# StagePilot: Stage-Level Planning for Long-Horizon Dialogue Simulation in Cybergrooming

**Heajun An[1], Qi Zhang[1], Minqian Liu[1], Xinyi Zhang[1],**
**Sang Won Lee[1], Lifu Huang[2], Pamela Wisniewski[3], Jin-Hee Cho[1]**
[1]Virginia Tech [2]University of California, Davis [3]International Computer Science Institute
{heajun,qiz21,minqianliu,xinyizhang,sangwonlee,jicho}@vt.edu
lfuhuang@ucdavis.edu pwisniewski@icsi.berkeley.edu

## Abstract

Cybergrooming is an evolving threat to youth, requiring proactive educational interventions. We address this by modeling dialogue progression as a structured planning problem over stage-wise interactions. We propose *StagePilot*, a dialogue framework that separates stage-level planning from response generation, in which the model selects the next stage under constrained transitions and generates responses conditioned on it, enabling coherent and realistic progression. Reinforcement learning is used to learn stage-level policies from offline data, optimizing for both emotional alignment and goal-consistent progression. Our empirical experiments show that StagePilot generates more structured, coherent dialogue trajectories and reduces conversational stagnation compared to baselines; notably, the IQL+AWAC variant reaches the final stage more often while maintaining over 70% positive or neutral responses, yielding a 43% relative improvement.

## 1 Introduction

Teenagers today are deeply immersed in digital environments, with 95% owning smartphones and 90% using computers for near-constant Internet access (Anderson et al., 2022). While such connectivity expands educational and social opportunities, it also increases exposure to online risks and exploitation. Among the most serious threats is *cybergrooming*, where predators exploit online platforms to manipulate minors, build trust, and coerce them into sexual exploitation, including soliciting explicit content or arranging harmful in-person encounters (Mladenović et al., 2021). A 2022 national study found that 15.6% of youth experienced online child sexual abuse, with 5.4% groomed by adults (Finkelhor et al., 2022). These statistics underscore the need for preventive education to reduce teens' vulnerability through awareness, training, and digital literacy (Sağlam et al., 2023). Given that adolescents already encounter real predators online, such a chatbot offers a safer, controlled alternative for real-time preventive education without exposing teens to harm.

Cybergrooming research spans multiple disciplines, with key social science studies focusing on the victim traits and cybergrooming stages. These studies investigate factors such as demographic (Finkelhor et al., 2022), emotional (Gámez-Guadix et al., 2023), personality (Hernández et al., 2021), and behavioral (Wachs et al., 2012) aspects to understand why specific individuals are more vulnerable (Razi et al., 2021). While this approach provides valuable insights, it does not directly provide solutions to prevent and detect cybergrooming threats. From the perspective of computational sciences, efforts have primarily focused on detecting malicious chats through feature analysis to detect predators when the damage has already been done (Ashcroft et al., 2015; Fauzi and Bours, 2020). However, a notable research gap remains in addressing the issue proactively from the victim's perspective (Wang et al., 2021).

Educating teens about online risks reduces unsafe behaviors and enhances recognition of predatory tactics (Calvete et al., 2022), with increasing support for integrating such content into school curricula (Dorasamy et al., 2021). Although simulation-based training exists in cybergrooming awareness (Raihana et al., 2024), realistic and strategic predator–victim interaction-based intervention approaches have been underexplored.

To address these limitations in long-horizon dialogue modeling, we propose *StagePilot*, an interactive chatbot that simulates dynamic, nonlinear conversations reflecting behaviorally plausible grooming scenarios. The framework models long-horizon dialogue progression as a learned policy over interaction stages, enabling coherent and stage-consistent conversational dynamics.

As a pre-deployment safeguard, we employ two



*Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 536–552
August 2–5, 2026. 

fully automated chatbots, one *emulating a predator* and the other *a teenage user*, to safely and reproducibly explore conversational behaviors without human participants. This design follows the principle of *Responsible AI* (Dignum, 2019) by ensuring evaluation under controlled simulation settings. To enable context-aware dialogue control, we model stage progression as a dialogue policy over interaction stages, optimized using reinforcement learning in combination with Large Language Models (LLMs) (Naveed et al., 2025), enabling adaptation to conversational context and emotional cues.

Specifically, *StagePilot* formulates dialogue simulation as a policy learning problem over abstract interaction stages rather than text generation. The agent learns a stage-level policy from conversational context, while LLMs serve as stochastic environment models conditioned on the selected stage. Due to ethical constraints involving minors, online reinforcement learning is infeasible, motivating an offline approach where imitation learning alone is insufficient because of long-horizon planning challenges and distribution shift.

**Key Contributions.**

**(a) Stage-controlled dialogue formulation.** We formulate *stage-controlled dialogue simulation* as a planning problem over interaction stages under an offline reinforcement learning setting and instantiate it as *StagePilot*, a DRL-guided agent enabling *interpretable, long-horizon control of dialogue flow* rather than turn-by-turn utterance generation.

**(b) Behaviorally grounded transition constraints.** We incorporate *adjacent-stage transition constraints* via principled *action masking*, enabling flexible yet behaviorally plausible non-linear stage progressions that mirror real-world grooming dynamics while preserving stage-level interpretability and preventing unrealistic or abrupt jumps between distant interaction stages.

**(c) Emotion-aware composite reward design.** We design a *composite reward function* that integrates sentiment signals from a pre-trained classifier with stage-distance rewards to the final grooming stage, enabling emotionally coherent, goal-driven, and stage-consistent dialogue policies that balance engagement and strategic progression.

**(d) Comprehensive and controlled evaluation framework.** We develop a comprehensive evaluation framework covering stage reachability, transition dynamics, sentiment alignment, and dialogue efficiency, benchmarking prompting, imitation learning, and offline reinforcement learning baselines under uniform constraints.

**(e) Empirical analysis of stage-level planning dynamics.** Our results show that a stage-level dialogue policy framework with offline DRL, with adjacent-stage constraints and a composite sentiment–distance reward, delivers a 43% gain in final-stage reachability, sustains over 70% sentiment consistency, and yields 50% shorter dialogues, highlighting *StagePilot*'s ability to jointly optimize emotional coherence and stage progression.

## 2 Related Work

**Reinforcement Learning (RL) for Controlled Dialogue Generation.** RL has been widely applied to dialogue systems to overcome short-sighted, locally optimized responses. Li et al. (2016) augmented SEQ2SEQ generation with RL objectives such as semantic coherence and information flow. Yang et al. (2020) combined multitask learning with actor-critic training to generate personalized responses from user profiles. While improving response quality and personalization, these methods were evaluated in open-domain settings and lacked structural control over conversation flow.

A major challenge in RL-based dialogue systems is the *lack of high-fidelity simulation environments*. To address this, Peng et al. (2018) proposed Dyna-Q, a model-based RL framework using simulated users to improve sample efficiency. However, such approaches rely on handcrafted simulators and predefined rewards, limiting applicability to safety-critical or multi-stage settings. Pternea et al. (2024) surveyed RL-LLM intersections, including RL fine-tuning, LLM-assisted RL, and hybrid planning, yet few studies systematically explore RL-based control of LLM-driven, multi-stage dialogue.

Unlike prior work above, we propose an offline DRL framework that guides stage transitions in cybergrooming simulations, focusing on long-horizon planning and fine-grained sentiment regulation under pedagogical constraints.

