Reward-on-the-Line: A Novel Offline Reinforcement Learning Method for Building Legal Conversational Agents

Xubo Lin[1], **Mingze Wang**[1], **Grace Hui Yang**[1], **Daniel Chen**[2]
[1]Georgetown University, USA
[2]Université Toulouse Capitole, France
{xl524, mw1222, grace.yang}@georgetown.edu, daniel.chen@iast.fr

Abstract

Offline reinforcement learning (RL) offers a promising path for training domain-specific conversational agents (CAs) using large-scale historical dialogue data, without the need for costly online interactions or human annotations. In the legal domain, vast amounts of publicly available courtroom transcripts provide a rich and underutilized resource for developing intelligent legal CAs. However, offline training suffers from distribution shift between the learned policy and the behavior policy embedded in the training data, which can degrade agent performance at deployment. We address this challenge with a novel offline RL method, *Reward-on-the-Line* (ROL), which calibrates rewards based on action-selection agreement among an ensemble of CAs. We apply ROL to the *U.S. Supreme Court* dataset to demonstrate its effectiveness in learning proactive, legally-informed dialogue strategies from historical court proceedings. To show the broader applicability of our approach, we also evaluate ROL on the *CraigslistBargain* negotiation dataset. Results in both domains confirm that ROL reduces distribution shift and improves agent performance in unseen dialogue scenarios.

Introduction

Building domain-specific conversational agents (CAs) through offline reinforcement learning (RL) (Levine et al. 2020; Jiang, Xu, and Liang 2017; Mbuwir et al. 2017; Verma et al. 2022; Tang and Yang 2020) has become an appealing approach. Rather than relying on costly live interactions or on-demand labeling, offline RL leverages previously collected data, such as transcripts from town hall meetings or courtroom dialogues. This paradigm can drastically reduce annotation requirements and exploit extensive historical conversational data that might otherwise remain underutilized.

However, deploying an offline RL–trained agent in real-world scenarios presents a major obstacle: *distribution shift*. The policy or value function learned from historical data may not align well with actual user queries or the linguistic norms of a new environment. For example, a customer service bot trained on British English dialogues might fail to interpret American slang, which illustrates a shift in language conventions between the training and test domains; or a legal CA trained on decades-old Supreme Court transcripts might use archaic phrasing that no longer reflects modern legal practice. Researchers have proposed methods such as importance sampling (Munos et al. 2016), policy-constrained offline approaches (Nair et al. 2018; Fujimoto, Meger, and Precup 2018), and Q-value penalization (Kumar et al. 2020; Verma et al. 2022) to better align the learned policy with unseen data. Yet, these techniques often demand extensive hyperparameter tuning, potentially yielding over-optimistic or over-pessimistic Q-value estimates that lead to trivial or excessively verbose responses. For instance, an over-pessimistic system might persistently offer unhelpful one-liners such as "I can't assist with that," regardless of the query. Conversely, an over-optimistic system might generate lengthy, multi-step instructions, even for simple user questions, resulting in unnecessarily cumbersome conversations.

In this paper, we propose a novel offline RL method, *Reward-on-the-Line* (ROL), which tackles distribution shift by using agreements among an ensemble of CAs to calibrate rewards in the training environment. Historically, *ensemble techniques* (e.g., bootstrap DQN (Osband et al. 2016)) introduced the notion of training multiple models to reduce overestimation bias and capture uncertainty. More recently, *Agreement-on-the-Line* (Baek et al. 2022) showed that if multiple models coincide on in-distribution data, they tend to preserve a linear correlation in their agreement levels under distribution shifts. We leverage and adapt these insights specifically for conversational agents, which grapple with domain shifts and costly reward labeling. Concretely, our method measures how frequently two or more Q-networks select the same action in the in-distribution (ID) domain and exploits the observed linear correlation in out-of-distribution (OOD) domains, even in the absence of labeled rewards. We use this correlation to *calibrate* the training rewards, effectively shifting the policy to match the test environment and minimizing the need for labor-intensive hyperparameter searches or new annotations. In conversational settings, this is particularly beneficial, as labeling dialogues in each new domain can be prohibitively expensive.

In summary, our main contributions are:

1. We introduce **Reward-on-the-Line (ROL)**, a novel offline reinforcement learning method that calibrates rewards based on ensemble agreement, enabling more robust policy learning under distribution shifts.

2. We focus on the legal domain and demonstrate how ROL can effectively leverage large-scale, publicly available courtroom transcripts to train legal conversational agents, without relying on costly annotations or live interactions.
3. To demonstrate the generality of our approach beyond legal applications, we also apply ROL to a negotiation dataset, showing that the method generalizes well across domains while maintaining strong performance in unseen scenarios.

Related Work

Conversational Agents

Classical Approaches. Historically, four paradigms have shaped conversational AI systems research: *sequence-to-sequence* (Seq2Seq) supervised generation (Vinyals and Le 2015), *retrieval-assisted* text generation (Lewis et al. 2020; Tang and Yang 2022), *knowledge graph*-driven question answering (KG-QA) (Zheng et al. 2018; Deng et al. 2025), and *RL*-based dialogue policies (Schulman et al. 2017). Early Seq2Seq models (Sutskever, Vinyals, and Le 2014) provided the foundation for many open-domain chatbots, while retrieval-assisted methods grounded responses in factual content to mitigate hallucination (Zhang et al. 2023). Knowledge graph-based systems (Chakraborty et al. 2019) tackled fact-oriented queries, and RL approaches (Kandasamy et al. 2017; Liu, Pan, and Luo 2020; Dhariwal et al. 2017; Tang, Kulkarni, and Yang 2021; Cai et al. 2024) aimed to optimize long-term utility in conversations.

Emergence of Large Language Models. The landscape of conversational AI has evolved substantially in recent years, propelled by advances in large language models (LLMs) such as GPT-3.5, GPT-4 (OpenAI 2023), Bard, Claude, and Llama (Touvron et al. 2023). These models often rely on Reinforcement Learning from Human Feedback (RLHF) (Christiano et al. 2023) to refine generation quality via carefully curated preference data. Although these models often produce more coherent and context-sensitive outputs than older supervised approaches, challenges, such as hallucination, bias, and alignment, remain focal points of current research (Bai et al. 2022). Retrieval-augmented generation and plugin interfaces further aid in grounding responses with external knowledge sources.

Unlike methods that rely heavily on large-scale human feedback for shaping reward models, we adopt an offline RL perspective to leverage historical conversational data without incurring the high cost of continual annotations. Specifically, we introduce a method based on ensemble agreements, Reward-on-the-Line, to adapt to domain shifts and reduce over-reliance on hyperparameter tuning.

