selection rate (PBSR). For each SPoRC validation example, we hold the recorded response fixed, order its candidate positions, and find the earliest position whose position-only A2 score reaches τ. PBSR is

the fraction of all examples for which this first crossing precedes the observed boundary; examples without a crossing do not count as premature selections. We select τ on a 300-point grid to bring validation PBSR close to the target ϵ, then freeze it for test and cross-dataset evaluation. Table 6 reports the resulting operating points.

Table 6: PBSR calibration and transfer for the position-only A2 scorer. Thresholds are selected on SPoRC validation data (approximately 670 examples) and evaluated without retuning on SPoRC test (897 examples) and NewsInterview (1,282 examples).

Target ϵ	Frozen τ	Val PBSR	SPoRC test PBSR	NewsInterview PBSR
0.05	0.7787	0.0522	0.0468	0.0078
0.10	0.7086	0.1017	0.0814	0.0179
0.15	0.5951	0.1549	0.1304	0.0601
0.20	0.4716	0.2043	0.1817	0.1482

At the 20% operating point, the frozen threshold transfers to NewsInterview with PBSR 0.1482. The NewsInterview rollout reuses $\tau = 0.47$ with the response-conditioned A2RH scorer, without target-arm recalibration; the accompanying A2RH check reports PBSR 0.0 on its validation and test splits and a rank-style `observed_keep_rate` of 0.94. In the rollout, four generated candidates are filtered by T and the highest-V eligible candidate is selected; if none passes, the highest-T candidate is used. The PBSR figures therefore describe the position-based calibration and transfer checks, not a bound on early interventions under this candidate-pool fallback policy. The same $\tau = 0.47$ is used across the NI scorer arms rather than recalibrating each arm on a separate target-domain development set.

A.3 Distillation Improves the Question Generator

We compare the base generator, constrained-selection SFT, and DPO with the SFT policy as its frozen reference. Table 7 separates candidate-level scores on source-domain boundaries from the independent Coverage@8 outcome on the NewsInterview development cohort. SFT raises the mean utility score from 0.594 to 0.714, and DPO raises it to 0.810. The independent utility evaluator follows the same order, from 0.452 to 0.536 to 0.630. Development-cohort Coverage@8 rises from 0.392 to 0.430 to 0.503. The DPO–SFT coverage difference is $+0.073$, with a paired 95% interval of $[0.025, 0.119]$. This cohort was used during method development, so the separate 348-conversation evaluation below supplies the main test of the decision rule.

Table 7: Generator training stages. T, V, and independent V_h are mean candidate-level scores on source-domain boundaries. Coverage@8 uses the NewsInterview development cohort (152 complete base rollouts; 150 each for SFT and DPO).

Generator	T	V	V_h	Coverage@8
Base	0.714	0.594	0.452	0.392
SFT	0.841	0.714	0.536	0.430
DPO	**0.889**	**0.810**	**0.630**	**0.503**

To isolate the effect of T-filtering during generator training, Table 8 compares the T-filtered generator with its no-T counterpart under each inference rule. The T-filtered generator has higher observed Coverage@8 in both settings: by $+0.0260$ under $\arg\max_V$ and $+0.0099$ under keep. These point estimates favor using T to filter generator-training data; paired uncertainty for the updated contrasts is unavailable.

A.4 NewsInterview Scorer Backbones and In-domain Adaptation

Table 9 reports source-domain scorer quality and within-backbone downstream comparisons for the two alternative scorer pairs. The Qwen3-8B scorer pair is the main system in Table 1, with source-domain T MRR 0.7920 and V AUROC 0.7429. For each pair, the generator, respondent, judge,

Table 8: Coverage@8 on 348 held-out NewsInterview conversations. Rows indicate whether the T-based filter is used during generator training; columns indicate the inference rule. The with-T generator is `m5s_dpo_refsft_lora`; the no-T generator is `m5s_dpo_refsft_lora_not`, trained with the T-based data filter removed ($n_{\text{pairs}} = 3981$); all other training settings are identical. Entries are point estimates; paired uncertainty for the updated contrasts is unavailable.

