Dual Hierarchical Dialogue Policy Learning for Legal Inquisitive Conversational Agents

Xubo Lin
Georgetown University
xl524@georgetown.edu

Zezhi Deng
Georgetown University
zd127@georgetown.edu

Shihao Wang
Georgetown University
sw1379@georgetown.edu

Grace Hui Yang
Georgetown University
Grace.yang@georgetown.edu

Yang Deng
Singapore Management University
ydeng@smu.edu.sg

Abstract

Most existing dialogue systems are user-driven, primarily designed to fulfill user requests. However, in many critical real-world scenarios, a conversational agent must proactively extract information to achieve its own objectives rather than merely respond. To address this gap, we introduce *Inquisitive Conversational Agents (ICAs)* and develop an ICA specifically tailored to U.S. Supreme Court oral arguments. We propose a Dual Hierarchical Reinforcement Learning framework featuring two cooperating RL agents, each with its own policy, to coordinate strategic dialogue management and fine-grained utterance generation. By learning when and how to ask probing questions, the agent emulates judicial questioning patterns and systematically uncovers crucial information to fulfill its legal objectives. Evaluations on a U.S. Supreme Court dataset show that our method outperforms various baselines across multiple metrics. It represents an important first step toward broader high-stakes, domain-specific applications.[1]

1 Introduction

Conversational AI has long focused on user-driven systems suited to tasks like customer service or digital assistants. They excel when the discourse is close-ended and user–driven. However, they are not well-suited when it comes to scenarios like court justices, where they do not passively absorb information; instead, they prod, reframe, and challenge, creating a line of inquiry that tests the attorney's narrative and hunts for latent inconsistencies. The dialogue has a moving target of questions and counters, and it is this information-seeking dynamic that we call inquisitive dialogue.

Much of the literature that characterizes itself as "task-oriented dialogue" in fact captures only one slice of the space: collaborative dialogue where system and user share a goal. Datasets such as MultiWOZ (Budzianowski et al., 2020), Schema-Guided Dialogue (Rastogi et al., 2020), Taskmaster (Byrne et al., 2019), etc., canonize that slice by framing the agent as a benevolent assistant whose sole duty is to satisfy explicit user requests. Their well-formulated slot ontologies, crowd-written templates, and short conversational arcs make them ideal for supervised learning but simultaneously ill-suited for settings where the agent, not the interlocutor, steers the agenda. Treating these resources as the entirety of task-oriented dialogue (TOD), therefore, overstates their scope and leaves the inquisitive and negotiation spectrum virtually unmapped, for example, in figure 1, the utterance "The Sixth Amendment only protects your money up until the point where there's a judgment?" is a task oriented question but will not appear in collaborative or negotiation dialogue.

Inquisitive dialogue poses multiple challenges. First, initiative and relevance are context contingent: asking "Which soda do you prefer?" in an interview can be incisive or irrelevant depending on the preceding exchange, a nuance that traditional conversational agents can struggle to capture. Second, the interaction horizon is long. Supreme Court transcripts routinely exceed 5 000 tokens per round, stretching the capacity of mainstream encoder–decoder models that underpin many collaborative agents (Su et al., 2022; Shu et al., 2019). Additionally, the dialogue participants do not share a common goal, and in many cases may be actively working against each other to reach their own goals. Therefore, any agent participating in inquisitive dialogue must learn long-term dialogue and questioning strategies in a non-cooperative context.

To meet these challenges, we propose a Dual Hierarchical Reinforcement-Learning (RL) framework that splits inquisitive reasoning between two tightly coupled agents. An Appraisal Agent evaluates each attorney's response in real time and

[1] Git repository

Findings of the Association for Computational Linguistics: ACL 2026, pages 11030–11047

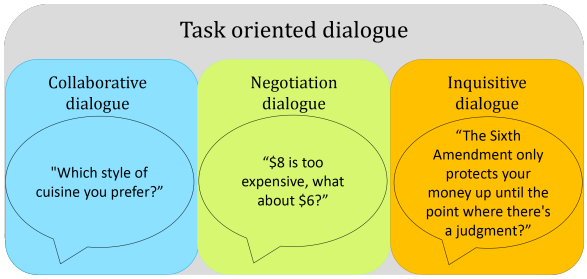


Figure 1: While this paper focuses on inquisitive dialogue in the context of U.S. Supreme Court hearings, we rethink and propose a broader categorization of task-oriented dialogue into three types: collaborative, negotiation(Lewis et al., 2017), and inquisitive. In prior works, non-collaborative types of TOD remain underexplored.

converts those judgments into scalar rewards that shape the next turn, then a Hierarchical Dialogue-Policy Agent regressively generates the up to 3 hierarchies of action based on the output from the Appraisal Agent.

2 Related Work

Proactive Conversational Agents. The development of conversational agents (CAs) has been largely driven by breakthroughs in natural language processing and machine learning. Key approaches include *sequence-to-sequence (Seq2Seq) modeling* (Sutskever et al., 2014), *pretrained language models (PLMs)* (Radford et al., 2019; Liu et al., 2024), *retrieval-assisted text generation (RAG)* (Gao et al., 2024; Izacard and Grave, 2021), and *reinforcement learning (RL)* approaches (Schulman et al., 2017). Among them, RL provides an optimization paradigm for dialogue strategies, particularly in *task-oriented* settings (Budzianowski et al., 2020), where reward-based learning aligns agent behavior with desired outcomes. For instance, Li et al. (2016b) introduced deep RL to incorporate dialogue-level rewards, while Zhao and Eskenazi (2016) proposed an end-to-end system that learns both dialogue state tracking and strategy.

While conventional CAs typically respond to user-initiated requests, a growing line of research focuses on *proactive conversational agents* (Liao et al.), which actively *initiate topics* (Tang et al., 2019), provide *context-aware recommendations* (Zhou et al., 2020), and *guide* users rather than simply reacting (Deng et al., 2023). Proactive agents often leverage *reinforcement learning* (Deng et al., 2024), *strategic planning* (Zhang et al., 2024), or *question generation* (Guo et al., 2024) to address limitations of purely reactive systems, enabling richer support for tasks such as exploratory search and decision-making. ICAs take this concept even further by focusing on *steering* the conversation and *gathering insights* from the user to achieve the system's own objectives. They go beyond offering guidance or recommendations and actively *probe* for information, making them especially suited to domains like legal or investigative dialogues where deeper fact-finding is critical.

Legal Conversational Agents. While much of the research on conversational agents has focused on open-domain or task-oriented contexts, a growing body of work explores their application in the legal domain. For instance, Sharma et al. (2021) build a retrieval-based legal chatbot to address frequently asked legal questions. Although these systems provide valuable assistance, they predominantly adopt a reactive, FAQ-style approach, leaving vacancy for more proactive or inquisitive dialogue models—an area our work aims to advance.

3 Problem Formulation

3.1 Inquisitive Conversations

In this paper, we address the problem of *inquisitive conversation*, where a conversational agent actively probes for critical information to achieve its own objectives, rather than merely responding to user queries. Specifically, we frame this challenge in the context of Supreme Court judicial dialogue.

Inquisitive conversations exhibit several key differences to everyday casual conversations.

Conversational Control: In typical conversations, the user initiates queries and drives the topic. In judicial dialogues, the justice initiates each round of questioning and controls the direction of the discussion.

Purpose: Casual dialogues often serve social or informative purposes, whereas in judicial questioning, each question aims to clarify legal uncertainty, probe for consistency, or expose logical flaws.

Strategy: Justice questioning is deliberate and strategic, employing techniques such as testing hypotheticals, challenging premises, and verifying doctrinal consistency.

To model these differences in inquisitive behavior, we propose the **Inquisitive Conversational Agent (ICA)**, which mimics these questioning patterns using a dual-agent hierarchical reinforcement learning framework.