Offline Reinforcement Learning

Offline RL (Levine et al. 2020) offers an feasible solution when direct human supervision or real-time interaction is unavailable. By learning from static, pre-collected data, offline RL methods circumvent many practical barriers of online exploration, but they also face the distribution shift problem: the target environment may differ substantially from the training distribution (Fujimoto, Meger, and Precup 2019; Robey, Hassani, and Pappas 2020). Methods like importance sampling (Munos et al. 2016), policy regulation (Munos et al. 2016; Everitt et al. 2017), and Q-value penalization (Xu, Zhan, and Zhu 2022; Kumar et al. 2020; Verma et al. 2022) have attempted to address these shifts, often at the cost of extensive hyperparameter tuning.

In conversational tasks, defining suitable rewards for agents trained by offline RL is another core challenge. Some approaches rely on automatic metrics (e.g., BLEU) (Zhou et al. 2017; Liu, Pan, and Luo 2020), while others employ or adapt RLHF (Christiano et al. 2023; Stiennon et al. 2022) or implicit reward functions (DPO) (Rafailov et al. 2023; Xu et al. 2024). Despite these efforts, ensuring robust performance on unseen dialogue scenarios, particularly those out of the training domain, remains an open problem.

ID vs. OOD Performance Correlation

Another body of work investigates correlations between classifiers' performance on in-distribution (ID) data and out-of-distribution (OOD) data. Miller et al. (Miller et al. 2021) showed how a fitted line's slope and intercept can quantify distribution shift in classification models; Nakkiran et al. (Nakkiran and Bansal 2020) and Jiang et al. (Jiang et al. 2022) investigated how *agreement* between neural networks can be used to gauge performance on unlabeled OOD data. Subsequent works (Chen et al. 2023; Baek et al. 2022) applied these insights to estimate model performance or reliability under distributional changes. Building on this finding, our method calibrates rewards by measuring ensemble agreements on ID and OOD data, thus mitigating the adverse effects of distribution shift for conversational agents in an offline RL setting.

The Reward-on-the-Line (ROL) Method

Offline RL techniques for conversational agents often face severe distribution shifts, where the data used for training differ substantially from real-world usage scenarios. Prior solutions adjust the agent's mathematical model (policy or value function) and can require heavy hyperparameter tuning, risking over- or underestimation of Q-values. In practical conversational systems, these errors can lead to poor or inconsistent responses. Meanwhile, collecting labeled data (e.g., conversation outcomes or user ratings) for every new domain is expensive, but obtaining unlabeled transcripts is often feasible.

In this section, we detail the proposed ROL method, which extends the *Agreement-on-the-Line* idea Agreementontheline to an offline RL setting for conversations. The core observation is that if multiple Q-networks often choose the same actions on one domain (the training set), they maintain a roughly *linear correlation* in their agreement levels another domain (the target set). When the models within an ensemble are highly consistent in choosing an action (or rating an utterance) in the training set, that typically indicates they share a similar notion of what is "good" for that context. This means that if we know how often two agents agree on training data (where the rewards are known), we can make a reasonable estimate of how often they will also agree in unlabeled test data. Moreover, because measuring "agreement"

only requires that the agents produce the same (or different) actions, we don't need the exact rewards at test time. This is especially useful for conversational agents, where obtaining fully labeled conversations in each new domain can be labor-intensive. In this paper, our method exploits this linearity and calibrate training rewards to improve the agent's performance in unseen scenarios.

Problem Setup

We consider an offline RL framework tailored for conversational agents. Each *state* s comprises:

$$s = (s_c, s_h),$$

where s_c is the conversation's context (e.g., domain-specific metadata or product details) and s_h is the dialog history (all utterances so far).

An *action* a corresponds to the next utterance from the role that our agent simulates. Depending on the application, actions (utterances) may also include additional attributes such as proposed prices or other relevant parameters. We denote the set of possible candidate utterances in a given state $s = (s_c, s_h)$ as

$$\Omega(s) = \{a \mid a \sim \text{LLM}(\text{concat}(s_h, s_c))\},$$

where an external generative model (e.g., an LLM) can be used to propose utterances for comparison or exploration.

We define a *reward* function $r(s, a)$ that provides feedback to the agent after it selects an utterance a in state s. In an offline RL setting, the agent's objective is to learn a value function $Q(s, a)$ or an optimal policy $\pi(a|s)$ that maximizes expected return $E\Big[\sum_{t=0}^{T-1} \gamma^t\, r(s_t, a_t)\Big]$. Here *offline* means we have only previously collected data rather than interactive access to the environment.

The Core ROL Method

Action Selection Agreement Among Q-Networks. In the conversational setting, each utterance a is evaluated by a Q-network $Q(s, a)$, which ideally ranks better (i.e., more rewarding) utterances higher. When we have an ensemble of Q-networks (e.g., trained with different initial seeds), we can measure how frequently these networks *agree* on the top-ranked action in each state s. Formally, two networks Q_i and Q_j *agree* on s if

$$\arg\max_a Q_i(s, a) \;==\; \arg\max_a Q_j(s, a).$$

Aggregating over a dataset D yields

$$Agr(Q_i, Q_j) \;=\; \frac{1}{|D|} \sum_{(s,a)\,\in\, D} \mathbf{1}\Big(\arg\max_a Q_i(s, a) \;==\; \arg\max_a Q_j(s, a)\Big). \tag{1}$$

Because this computation merely checks whether two networks select the same best action, it does not rely on labeled rewards in that dataset.

Linear Correlation Between ID and OOD Data. Let Agr_{ID} be the ensemble agreement on action selection for the *in-distribution* data (D_{ID}), where the agent was originally trained (and rewards are known), and let Agr_{OOD} be the agreement on *out-of-distribution* data (D_{OOD}), which may have no labels. Empirical findings Agreementontheline show that:

$$Agr_{OOD} \;\approx\; k \cdot Agr_{ID} \;+\; b, \tag{2}$$

where k (slope) and b (intercept) capture how ID agreement translates to OOD agreement. By fitting (k, b) via least squares on the measurable portion of OOD data (potentially unlabeled in terms of rewards but used for action-level agreement), we can calibrate the ID rewards.

Reward Calibration. Although the linear parameters (k, b) arise from measuring how different Q-networks agree on action selections, we apply these parameters directly to the training rewards rather than to the Q-values themselves. In practice, this step relies on the observation that in our offline RL setting, the Q-function is closely tied to the immediate (or near-immediate) reward for each utterance. Because the dialogue horizon is finite and dominated by reward signals at each turn, the Q-values for a single state-action pair (s, a) typically reflect only a small sum (or scalar multiple) of immediate rewards. Consequently, scaling $Q(s, a)$ by k often has an effect similar to scaling $r(s, a)$ by k. Likewise, the intercept b can be seen as a per-episode (or per-utterance) offset that meaningfully shifts the final payoff.