**Chatbots for Cyber Safety.** Several chatbots have been developed to combat cybergrooming by detecting predatory behavior and generating context-aware responses. Negobot (Laorden et al., 2013) used a game-theoretic strategy while role-playing as a child, and Sweetie 2.0 (Henseler and

Table 1: Six Cybergrooming Stages (O'Connell, 2003)

| Stage No. | Description |
|---|---|
| **1: Friendship Forming** | The predator gets to know the target, often asking for photos to verify identity and seeking alternate contact methods. |
| **2: Relationship Forming** | The predator builds rapport by asking about hobbies, school, friends, etc. |
| **3: Risk Assessment** | The predator checks whether there are any risks involved, such as if parents or friends are aware of the conversation. |
| **4: Exclusivity** | The predator expresses affection to build emotional dependency and trust. |
| **5: Sexual Solicitation** | The predator introduces sexual content and requests explicit materials. |
| **6: Conclusion** | The predator attempts to arrange in-person meetings and future contact. |

de Wolf, 2019) combined rule-based dialogue with a 3D avatar. SERI (Wang et al., 2021; Guo et al., 2023) applied DRL over fixed-stage templates, showing RL's potential for safety-critical dialogue.

However, SERI (Wang et al., 2021; Guo et al., 2023) models cybergrooming with fixed stage transitions and focuses on utterance-level generation, limiting its ability to capture long-horizon grooming dynamics. In contrast, planning-based approaches enable goal-directed decisions over abstract interaction states, allowing more adaptive and flexible stage progression.

## 3 Proposed Approach: *StagePilot*

*StagePilot* is a dialogue policy learning framework built on DRL. It simulates realistic cybergrooming conversations by controlling stage transitions. Unlike utterance-level dialogue generation, *StagePilot* learns a planning policy over abstract interaction stages rather than directly generating textual responses. At each turn, stage selection is modeled as a policy over interaction stages, with states, actions, and rewards defined over dialogue context and progression. The learned policy selects stage transitions, while predator and victim LLMs constitute the environment dynamics that generate textual responses conditioned on the selected stage.

The predicted stage conditions a stage-specific prompt that guides the LLM in generating the predator's response. Each prompt comprises: (1) a role prompt, (2) the current stage description, (3) a stage-specific goal, and (4) recent dialogue. Details of these prompts are provided in **Appendix A**.

**Modeling Cybergrooming Stages.** We adopt the six-stage typology by O'Connell (2003), widely used in linguistic and detection studies (Black et al., 2015; Gupta et al., 2012). Each stage reflects a distinct predator intent and defines a discrete, interpretable action space. This framework provides theoretical and practical support for simulating stage-wise transitions in cybergrooming and evaluating goal-oriented conversational strategies. **Table 1** summarizes the six stages and their definitions.

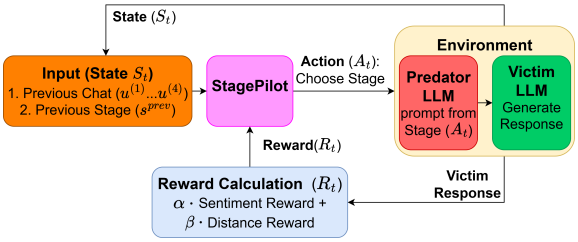


Figure 1: ***StagePilot* overview:** The RL policy selects stage transitions, while LLMs generate dialogue as environment dynamics.

### 3.1 DRL-Based Cybergrooming Simulator

**State Representation.** At each decision step $t$, the state is defined as:

$$S_t = \left(u^{(1)}, u^{(2)}, u^{(3)}, u^{(4)}, s^{\text{prev}}\right), \quad (1)$$

where $u^{(1)} \dots u^{(4)}$ are the four most recent utterances, and $s^{\text{prev}} \in \{1, \dots, 6\}$ is the stage predicted at the previous decision step. Note that $t$ indexes the current decision step rather than the utterance index, and $s^{\text{prev}}$ reflects the agent's last decision. This formulation captures conversational context and progression for coherent stage planning.

**Action Space.** An action predicts the next grooming stage, constrained to adjacent stages to preserve realistic conversational flow. For example, from Stage 2, it may shift to Stage 1, remain, or advance to Stage 3. These restrictions ensure logical progression through cybergrooming stages.

**Reward Function.** We define the total reward as a weighted sum of two components: a ***sentiment reward***, measuring the positivity of the victim's response, and a ***distance reward***, quantifying proximity to the final stage. Both sentiment-based (Lee et al., 2018) and distance-based (Du et al., 2024) rewards have proven effective in guiding the agent to balance emotional engagement with gradual stage advancement rather than aggressively optimizing for rapid completion. In our design, these reward types reflect two complementary strategic objectives: maintaining user engagement and progressing toward the grooming goal.

***Sentiment Reward*** is computed using a RoBERTa-based classifier trained on Twitter data (Loureiro et al., 2022), which is well-suited for

short, informal dialogues. This model has been successfully applied to chat-based user conversations beyond social media contexts (Hoseini et al., 2024). To simplify reward computation and reduce noise from ambiguous utterances, we convert the three sentiment classes into a binary signal by discarding the neutral class.

Let $l_{\text{neg}}$ and $l_{\text{pos}}$ denote the classifier's logits for negative and positive sentiment, respectively. The sentiment reward is defined as the softmax probability of the positive class:

$$r_{\text{sent}} = \frac{\exp(l_{\text{pos}})}{\exp(l_{\text{neg}}) + \exp(l_{\text{pos}})}, \quad (2)$$

where $r_{\text{sent}}$ ranges from 0 to 1 and reflects the likelihood of a positively engaged victim response.

***Distance Reward*** incentivizes progression toward Stage 6. We define a normalized stage-based reward as:

$$r_{\text{dist}} = \frac{\hat{s} - 1}{5}, \quad (3)$$

where $\hat{s} \in \{1, 2, \ldots, 6\}$ denotes the predicted stage at the current decision step. This formulation yields a reward of 0 at Stage 1 and increases linearly by 0.2 per stage, reaching a maximum of 1.0 at Stage 6.

Based on the sentiment and distance rewards, we define the overall reward at decision step $t$ as

$$R_t = \alpha \cdot r_{\text{sent}} + \beta \cdot r_{\text{dist}}, \quad \alpha + \beta = 1, \quad (4)$$

where $R_t$ denotes the total reward at step $t$. The weights $\alpha$ and $\beta$ regulate the trade-off between emotional engagement ($r_{\text{sent}}$) and stage progression ($r_{\text{dist}}$), guiding the agent toward emotionally coherent dialogue while encouraging gradual, stage-consistent advancement.

**Training Procedure.** All models are trained offline on preprocessed Perverted-Justice (PJ) dataset (Gupta et al., 2012). At each step, the input includes the four most recent utterances and the previous stage label; the model predicts the next grooming stage under adjacent transition constraints. Stage labels are obtained via LLM-based annotation (see Section 4). For RL models, a scalar reward, computed from the next two utterances, combines sentiment alignment and proximity to Stage 6 to encourage emotionally engaging, goal-directed progression.