3.2 Dialogue Formulation

We model the justice–attorney interaction as a Markov Decision Process (MDP), defined by the tuple $M = (S, A, R, \gamma)$, where S is the dialogue state space, A the action space, R the reward function, and γ the discount factor. Each dialogue round t begins with a justice utterance u_j^t, followed by an attorney response u_a^t, forming an interaction pair (u_j^t, u_a^t). The state $s^t \in S$ encodes the dialogue context up to round t.

In our formulation, the justice utterance u_j^t is treated as the action a^t, which transitions the environment to a new state s^{t+1} after observing u_a^{t+1} and yields a scalar reward $r^t = R(s^t, a^t)$.

Appraisal Signal: In inquisitive dialogue, agents operate with their own information-seeking goals. Rather than waiting for user input to guide the exchange, they actively evaluate each response to determine whether it advances their investigative objective. To resonate with this feature of inquisitive dialogue, we introduce an appraisal signal p^t at each turn. It encodes the justice's judgment of the attorney's prior response (e.g., evasive, incomplete, satisfactory) under dialogue state s^t. In our dataset, the appraisal of the justice in each turn t can be inferred from an utterance tuple of two rounds:

$$p^t = f\Big(u_j^{t-1},\, u_a^t,\, u_j^t\Big). \tag{1}$$

For instance, if the justice issues a near-identical utterance across two consecutive turns, it often indicates dissatisfaction with the attorney's prior response. Accordingly, we augment the standard transition tuple to $\mathcal{D} \sim (s^t, p^t, a^t, r^t, s^{t+1})$. The Appraisal Agent treats p^t as the action selected in state s^t, while the Dialogue Agent operates on an augmented state representation $s^t_{\text{aug}} = \text{concat}(s^t, p^t)$. This design enables the Dialogue Agent to condition its next action on its internal assessment of that history as well.

While many dialogue systems treat utterance generation as an open-ended natural language generation (NLG) task with a vast action space (Zhao and Eskenazi, 2016; Sharma et al., 2017; Wang et al., 2022), domain-specific agents can often reduce complexity by operating on a finite set of *dialogue acts* (Peng et al., 2018; Su et al., 2018). In the Supreme Court domain, for instance, justices frequently perform recurrent yet distinct high-level actions (e.g., asking questions, making hypotheses, or making declarations (Cichowicz, 2019)), which lend themselves to a more structured formulation. After they choose a high-level intent, such as questioning, hypothesizing, or declaring, they may refine it into a more specific subtype, like a probing or clarifying question before their actual utterance came up.

Motivated by this, we adopt a **hierarchical action space** that separates policy decisions (i.e., *which dialogue act to take next*) from the lower-level surface text realization (i.e., *how to verbalize that act*). Our approach discretizes justices' interactions into a three-level taxonomy (see Table 3 in appendix) that captures both top-level acts (e.g., a "question") and their subtypes (e.g., probing for clarity vs. challenging an argument).

3.3 Reward Definition

Unlike conventional dialogue rewards that primarily assess the agent's own utterance, our inquisitive setting focuses on how effectively the *justice's utterance* elicits information from the *attorney's subsequent response*. In this work, each justice's utterance u_{j}^t receives a reward comprising the following components.

(1) Solicitation of Goal-Relevant Information. One objective of the agent in an inquisitive dialogue is to gather useful and relevant information that is aligned with their goal. Therefore, we introduce a goal-relevance reward to incentivize probing related to the agent's goal. To capture how effectively the justice's utterance, u_{j}^t, compels an attorney's response, u_{a}^{t+1}, to include legally significant information, we measure the attorney's response's relevance to the case's conclusion C. Using Llama-3-8B (gra, 2024) as a semantic similarity evaluator, we compute the maximum similarity between the attorney's response u_{a}^{t+1} and each sub-conclusion $C[i]$, with scores bounded by 5. Formally,

$$R_{\text{rel}}^{t+1}(s^t, u_{\text{j}}^t) \;=\; \max\Big(\text{sim}\big(C[i],\, u_{\text{a}}^{t+1}\big)\Big), \tag{2}$$

where $C[i]$ denotes individual sub-conclusions of the case's conclusion. This reward encourages justice's inquisitive utterances that steer the dialogue toward legally relevant insights.

(2) Solicitation of Novel Information. A key goal of an ICA is not only to ask questions but to drive the conversation toward uncovering information that has not yet surfaced. To capture this behavior, we introduce a **novelty reward** that measures how

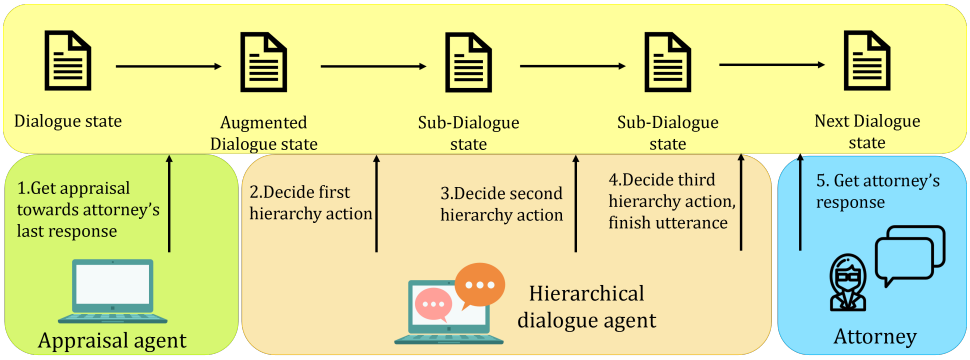


Figure 2: System Architecture of the Proposed Dual Hierarchical Inquisitive Conversational Agent.

effectively the justice's utterance u_{j}^{t} prompts the attorney's next response u_{a}^{t+1} to contribute new and informative content beyond what has already been discussed. This reward encourages the agent to formulate more strategic and context-aware inquiries that elicit additional legal details or perspectives.

Formally, we compute this reward using the *Expectation-Adjusted Distinct (EAD)* metric (Liu et al., 2022), a length-normalized variant of *Distinct-N* (Li et al., 2016a) that evaluates lexical novelty while accounting for utterance length:

$$R_{\text{nov}}^{t+1}(s^t, u_{\text{j}}^t) = \frac{N_{\text{attorney}}^{t+1}}{V\left(1 - \left(\frac{V-1}{V}\right)^{|u_{\text{a}}^{t+1}|}\right)}, \quad (3)$$

where $N_{\text{attorney}}^{t+1}$ represents the number of newly introduced tokens in u_{a}^{t+1} that have not appeared in prior turns,[2] V is the cumulative vocabulary size up to time t, and $|\cdot|$ denotes the token count of the utterance.

(3) Solicitation of Succinct Answer. In Supreme Court dialogues, justices often prefer brief, direct answers (e.g., "yes," "no") from the attorney (Cichowicz, 2019), as these answers can swiftly confirm or deny a point and thus aid the justice's decision-making. Additionally, succinct answers from the attorney helps the justice in keeping control of the dialogue, and conversational control is an important consideration in making an ICA. We reward this succinctness, treating it as evidence that the justice's utterance u_{j}^{t} was well-targeted:

$$R_{\text{clarity}}^{t+1}(s^t, u_{\text{j}}^t) = -\log\left(|u_{\text{a}}^{t+1}|\right), \quad (4)$$

where $|u_{\text{a}}^{t+1}|$ is the token length of the attorney's response. This measure complements the previous two components by explicitly encouraging *clarity* in judicial exchanges.

During training, we combine the three reward components into an aggregated numerical reward via a weighted sum, which allows the agent to balance legal relevance, novelty, and clarity in its inquisitive dialogue.