Formally, if

$$Q(s, a) \;\approx\; \alpha\, r(s, a) \;+\; \delta,$$

for some constants α and δ (for instance, due to a bounded horizon or because future returns scale uniformly across actions), then calibrating Q by factors (k, b) is nearly equivalent to calibrating r by those same factors. Once we solve

$$(k, b) \;=\; \arg\min_{k,b} \sum \Big(Agr_{OOD} - k\, Agr_{ID} - b\Big)^2, \tag{3}$$

we therefore update each ID reward $r_{ID}(s, a)$ as:

$$r_{ROL}(s, a) = \begin{cases} k \cdot r_{ID}(s, a), & \text{if } a \text{ is not the last action,} \\ k \cdot r_{ID}(s, a) + b, & \text{if } a \text{ is the last action.} \end{cases} \tag{4}$$

Here, k rescales how "valuable" each training example appears under OOD conditions, and b offsets the reward, particularly relevant for final utterances in a conversation, which can carry a larger payoff (e.g., a concluding decision). Our approach provides a practical solution: rather than modifying each Q-network directly, we rescale the training rewards to ensure that the learned policy aligns more closely with OOD conditions.

ROL Implementation Details

Figure 1 illustrates our *Reward-on-the-Line* pipeline. It consists of six steps:

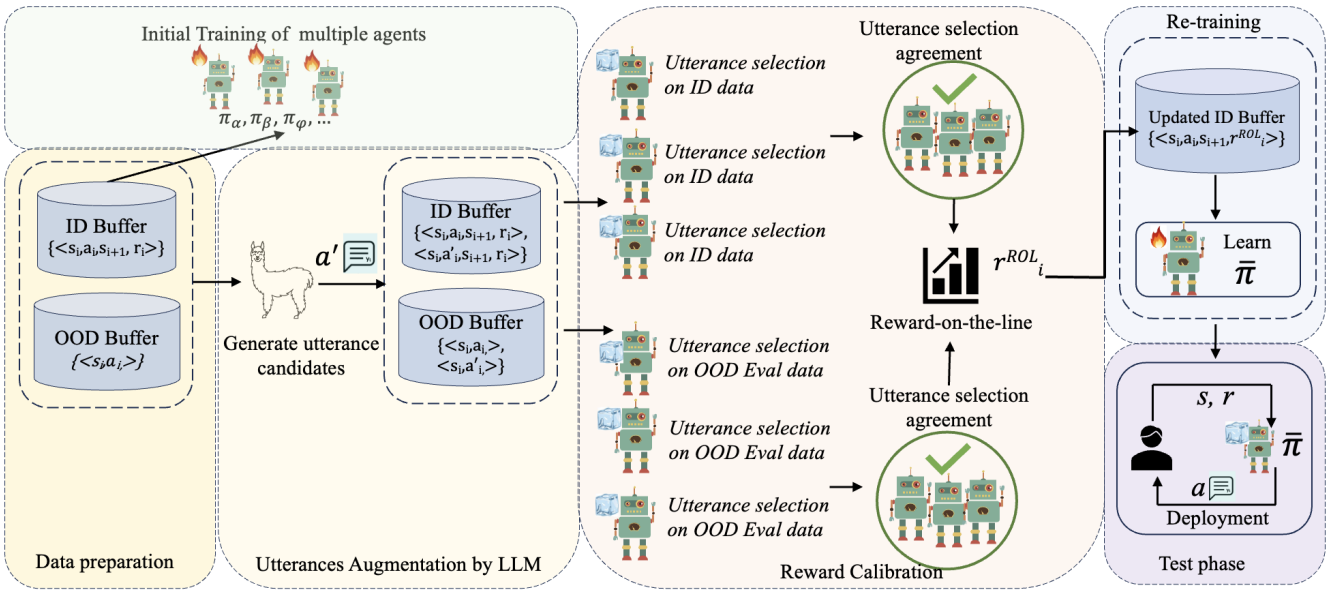

Figure 1: The Reward-on-the-Line (ROL) Method. It contains six steps, namely (1) data preparation, (2) initial training to create multiple agents, (3) utterance candidates generation by LLM, (4) calibrating rewards based on agents' agreements over utterance selection, (5) re-training of the agent, and (6) deployment.

(1) Data Preparation. We gather conversation transcripts and split the dataset into multiple domains (see Table 2 and Table 1). For each domain D_{ID}, the remaining domains serve as out-of-distribution data D_{OOD}. Although D_{OOD} labels exist when that domain acts as its own ID, we hold them out for the OOD condition. The ID dataset has known rewards (e.g., a final negotiation outcome or legal decision), while OOD data typically lacks labels.

(2) Initial Training (Ensemble of Q-Networks). Using an offline Double DQN (DDQN) hessel2018double,DQN, we train multiple Q-networks on $D_{train} \subset D_{ID}$. Each network starts from a different initialization but shares the same architecture. The result is an ensemble of conversational agents that can each evaluate utterances via $Q(s, a)$.

(3) Generating Additional Utterance Candidates. In offline RL, each state s typically has only one historically chosen utterance a, here selected from *successful* conversations (e.g., those reaching a low neogociated price or yielding a decisive court ruling). To enrich the comparison space, we use Llama3 touvron2023llama to generate additional candidate utterances. These candidates are then inserted into both the ID and OOD buffers, allowing the Q-networks to learn more diverse preference rankings.

(4) Measuring Agreement and Calibrating Rewards. We measure Agr_{ID} on D_{train} and Agr_{OOD} on the unlabeled $D_{val} \subset D_{OOD}$. By fitting a line $Agr_{OOD} = k \cdot Agr_{ID} + b$, we obtain calibration parameters k and b. We then replace each ID reward $r_{ID}(s, a)$ with $r_{ROL}(s, a)$ (Eq. 12) to align the agent's policy closer to OOD conditions.

(5) Re-training with Calibrated Rewards. We re-run the offline RL training procedure (DDQN) using r_{ROL} instead of r_{ID}. States and actions are embedded using BERT devlin2019bert, and we adopt a 3-layer feed-forward Q-network for scoring each utterance. Empirically, we use $\gamma = 1$ (no discounting) that suits dialogues where final utterances can have major impact (e.g., a concluded sale price). The learning rate α is set to 1, though it can be tuned if the domain's reward structure is noisier or more variable.

(6) Deployment. Finally, we deploy the agent to a test dataset, whether in the same domain or out-of-distribution. Because the training reward structure is rescaled to match potential OOD conditions, the agent handles distribution shifts more robustly, yet still performs well on ID data without additional labels.