Generator training	$\arg\max_V$	keep
With T	0.5042	0.5676
Without T	0.4782	0.5577

cohort of 348 conversations, and eight-question budget are fixed, while the two rules produce their own rollouts. The constrained-minus-utility-only gap is positive for all three scorer pairs, but the Qwen3-14B interval crosses zero. The 14B utility scorer has the highest YIELD AUROC of the three, yet its downstream gap is the smallest. Thus, better source-domain classification does not by itself predict a larger improvement from the decision rule.

Table 9: NewsInterview alternative scorer backbones. T MRR is measured on SPoRC and V AUROC on YIELD. All intervals are for the within-backbone constrained-minus-utility-only contrast. The main Qwen3-8B pair appears in Table 1.

Scorer backbone	T MRR	V AUROC	Constrained	V-only	Difference [95% CI]
Llama-3.1-8B	0.8511	0.7371	0.5625	0.4654	+0.0971 [0.0629, 0.1306]
Qwen3-14B	0.7536	0.7466	0.5266	0.4936	+0.0330 [−0.0001, 0.0650]

We also tested whether training the scorers on a small NewsInterview development split improves the outcome. For the Qwen3-8B T scorer, training a new adapter on the 112-conversation training split reduces MRR on 64 held-out ranking groups from 0.8581 to 0.7943 (difference -0.0638 [-0.1159, -0.0153]). Continuing from the source-trained adapter yields 0.8841, a numerical gain of 0.0260 [-0.0046, 0.0579] over zero-shot transfer. The corresponding V AUROCs on 193 held-out pairs are 0.7217, 0.7935, and 0.8217 for source-trained, new-adapter, and continued-adapter models. These small held-out sets assess adaptation of the scorers. They do not replace the 348-conversation policy evaluation.

Table 10 tests four adapted scorer pairings on the 348-conversation cohort. Every within-pair constrained-minus-utility-only interval excludes zero, and constrained-policy coverage ranges from 0.5501 to 0.5667 under the adaptation analysis. These rollouts have separate coverage accounting, so their absolute values are not substituted for the main Qwen3-8B policy result. The table reports each pair's own utility-only rollout rather than reusing one utility-only outcome across altered scorers.

Table 10: NewsInterview scorer-adaptation ablation on 348 conversations under adaptation-analysis accounting. ZS denotes the source-trained scorer, scratch denotes a new in-domain adapter, and FT denotes continuation from the source-trained adapter. The generator stays fixed.

T / V scorers	Constrained	V-only	Difference [95% CI]
FT / FT	0.5621	0.5157	+0.0464 [0.0139, 0.0788]
ZS / FT	0.5667	0.5091	+0.0576 [0.0255, 0.0904]
Scratch / ZS	0.5550	0.5175	+0.0375 [0.0055, 0.0707]
ZS / scratch	0.5501	0.5075	+0.0427 [0.0108, 0.0730]

A.5 Sensitivity to the Information-Item Judge

We evaluate the same T-ni × V-best and $\arg\max_V$ utility-only rollouts with Qwen3.5-397B-A17B and DeepSeek-V4-Pro as information-item judges. For PAUSE, Coverage@8 is 0.5676 with Qwen and 0.5652 with DeepSeek. Thus, the judged coverage of each fixed rollout is stable within paired uncertainty.

A.6 CLARIQ THRESHOLD AND BACKBONE DIAGNOSTICS

For the held-out ClariQ experiment, each T-arm has its own threshold calibrated on dev-25 to target UAR=10%. We scan 400 thresholds from 0.001 to 0.999, select the one whose development UAR is closest to the target (breaking ties with the lower threshold), and freeze it for test-50. Here, UAR is the fraction of no-ask topics (`clarification_need`=1) on which the policy asks. Table 11 reports the calibration outcomes; test ask rates and UA are also shown, with UA computed over all test topics.

Table 11: ClariQ held-out threshold calibration on train-162/dev-25/test-50. Each threshold is calibrated independently for its T-arm on dev-25 at target UAR=10%, then frozen for test. Test UA is the fraction of all test topics on which the policy asks when no clarification is needed.