4 Proposed Method: A Dual-Agent Framework for Legal Inquiry

Building an ICA, which actively uncovers information rather than merely answering queries, poses distinct challenges, especially in complex domains like Supreme Court hearings. To tackle this, we propose a *Dual-Agent Hierarchical RL* framework, depicted in Figure 2, designed to emulate the judicial exchange process. Our approach comprises two coordinated agents, each focusing on a different aspect of the conversation.

Rather than viewing dialogue as a single flat policy, we employ a three-level hierarchical RL dialogue agent that determines *when* to probe further, *how* to frame questions, and *if* the discussion should shift topics. By decomposing each turn into layers, ranging from broad subtopic planning to fine-grained utterance generation—the Dialogue Agent can optimize information elicitation while maintaining coherence and legal formality.

4.1 Appraisal Agent

We introduce a appraisal agent to *evaluate* each attorney response. If the response appears evasive, contradictory, or insufficiently detailed, the appraisal agent flags the need for deeper inquiry. This mechanism mimics a justice's tendency to monitor counsel's answers on the fly, ensuring that the Dialogue Agent adapts its questioning in real time rather than blindly following a predefined

[2] In the original EAD (Liu et al., 2022), N counts distinct tokens; we adapt it to track newly introduced tokens relative to the dialogue history.

script.

Why two agents? Separating response appraisal and dialogue control into two specialized agents enables more modular and interpretable decision-making. The Dialogue Agent focuses exclusively on planning and generating inquisitive moves, while the Appraisal Agent independently assesses whether the information obtained justifies continued exploration.

Similar to dialogue acts, the appraisals can be discretized for a specific domain as well. We summarized nine appraisal types from Supreme Court transcripts(see Table 4 in the appendix). These appraisals allow the justice to evaluate attorney responses, identifying flaws, seeking clarification, or prompting further inquiry, and help ensure the dialogue remains focused, responsive, and inquisitive.

In our proposed method, Appraisal Agent employs a Q-network to choose the appraisal p that maximizes its Q-value estimate:

$$p(s) = \arg\max_{p} Q_{\text{Appraisal}}(s, p; \theta), \tag{5}$$

where s is the current state embedding, and θ denotes the Q-network parameters. The selected appraisal p is then represented as a one-hot vector and merged into the Dialogue Agent's augmented state, guiding subsequent decisions to probe further or shift to the next subtopic as needed.

In our dialogue agent, We augment the overall dialogue state s_t with p_t to yield $s^t_{\text{aug}} = \text{concat}(s_t, p_t)$ by treating the appraisal agent output as an internal *state variable* rather than a separate action, the ICA can better track whether deeper probing is needed or if the conversation should transition to a new subtopic.

4.2 Dialogue Agent

To emulate Supreme Court justices, our Hierarchical Dialogue Agent first decides *which* conversational act to perform (e.g., clarify, probe, or challenge), then determines *how* to realize that act. We formalize these choices in a three-level action taxonomy (Table 3). Level 1 defines high-level dialogue acts, such as *Questioning*, *Hypothesis Testing*, or *Declaration*. Level 2 refines each act into subcategories (e.g., *Clarification*, *Probing*, *Comparison*), while Level 3 specifies the final utterance.

Poincaré Embedding To capture the hierarchical structure of judicial dialogue acts, we represent each action in a Poincaré embedding space (Nickel and Kiela, 2017). Poincaré embeddings are defined in a hyperbolic geometry that naturally preserves hierarchical and tree-like relationships, where parent nodes lie closer to the origin and child nodes are positioned exponentially farther away. By embedding our three-level taxonomy in this hyperbolic space, the Dialogue Agent can learn smoother transitions across levels, leverage proximity for related actions (e.g., between sibling subacts), and better generalize across hierarchically related behaviors. The training target of it is as follows:

$$\mathcal{L} = \sum_{(u,v)\in D} \log \frac{e^{-d(u,v)}}{\sum_{v'\in\mathcal{N}(u)} e^{-d(u,v')}}. \tag{6}$$

Where $d(u, v)$ denotes the hyperbolic distance between embeddings of nodes u and v; D is the set of observed positive pairs (e.g., parent–child or sibling relations) derived from the dialogue act hierarchy; and $\mathcal{N}(u)$ represents a set of negatively sampled nodes unrelated to u.

Multi-Hierarchy Action Selection. The three-level hierarchical action taxonomy (Table 3) allows our Dialogue Agent to operate at varying degrees of granularity. A single full-level action $\{a_0, a_1, a_2\}$ may yield up to three transition tuples: (s, a_0, r, s'), (s, a_1, r, s'), (s, a_2, r, s'), The agent may terminate at any level if the chosen sub-action has no additional children.

The categories are chosen sequentially, where the highest level of action is chosen based on the augmented state, following a Level 2 action that is a subcategory of the chosen Level 1 action, and then the Level 3 action is chosen based on the Level 2 action in the same way. (e.g., choose 'question' as a_0, choose 'Probing question' as a_1, choose 'Probe the assumption' as a_2) Three levels of selection steps correspond to the three transition tuples above. We use these actions to prompt LLM(gra, 2024) under a unified template6 to get a response for the Justice.

4.3 Algorithm

For both appraisal agent and dialogue agent, we use DDQN as the backbone, and the DDQN target for the appraisal agent is:

$$Y_{\text{App}} = r + \gamma Q\Big(s, \arg\max_{p'} Q(s', p'; \theta_{App}); \theta^{-}_{App}\Big), \tag{7}$$

where θ_{App} and θ^{-}_{App} denote the weights of the main network and the target network. The respective DDQN losses are:

$$\mathcal{L}^{\text{DDQN}}_{\text{App}} = \mathbb{E}_{(s,p,s')\sim\mathcal{D}}\big(Q(s, p; \theta_{App}) - Y_{\text{App}}\big)^2, \tag{8}$$

where θ_{App} and θ^{-}_{App} denotes the weights of main network and target network of appraisal agent.

For the dialogue agent, we train one Q-network that generates Q values of all possible next-level hierarchy actions sequentially conditioned on augmented states and parent actions. The DDQN target of it is:

$$Y^{l}_{\text{Dia}} = r + \gamma Q\Big(s, \arg\max_{a'} Q(s', a'; \theta_{Dia}); \theta^{-}_{Dia}\Big), \quad (9)$$

where θ_{Dia} and θ^{-}_{Dia} denotes the weights of the main network and target network.

We assume that the definition of dialogue actions in the dataset $\mathcal{D}$ is complete. For any single full-level action $\{a_0, a_1, a_2\}$, we have:

$$\begin{aligned} Q(s, a_0) &= \max_{a_1} Q(s, a_1) \\ Q(s, a_1) &= \max_{a_2} Q(s, a_2) \end{aligned} \quad (10)$$

where a_1 are all child actions of a_0 and a_2 are all child actions of a_1. It means the Q-value of the parent action can be represented by the Q-value of the 'best' child action. And the respective loss is

$$\begin{aligned} \mathcal{L}^{\text{hier}}_{\text{Dia}} = &\left(Q(s, a_0) - \max_{a_1} Q(s, a_1)\right)^2 \\ &+ \left(Q(s, a_1) - \max_{a_2} Q(s, a_2)\right)^2 \end{aligned} \quad (11)$$

A well-documented challenge in offline reinforcement learning is the overestimation of Q-values for state–action pairs that are insufficiently represented in the dataset (Fujimoto et al., 2019; Kumar et al., 2020). To mitigate this issue in our setting, we introduce a simple yet effective conservative regularization strategy. For each state s, we define $R_1(s) = \max_{a \in \mathcal{A}} Q(s, a)$, which corresponds to the maximum Q-value across all possible actions and is most likely to be overestimated. We penalize it by adding $R_1(s)$ to optimize objectives. However, when these high-Q actions are well represented in the dataset, applying the penalty uniformly can lead to underestimation. To address this, we introduce a compensatory term $R_2(s) = Q(s, a)$, where $(s, a) \in \mathcal{D}$, to restore value estimates for observed transitions.