Experimental Setup

Datasets

We evaluate our method on two publicly available datasets. Our primary focus is the *U.S. Supreme Court* dataset, which contains extensive public courtroom transcripts that offer a rich and underutilized resource for training legal conversational agents. This dataset enables us to explore the development of intelligent dialogue systems tailored for legal discourse. To demonstrate the generality and adaptability of our offline RL approach beyond the legal domain, we also include experiments on the *CraigslistBargain* negotiation dataset.

U.S. Supreme Court Dataset. Our work focuses on a subset of appellate cases from www.Oyez.org (spanning 2010–2020). Each case in our subset includes a *context* (background facts, parties' positions, and metadata such

	Turns/case		Words/utter		Words/case		#Cases
Domain	avg	max	avg	max	avg	max	
Regulatory	192.6	517	47.7	881	9183.2	24935	589
Civil Rights	206.0	364	44.0	742	9074.5	19505	337
Criminal	200.6	409	45.0	721	9035.7	15410	418
IP	173.4	256	52.5	583	9101.2	10371	19
Commerce	199.3	365	46.4	841	9252.2	16675	107
Labor	262.3	349	54.9	476	14407.3	17687	101
Immigration	176.4	307	55.3	535	9762.3	17443	16
Environment	262.3	349	54.9	476	14407.3	17687	3
Others	200.4	391	47.6	475	9541.2	12079	18
Overall	198.6	517	45.9	881	9121.5	24935	1608

Table 1: Supreme Court Dataset Stats.

	Turns/case		Words/utter		Words/case		#Cases
Domain	avg	max	avg	max	avg	max	
Phone	8.1	22	13.2	103	106.6	367	570
Bike	8.2	19	13.6	102	111.5	380	994
Housing	8.9	44	14.5	103	128.7	479	1137
Furniture	8.0	28	13.4	107	107.5	466	1477
Car	9.2	27	14.0	117	129.0	459	726
Electronics	8.1	18	12.6	88	101.6	322	725
Overall	8.4	44	13.6	117	114.4	479	5629

Table 2: CraigslistBargain Dataset Stats.

as the final ruling) and a *transcript* (the complete oral argument). For example, one case addresses whether *"the Bankruptcy Code's automatic stay provision ... requires an entity that is passively retaining possession of property ... to return that property immediately upon the filing of the bankruptcy petition."* In that discussion, the City of Chicago had impounded the petitioner's vehicle, and Justices pressed counsel on how quickly such property must be restored. As shown in Table 1, these appeal cases come from diverse domains such as Regulatory, Civil Rights, and Criminal, reflecting the breadth of issues heard by the Court.

CraigslistBargain Dataset. We use the negotiation dialogues collected in (He et al. 2018), where buyers and sellers discuss second-hand products on www.Craigslist.com. Each entry contains both the conversation transcript and metadata like product name, description, and the respective price expectations of the buyer and seller. For instance, one dialogue might feature *a phone listed at $200 and a buyer initially offering $140, ultimately converging on a mutually agreeable price $160* after runs of bids and counteroffers. As shown in Table 2, these dialogues encompass multiple of price negotiations, e.g., Electronics, Furniture, Vehicles.

Data Splits for Both Datasets. After gathering the conversation transcripts, we further divide each dataset into domains (see Tables 1 and 2). For both datasets, each of their domains serves as an *in-distribution* dataset (D_{ID}), while the remaining domains act as *out-of-distribution* data (D_{OOD}). Although the conversations in D_{OOD} have labels when considered as their own D_{ID}, we withhold those labels for the OOD condition. The agents thus sees reward labels only for D_{ID}, ensuring we fairly evaluate its performance on unseen domains.

Dataset-Specific Reward Functions

We use domain-specific rewards for our two datasets:

US Supreme Court Rewards For this dataset, our agent role-plays the *justice*. Each justice utterance $U_{\text{jus}}(t)$ receives a reward comprising five components: (i) relevance to the final decision, (ii) relevance to cited laws, (iii) relevance to the case's focal object, (iv) the attorney's response polarity, and (v) the justice's subsequent utterance's continued focus on that focal object. Intuitively, if the attorney's utterance is highly polarized, it suggests the justice's question was clear and on-point, eliciting a definitive response. Moreover, if the justice's next utterance $(t+1)$ remains aligned with the case object, it indicates the current justice utterance kept the conversation on track. Formally, for a justice's utterance $U_{\text{jus}}(t)$:

$$\begin{aligned} r_{\text{jus}}(t) \;=\; & \alpha_d \,\text{sim}\big(U_{\text{jus}}(t), C_{\text{dec}}\big) + \alpha_l \,\text{sim}\big(U_{\text{jus}}(t), C_{\text{law}}\big) \\ & + \alpha_o \,\text{sim}\big(U_{\text{jus}}(t), C_{\text{obj}}\big) + \alpha_p \,\text{pol}\big(U_{\text{att}}(t)\big) \\ & + \alpha_{\text{next}} \,\text{sim}\big(U_{\text{jus}}(t+1), C_{\text{obj}}\big), \end{aligned} \tag{5}$$

where $U_{\text{jus}}(t)$ is the justice's utterance at time t, $U_{\text{att}}(t)$ is the attorney's utterance at time t, $C_{\text{dec}}, C_{\text{law}}, C_{\text{obj}}$ represent the decision, laws, and object, $\text{sim}(\cdot,\cdot)$ is an embedding-based similarity function, $\text{pol}(\cdot)$ measures sentiment polarity VADER (Hutto and Gilbert 2014), and $\alpha_p, \alpha_d, \alpha_l, \alpha_o, \alpha_{\text{next}}$ are weighting coefficients.

CraigslistBargain Rewards Here, our agent role-plays the *buyer*. A negotiation dialogue concludes either with an agreed-upon transaction price P_t or a "quit" (no deal). For each buyer utterance, the reward has two components: (i) a *buyer utility* term, Util_{buy}, capturing the benefit when a deal is made, and (ii) the sentiment polarity of the seller's next utterance, $U_{\text{sell}}(t+1)$. Because Util_{buy} varies numerically across dialogues, we normalize it by the dialogue length L_{dlg}:

$$r_{\text{buy}}(t) = \frac{\text{Util}_{\text{buy}} \;+\; \text{pol}\big(U_{\text{sell}}(t+1)\big)}{L_{\text{dlg}}}, \tag{6}$$

where $\text{Util}_{\text{buy}} = \frac{P_s - P_t}{P_s - P_b}$, P_b and P_s are the buyer's and seller's target prices, $\text{pol}(\cdot)$ measures sentiment polarity, and L_{dlg} is the conversation's total turn count. If negotiation fails, we set $\text{Util}_{\text{buy}} = 0$. The seller's utility is $\text{Util}_{\text{sell}} = 1 - \text{Util}_{\text{buy}}$.