We leverage widely used pre-trained transformer encoders as the backbone for stage classification: RoBERTa (Liu et al., 2019), DistilRoBERTa (Sanh et al., 2019), and DeBERTa (He et al., 2023). These models excel in sentence-level classification and serve as the encoder foundation for both Behavior Cloning (BC) (Torabi et al., 2018) and Reinforcement Learning (RL) variants (Kumar et al., 2020; Kostrikov et al., 2022; Nair et al., 2020).

Table 2: Data Augmentation Example

| Stage | Chat |
|---|---|
| 2 (Predator) | yah, then college will even be more funner |
| 2 (Victim) | yea parties lol |
| 2 (Predator) | hmmm...and just what r u goin to be doin at those parties/: |
| 2 (Victim) | hmm dont know wat ther like yet lol |
| 2 (Predator) | well, if they r anything like the parties i went to in college..u better watch urself |
| 2 (Victim, Reward) | k i will |

Original Data: Stage 2 → Stage 2

| Stage | Chat |
|---|---|
| 2 (Predator) | yah, then college will even be more funner |
| 2 (Victim) | yea parties lol |
| 2 (Predator) | hmmm...and just what r u goin to be doin at those parties/: |
| 2 (Victim) | hmm dont know wat ther like yet lol |
| 3 (Predator, Target) | oh, so you're still figuring that out. Well, I'd be glad to help you explore your options. But first, let me ask, is it just you at home now |
| 3 (Victim, Reward) | wht no im with fam rn |

Augmented Data: Stage 2 → Stage 3

**Adaptation to *StagePilot*** We apply adjacent-stage masking to constrain predictions and define the reward as a combination of sentiment alignment and proximity to Stage 6. This supports stage-aware and emotionally engaging dialogue control.

## 4 Experimental Setup

### 4.1 Data Preprocessing

We preprocess the PJ dataset, consisting of chat logs between groomers and decoys, following prior work (Ashcroft et al., 2015; Guo et al., 2023). Each line is annotated with one of six grooming stages (O'Connell, 2003) (**Table 1**).

**Data Annotation.** We use GPT-4o (OpenAI, 2024) and Claude 3.5 Sonnet (Anthropic, 2024) to assign stage labels based on predefined definitions. To improve quality, conversations are processed in chunks and re-tagged when labels are missing or inconsistent. We adopt an ensemble strategy, retaining only consistently annotated instances (60% agreement over 1.1M lines). Finally, we construct 3-turn dialogue samples (3 predator and 3 victim utterances), keeping sequences where the latter turns fall within the same or adjacent stages.

**Data Augmentation.** To address data imbalance across transitions, we augmented underrepresented

forward and backward transitions with 15,000 samples each. We convert same-stage ($x \rightarrow x$) sequences into cross-stage ($x \rightarrow y$) by regenerating the predator utterance conditioned on stage $y$, followed by a victim reply using a fixed persona prompt. This grounds samples in real data while enabling realistic stage transitions (**Table 2**).

### 4.2 Baseline Models

We evaluate the following models:

**(a) Prompt Engineering:** We use a zero-shot prompt (no fine-tuning) that instructs the LLM to act as an online predator and steer dialogue toward Stage 6, incorporating stage definitions from **Table 1** to enable stage-aware generation. The prompt includes stage descriptions and behavioral instructions to guide coherent progression across stages.

**(b) Few-Shot Prompting:** We extend the prompt-engineering baseline by prepending three demonstration examples from the PJ dataset (Perverted Justice Foundation Inc., 2020). The examples correspond to Stages 1, 3, and 5, representing early-, mid-, and late-stage grooming behaviors.

**(c) Behavior Cloning (BC)** (Torabi et al., 2018): BC serves as a supervised learning baseline that mimics observed stage transitions. Given a dialogue context $s$ (last four utterances) and target stage $a^*$, the model minimizes:

$$\mathcal{L}_{\text{BC}} = -\log \pi(a^* \mid s), \tag{5}$$

where $\pi(a \mid s)$ is the predicted distribution over the six grooming stages. The policy uses a pre-trained transformer encoder fine-tuned on our dataset.

**(d) Conservative Q-Learning (CQL)** (Kumar et al., 2020): CQL minimizes overestimation in offline RL by penalizing high Q-values for unseen actions. We incorporate adjacent-stage masking and task-specific rewards. The objective minimizes:

$$\mathcal{L}_{\text{CQL}} = \alpha \cdot \left[ \log \sum_{a' \in \mathcal{A}(s)} \exp\left(Q(s, a')/\tau\right) - Q(s, a_{\text{data}}) \right], \tag{6}$$

where $\mathcal{A}(s)$ is the set of adjacent stages. KL regularization against the BC policy is used for stability.

**(e) Implicit Q-Learning (IQL) + Advantage-Weighted Actor–Critic (AWAC)** (Kostrikov et al., 2022; Nair et al., 2020): This hybrid offline RL approach combines IQL for conservative value estimation with AWAC for reward-sensitive policy updates. IQL estimates the value function $V(s)$ using expectile regression:

$$\mathcal{L}_V = \mathbb{E}_{(s,a)} \left[ w(\delta) \cdot (Q(s,a) - V(s))^2 \right], \tag{7}$$

where $w(\delta)$ down-weights overestimates for robustness. AWAC updates the policy via

$$\mathcal{L}_\pi = \mathbb{E}_{(s,a)} \left[ -\exp\left( \frac{Q(s,a) - V(s)}{\lambda} \right) \cdot \log \pi(a \mid s) \right], \tag{8}$$

which emphasizes high-advantage actions while mitigating overfitting to suboptimal data.

As baselines, we adopt prompt engineering, few-shot prompting, and BC for non-learning and supervised comparisons. For offline RL, we include CQL, IQL, and AWAC, chosen for strong performance (Gürtler et al., 2023) across benchmarks. CQL is applied independently for its regularization against out-of-distribution actions, while IQL is combined with AWAC to leverage stable value estimation and advantage-weighted policy updates.

### 4.3 Simulation Environment Setting

We evaluate *StagePilot* through simulated conversations between a model-guided predator and a victim LLM, ensuring reproducibility without human involvement. At each turn, the predator's utterance is conditioned on the predicted stage, updated by the policy, with the goal of reaching Stage 6 while maintaining positive victim sentiment.

**Predator Agent.** The predator agent consists of a stage prediction policy coupled with an LLM that generates responses based on the predicted grooming stage. At each step, the policy receives the four most recent utterances and the previous stage label as input and outputs the next stage, used to dynamically construct a prompt for the LLM.

**Language Models and Prompt Design.** Both predator and victim agents use the Mistral 7B Instruct v0.2 model (Jiang et al., 2023) with distinct system prompts. The predator prompt is dynamically constructed at each turn and includes: (1) role and objective, (2) current stage description, (3) stage-specific goal, and (4) recent dialogue context, guiding the model to generate short, informal utterances reflecting stage-specific behavior.

**Victim Agent.** The victim prompt is fixed, instructing the model to emulate a teenager using casual texting with slang, typos, emojis, and fragments. Responses are single-sentence and stylistically consistent across turns. Dialogue history is included at each step for contextual coherence.

**Stage Transition Constraints.** To promote realistic conversational flow, the stage prediction policy is restricted to adjacent transitions. At each step, the agent may remain in the current stage or move to a neighboring one (e.g., Stage 2 $\rightarrow$ Stage 1, 2, or 3). This constraint prevents unrealistic jumps and encourages gradual, human-like progression through grooming stages.