Appropriateness scorer	Frozen τ	Dev UAR	Dev ask rate	Test ask rate	Test UA
T_{ZS}	0.3212	0.000	0.560	0.580	0.040
T_{ft}	0.9415	0.000	0.600	0.600	0.000
T_{scratch}	0.8964	0.000	0.880	0.840	0.020

The selected PAUSE result uses $T_{\mathrm{scratch}} \times V_{\mathrm{ZS}}$ at $\tau = 0.8964$, giving nDCG@20 of 0.1504 on test-50 (Table 3). Table 12 retains the threshold sweep for the earlier task-specific ClariQ scorer configuration. Thresholds from 0.20 through 0.90 produce the same always-ask decisions on those topics. The earlier development-calibrated value of 0.9615 asks on 39 topics, removes the four unwanted asks, and gives the highest observed nDCG@20 among the displayed thresholds; raising it to 0.975 or 0.99 omits more questions and reduces nDCG@20.

Table 12: ClariQ threshold sensitivity for the earlier task-specific scorer configuration on the 50 test topics. UA is the unwanted-ask rate over all topics.

Gate threshold τ_{C}	nDCG@20	Ask rate	UA
0.20–0.90	0.1390	1.00	0.08
0.9615	**0.1456**	0.78	0.00
0.975	0.1244	0.52	0.00
0.990	0.1123	0.02	0.00

The ClariQ backbone runs in Table 13 also record the utility-only and answer-informed controls (Table 14). Across all five scorer configurations, the observed joint-policy nDCG@20 exceeds that of its own always-ask utility ranker, but the differences are small and all configurations are evaluated on the same 50 test topics. Ask rates vary substantially, from 0.32 for six-epoch Llama to 0.88 for six-epoch Qwen3-14B. The answer-informed selector is much stronger than either learned ranker within each configuration, reinforcing that candidate choice remains an important limitation.

The gold-need and answer-informed rows in Table 3 retain diagnostics for the earlier task-specific scorer configuration. Under that configuration, replacing the learned need-to-ask decision with gold need labels raises nDCG@20 from 0.1456 to 0.1472. Keeping its learned gate but selecting questions with answer-informed retrieval scores raises nDCG@20 to 0.1914, pointing to question ranking as a remaining bottleneck. A separate value-of-information oracle also optimizes whether to ask and which question to choose, reaching 0.2246 nDCG@20 at an ask rate of 0.64. These diagnostics use information unavailable to the reported training-mix policy.

A.7 CROSS-TASK SCORER INITIALIZATION ABLATION ON CLARIQ

Table 15 reports an exploratory initialization ablation. Each ClariQ scorer was either trained under the original task-specific recipe or continued from its source-domain adapter before task-specific fine-tuning. Continuing only the appropriateness scorer changes joint nDCG@20 from 0.1456 to 0.1465. Continuing only the utility scorer lowers it to 0.1329, and continuing both gives 0.1371. The observed changes are measured on the same 50 test topics and are not accompanied by paired uncertainty estimates. They therefore identify a possible transfer limitation, not a statistically estab-

Table 13: ClariQ scorer-backbone sensitivity for the earlier task-specific policy on the same 50 test topics. T AUROC measures need-to-ask prediction, and V Spearman measures candidate ranking. The Qwen3-8B V correlation was not reported in the experiment record.

Scorer backbone	Epochs	T AUROC	V Spearman	Joint nDCG@20
Qwen3-8B	2	0.9788	n.r.	**0.1456**
Llama-3.1-8B	2	0.8899	-0.051	0.1330
Llama-3.1-8B	6	0.8844	$+0.076$	0.1272
Qwen3-14B	2	**0.9890**	-0.040	0.1255
Qwen3-14B	6	0.9796	$+0.067$	0.1392

Table 14: Additional ClariQ scorer-backbone outcomes on the same 50 test topics. The answer-informed column is a diagnostic upper bound with the learned T gate, not a realizable policy.

Scorer backbone	Epochs	Joint nDCG@20	V-only nDCG@20	Answer-informed nDCG@20 / ask
Qwen3-8B	2	0.1456	0.1390	0.1914 / 0.78
Llama-3.1-8B	2	0.1330	0.1230	0.1560 / 0.64
Llama-3.1-8B	6	0.1272	0.1203	0.1573 / 0.32
Qwen3-14B	2	0.1255	0.1175	0.2003 / 0.86
Qwen3-14B	6	0.1392	0.1275	0.2003 / 0.88

lished deterioration. They also show why scorer adaptation should be evaluated at the downstream decision level rather than inferred from a training objective alone.