Evaluation Metrics

We adopt an evaluation paradigm similar to CHAI (Verma et al. 2022), generating multiple candidate utterances per state with Llama 3 and having our trained agents rank them by Q-value. For the *Supreme Court* dataset, we use the Justices' actual utterances as reference responses, discarding

filler or procedural remarks (e.g., "The court will be in recess") that offer no substantive legal content. In *Craigslist-Bargain*, we select utterances from successful cases where the final transaction price closely matches the buyer's target. Our experiment uses the following metrics, each capturing a unique aspect of response quality:

(1) Top Selection (TS) We evaluate whether each agent's single highest-ranked utterance closely aligns with the reference, using their longest common subsequence rather than exact sentence matching. This metric reflects how well the agent's very first choice matches with the reference content.

(2) ROUGE-2 ROUGE-2 $= \frac{2\times P_2 \times R_2}{P_2+R_2}$, where P_2 and R_2 are the precision and recall of the bigram overlap between generated and reference utterances. ROUGE-2 is an F1-like measure balancing both precision and recall—allowing partial matches to be credited for overlapping content.

(3) Conformity Score (CS) In multi-turn dialogues, a good conversation goes beyond surface-level matching, encompassing style and argument flow. We prompt an LLM (ChatGPT o1) to compare the utterances of ROL and a baseline side by side on how well it adheres to the expected style (e.g., legal language or negotiation tone). CS is the ratio of "ROL wins" to "ROL losses" across all pairwise comparisons.

(4) Progression Score (PS) evaluates whether u_i effectively advances the discussion or negotiation forward rather than stalling or digressing. PS is the ratio of ROL's "wins" to "losses," based on an LLM's side-by-side comparisons of ROL and baseline utterances.

(5) Goal Relevance Score (GS) checks whether u_i remains relevant to the ultimate outcome of the conversation (e.g., a court decision or a successful deal). We prompt the LLM to compare ROL's utterance against the baseline, and GS is the ratio of "ROL wins" to "ROL losses."

Baseline Systems for Comparison

We compare our method, ROL, with several established offline RL methods applied to dialogue tasks: Double-DQN (DDQN) (van Hasselt, Guez, and Silver 2016), Conservative Q-Learning (CQL) (Kumar et al. 2020), GoChat (Liu, Pan, and Luo 2020), and PPO (Dhariwal et al. 2017). Although PPO is on-policy, we adapt it for offline usage by training a reward model in a manner similar to RLHF (Christiano et al. 2023). In all experiments, each agent is trained on a single domain (D_{ID}) and tested on the remaining unseen domains (D_{OOD}), following the aforementioned data-split scheme, so that all agent's performance is evaluated on unseen environments.

To promote a fair comparison, we tune the hyperparameters of each baseline via small grid searches, selecting configurations that achieve stable performance on a validation set. For the U.S. Supreme Court dataset, we use a learning rate of 3×10^{-4}, a batch size of 128, and 5,000 training steps per round. For CraigslistBargain, we set the learning rate to 1×10^{-6}, keep the batch size at 128, and run 10,000 steps per round. Each baseline agent is trained on the same offline trajectories and uses the same feed-forward neural network architecture for its Q-value (or equivalent) component. All experiments are conducted on three NVIDIA 2080 Ti GPUs. Training each agent to convergence required approximately 30 hours on the U.S. Supreme Court dataset and 10 hours on CraigslistBargain. Note these resource requirements may increase for larger datasets.

	TS	ROUGE-2	CS	PS	GS
PPO	0.71	0.76	0.89	1.03	1.22
GoChat	0.73	0.69	0.78	1.63	1.06
DDQN	0.73	0.76	1.17	1	1.68
CQL	0.71	0.76	1.11	0.95	2.98
ROL (this paper)	0.75	0.89	-	-	-

Table 3: Results on US Supreme Court.

	TS	ROUGE-L	CS	PS	GS
PPO	0.67	0.77	1.05	1.08	0.68
GoChat	0.56	0.69	1.20	0.89	1.14
DDQN	0.66	0.78	1.00	0.98	1.04
CQL	0.70	0.79	1.08	0.95	1.21
ROL (this paper)	0.72	0.82	-	-	-

Table 4: Results on CraigslistBargain.

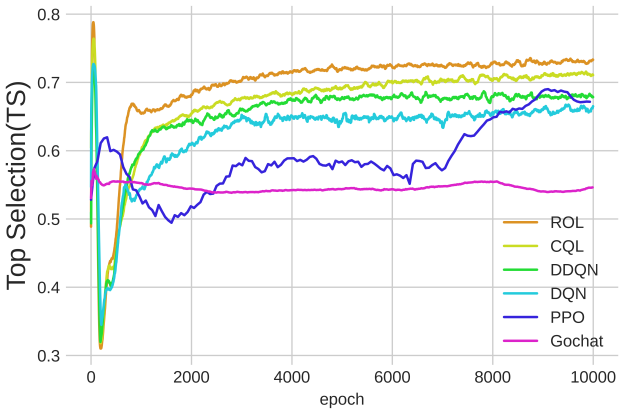

Figure 2: Learning curves on CraigslistBargain (10,000 epochs). ROL converges faster and outperforms all baselines.

Experimental Results

Main Results

Table 3 and Table 4 summarizes the main effectiveness results for our proposed method and baseline agents on the dateset of US Supreme Court and CraigslistBargain, respectively. As shown in the tables, our ROL approach consistently achieves the highest scores for nearly all metrics, often edging out the strongest baselines by noteworthy margins. On the *US Supreme Court* dataset, ROL pushes TS Score to 0.75 and ROUGE-L to 0.80, surpassing competitive approaches like CQL and DDQN. These gains are not trivial in specialized legal dialogues, where even a few

points can mean more precise, context-driven exchanges. For *CraigslistBargain*, ROL likewise leads with TS of 0.72 versus 0.70 from the nearest rival, and attains a ROUGE-L of 0.82, outperforming the 0.79 baseline. Moreover, ROL's style conformity, progressiveness in dialogue, and goal relevance also compare favorably to the strongest baselines. ROL maintains robust performance across a wide range of conversational metrics.

Beyond the aggregate results, our method, ROL, excels in certain domains. For instance, it surpasses PPO by nearly 9% on *Bike* and 5% on *Furniture* in OOD performance, and maintains a 6% advantage over CQL on *Electronics*. One likely reason is that these product types exhibit greater price variance or subtler negotiation cues, which challenge methods lacking explicit reward calibration. By mitigating domain shifts through reward adjustments, ROL more effectively captures nuances like price anchoring and stylistic differences.