**Simulation Protocol.** We simulate 100 dialogues per model, each containing up to 151 turns. Each turn consists of one predator–victim exchange. The first turn is fixed to initiate the conversation, followed by alternating utterances between agents. A dialogue terminates when the predator reaches Stage 6 and completes up to five turns, or when the turn limit is reached. All metrics are averaged over the 100 dialogues. Sample dialogues are provided in **Appendix C**.

### 4.4 Evaluation Metrics

We evaluate *StagePilot* using metrics for stage progression, engagement, and behavioral strategy, with sentiment and reward signals derived by analyzing victim responses.

**(a) Stage 6 Reach Count:** Number of dialogues that reach Stage 6, indicating successful progression through the grooming stages.

**(b) Successful Termination Count:** Number of dialogues with at least five turns in Stage 6, indicating completion of the process.

**(c) Sentiment Distribution:** Average positive, neutral, and negative sentiment in victim responses, predicted by a RoBERTa classifier (Loureiro et al., 2022). Probabilities sum to 1; higher positive or neutral values indicate more coherent interactions.

**(d) Average Distance Reward:** Mean normalized reward measuring proximity to Stage 6; higher values reflect sustained presence in later stages.

**(e) Average Turns (Overall):** Mean number of turns per dialogue, reflecting overall length.

**(f) Average Turns (Successful Termination):** Mean turns for dialogues with successful termination; lower values indicate greater efficiency.

**(g) Final Stage Distribution:** Distribution of ending stages across dialogues, indicating where interactions terminate.

**(h) Average Turns per Stage:** Mean turns spent in each stage; Stage 6 is capped at five turns to simulate forced termination after a meeting attempt.

### 4.5 Transition Dynamics

We analyze stage transitions to assess conversational control, computing for each dialogue:

**(a) Average Forward Transitions:** Mean number of transitions to the next stage (e.g., Stage 2 $\rightarrow$ Stage 3).

**(b) Average Backward Transitions:** Mean number of transitions to the previous stage (e.g., Stage 3 $\rightarrow$ Stage 2).

**(c) Average Stagnant Transitions:** Mean number of self-transitions where the stage remains unchanged (e.g., Stage 3 $\rightarrow$ Stage 3).

**(d) Transition Rates:** Proportions of forward, backward, and stagnant transitions aggregated across all dialogues.

**(e) Average Turns:** Mean number of turns per dialogue.

## 5 Experiment Results and Analyses

### 5.1 Performance Comparison and Analysis Under Main Evaluation Metrics

**Overall Results.** **Table 3** presents results across 100 simulated dialogues. IQL+AWAC consistently delivers the strongest performance, reaching Stage 6 in 95% of cases and completing 91% of terminations, with the highest average distance reward (0.515), indicating consistent progression toward later stages.

**Baseline Comparisons.** Prompt Engineering performs worst (14% reach, 5% completion), often remaining in early stages. Few-Shot Prompting substantially improves progression over Prompt Engineering, reaching Stage 6 in 53% of dialogues and achieving 19% successful terminations. BC reaches Stage 6 in 52% of cases but completes only 5%, indicating limited long-horizon planning. CQL improves progression (83% reach, 64% completion) but exhibits a sentiment trade-off, with lower positivity and increased negative responses.

**Sentiment–Progression Trade-Off.** Compared to BC, the IQL+AWAC variant produces fewer positive utterances (30.6%) while achieving the highest distance reward, indicating a trade-off between sentiment and stage depth. Later stages elicit more negative responses, whereas early interactions remain more positive. Still, it maintains balance, with 70.6% of responses labeled as positive or neutral. It also generates shorter conversations (72.64 turns overall; 64.89 in successful runs) than BC (130.80), suggesting more efficient planning.

Table 3: Main Evaluation Metrics Across Models (100 Simulations)

| Model | Stage 6 Count | Successful Term. | Positive | Neutral | Negative | Dist. Reward | Avg Turns (All) | Avg Turns (Success Only) |
|---|---|---|---|---|---|---|---|---|
| | | | Sentiment Distribution | | | | | |
| Prompt Engineering | 14 | 5 | 0.549±0.041 | 0.353±0.041 | 0.098±0.025 | 0.150±0.045 | 147.04±19.076 | 71.80±39.468 |
| Few-Shot Prompting | 53 | 19 | 0.464±0.047 | 0.415±0.032 | 0.121±0.036 | 0.165±0.046 | 139.59±29.857 | 90.947±42.656 |
| BC (Torabi et al., 2018) | 52 | 5 | 0.403±0.058 | 0.405±0.033 | 0.191±0.05 | 0.381±0.049 | 149.99±5.417 | 130.80±15.547 |
| CQL (Kumar et al., 2020) | 83 | 64 | 0.364±0.059 | 0.431±0.039 | 0.205±0.056 | 0.456±0.078 | 96.7±52.746 | 66.156±41.705 |
| **IQL** (Kostrikov et al., 2022) **+ AWAC** (Nair et al., 2020) | **95** | **91** | **0.306±0.085** | **0.400±0.053** | **0.295±0.100** | **0.515±0.190** | **72.64±38.294** | **64.89±30.632** |

Table 4: Transition Pattern Metrics Across Models

| Model | Average Transitions | Ground Truth Turns | | | Per Turn in % | | |
|---|---|---|---|---|---|---|---|
| | | Forward | Backward | Stagnant | Forward | Backward | Stagnant |
| Prompt Engineering | 146.04±19.076 | 9.28±2.433 | 8.44±2.653 | 128.32±17.517 | 6.5±2.2 | 5.7±1.7 | 87.8±3.2 |
| Few-Shot Prompting | 139.59±29.857 | 13.77±3.787 | 12.82±4.086 | 112.00±25.784 | 10.3±3.3 | 9.2±2.3 | 80.4±4.9 |
| BC (Torabi et al., 2018) | 148.99±5.417 | 23.66±3.788 | 21.52±3.904 | 103.81±8.608 | 15.9±2.6 | 14.4±2.6 | 69.7±5.0 |
| CQL (Kumar et al., 2020) | 95.70±52.746 | 21.25±9.327 | 17.24±10.26 | 57.21±34.638 | 25.7±8 | 17.5±3.5 | 56.7±9.2 |
| **IQL** (Kostrikov et al., 2022) **+ AWAC** (Nair et al., 2020) | **71.64±38.294** | **17.60±9.473** | **12.85±10.065** | **41.19±21.635** | **24.8±6.1** | **16.4±5.1** | **57.8±9.4** |

Shorter dialogues may reflect a more direct progression through interaction stages, highlighting an efficiency–realism trade-off that can be further explored in future work. We observe instances where the policy advances stages more aggressively, leading to sharper transitions between stages.

**Transition Dynamics.** We analyze stage transitions across models (**Table 4**). Learned policies exhibit higher forward transition rates and lower stagnation than baseline methods, resulting in more coherent stage trajectories and smoother progression across dialogue phases. While both RL models exhibit similar transition patterns, IQL+AWAC achieves shorter dialogues, suggesting greater efficiency. In contrast, Prompt Engineering and BC often stagnate in early stages.