Table 15: Exploratory ClariQ cross-task initialization ablation on 50 test topics. Task-specific denotes the original ClariQ-trained scorer; continued denotes initialization from the source-domain adapter followed by ClariQ fine-tuning.

T initialization	V initialization	Joint nDCG@20
Task-specific	Task-specific	0.1456
Continued	Task-specific	0.1465
Task-specific	Continued	0.1329
Continued	Continued	0.1371

A.8 Illustrative Failure Modes

The aggregate gains do not imply that every conversation benefits. In one inspected NewsInterview conversation (ID 104840), utility-only selection starts with an incorrect guest identity and subsequently fails to acquire any of eight reference items, while the constrained policy covers seven. Conversely, in conversation 81736, the constrained policy drifts to topics unsupported by the reference transcript and covers none of the items, while utility-only selection covers all of them. These cases are selected illustrations, not frequency estimates: contextual fit can prevent a poor opening, but it does not itself guarantee grounding to the respondent's available information.

B Training and Implementation Details

B.1 Model Training

Table 16 summarizes the main Qwen3-8B training runs. Batch/accumulation lists the per-device microbatch size and the number of gradient-accumulation steps. Sequence lengths are token limits. The alternative Llama-3.1-8B and Qwen3-14B scorer runs use the corresponding scorer recipe, except for the explicitly labeled longer-epoch diagnostics.

The source-domain appropriateness and utility scorers are trained as separate label-token predictors in bfloat16. Both use LoRA rank 16, scaling factor 32, and dropout 0.05 on the query, key, value, output, gate, up, and down projection layers. They use AdamW with weight decay 0.01, a cosine

Table 16: Hyperparameters for the main trained components. Batch/accumulation gives the per-device microbatch size and gradient-accumulation steps. Maximum length is in tokens.

Component	Epochs	Learning rate	Batch/acc.	Max tok.
SPoRC T scorer	2	10^{-4}	2×8	900
YIELD V scorer	2	10^{-4}	2×8	768
Generator SFT	2	10^{-4}	8×2	1024
Generator DPO	1	5×10^{-6}	2×8	1024
ClariQ T scorer	2	10^{-4}	8×2	320
ClariQ V scorer	2	10^{-4}	8×2	384

learning-rate schedule with 3% warmup, and gradient clipping at norm 1.0. The SPoRC scorer is selected by validation mean reciprocal rank, and the YIELD scorer by validation average precision.

For the ClariQ scratch variants, new scorer adapters are trained on ClariQ data from the base model; the FT variants instead continue from the source-domain adapters. The main PAUSE cell uses the scratch need-to-ask scorer together with the YIELD-trained V scorer in zero-shot transfer, without ClariQ training for V. The need-to-ask scorer reads the initial request, and the ClariQ-trained question-utility scorer reads the request together with a candidate clarification question. Its utility target is positive when the candidate's answer-conditioned retrieval value reaches the global median computed on training examples. The ClariQ-trained scorers use rank-16 LoRA modules, bfloat16 computation, gradient checkpointing, AdamW with weight decay 0.01, OneCycle learning-rate scheduling with a 3% warmup fraction, and gradient clipping at norm 1.0. Checkpoints are selected by validation AUROC for T and validation Spearman correlation for V. The longer six-epoch backbone runs are diagnostic variants, not the main two-epoch configuration.

B.2 Generator Distillation and Evaluation

For generator training, we sample eight questions per boundary from the base Qwen3-8B model at temperature 0.9 and top-p 0.95, with at most 96 new tokens. The source comprises 4,000 training boundaries. We keep the top $\lceil 0.875K \rceil$ candidates by appropriateness, require $T \geq 0.0952$ and at least 15 characters, and choose the remaining candidate with the highest utility for SFT. An episode-level split yields 3,799 training and 201 validation targets. DPO uses 3,950 pairs: the chosen and rejected questions are the highest- and lowest-utility distinct candidates within the same filtered set, with a utility margin of at least 0.10 and chosen utility of at least 0.15.