DDQN is the backbone of our ROL approach, it has a large difference between the upper and lower bounds of its performance on OOD's metrics. Compared to DDQN, both ROL and DDQN's lower bounds are on cars, but it's 0.6 better than the DDQN.

PPO, on the other hand, made over-pessimistic measurements to both datasets. Although PPO showed the smallest ID OOD gap, it got low TS value on both datasets. CQL, another offline RL method that aims to increase performance on unknown distribution, showed stable performance across all product types, ranging from 0.65 to 0.73 on the OOD metric. It performs on par with the CQL, another approach that aims to mitigate the effect of distribution shift. When compared to CQL, because ROL is able to adjust the degree of relaxation based on the relationship between ID and OOD, it outperforms CQL on the overall OOD metric.

Our experimental evaluation results on both datasets show that ROL got the best TS value on OOD data, indicating that the ROL method mitigates the impact of distribution shift under the experiment setting.

Figure 2 plots the learning curves of the TS score for multiple RL-based conversational agents for 10,000 epochs on the CraigslistBargain dataset, scaling each curve by approximate training times. Our ROL approach (orange line) converges much more quickly and achieves higher TS Score than all baselines, highlighting the advantage of reward calibration via ensemble agreement.

Impact of Distribution Shifts on Learning Effectiveness

To examine how distribution shift affects RL agents' learning, we train a DDQN agent on $\langle s_i, a_i, s_{i+1}, r_i \rangle$ in D_{ID}, then introduce controlled distribution shifts by sampling a standard normal vector Δ of the same dimension as a_i, creating shifted tuples $\langle s_i, a_i + \beta\,\Delta, a_i' \rangle$. Varying β quantifies the shift. Figure 3 plots the resulting TS Scores (on the CraiglistBargin successful cases) vs. β. What we observe is that as β grows, TS scores decline sharply, especially beyond $\beta = 0.003$. It suggests over-penalization and destabilized training. Conversely, small or zero β (say $\beta < 0.001$) maintains robust performance, reinforcing the need for careful reward calibration. This observation aligns with our theoretical proof (see Appendix).

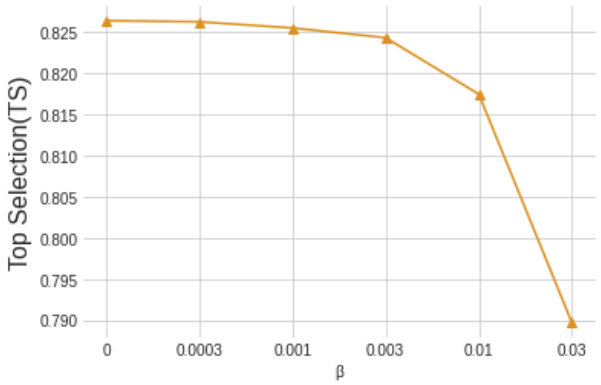

Figure 3: Impact of amount of distribution shift (by varying coefficient β) on the agent's performance metric.

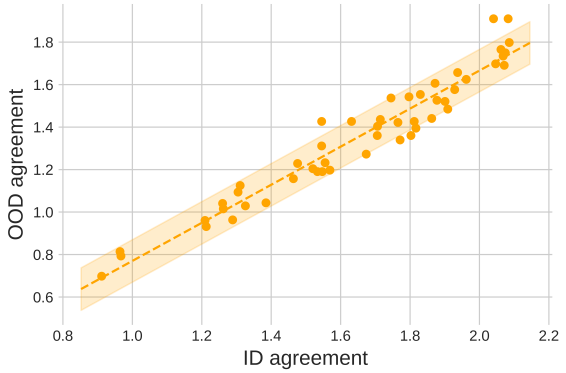

Figure 4: Scatter plot of ID vs. OOD action-selection agreements after probit scaling Φ^{-1} (the inverse CDF of $\mathcal{N}(0,1)$). Points are sampled for image readability. The correlation coefficient is 0.96, indicating a near-linear relationship.

Why ROL Works?

Figure 4 plots OOD action-selection agreement against ID agreement. Each point represents a different training run or hyperparameter setting, plotting OOD action-selection agreement against ID agreement. The near-perfect linear trend (dashed line and shaded region) suggests that ID agreement translates proportionally to OOD agreement, revealing a slope k that can be linked to quantifiable distribution shift. By leveraging this slope in our ROL framework, we recalibrate reward labels to reduce mismatch between training (ID) and testing (OOD) conditions. Effectively, ROL manages to keep β (the amount of distribution shift) low or near zero, thus is able to preserve strong offline performance even on OOD data.

Proof

Proof Setup. We adopt the binary classification framework from Miller's "accuracy-on-the-line" work (Miller et al. 2021), where each label y is uniformly distributed over $\{-1, 1\}$. Let D_{binary} be our dataset, and let θ be a linear model. We define the *accuracy* of θ on D_{binary} as the fraction of examples (x, y) for which $\theta^\top x$ and y have the same sign. Formally:

$$acc_{D_{\text{binary}}}(\theta) = \frac{\left|\{(x,y) \in D_{\text{binary}} : \text{sign}(\theta^\top x) = y\}\right|}{|D_{\text{binary}}|}. \tag{7}$$

In our setting, each data point consists of two candidate utterances, (u_{i1}, u_{i2}), with an associated $\text{pref}(u_{i1}, u_{i2})$ that is uniformly distributed. We define the accuracy of a linear model θ as the fraction of pairs whose preference is correctly predicted by $\text{sign}\big(\theta^\top(u_{i1} - u_{i2})\big)$:

$$\text{acc}_D(\theta) = \frac{1}{|D|}\Bigg|\Bigg\{(u_{i1}, u_{i2}) \in D : \text{sign}\left(\theta^\top(u_{i1} - u_{i2})\right) = \text{pref}(u_{i1}, u_{i2})\Bigg\}\Bigg|. \tag{8}$$

ID & OOD Dataset Construction. Let the embedding dimension be d. We construct the in-distribution (*ID*) and out-of-distribution (*OOD*) datasets as:

$$\begin{aligned} x_{\text{ID}} &\sim \mathcal{N}\big(\mu \cdot y,\ \sigma^2 I_{d\times d}\big), \\ x_{\text{OOD}} &\sim \mathcal{N}\big(\big(\alpha\mu + \beta\Delta\big) \cdot y,\ (\gamma\sigma)^2 I_{d\times d}\big), \end{aligned} \tag{9}$$

where x_{ID} and x_{OOD} are embedded utterances in R^d, $\mu \in R^d$ is the central embedding, $\Delta \in R^d$ is a unit vector indicating the shift direction, and $\alpha, \beta, \gamma > 0$ are scalar parameters. The label $y \in \{-1, 1\}$ is uniformly distributed. We denote the model trained on the ID set by θ_{ID}. The predicted preference of a candidate pair is $sign(\theta^x)$.