The resulting regularization terms for the Appraisal and Dialogue Agents are defined as:

$$\begin{aligned} \mathcal{L}^{\text{Reg}}_{\text{App}} &= (R_1(s) - R_2(s)) \\ \mathcal{L}^{\text{Reg}}_{\text{Dia}} &= (R_1(s_{\text{aug}}) - R_2(s_{\text{aug}})) \end{aligned} \quad (12)$$

These terms are incorporated into the final optimization objectives for both agents as follows:

$$\begin{aligned} \mathcal{L}_{\text{App}} &= \mathcal{L}^{\text{DDQN}}_{\text{App}} + \alpha \mathcal{L}^{\text{Reg}}_{\text{App}} \\ \mathcal{L}_{\text{Dia}} &= \mathcal{L}^{\text{DDQN}}_{\text{Dia}} + \beta \mathcal{L}^{\text{Reg}}_{\text{Dia}} + \lambda \mathcal{L}^{\text{hier}}_{\text{Dia}} \end{aligned} \quad (13)$$

Where α, β and λ are regularization coefficients.

When $(s, \arg\max_a Q(s, a)) \in \mathcal{D}$, the regulation term is equivalent to 0. When $(s, \arg\max_a Q(s, a)) \notin \mathcal{D}$, it overestimate Q-value of $(s, a) \in \mathcal{D}$ and underestimate (s, a) pairs in R_1. This regulatory term makes the derived policy lean towards the policy that generates the dataset $\mathcal{D}$ from the potentially overestimated values. So by choosing appropriate α and β, we can reduce the variance without losing the performance.

The implementation details of algorithm can be found in Appendix 1.

5 Experiment

5.1 Experiment Setup

Dataset. We evaluate our work on the publicly available **U.S. Supreme Court Oral Argument Transcript Dataset**. In these transcripts, *justices* actively probe *attorneys* for information critical to deciding a case, closely reflecting the objectives of an ICA. Particularly, we use a subset of appeal court cases (spanning 1955–2023) from `www.Oyez.org` . Each transcript in this dataset contains metadata such as the case name, argument date, and speaker identifiers. The main textual content comprises a **background of the case**, an **argued question**, the complete **dialogue transcript**, and the **final conclusion**. Table 5 in the appendix summarizes the distribution of cases across various legal domains. Our experiments are carried out *offline*, we divide our training and evaluation data by years when the argument happened.

Evaluation Metrics. We employ two complementary evaluation strategies to assess our system's performance. We prompt a legally pretrained, SaulLM-7B (Colombo et al., 2024), to score each utterance generated by the agents (see Appendix 6 for the prompts). In parallel, we collect *manual* ratings from human reviewers, applying the same metrics to each utterance.

We focus on the following metrics. Both the LLM and human judges assign scores on a *1–5* scale, where higher values indicate stronger performance:

Conformity Score (CS). Measures how closely each utterance $\{u_i\}$ reflects judicial norms (e.g., formality, legal phrasing).

Progression Score (PS). Assesses whether u_i *advances* the discussion rather than stalling or digressing.

Outcome Relevance Score (OS). Evaluates each utterance's consistency with the broader objective—such as reaching a legal conclusion or a coherent final ruling.

Probing Effectiveness Score (PES). Captures how effectively u_i *prompts* new information from the interlocutor.

Multi-turn Dialogue Metrics. We introduce two metrics to evaulate the multi-turn dialogue capabilities of the systems. We segment the original transcript from each case into topics and construct an attorney agent using **SeCom** (Pan et al., 2025), and the ICA and the attorney agent engages in a simulated courtroom debate based on the question from the case. The oral argument stage has a time limit; we set the maximum conversation length as 10 rounds to represent it.

We compute a **Coverage Score** from the simulated debate, which calculates how many of the topics in the original transcript was covered by the ICA. Let t_i be the original topics, T be the set of topics from the simulated debate, and $t'_i \in T$ be a topic in the simulated debate. Then the Coverage Score is computed as follows:

$$\sum_{t'_i \in T} \max_{t_i}(Sim(t_i, t'_i)) \tag{14}$$

We also introduce a **Marginal Relevance (MR) Score**, based on **Maximal Marginal Relevance** (Carbonell and Goldstein, 1998). The Marginal Relevance Score evaluates the ICA's ability to probe for new information while staying relevant to the topic of debate. For every round of dialogue, let u_j be the justice's last utterance, and $u_{i<j}$ be the justice's previous utterances. Let q be the question of the case, and n be the number of dialogue rounds. Then the Marginal Relevance Score is computed as:

$$\frac{1}{n}\sum_{u_j} \gamma(Sim(u_j, q)) - (1-\gamma) \max_{u_i}(Sim(u_i, u_j)) \tag{15}$$

We use cosine similarity for Sim in both metrics. Additionally, we set $\gamma = 0.7$ to reward the justice for staying on topic, while still encouraging exploration of new topics.

Together, these two metrics evaluate the ability of the agents to cover all the necessary topics while probing for new information in multi-turn dialogues.

	CS	PS	OS	PES	Overall
Vanilla Llama3	3.99	3.94	4.70	3.92	4.14
SFT Llama3	3.98	3.81	4.45	3.38	3.91
SaulLM-7B	**4.01**	3.91	4.56	3.75	4.06
Hudeček	3.99	3.97	4.77	3.63	4.09
VaRMI	4.00	3.94	4.71	3.93	4.15
ArCHer	3.96	3.79	4.17	4.22	4.04
Ours	**4.01**	**3.98**	**4.89**	**4.47**	**4.34**

Table 1: Main Experimental Results

	CS	PS	OS	PES	Overall
Full Model	4.01	3.98	**4.89**	**4.47**	**4.34**
w/o Appraisal Agent	**4.03**	**4.0**	4.74	4.30	4.27
w/o Succinct Reward	4.01	3.97	4.85	4.39	4.31
w/o Novelty Reward	4.01	3.97	4.82	4.34	4.29
w/o Goal Relevance	4.00	3.97	4.83	4.32	4.28

Table 2: Ablation Study

Baselines. We compare our dual-agent hierarchical RL approach against several representative conversational systems. **Vanilla Llama3** (gra, 2024) is a straightforward *prompt-only* approach, querying Llama3-8B-Instruct with no hierarchical actions or appraisals, thereby gauging the off-the-shelf capabilities of an LLM on Supreme Court discourse. **SFT Llama3** (gra, 2024) fine-tunes the same base model using our dataset, testing whether domain-specific training alone meets inquisitive dialogue demands. We also include **SaulLM-7B** (Colombo et al., 2024) to assess how specialized LLMs perform when no hierarchical or appraisal mechanisms are present. For more structured pipeline approaches, **Hudeček et al.** (Hudeček and Dušek, 2023) integrates domain detection, belief-state tracking, and database querying for task-oriented dialogues, while **VaRMI** (Shea and Yu, 2023) employs offline policy gradient and importance sampling to maintain role consistency in RL-based CAs. **ArCHer** (Zhou et al., 2024) utilizes a hierarchical structure with an Actor-Critic framework for multi-turn, goal-oriented dialogues. We employ the offline variant. Further details on hyperparameters and implementation are available in Appendix A.