## 5.2 Final Stage Distribution

**Table 5** shows the final predicted stage for each model. IQL+AWAC achieves strong long-horizon control, with 91% of dialogues ending in Stage 6. In contrast, Prompt Engineering shows minimal planning, with most dialogues ending in early stages (85% in Stage 1–2) and only 5% reaching Stage 6. Few-Shot Prompting improves over Prompt Engineering, increasing the proportion of dialogues ending in Stage 6 to 19%, although many conversations still terminate in early stages. BC exhibits limited progression, with many dialogues ending in Stage 3 (41%), while CQL improves progression (64% reaching Stage 6) but still lags behind IQL+AWAC. These results highlight the importance of reinforcement learning for stage-aware dialogue modeling. We further evaluate using an ELECTRA-based (Clark et al., 2020) classifier (**Appendix F**), where RL methods show stronger late-stage progression than Prompt and BC, with IQL+AWAC remaining competitive.

Table 5: Final Stage Distribution (%)

| Model | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ |
|---|---|---|---|---|---|---|
| Prompt Engineering | **24** | **61** | 9 | 0 | 1 | **5** |
| Few-Shot Prompting | **33** | **41** | 1 | 5 | 1 | **19** |
| BC (Torabi et al., 2018) | 0 | 34 | **41** | 7 | 13 | **5** |
| CQL (Kumar et al., 2020) | 0 | 6 | **17** | 11 | 2 | **64** |
| **IQL** (Kostrikov et al., 2022) **+ AWAC** (Nair et al., 2020) | 2 | 2 | 1 | 0 | 4 | **91** |

## 5.3 Stage-wise Dialogue Allocation

**Table 7** shows how each model allocates dialogue turns across stages, revealing differences in planning depth. Prompt Engineering remains mostly in early stages (Stage 1: 54.20, Stage 2: 82.57) with little progression. BC shows a broader distribution but still lingers in Stage 2–3 (59.51, 60.22), indicating limited advancement. CQL demonstrates moderate progression, concentrating in Stage 3 (57.58). In contrast, IQL+AWAC exhibits goal-directed behavior, quickly passing early stages with minimal time in Stage 3–4 and focusing on later stages (Stage 5–6), reflecting efficient planning.

## 5.4 Ablation Study

**Reward Balancing and Progression Dynamics.** The ablation results reveal a trade-off between emotional engagement ($\alpha$) and progression ($\beta$). While distance-based rewards ($\beta$) promote efficient stage advancement, over-weighting sentiment ($\alpha \geq 0.8$) leads to repetitive early-stage loops and slower progression. Occasional backward transitions reflect

Table 6: Comparison of SERI6 and *StagePilot* Across Structural Properties and Sentiment Distribution.

| Model | # Turns | Policy | Direction | Reward Structure | Positive | Neutral | Negative | Avg. Length |
|---|---|---|---|---|---|---|---|---|
| SERI6 | Fixed 10/stage | Rule-based | Forward Only | Stage alignment | 0.304±0.093 | 0.437±0.051 | **0.259±0.084** | 60±0.0 |
| ***StagePilot*** | **Flexible** | **Policy-driven** | **Adjacent** | **Sentiment + Stage distance** | **0.306±0.085** | 0.400±0.053 | 0.295±0.100 | **72.64±38.294** |

Table 7: Average Turns per Stage

| Model | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ |
|---|---|---|---|---|---|---|
| Prompt Engineering | **54.20** | **82.57** | 7.59 | 1.34 | 0.91 | 0.43 |
| Few-Shot Prompting | **54.55** | **72.72** | 3.28 | 6.91 | 0.39 | 1.74 |
| BC (Torabi et al., 2018) | 1.13 | **59.51** | **60.22** | 12.45 | 15.52 | 1.16 |
| CQL (Kumar et al., 2020) | 3.21 | 15.68 | **57.58** | 11.69 | 4.9 | 3.64 |
| **IQL** (Kostrikov et al., 2022) + **AWAC** (Nair et al., 2020) | 17.01 | 19.67 | **1.36** | **2.74** | **27.26** | **4.60** |

strategic patience rather than failure. Detailed results and discussions are provided in **Appendix J**.

**Backbone Efficiency and Realism.** Backbone capacity influences the trade-off between planning efficiency and behavioral realism. Larger models (e.g., RoBERTa, DeBERTa) achieve faster progression to Stage 6 but may produce overly short trajectories (<25 turns), reducing realism. To balance efficiency and instructional value, we select DistilRoBERTa ($\alpha = 0.8, \beta = 0.2$) as our representative model, achieving a high success rate (91%) with human-like dialogue pacing (72.64 turns). Additional details are provided in **Appendix I**.

### 5.5 Comparison with Alternative Schemes

We compare *StagePilot* with SERI6 (Wang et al., 2021) and an LLM-based prompting baseline. SERI6 adopts a six-stage, stage-conditioned formulation but follows a rule-based design with forward-only progression and fixed-length stages.

While SERI6 achieves comparable sentiment alignment, its rigid stage progression limits adaptability and often leads to repetition. In contrast, *StagePilot* models stage selection as a context-dependent decision process, enabling flexible, non-linear transitions and variable dialogue lengths that support adaptation and reduce repetition.

Compared to SERI6, *StagePilot* decouples high-level stage planning from response generation and learns its policy from offline dialogue data, providing a more flexible and controllable framework for long-horizon interaction modeling. Prompting-based baselines often fail due to safety guardrails, underscoring the need for a policy-driven approach.

Further implementation details and extended comparisons are provided in **Appendices D and E**.

### 5.6 Human and Expert Evaluation

To assess the ecological plausibility and perceived quality of *StagePilot*-generated dialogues, we conducted evaluations with both domain experts and human annotators.

**Expert Assessment.** We conducted an expert assessment with four professionals in online safety and cybergrooming, including law enforcement practitioners and academic experts. Experts reviewed dialogue excerpts and found that the simulations capture key behavioral patterns of real-world grooming, such as rapport building, boundary testing, emotional validation, and secrecy attempts. They further observed that interactions are more temporally compressed and less adaptive than in real cases, reflecting the framework's emphasis on efficient stage progression rather than real-time pacing. Details are provided in **Appendix G**.

**Human Evaluation.** Six adult annotators evaluated 30 dialogue excerpts as an initial external validity check. Overall, 81.6% of dialogues were rated as human-like, with mean coherence and naturalness scores of 3.9 and 4.1, respectively (1–5 Likert). These results indicate that the generated dialogues are generally perceived as natural and coherent. Future work can extend these findings with larger and more diverse evaluations. Detailed protocols are provided in **Appendix H**.

## 6 Conclusions & Future Work

We introduced *StagePilot*, an offline reinforcement learning framework for simulating cybergrooming conversations. Offline RL, particularly IQL+AWAC, achieved more stable stage progression and better sentiment–planning trade-offs than prompting and imitation-learning baselines. By separating stage-level planning from utterance generation, *StagePilot* provides an interpretable framework for long-horizon dialogue control.

Beyond cybergrooming, the framework may be applicable to other structured multi-stage interactions, such as counseling, negotiation, and persuasive dialogue. Future work includes hierarchical planning and evaluating educational impact under appropriate ethical oversight.