In this construction, a larger norm $\|\mu\|$ increases the distance between the two means $\mathcal{N}(\mu \cdot y,\ \sigma^2 I)$ and $\mathcal{N}(-\mu \cdot y,\ \sigma^2 I)$. Meanwhile, a larger σ raises the likelihood of sampling data far from each mean, creating more overlap. If $\mu = 0$, both labels share the same distribution and thus cannot be separated. Conversely, if μ is large but σ is relatively small, most points become distinguishable by their proximity to one mean versus the other. Concretely, for any scalar $\lambda \neq 0$ and index i, consider:

$$\begin{aligned} &P(\mathcal{N}(\mu_i \cdot y \cdot \lambda; (\sigma\lambda)^2 I_{d\times d}) > \mathcal{N}(-\mu_i \cdot y \cdot \lambda; (\sigma\lambda)^2 I_{d\times d})) \\ =&P(\mathcal{N}(2\mu_i \cdot y \cdot \lambda; 2(\sigma\lambda)^2 I_{d\times d}) > 0) \\ =&\frac{\sqrt{2}\sigma}{2\mu_i \cdot y} \\ =&P(\mathcal{N}(\mu_i \cdot y; \sigma^2 I_{d\times d}) > \mathcal{N}(-\mu_i \cdot y; \sigma^2 I_{d\times d})) \end{aligned} \tag{10}$$

This implies scaling μ and σ by λ does not alter the relative probability of one mean exceeding the other, showing that larger $\|\mu\|$ or smaller σ helps separate the two classes more effectively.

Clarity of a Dataset. We define the *clarity* of a dataset as $\frac{\|\mu\|}{\sigma}$. Intuitively, a larger $\|\mu\|$ and smaller σ give a more separable dataset, raising the maximum achievable accuracy for a linear classifier. In subsequent analyses, we keep this clarity fixed across different pairs of datasets to isolate the effect of distribution shift from any inherent dataset "difficulty."

Accuracy under Distribution Shift. Let $acc_D(\theta)$ be the accuracy of a linear classifier θ on dataset D. From Miller (Miller et al. 2021), we have

$$acc_D(\theta) \ = \ \Phi\Big(\frac{\theta^\top \mu}{\|\theta\|\,\sigma}\Big), \tag{11}$$

where Φ is the CDF of the standard normal distribution. A fully converged classifier θ maximizes this expression, so θ is parallel to μ.

Now consider two datasets, D_{ID} and D_{OOD}, both with the same *clarity* (i.e., $\frac{\|\mu\|}{\sigma}$ remains fixed), but D_{OOD} includes a distribution shift $\beta\,\Delta$. For θ_{ID} trained on D_{ID}:

$$acc_{D_{OOD}}(\theta_{ID}) = \Phi\left(\frac{\theta_{ID}^{(}\alpha \cdot \mu + \beta \cdot \Delta)\|\mu\|}{\|\theta_{ID}\|\sigma\|\alpha \cdot \mu + \beta \cdot \Delta\|}\right) \tag{12}$$

Hence, the ratio of $\Phi^{-1}\big(acc_{D_{OOD}}(\theta_{ID})\big)$ to $\Phi^{-1}\big(acc_{D_{ID}}(\theta_{ID})\big)$ is

$$\frac{\Phi^{-1}\big(acc_{D_{OOD}}(\theta_{ID})\big)}{\Phi^{-1}\big(acc_{D_{ID}}(\theta_{ID})\big)} \ = \ \frac{\mu^\top(\alpha\,\mu + \beta\,\Delta)}{\|\mu\|\ \|\alpha\,\mu + \beta\,\Delta\|}. \tag{13}$$

Note that $\Phi^{-1}\big(acc_{D_{\text{ID}}}(\theta_{\text{ID}})\big)$ does not depend on β. As β grows, the ratio increases, implying a larger value of $\Phi^{-1}\big(acc_{D_{\text{OOD}}}(\theta_{\text{ID}})\big)$, however this leads to *lower* actual accuracy on D_{OOD}. Geometrically, a bigger β shifts test points in a direction the ID-trained classifier cannot fully capture, therefore reduces OOD performance.

In summary, under the stated assumptions, we have shown that larger distribution shifts generally reduce accuracy in out-of-distribution domains. If we can manage to measure the linear relationship between ID and OOD agreements, such as demonstrated in this paper, we can quantify the shift via its slope (as seen in Figure 4). The greater this shift, the more we must adjust the rewards, which motivates the calibration approach in our ROL framework.

Conclusion

This paper introduced Reward-on-the-Line (ROL), a novel offline reinforcement learning method for training high-quality, domain-specific conversational agents using static dialogue corpora. Our primary focus is on the legal domain, where large collections of publicly available courtroom transcripts remain an underused resource for building intelligent dialogue systems. These transcripts, rich in legal reasoning and structured argumentation, offer a unique opportunity to develop proactive legal agents that can engage in complex legal conversations without the need for expensive online interactions or manually labeled supervision.

ROL addresses a central challenge in offline reinforcement learning, which is the distribution shift between training data and deployment environments. By leveraging agreement across multiple independently trained value-based agents, our method calibrates the rewards in the training data to better reflect performance expectations in unseen settings. This calibration process is guided by a measurable correlation between agent agreement in known and unknown domains, enabling robust learning without extensive hyperparameter tuning or manual reward modeling. This is particularly valuable in the legal domain, where labeling conversational outcomes is time-consuming, but high-quality historical data is abundant.

Our experiments on the U.S. Supreme Court dataset demonstrate that ROL enables the learning of legally coherent and goal-driven dialogue strategies that outperform strong baseline methods across multiple evaluation metrics, including top utterance selection and legal goal relevance. To further demonstrate the generality of our method, we applied it to the CraigslistBargain dataset, where similar improvements were observed in negotiation dialogues. While legal dialogue remains our primary target, this second experiment confirms that the underlying approach generalizes across different task domains and conversational dynamics.

In summary, the proposed method offers a practical and principled framework for training conversational agents directly from existing transcripts, without the need for new labels or live deployment. It opens new possibilities for applying offline reinforcement learning to specialized domains such as law, where high-quality dialogue data is plentiful but supervision is limited. We believe this work contributes to the broader goal of making intelligent, domain-aware conversational systems more accessible and scalable.