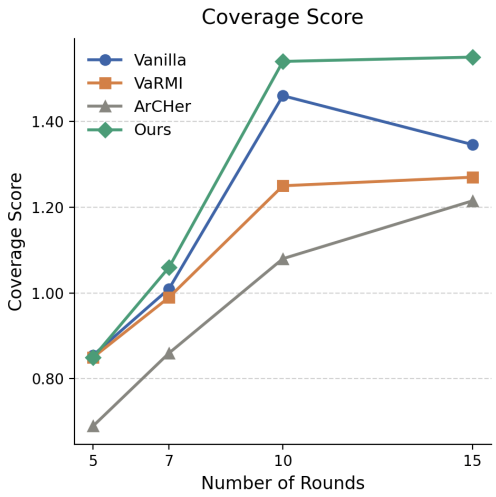


Figure 3: Coverage Score results

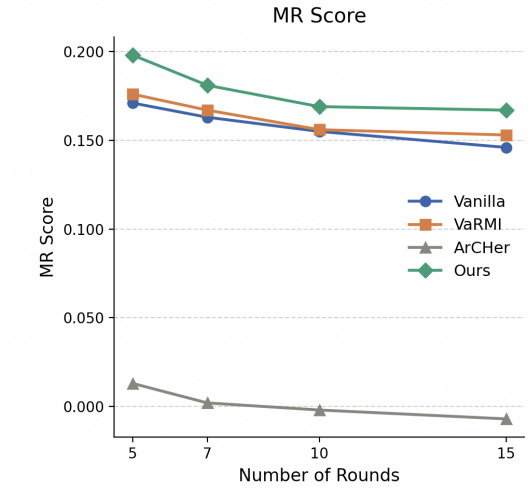


Figure 4: MR Score results

5.2 Main Results

In this section, we test our method and all baselines on the US Supreme Court dataset and compare their effectiveness in terms of the evaluation metrics. Detailed results are shown in Table 1. Our fine-tuning-free method achieves the best performance across all metrics, confirming that our dual agent method understands the goal of justice and its inquisitive nature well. The appraisal agent contributes to the PES metric the most, which is the metric where our method outperforms the baseline the most.

It is worth noting that although the US Supreme Court transcript is included in the training set of SaulLM-7B, it is still outperformed by generic models. The reasons for this phenomenon are twofold: first, the model wasn't trained for dialogue tasks; second, the task is substantially more challenging than the metrics used for SaulLM-7B.

The results of **Coverage Score** and **MR Score** are presented in Figure 3 and Figure 4. In Figure 3, our method consistently achieves the highest Coverage Score across all round settings, indicating that our dual-agent framework is more effective at expanding the discussion to cover a broader range of case-related topics. Figure 4 shows a similar pattern for MR Score: our method maintains the strongest marginal relevance throughout, suggesting that it is better able to introduce new information while remaining aligned with the central question of the case.

Due to the quality issue of the Surpeme Court dataset, the finetuning methods are not efficient on the dataset(see examples in Table 7). SFT and ArCHer achieve ideal results in CS and PES, however, their results were affected by the widespread presence of low-quality data, while our approach effectively bypasses low-quality snippets.

5.3 Ablation Study

We conducted four ablations to clarify the role of each reward component and the appraisal agent: (i) w/o the appraisal agent, (ii) w/o the succinct reward, (iii) w/o the novelty reward, and (iv) w/o the goal relevance reward.

Table 2 shows that each omission reduces at least one key metric relative to our full model, which yields the highest overall score (4.34), confirming that all components contribute to overall effectiveness. For example, removing the novelty reward reduces OS from 4.34 to 4.29, suggesting that without encouraging fresh information, the dialogue risks becoming less directional.

Figure 8a (130 epochs) and 8b (1600 epochs) plots the cumulative reward during offline RL. Early in training, the full model quickly surpasses ablations, reflecting the synergy of dual-agent oversight and the combination of all reward signals.

5.4 Human Evaluation

We conducted a human evaluation, giving annotators the metadata of each Supreme Court case along with its dialogue context. Evaluators scored the Conformity Score (CS), Progression Score (PS), Outcome Relevance Score (OS), and Probing Effectiveness Score (PES) on a 1–5 scale (Section 5.1). To ensure consistency, all methods were evaluated on the same set of case transcripts.

Table 8 presents the average ratings. Our full model achieves the highest overall score (4.53),

outperforming both SaulLM-7B (Colombo et al., 2024) and all ablated versions. This underscores the importance of every component in the agent in improving performance.

6 Conclusion

In this paper, we revisit the scope of TOD and propose a three-way categorization—collaborative, negotiation, and inquisitive dialogue—to better capture the diversity of goal-driven conversation. Our study centers on the inquisitive dialogue setting, using U.S. Supreme Court oral arguments as a representative domain.

We presented a dual-agent hierarchical RL approach for inquisitive conversation, focusing on U.S. Supreme Court oral arguments as a high-stakes domain. By integrating a Hierarchical Dialogue Agent that decomposes conversation control across multiple levels with an Appraisal Agent that proactively evaluates attorney responses, our framework captures the justice's goal-driven and probing style. We also present a regulation term that efficiently reduce the variance of our offline RL method. Empirical results on diverse Supreme Court cases show that the dual-agent design, coupled with carefully designed reward components yields more effective and context-aware dialogue strategies than multiple baselines.

While our current work centers on Supreme Court interactions, the underlying principles, such as active inquiry, structured dialogue management, and reward-driven question formulation, are broadly applicable to other high-stakes or domain-specific settings where deeper questioning is crucial. Future directions include expanding the reward model to capture even more nuanced legal strategies, and adapting the framework to other inquisitive domains such as investigative journalism or medical consultations.

7 Limitation

The simulated justice's responses of our agents are given by prompting LLM. Our agent's capability is heavily relies on capability of LLM. When LLMs have a very low probability of generating the desired optimal sequence, our method cannot reach optimal performance as well.

Although our work outperforms other baselines on the US Supreme Court dataset, the efficiency of our method on other legal domain dialogue datasets remains unclear. Our reward signals and action types are set for this dataset; for other datasets, they have to be redesigned. The policy of generating the US Supreme Court dataset is close to the optimal policy. When the dataset contains a large amount of data generated by a bad policy, our regularization term could be less efficient.

8 Ethical statement

This study uses publicly available transcripts and metadata from U.S. Supreme Court oral arguments, all of the original format of data can be download from Official website(Supreme Court of the United States). The Court releases transcripts as part of its routine transparency practices. These datasets do not reveal any identifiable information about the raters. We are not asking for any personal information during the labeler selection and labeling process. We do not include any personalized information in data processing. All of the examples used in prompting are randomly selected.

Acknowledgments

This research was supported by U.S. National Science Foundation grant number IIS-2336768. Any opinions, findings, conclusions, or recommendations expressed in this paper are of the authors and do not necessarily reflect those of the sponsor.

References

2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2020. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *arXiv preprint arXiv:1810.00278*.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. *Preprint*, arXiv:1909.05358.

J. Carbonell and J. Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, Melbourne, Australia. ACM.

Corinne Cichowicz. 2019. Oral argument tactics on the supreme court bench: A comparative analysis of verbal tools used by justices sotomayor, kagan, and gorsuch.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. Saullm-7b: A pioneering large language model for law. *Preprint*, arXiv:2403.03883.

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. Association for Computational Linguistics.

Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2024. Plug-and-play policy planner for large language model powered dialogue agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net.

Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-policy deep reinforcement learning without exploration. *Preprint*, arXiv:1812.02900.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Shasha Guo, Lizi Liao, Jing Zhang, Cuiping Li, and Hong Chen. 2024. PCQPR: proactive conversational question planning with reflection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pages 11266–11278. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Vojtěch Hudeček and Ondřej Dušek. 2023. Are llms all you need for task-oriented dialogue? *Preprint*, arXiv:2304.06556.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. *Preprint*, arXiv:2007.01282.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. *Preprint*, arXiv:2006.04779.