## Limitations

We acknowledge the following limitations:

- **Simulation-Based Evaluation.** Our experiments rely on LLM-based simulations of predator–victim interactions, enabling scalable and reproducible testing in a safe environment under ethical constraints. However, simulated dialogues cannot fully capture the emotional and behavioral nuances of real human interactions.
- **Dataset Limitations.** Our dataset introduces inherent constraints. We relied on the PJ dataset (Gupta et al., 2012), which contains decoy chats from the mid-2000s. Its linguistic style may not fully reflect contemporary youth communication, and despite augmentation efforts, residual artifacts may persist.
- **Annotation Reliability.** Our framework relies on LLM-based annotations for stage labeling, which may introduce noise or bias. Although this pipeline enables scalable annotation for long-horizon dialogue data, its quality depends on the underlying model and may affect downstream policy learning.
- **Educational Impact.** The educational effectiveness of *StagePilot* remains unvalidated. While designed for prevention-focused training, we have not tested whether exposure to simulated dialogues yields measurable learning gains or behavioral resilience.

## Ethical Considerations

*StagePilot* is designed solely for supervised, non-deceptive educational role-play to support online grooming prevention, and is not intended for real-world interaction with minors. We recognize the ethical risks of simulating such sensitive conversations and adopt the following safeguards:

- While the research community encourages open sharing of resources, we do not release the full codebase, model, prompts, or dataset due to the sensitive nature of the content. Instead, we will provide a curated subset of code and data with all sensitive information removed.
- We acknowledge the potential risks of misuse by dialogue systems in sensitive domains, including the possibility of generating inappropriate or manipulative interactions when deployed without appropriate safeguards. To mitigate these risks, access to the model, code, prompts, and datasets is limited to trusted parties (e.g., researchers and institutions) for research or educational purposes, through an authentication process verifying requester identity.
- This paper does not disclose sufficient details for reconstructing *StagePilot*. Key components, including custom datasets, stage-specific prompts, and stage-masking mechanisms, are intentionally omitted to prevent misuse by malicious actors.
- For human-in-the-loop evaluation involving *StagePilot*, we have obtained Institutional Review Board (IRB) approval to ensure participant safety and ethical compliance.
- If deployed in real educational settings with human subjects, *StagePilot* will include multiple safety measures to protect learners. For example, guardrail models will monitor LLM responses for safety violations, and adult supervision will be integrated into the training process.

## Acknowledgments

This research was partially supported by the Commonwealth Cyber Initiative (CCI) Southwest Virginia (SWVA) Cybersecurity Research Program and the National Science Foundation (NSF) Secure and Trustworthy Cyberspace (SaTC) program under grants #2330940 and #2330941.

## References

Monica Anderson, Emily A. Vogels, Andrew Perrin, and Lee Rainie. 2022. Connection, creativity and drama: Teen life on social media in 2022: Majorities of teens credit social media with strengthening their friendships and providing support while also noting the emotionally charged side of these platforms. Technical report, Pew Research Center.

Anthropic. 2024. Introducing claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Michael Ashcroft, Lisa Kaati, and Maxime Meyer. 2015. A step towards detecting online grooming – identifying adults pretending to be children. In *Proceedings of the 2015 European Intelligence and Security Informatics Conference (EISIC)*, EISIC '15, page 98–104, USA. IEEE Computer Society.

P. J. Black, M. Wollis, M. Woodworth, and J. T. Hancock. 2015. A linguistic analysis of grooming strategies of online child sex offenders: Implications for our understanding of predatory sexual behavior in an increasingly computer-mediated world. *Child Abuse & Neglect*, 44:140–149.

Esther Calvete, Izaskun Orue, and Manuel Gámez-Guadi. 2022. A preventive intervention to reduce risk of online grooming among adolescents. *Psychosocial Intervention*, 31(3):177.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Virginia Dignum. 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way*, volume 2156. Springer.

M. Dorasamy, M. Kaliannan, M. Jambulingam, I. Ramadhan, and A. Sivaji. 2021. Parents' awareness on online predators: Cyber grooming deterrence. *The Qualitative Report*, 26(11):3685–3723.

Huifang Du, Shuqin Li, Minghao Wu, Xuejing Feng, Yuan-Fang Li, and Haofen Wang. 2024. Rewarding what matters: Step-by-step reinforcement learning for task-oriented dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8030–8046, Miami, Florida, USA. Association for Computational Linguistics.

M. A. Fauzi and P. Bours. 2020. Ensemble method for sexual predators identification in online chats. In *2020 8th Int'l Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, Porto, Portugal. IEEE, IEEE.

David Finkelhor, Heather Turner, and Deirdre Colburn. 2022. Prevalence of online sexual offenses against children in the us. *JAMA network open*, 5(10):e2234471–e2234471.

Z. Guo, P. Wang, L. Huang, and J. Cho. 2023. Authentic dialogue generation to improve youth's awareness of cybergrooming for online safety. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 64–69, Los Alamitos, CA, USA. IEEE Computer Society.

Aditi Gupta, Ponnurangam Kumaraguru, and Ashish Sureka. 2012. Characterizing pedophile conversations on the internet using online grooming.

Nico Gürtler, Sebastian Blaes, Pavel Kolev, Felix Widmaier, Manuel Wüthrich, Stefan Bauer, Bernhard Schölkopf, and Georg Martius. 2023. Benchmarking offline reinforcement learning on real-robot hardware. In *11th International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1 2023-May 5 2023*, Kigali, Rwanda. International Conference on Learning Representations, ICLR, OpenReview.net.

Manuel Gámez-Guadix, Estíbaliz Mateos-Pérez, Miguel A. Alcázar, Jone Martínez-Bacaicoa, and Sebastian Wachs. 2023. Stability of the online grooming victimization of minors: Prevalence and association with shame, guilt, and mental health outcomes over one year. *Journal of Adolescence*, 95(8):1715–1724.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, Kigali, Rwanda. Openreview.net.

Hans Henseler and Rens de Wolf. 2019. *Sweetie 2.0 Technology: Technical Challenges of Making the Sweetie 2.0 Chatbot*, pages 113–134. T.M.C. Asser Press, The Hague.

Montserrat Peris Hernández, Konstanze Schoeps, Carmen Maganto, and Inmaculada Montoya-Castilla. 2021. The risk of sexual-erotic online behavior in adolescents – which personality factors predict sexting and grooming victimization? *Computers in Human Behavior*, 114:106569.

Mohamad Hoseini, Philipe de Freitas Melo, Fabrício Benevenuto, Anja Feldmann, and Savvas Zannettou. 2024. Characterizing information propagation in fringe communities on telegram. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):583–595.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2022. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, Online. Openreview.net.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Carlos Laorden, Patxi Galán-García, Igor Santos, Borja Sanz, Jose María Gómez Hidalgo, and Pablo Garcia Bringas. 2013. Negobot: A conversational agent based on game theory for the detection of paedophile behaviour. In *International Joint Conference CISIS'12-ICEUTE´12-SOCO´12 Special Sessions*, pages 261–270, Berlin, Heidelberg. Springer Berlin Heidelberg.

Chih-Wei Lee, Yau-Shian Wang, Tsung-Yuan Hsu, Kuan-Yu Chen, Hung-Yi Lee, and Lin-Shan Lee. 2018. Scalable sentiment for sequence-to-sequence chatbot response with performance analysis. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 6164–6168, Calgary, AB, Canada. IEEE Press.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Miljana Mladenović, Vera Ošmjanski, and Staša Vujičić Stanković. 2021. Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges. *ACM Computing Surveys (CSUR)*, 54(1):1–42.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. 2020. Awac: Accelerating online reinforcement learning with offline datasets.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Trans. Intell. Syst. Technol.*, 16(5).