The generator SFT adapter uses LoRA rank 32, scaling factor 64, and dropout 0.05 on the same seven projection layers. It is trained in bfloat16 with gradient checkpointing, AdamW weight decay 0.01, and a OneCycle schedule with a 3% warmup fraction. The loss is computed only on completion tokens. DPO continues from that adapter, uses $\beta = 0.1$, a frozen copy of the SFT model as reference, AdamW weight decay 0, and the same schedule and completion-token masking. Both generator stages clip gradients at norm 1.0 and use seed 2026.

During the NewsInterview evaluation, the trained generator samples four candidates per turn at temperature 0.8 and top-p 0.95, with sampling seed 2027. The respondent and information-item judge use temperature 0. Dialogue outcomes are measured after eight questions, and paired differences are summarized with 5,000 conversation-clustered bootstrap resamples using seed 2026. ClariQ uses the fixed retrieval lookup for its selected question or no-ask action. Uncertainty intervals resample the 50 test topics with the same bootstrap count and seed.

C Development-Set Threshold Selection

The two tasks use different appropriateness scores, so their decision thresholds are calibrated separately. Table 17 gives the selected development-set operating points. In each case, we scan a fixed threshold grid to match a target intervention-error rate, rather than choosing the threshold that maximizes downstream Coverage@8 or nDCG@20.

Table 17: Development-set calibration of the appropriateness gates. PBSR is the pre-boundary selection rate. Calibration UAR is the ask rate among examples labeled as not needing clarification. The NewsInterview rollout uses the rounded threshold 0.47.

Task	Target	Selected τ	Dev rate
NewsInterview	PBSR 0.20	0.4716	0.2043
ClariQ, task-specific (T+V)	UAR 0.10	0.9615	0.0000
ClariQ, $T_{\mathrm{ClariQ}} \times V_{\mathrm{YIELD}}^{\mathrm{ZS}}$	UAR 0.1000	0.8964	0.0000.

C.1 NEWSINTERVIEW

For the multi-turn policy, we calibrate a pre-boundary selection rate (PBSR) on the SPoRC validation split. Within each validation example, candidate intervention positions are scanned in dialogue order, and an intervention is premature if the earliest position whose score meets or exceeds τ precedes the observed boundary. We evaluate 300 thresholds between 0.001 and 0.999 and choose the one whose validation PBSR is closest to the 0.20 operating target. This gives $\tau = 0.4716$ and validation PBSR 0.2043. The NewsInterview rollout uses the rounded value $\tau_{\mathrm{N}} = 0.47$, fixed before the 348-conversation evaluation.

The validation calibration uses a position-only turn-fit scorer, whereas the NewsInterview rollout applies the rounded threshold to the response-conditioned appropriateness scorer. The PBSR reported in Table 17 therefore describes the calibration scorer, not a separately measured PBSR for the deployed scorer. The generator-training filter $\tau_{\mathrm{tr}} = 0.0952$ is also distinct: it is the tenth percentile of observed-response appropriateness scores in a held-out SPoRC candidate-pool analysis and is used only to construct SFT targets and DPO pairs.

C.2 CLARIQ

For the earlier task-specific single-turn policy, $T_{\mathrm{C}}(s)$ scores whether the initial request needs a question. We evaluate 400 thresholds between 0.001 and 0.999 on its development examples and choose the threshold whose conditional unwanted-ask rate is closest to 0.10, yielding $\tau_{\mathrm{C}} = 0.9615$. The measured conditional rate is 0.00 because attainable rates are discrete. This threshold is held fixed for the earlier joint $T_{\mathrm{C}} + V_{\mathrm{C}}$ policy. The replacement PAUSE cell in Table 3 instead uses $T_{\mathrm{ClariQ}} \times V_{\mathrm{YIELD}}^{\mathrm{ZS}}$, with $\tau = 0.8964$ calibrated on dev-25 at target UAR=0.10. Test-set UA uses all 50 topics as its denominator, whereas calibration UAR uses only no-ask development examples.