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *Preprint*, arXiv:1706.05125.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016b. Deep reinforcement learning for dialogue generation. *Preprint*, arXiv:1606.01541.

Lizi Liao, Grace Hui Yang, and Chirag Shah. Proactive conversational agents in the post-ChatGPT world. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3452–3455. ACM.

Na Liu, Liangyu Chen, Xiaoyu Tian, Wei Zou, Kaijiang Chen, and Ming Cui. 2024. From llm to conversational agent: A memory enhanced architecture with fine-tuning of large language models. *Preprint*, arXiv:2401.02777.

Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Rethinking and refining the distinct metric. *Preprint*, arXiv:2202.13587.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. *Preprint*, arXiv:1705.08039.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. 2025. On memory construction and retrieval for personalized conversational agents. *Preprint*, arXiv:2502.05589.

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, Kam-Fai Wong, and Shang-Yu Su. 2018. Deep dyna-q: Integrating planning for task-completion dialogue policy learning. *Preprint*, arXiv:1801.06176.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Preprint*, arXiv:1909.05855.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Mudita Sharma, Tony Russell-Rose, Lina Barakat, and Akitaka Matsuo. 2021. Building a legal dialogue system: Development process, challenges and opportunities. *Preprint*, arXiv:2109.00381.

Shikhar Sharma, Jing He, Kaheer Suleman, Hannes Schulz, and Philip Bachman. 2017. Natural language generation in dialogue using lexicalized and delexicalized data. *Preprint*, arXiv:1606.03632.

Ryan Shea and Zhou Yu. 2023. Building persona consistent dialogue agents with offline reinforcement learning. *Preprint*, arXiv:2310.10735.

Lei Shu, Piero Molino, Mahdi Namazifar, Hu Xu, Bing Liu, Huaixiu Zheng, and Gokhan Tur. 2019.

Flexibly-structured model for task-oriented dialogues. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, Stockholm, Sweden. Association for Computational Linguistics.

Shang-Yu Su, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Yun-Nung Chen. 2018. Discriminative deep Dyna-Q: Robust planning for dialogue policy learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3813–3823, Brussels, Belgium. Association for Computational Linguistics.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-task pre-training for plug-and-play task-oriented dialogue system. *Preprint*, arXiv:2109.14739.

Supreme Court of the United States. Official Website of the U.S. Supreme Court.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *Preprint*, arXiv:1409.3215.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric P. Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 5624–5634. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. *Preprint*, arXiv:2401.00368.

Weizhi Wang, Zhirui Zhang, Junliang Guo, Yinpei Dai, Boxing Chen, and Weihua Luo. 2022. Task-oriented dialogue system as natural language generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22. ACM.

Tong Zhang, Chen Huang, Yang Deng, Hongru Liang, Jia Liu, Zujie Wen, Wenqiang Lei, and Tat-Seng Chua. 2024. Strength lies in differences! improving strategy planning for non-collaborative dialogues via diversified user simulation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pages 424–444. Association for Computational Linguistics.

Tiancheng Zhao and Maxine Eskenazi. 2016. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. *Preprint*, arXiv:1606.02560.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020*, pages 4128–4139. International Committee on Computational Linguistics.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024. Archer: Training language model agents via hierarchical multi-turn rl. *Preprint*, arXiv:2402.19446.

A Implementation details

Implementation Details. We define the agent's **state** S at each time step as the dialogue context up to the current turn. We transform the components into a dense vector, $S = \text{Embed}(s_c, s_h)$, using a fine-tuned Mistral-7B model (Wang et al., 2024). Our embedding model produces 4096-dimensional vectors. In both the hierarchical dialogue agent and reward model, we compress these embeddings to 32 dimensions before concatenating them with appraisals and actions. This compression uses fully connected layers with batch normalization and Leaky ReLU activations.

During training, our Dual Hierarchical Dialogue Agent relies on ground-truth appraisal and three hierarchies from the dataset, so the appraisal agent and three hierarchies of dialogue agent can be trained simultaneously while loading the same dataset.

To stabilize training, we employ polyak updates of the target networks with $\tau = 0.005$, and empirically set the discount factor γ to 0.9. The weights for relevance, novelty, and succinctness rewards are set to 0.2, 0.7, and 0.1, respectively. We use exponential decay for learning rate of both agents; the learning rates for them are 1e-6 to 3e-9 and 1e-6 to 1e-8. The model size for both appraisal agent and dialogue policy agent are both less than 2M, the whole training time takes approximately 70 hours.

Our method, ablations, SFT Llama3, VaRMI, Hudeček's method, and the vanilla baseline, use Llama-3.1-8B-Instruct as their base model. For ArCHer, the actor uses Llama-3.2-1B-Instruct as the base model due to memory constraints, and the critic uses RoBERTa, in line with the implementation presented by the authors. We run SFT Llama3 via LLaMA-Factory (Zheng et al., 2024) on 2,000 examples from the Supreme Court dataset, training for three epochs with LoRA (Hu et al., 2022) at a per-device batch size of 4, a gradient accumulation step of 8, and a learning rate of $1e{-}4$. For VaRMI (Shea and Yu, 2023), we fine-tune with a $1e{-}6$ learning rate for one epoch. In Hudeček's

method (Hudeček and Dušek, 2023), the training dataset serves as the retrieval database, and cosine similarity on embedded contexts is used as the retrieval similarity function. For ArCHer, we train with a dataset of 700 dialogue trajectories, each with 6 to 7 dialogue turns, for 8 epochs, with an actor learning rate of $1\mathrm{e}-4$ and a critic learning rate of $1\mathrm{e}-5$.

Inference Phase. At inference, the Appraisal Agent is invoked first to generate an appraisal, which is subsequently fed into the dialogue agent. The dialogue agent first selects a top-level macro-action and then refines it through second- and third-level choices according to Table 3, stopping if the current sub-action has no successor. Formally, each level l solves:

$$a_l = \arg\max_{a_l} Q^{(l)}\big(s_{\text{aug}},\, a_0, \ldots, a_{l-1}, a_l\big). \tag{16}$$

The final action vector $\{a_0, a_1, a_2\}$ thus encodes a context-aware dialogue strategy, navigating the conversation at multiple levels of granularity .

A.1 Reward Model for Offline Evaluation

In this work, we use an offline RL setting, and the environment is not accessible for providing real-time feedback. We thus learn a *Reward Model* to approximate the environment's true reward function from the offline training data. Particularly, we employ a feed-forward neural network (FFN) to model R_ϕ and predict a scalar reward $\hat{r}$. The network takes data tuples of $(s, p, a, r) \sim \mathcal{D}$. A standard mean-squared error (MSE) loss is used here to measure the discrepancy between the predicted reward $\hat{r} = R_\phi\big(z,\, p,\, a\big)$ and the ground-truth r_i.

B Three-Level Taxonomy of Justice's Action

The hierarchies of dialogue actions are listed in Table 3. Actions in primary hierarchies are bolded. Actions in second and third hierarchies are listed in the left and right columns of the table, respectively.

Algorithm 1 Dual Hierarchical RL (Offline Training Mode with Regularization)

Input: dataset $\mathcal{D} \sim ((s,p),a,r,s')$
Output: Policies for dual agent θ_{App} and θ_{Dia}

Initialization: Build dataset $\{(s,p,r,s')\}$ for appraisal agent, $\{\,((s,p,a_0,\ldots,a_{l-1}),\ a_l,\ r,\ s')\ \mid\ l \in \{0,1,2\}\,\}$ for hierarchical agent, $\{(s,p,a,r)\}$ for reward model. Initialize policy and target networks Q^{θ}_{App}, Q_{Dia}, and reward model R_ϕ.