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

Baek, C.; Jiang, Y.; Raghunathan, A.; and Kolter, Z. 2022. Agreement-on-the-Line: Predicting the Performance of Neural Networks under Distribution Shift.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; El-Showk, S.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.

Cai, Q.; Zhao, X.; Pan, L.; Xin, X.; Huang, J.; Zhang, W.; Zhao, L.; Yin, D.; and Yang, G. H. 2024. AgentIR: 1st Workshop on Agent-based Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, 3025–3028. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704314.

Chakraborty, N.; Lukovnikov, D.; Maheshwari, G.; Trivedi, P.; Lehmann, J.; and Fischer, A. 2019. Introduction to neural network based approaches for question answering over knowledge graphs. *arXiv preprint arXiv:1907.09361*.

Chen, J.; Liu, F.; Avci, B.; Wu, X.; Liang, Y.; and Jha, S. 2023. Detecting Errors and Estimating Accuracy on Unlabeled Data with Self-training Ensembles. arXiv:2106.15728.

Christiano, P.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2023. Deep reinforcement learning from human preferences. arXiv:1706.03741.

Deng, Y.; Liao, L.; Lei, W.; Yang, G. H.; Lam, W.; and Chua, T.-S. 2025. Proactive Conversational AI: A Comprehensive Survey of Advancements and Opportunities. *ACM Trans. Inf. Syst.*, 43(3).

Dhariwal, P.; Hesse, C.; Klimov, O.; Nichol, A.; Plappert, M.; Radford, A.; Schulman, J.; Sidor, S.; Wu, Y.; and Zhokhov, P. 2017. OpenAI Baselines. https://github.com/openai/baselines.

Everitt, T.; Krakovna, V.; Orseau, L.; and Legg, S. 2017. Reinforcement Learning with a Corrupted Reward Channel. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, 4705–4713. AAAI Press. ISBN 9780999241103.

Fujimoto, S.; Meger, D.; and Precup, D. 2018. Off-Policy Deep Reinforcement Learning without Exploration. In *International Conference on Machine Learning*.

Fujimoto, S.; Meger, D.; and Precup, D. 2019. Off-Policy Deep Reinforcement Learning without Exploration. arXiv:1812.02900.

He, H.; Chen, D.; Balakrishnan, A.; and Liang, P. 2018. Decoupling Strategy and Generation in Negotiation Dialogues. arXiv:1808.09637.

Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, 216–225.

Jiang, Y.; Nagarajan, V.; Baek, C.; and Kolter, J. Z. 2022. Assessing Generalization of SGD via Disagreement. arXiv:2106.13799.

Jiang, Z.; Xu, D.; and Liang, J. 2017. A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*.

Kandasamy, K.; Bachrach, Y.; Tomioka, R.; Tarlow, D.; and Carter, D. 2017. Batch Policy Gradient Methods for Improving Neural Conversation Models. arXiv:1702.03334.

Kumar, A.; Zhou, A.; Tucker, G.; and Levine, S. 2020. Conservative Q-Learning for Offline Reinforcement Learning. *CoRR*, abs/2006.04779.

Levine, S.; Kumar, A.; Tucker, G.; and Fu, J. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *ArXiv*, abs/2005.01643.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Liu, J.; Pan, F.; and Luo, L. 2020. GoChat: Goal-oriented Chatbots with Hierarchical Reinforcement Learning. arXiv:2005.11729.

Mbuwir, B. V.; Ruelens, F.; Spiessens, F.; and Deconinck, G. 2017. Battery energy management in a microgrid using batch reinforcement learning. *Energies*, 10(11): 1846.

Miller, J.; Taori, R.; Raghunathan, A.; Sagawa, S.; Koh, P. W.; Shankar, V.; Liang, P.; Carmon, Y.; and Schmidt, L. 2021. Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization.

Munos, R.; Stepleton, T.; Harutyunyan, A.; and Bellemare, M. G. 2016. Safe and Efficient Off-Policy Reinforcement Learning. *CoRR*, abs/1606.02647.

Nair, A.; McGrew, B.; Andrychowicz, M.; Zaremba, W.; and Abbeel, P. 2018. Overcoming Exploration in Reinforcement Learning with Demonstrations. arXiv:1709.10089.

Nakkiran, P.; and Bansal, Y. 2020. Distributional Generalization: A New Kind of Generalization. arXiv:2009.08092.

OpenAI, R. 2023. GPT-4 technical report. *arXiv*, 2303–08774.

Osband, I.; Blundell, C.; Pritzel, A.; and Roy, B. V. 2016. Deep Exploration via Bootstrapped DQN. In *Advances in Neural Information Processing Systems (NeurIPS)*, 4026–4034.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290.

Robey, A.; Hassani, H.; and Pappas, G. J. 2020. Model-Based Robust Deep Learning: Generalizing to Natural, Out-of-Distribution Data. arXiv:2005.10247.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D. M.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. 2022. Learning to summarize from human feedback. arXiv:2009.01325.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural Networks. arXiv:1409.3215.

Tang, Z.; Kulkarni, H.; and Yang, G. H. 2021. High-Quality Dialogue Diversification by Intermittent Short Extension Ensembles. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1861–1872. Online: Association for Computational Linguistics.

Tang, Z.; and Yang, G. H. 2020. Corpus-Level End-to-End Exploration for Interactive Systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03): 2527–2534.

Tang, Z.; and Yang, G. H. 2022. A Re-classification of Information Seeking Tasks and Their Computational Solutions. *ACM Trans. Inf. Syst.*, 40(4).

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

van Hasselt, H.; Guez, A.; and Silver, D. 2016. Deep Reinforcement Learning with Double Q-Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Verma, S.; Fu, J.; Yang, M.; and Levine, S. 2022. CHAI: A CHatbot AI for Task-Oriented Dialogue with Offline Reinforcement Learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Vinyals, O.; and Le, Q. 2015. A Neural Conversational Model. arXiv:1506.05869.

Xu, H.; Zhan, X.; and Zhu, X. 2022. Constraints Penalized Q-learning for Safe Offline Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8): 8753–8760.

Xu, S.; Fu, W.; Gao, J.; Ye, W.; Liu, W.; Mei, Z.; Wang, G.; Yu, C.; and Wu, Y. 2024. Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study. arXiv:2404.10719.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zheng, W.; Yu, J. X.; Zou, L.; and Cheng, H. 2018. Question Answering Over Knowledge Graphs: Question Understanding Via Template Decomposition. *Proc. VLDB Endow.*, 11: 1373–1386.

Zhou, L.; Small, K.; Rokhlenko, O.; and Elkan, C. 2017. End-to-End Offline Goal-Oriented Dialog Policy Learning via Policy Gradient. arXiv:1712.02838.