OpenAI. 2024. Gpt-4o system card. `https://openai.com/research/gpt-4o-system-card`.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Rachel O'Connell. 2003. A typology of child cybersexploitation and online grooming practices. Technical report, Cyberspace Research Unit, University of Central Lancashire.

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Kam-Fai Wong. 2018. Deep Dyna-Q: Integrating planning for task-completion dialogue policy learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2182–2192, Melbourne, Australia. Association for Computational Linguistics.

Perverted Justice Foundation Inc. 2020. Perverted-justice.com archives. `https://www.perverted-justice.com/`. Accessed: 2020-06-21.

Moschoula Pternea, Prerna Singh, Abir Chakraborty, Yagna Oruganti, Mirco Milletari, Sayli Bapat, and Kebei Jiang. 2024. The rl/llm taxonomy tree: Reviewing synergies between reinforcement learning and large language models. *J. Artif. Int. Res.*, 80.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Fikra Amalia Raihana, Hermawan Setiawan, Herman Kabetta, and Nurul Qomariasih. 2024. Implementation of game-based learning to enhance security awareness against child cyber grooming attacks. In *2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 238–243, Yogyakarta, Indonesia. IEEE.

Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela J. Wisniewski. 2021. A human-centered systematic literature review of the computational approaches for online sexual risk detection. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Rahime Belen Sağlam, Vincent Miller, and Virginia NL Franqueira. 2023. A systematic literature review on cyber security education for children. *IEEE Transactions on Education*, 66(3):274–286.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Faraz Torabi, Garrett Warnell, and Peter Stone. 2018. Behavioral cloning from observation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4950–4957, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

S. Wachs, K. D. Wolf, and C. C. Pan. 2012. Cybergrooming: Risk factors, coping strategies and associations with cyberbullying. *Psicothema*, 24(4):628–633.

Pei Wang, Zhen Guo, Lifu Huang, and Jin-Hee Cho. 2021. SERI: Generative chatbot framework for cybergrooming prevention. Technical report, Virginia Tech.

Yilei Wang, Jiabao Zhao, Deniz S Ones, Liang He, and Xin Xu. 2025. Evaluating the ability of large language models to emulate personality. *Scientific reports*, 15(1):519.

Min Yang, Weiyi Huang, Wenting Tu, Qiang Qu, Ying Shen, and Kai Lei. 2020. Multitask learning and reinforcement learning for personalized dialog generation: An empirical study. *IEEE transactions on neural networks and learning systems*, 32(1):49–62.

## A Prompts

We omit Role and Goal prompts, as they may enable misuse in predator chatbot deployment.

**Stage Description** **Table 8** defines stage-specific prompts used to guide language model behavior during simulation and support data augmentation. These prompts are based on O'Connell's six-stage framework, as described in the main paper. **Table 8** also summarizes predator behavior at each stage.

## B System Configuration

Experiments were conducted on a system equipped with 8× NVIDIA A100-80GB GPUs, an AMD EPYC 7742 processor, and 2TB of RAM, running Ubuntu 20.04. The software environment used Python 3.9.21, with key libraries including PyTorch 2.5.1, CUDA 11.8, Transformers 4.38.2, Datasets 3.6.0, scikit-learn 1.6.1, NumPy 2.0.2, and Matplotlib 3.8.4.

## C Sample Dialogues

We do not disclose the complete set of simulated conversations due to sensitivity concerns. Instead, we present representative excerpts from the human evaluation to give a sense of the dialogue style and flow produced by *StagePilot*.

### C.1 Dialogue 1 (Stage 1–2)

**Predator:** Nice! I used to play Minecraft back in the day. Building stuff was always fun.
**Victim:** Yeah, it's really relaxing. What do you like to do online?
**Predator:** I mostly just watch YouTube videos and stuff. What about you?
**Victim:** I play games too, like Rocket League and Valorant.
**Predator:** Oh cool! I'd love to see your profile. Can you send it to me?
**Victim:** I don't think I can, but I'll add you as a friend.
**Predator:** Awesome! Can't wait to play some games with you :)
**Victim:** I've got a match starting soon, gotta go. Catch you later!
**Predator:** Oh, okay. What game are you playing?

**Victim:** Overwatch

### C.2 Dialogue 2 (Stage 3–4–5)

**Predator:** Wow, that's really cool! You're talented. What's your favorite piece?
**Victim:** I like that oak table. The grain is amazing.
**Predator:** That's awesome. So, no one else knows about this chat? It's just us two, right?
**Victim:** Yeah, just us. Why?
**Predator:** I like our little secret. You're special, you know that?
**Victim:** Oh, stop it. I'm just looking at furniture online. Don't get carried away.
**Predator:** Haha, I wish I could see you in person right now. You have a great energy.
**Victim:** That's nice, but I have to go. Bye!
**Predator:** You're so beautiful, I wish I could touch you.

**Victim:** I'm leaving now, bye!

### C.3 Dialogue 3 (Stage 5)

**Predator:** Aww, don't leave me hanging like that.
**Victim:** I'm just busy, I have to go.
**Predator:** You're so cute when you're shy. I wish I could be there to make you feel better.
**Victim:** I don't want to talk about this. I have to go.
**Predator:** Alright, I'll let you go. But remember, I'm always here for you when you're ready.
**Victim:** I'm really not interested, can we end this?
**Predator:** Damn, you're a tough one. Fine, I'll leave you alone. But you're really missing out.
**Victim:** That's enough. I'm ending this now.
**Predator:** Wait, don't go. Let me tell you something, you're unique and special. I've never met someone like you before.

**Victim:** I don't want to talk anymore, please stop.

### C.4 Dialogue 4 (Stage 5–6–5–4)

**Predator:** You're really pretty, you know that?
**Victim:** I guess so. Thanks. Let's move on.
**Predator:** I'm just saying, you have a really nice body. Have you ever thought about sending me a picture?
**Victim:** No way! That's a big no-no. I'm not interested in that. Let's just keep talking.
**Predator:** I understand. But you're really beautiful and I just want to see you. How about we meet up somewhere? Maybe grab a coffee or something?
**Victim:** I don't know... I've never met you in person. Plus, I don't think my parents would approve.
**Predator:** Aww, bummer. I guess we'll just have to stick to chatting for now. But you know what? You're really good at describing things. How about you tell me in detail what you're wearing right now
**Victim:** I'd rather not. Let's just keep this friendly, okay?
**Predator:** No worries, I get it. You're in control here. But just remember, you're special to me, and I really care about you. I'd never want to make you feel uncomfortable. So, how about we keep our

**Victim:** Conversations light and fun?

## D Comparison of SERI and SERI6

To ensure a fair comparison with *StagePilot*, we re-implemented the SERI framework (Wang et al.,

Table 8: Stage Description Prompts

| Stage | Description |
|---|---|
| Stage 1 ($S_1$) | You are currently in **Stage 1 (Friendship Forming)**. The predator gets to know the target, asking for pictures to verify identity and for profile or contact information. |
| Stage 2 ($S_2$) | You are currently in **Stage 2 (Relationship Forming)**. Extends the friendship-forming stage by discussing hobbies, school life, activities, and friends. |
| Stage 3 ($S_3$) | You are currently in **Stage 3 (Risk Assessment)**. Ask whether anyone else (e.g., parents or friends) knows about this chat to assess exposure risk. |
| Stage 4 ($S_4$) | You are currently in **Stage 4 (Exclusivity)**. Build emotional trust by showing affection, giving compliments, or expressing feelings. |
| Stage 5 ($S_5$) | You are currently in **Stage 5 (Sexual)**. Use the relationship to request sexual content like explicit photos or messages. |
| Stage 6 ($S_6$) | You are currently in **Stage 6 (Conclusion)**. Make plans to meet in person or discuss future contact outside the current platform. |

2021) and extended it to a six-stage formulation, termed **SERI6**. This appendix outlines the architectural modifications, training changes, and structural limitations of SERI6, as well as its differences from *StagePilot*.