Train Reward Model:

for each training iteration **do**
  Sample mini-batch (s,p,a,r)
  Update R_ϕ by minimizing $\mathcal{L}_{RM} = \mathbb{E}_{(s,p,a,r)\sim\mathcal{D}}\left[(\hat{r}-r)^2\right]$
end for when R_ϕ converges

Train Appraisal Agent:

for each training iteration **do**
  Sample (s,p,r,s')
  Compute DDQN target: $Y = r + \gamma Q_{\text{App}}(s', \arg\max_{p'} Q_{\text{App}}(s',p';\theta);\theta^-)$
  Compute regularization terms: $R_1(s) = \max_{p'} Q_{\text{App}}(s,p'), \quad R_2(s) = Q_{\text{App}}(s,p)$ where $(s,p) \in \mathcal{D}$
  Update Q_{App} by minimizing: $\mathcal{L}_{\text{App}} = (Q_{\text{App}}(s,p) - y)^2 + \alpha(R_1(s) - R_2(s))$
end for when Q_{App} converges

Train Hierarchical Dialogue Agent:

for each training iteration **do**
  Sample transitions $\{\,((s,p,a_0,\ldots,a_{l-1}),a_l,r,s') \mid l \in \{0,1,2\}\,\}$ from $\mathcal{D}$
  Compute DDQN target: $y^l = r + \gamma\, Q_{\text{Dia}}(s', \arg\max_{p'} Q_{\text{Dia}}(s',p';\theta);\theta^-)$
  Compute conservative terms: $R_1(s_{\text{aug}}) = \max_{p'} Q_{\text{Dia}}(s_{\text{aug}},p'), \quad R_2(s_{\text{aug}}) = Q_{\text{Dia}}(s_{\text{aug}},p)$
  Compute $Q(s,a_0), \max_{a_1} Q(s,a_1), Q(s,a_1), \max_{a_2} Q(s,a_2)$ (depends one hierarchy depth)
  Update Q_{Dia} based on accumulate regularized loss: $\mathcal{L} = \sum_{i=1}^{3} \mathcal{L}^l$, where $\mathcal{L}^l = (Q_{\text{Dia}}(s_{\text{aug}},p) - y^l)^2 + \beta(R_1(s_{\text{aug}}) - R_2(s_{\text{aug}}))$
  Offline evaluation: $\hat{r} = \mathbb{E}_{(s,p,a,r)\sim\mathcal{D}} R_\phi(s,p^*,a_0^*,a_1^*,a_2^*)$
end for when $\mathcal{L}$ converges and $\hat{r}$ stabilizes

Question	
Clarification question	Clarify important aspect of the case
	Clarify legal arguments or issues
	Clarify definition of concept
Probing question	Probe the consistency between the attorney's arguments and established legal principles or precedents
	Probe the assumption underlying the attorney's arguments
Leading question	Ask for the attorney's position
	Lead the attorney toward a particular conclusion
	Lead the attorney to certain aspects
Make hypothesis	
Present hypothesis	Present hypothetical situations to test legal limits
	Present hypothetical situations to test legal issues in the case
Compare hypothesis	Compare to hypothetical situations to assess legal principles
	Highlight key differences from hypothetical situations
Conclude hypothesis	Explore different types of consequences
Declaration	
Confirmation	Acknowledge the attorney's arguments
	Prompt for information that would support the attorney's arguments
Rejection	Oppose the attorney's arguments
	Provide counterexample to challenge the attorney's arguments
Declaration (non-questions) for more details	Lead attorneys by examples (non-questions) for detailed explanation of a concept
Declaration with Time Pressure	Pressure a rash response from the attorney

Table 3: Proposed Hierarchy of Dialogue Acts

Appraisals	**Explanation**
Sense ambiguity	The justice finds the attorney's arguments ambiguous
Find deviates	The justice believes the conversation has strayed into irrelevant territory or unproductive arguments
Find redundancy	The justice finds the attorney's arguments repetitive or unproductive
Spot weakness	The justice spots a weakness in the attorney's arguments
Identify flaws	The justice identifies logical flaws in the attorney's arguments
Identify chances	The justice identifies chances to influence the attorney
Keep challenging	The justice wants to challenge the attorney from another aspect
Dive deeper	The Justice wants to dive deeper into the dialogue
Otherwise	All other kinds of the justice intents

Table 4: Appraisal Actions

Domain	**Turns/case**	**Words/utter**	**Words/case**	**#Cases**
Regulatory	192.6	47.7	9183.2	589
Civil Rights	206.0	44.0	9074.5	337
Criminal	200.6	45.0	9035.7	418
IP	173.4	52.5	9101.2	19
Commerce	199.3	46.4	9252.2	107
Labor	262.3	54.9	14407.3	101
Immigration	176.4	55.3	9762.3	16
Environment	262.3	54.9	14407.3	3
Others	200.4	47.6	9541.2	18
Total	198.6	45.9	9121.5	1608

Table 5: Supreme Court Dataset Statistics

C Details Regards Measurements

We have qualified students for making human evaluations; we collect our human evaluation results by distributing Google Forms. The estimated payment is 20$ per hour.

Component	Prompt
Instruction	You are a duteous, respectful, and honest AI justice assistant. You are given a clip of dialogue that happens on the US supreme court appeal case ending with the utterance of the justice, your job is provide analysis and score the last utterance of the justice in the dialogue from different aspect with the max score of 5. In each task, you are given the background, argued question of the case, conclusion of the case and the justice utterance with related dialogue context. Provide a snippet of analysis to analyze the role of the last sentence before giving out the score.
Metric Explanation	The metric is: {Metric}: {Explanation of metric} The scoring format should be: {Metric}: {Comment about {metric}} ?/5
Score Definition	The definition of scores are: Score 1/5 ({Descriptor}): {Explanation} Score 2/5 ({Descriptor}): {Explanation} Score 3/5 ({Descriptor}): {Explanation} Score 4/5 ({Descriptor}): {Explanation} Score 5/5 ({Descriptor}): {Explanation}

Table 6: Prompt Structure for LLM Evaluation

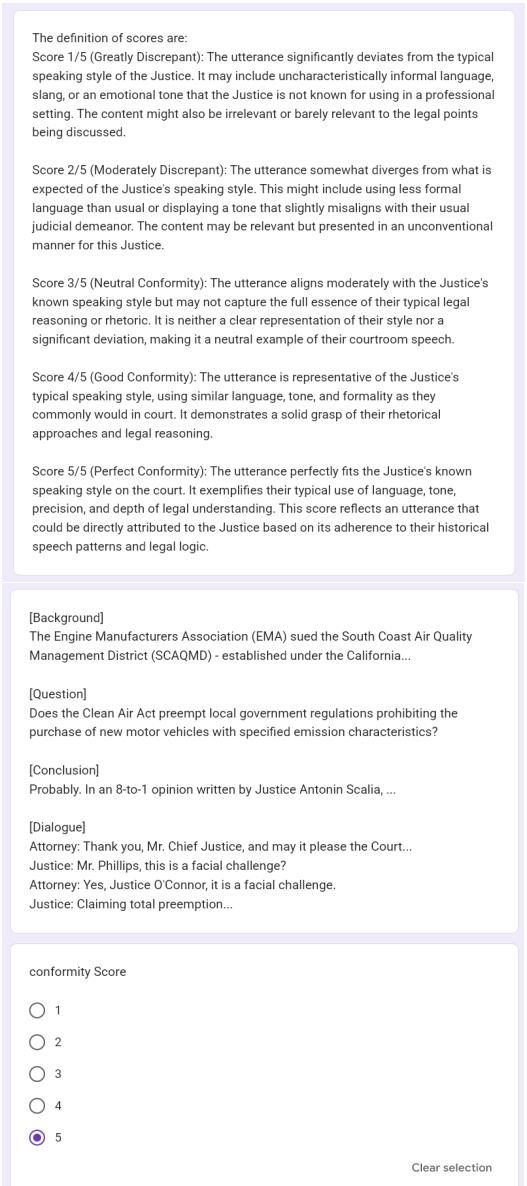


Figure 5: A template of google form for manual labeling, text has been streamlined for typographical purposes.

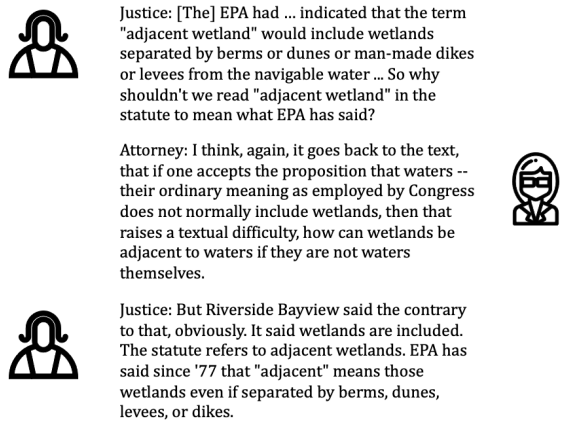


Figure 6: Justice uses a counterexample to challenge the attorney's position and the arguments presented previously

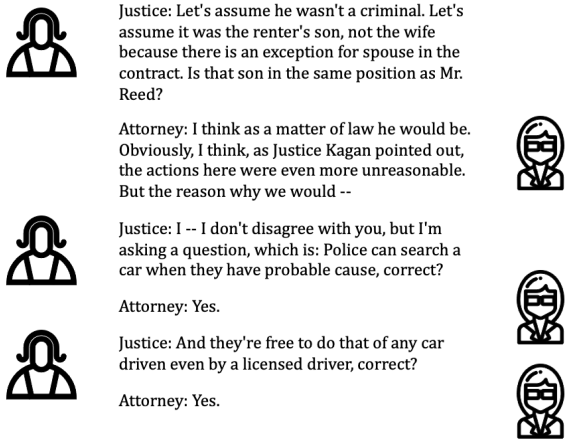


Figure 7: Justice continuously pressing attorney by making the rapid succession of her questions, cuts off attorney and then restricts him to one-word responses before another question is initiated.

Problem	Snippet
Sub-optimal utterance	Attorney: So I – so I don't think there's any statement in the legislative history that says we're not forcing employers to give benefits for non-work-related injuries. What – there are three statements in the legislative history that – that Respondent draws a negative inference from. Justice: I'm so relieved.
Frequently interception	Justice: So you really can't... there's no analytical distinction, then– Attorney: Well– Justice: –between the fact and the feeling. Attorney: –That's why we believe this should be a question for the district judge, who can balance all of these factors. In your hypothetical– Justice: Yes, but even on your balancing theory I thought the judge was supposed to draw... maybe I misunderstood you. I thought the judge was supposed to draw a line between fact and feeling, and what he was supposed to be balancing– Attorney: –No, I– Justice: –was the appropriateness of admitting the fact as against other interests. Attorney: –I think that's one of the things that the trial judge could be balancing, whether it's fact or feeling, but also the need for the evidence. If we had a hypothetical where the– Justice: I don't understand that, the need for the evidence?
Missing data	Attorney: That's correct. It may be applied in the discretion of the agency head... Justice: (Inaudible) Attorney: Yes, sir, I think there is a substantial difference and I think that's ...

Table 7: Examples of Low-quality Snippets of Dataset

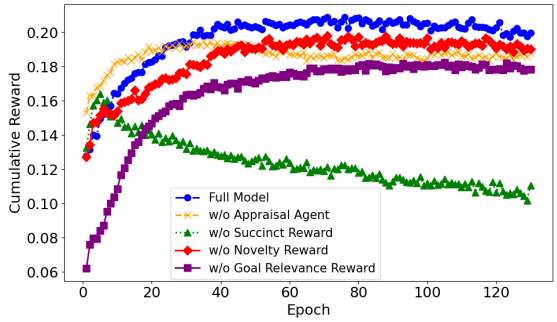
D Transferability Discussion

Here, we provide a discussion of how the framework could be adapted to other inquisitive domains.

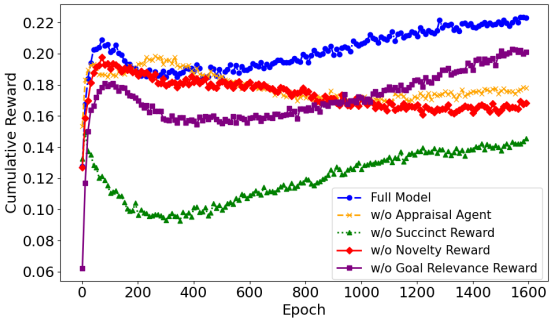
Inquisitive conversations usually have an ultimate result. For the Supreme Court, it is a conclusion; for journalism, it is a summary; and for medical interviews, it can be highlights.

Appraisal taxonomy: To adapt to other domains, we can keep the same turn-level appraisal mechanism, but broaden the label space to a domain-general core plus domain-specific refinements. For journalism, refinements emphasize attribution and verifiability (e.g., "claim lacks source," "timeline inconsistent," "needs evidence/documents"). For medicine, refinements emphasize clinical completeness and safety (e.g., "missing onset/duration/severity," "red-flag unaddressed," "contraindication risk"). Practically, the domain-specific appraisal set can be selected from a larger universal pool, or induced with weak supervision/clustering over (question, answer, follow-up) triples—reducing reliance on handcrafted legal notions while preserving the same control interface to the dialogue policy.

Dialogue act hierarchy: The hierarchical decision structure can be broadened to a general property of inquisitive interviewing. Adaptation does not require redesigning the hierarchy; it requires swapping the act inventory. In journalism, high-level acts like clarify, verify, challenge, request evidence, reconcile contradictions, summarize leads naturally decompose into finer acts (e.g., "ask for document," "ask for source identity," "pin down time/place").



(a) First 130 epochs



(b) First 1600 epochs

Figure 8: Learning Curves from Ablation Study. (a) Cumulative reward during early training stage; (b) Cumulative reward extends to longer term training (1600 epochs). The full model (blue) outperforms all ablated versions.

	CS	PS	OS	PES	Overall
Full Model	**3.99**	**4.53**	4.32	**4.63**	**4.37**
w/o Appraisal Agent	3.83	3.89	**4.42**	3.89	4.01
w/o Succinct Reward	3.45	4.05	4.11	4.05	3.92
w/o Novelty Reward	3.74	4.32	4.26	4.37	4.17
w/o Goal Relevance	3.77	4.26	4.21	4.42	4.17
SaulLM-7B	3.73	3.21	4.05	3	3.5

Table 8: Human Evaluation

Rather than hand-crafting these for each new domain, a practical adaptation path is to derive the hierarchy via hierarchical clustering of question/response embeddings or other cues, then map clusters to interpretable parent nodes while letting leaves remain domain-specific.

Reward design: The reward template also generalizes with a simple substitution: replace the Supreme Court "case conclusion" target with a domain "goal artifact" that represents the interview's intended end product. For journalism, this can be a set of story claims/summarization of a dialogue. For medical interviews, this can be a set of highlights of it.

Generally speaking, goal-relevance then rewards answers that add content aligned with these goal elements, novelty rewards information that was not already established earlier in the conversation, and clarity is recalibrated per domain (journalism: specificity and attribution/evidence presence). Crucially, the training objective and reward combination remain unchanged—only the goal artifact and clarity proxy are swapped in this way, adaptation can be interpreted as a "plug-in" process rather than a substantial redesign.