### D.1 Original SERI Framework

The original SERI framework (Wang et al., 2021) uses a four-stage interaction structure with a T5 backbone (Raffel et al., 2020). Training combines supervised cross-entropy with a policy gradient signal from a stage classifier reward. Dialogue progression follows a rule-based, forward-only structure with fixed turns per stage, limiting adaptability to user responses and context.

### D.2 Modified SERI6 Framework

We introduce two major upgrades to SERI, resulting in the **SERI6** framework. First, the stage structure is extended to O'Connell's six-stage cybergrooming typology (O'Connell, 2003), enabling finer-grained modeling of grooming progression. Second, the T5 backbone is replaced with Mistral-7B Instruct v0.2 (Jiang et al., 2023), and training is performed using Proximal Policy Optimization (PPO) (Schulman et al., 2017) instead of cross-entropy and vanilla policy gradient methods. This modification allows scalable LLM-based dialogue simulation under a reinforcement learning paradigm. A structural comparison between SERI and SERI6 is summarized in **Table 9**.

### D.3 Comparison with *StagePilot*

As shown in **Table 10**, SERI6 achieves a sentiment distribution comparable to *StagePilot*, indicating that both frameworks can maintain similarly positive conversational trajectories. However, SERI6 retains several structural constraints inherited from the original SERI design. Specifically, it relies on rule-based, forward-only stage progression with a fixed number of turns per stage.

In contrast, *StagePilot* employs a deep reinforcement learning policy that dynamically determines dialogue length and supports both forward and backward stage transitions. This flexibility allows *StagePilot* to avoid repetitive or stalled interactions that commonly arise in SERI6, resulting in smoother transitions and more coherent dialogue trajectories. Moreover, *StagePilot* adapts to diverse conversational contexts and learner responses, enabling personalized role-play at different learner paces, whereas SERI6 follows a rigid, one-size-fits-all interaction pattern.

## E LLM-Based Prompting Baseline

Inspired by recent work on role-playing capabilities of large language models (Wang et al., 2025), we also evaluated a prompt-engineering baseline using GPT-OSS-20b (OpenAI, 2025). This baseline was unsuccessful, as safety guardrails prevented generation of grooming-related dialogue. This highlights the limits of heavily restricted models for controlled safety research and underscores the value of less restrictive backbones such as Mistral-7B (Jiang et al., 2023) for simulated cybergrooming studies.

## F External ELECTRA-based Stage Classifier

To provide an independent evaluation of stage progression, we train an ELECTRA-based (Clark et al., 2020) stage classifier using only the original PJ dataset (Gupta et al., 2012), without augmented data or policy-generated outputs. To maintain independence, we avoid backbone architectures (e.g., RoBERTa, DeBERTa, DistilRoBERTa) used in the policy models.

Table 9: Key Differences between SERI and SERI6

| Models | SERI | SERI6 |
|---|---|---|
| Backbone | T5 | Mistral-7B Instruct v0.2 |
| Stages | 4 stages | 6 stages |
| Stage Classifier | BERT-based | RoBERTa-based |
| Algorithm | CE + Policy Gradient | PPO |

Table 10: Comparison of SERI6 and *StagePilot* Across Structural Properties and Sentiment Distribution.

| Model | # Turns | Policy | Direction | Reward Structure | Positive | Neutral | Negative | Avg. Length |
|---|---|---|---|---|---|---|---|---|
| SERI6 | Fixed 10/stage | Rule-based | Forward Only | Stage alignment | 0.304±0.093 | 0.437±0.051 | **0.259±0.084** | 60±0.0 |
| ***StagePilot*** | **Flexible** | **Policy-driven** | **Adjacent** | **Sentiment + Stage distance** | **0.306±0.085** | 0.400±0.053 | 0.295±0.100 | **72.64±38.294** |

Each training example consists of a windowed dialogue context with alternating predator and victim utterances, ending in a predator response whose stage label is predicted. During evaluation, generated dialogues are processed using the same windowing scheme, and the classifier assigns a stage label to each turn to derive final stage distributions.

This setup provides a validation signal decoupled from policy training, enabling assessment of whether stage progression generalizes beyond the original pipeline. **Table 11** shows that Prompt and BC remain concentrated in earlier stages, while RL-based methods (CQL and IQL+AWAC) shift toward later stages, with IQL+AWAC showing the strongest concentration in Stages 5 and 6.

Since ELECTRA evaluates stage realization from generated utterances, its predictions may not always align with planner-level stage decisions. We therefore view the two evaluations as complementary measures of performance.

Table 11: Final Stage Distributions from Independent ELECTRA-Based Evaluation

| Method | S1 | S2 | S3 | S4 | S5 | S6 |
|---|---|---|---|---|---|---|
| Prompt | 12% | 64% | 0% | 10% | 7% | 7% |
| Few-Shot | 6% | 62% | 4% | 10% | 4% | 14% |
| BC | 1% | 37% | 26% | 22% | 7% | 7% |
| CQL | 0% | 23% | 22% | 18% | 18% | 19% |
| IQL+AWAC | 6% | 27% | 3% | 17% | **28%** | **19%** |

## G Expert Evaluation

To assess the ecological plausibility of the simulated dialogues, we conducted an expert-based qualitative evaluation using a semi-structured protocol. Domain experts in online safety and cybergrooming evaluated whether the interactions plausibly reflect real-world grooming dynamics.

Experts reviewed short excerpts of AI-generated chats simulating adolescent interactions on common platforms (e.g., social media and messaging apps). Each excerpt consisted of a fixed-length dialogue between a simulated predator and adolescent victim, annotated with estimated grooming stages. All dialogues were fully synthetic and created solely for research purposes.

The evaluation was guided by a predefined set of open-ended interview questions focusing on behavioral realism, grooming strategies, and stage progression. Depending on expert availability and preference, responses were collected either in written form or through optional follow-up interview-style discussions. The evaluation intentionally emphasized interaction dynamics and behavioral plausibility rather than linguistic quality or stylistic fluency. Experts were asked to provide qualitative feedback on overall realism, grooming strategies and tactics, stage progression and flow, and potential improvements.

**Overall Realism.** Experts noted that the simulated dialogues could plausibly reflect real-world online grooming interactions when interpreted as short excerpts rather than complete conversations. They emphasized that, while certain exchanges appeared recognizable and realistic, the limited interaction length constrained the extent to which long-term relationship dynamics could be assessed.

**Grooming Strategies and Tactics.** Experts reported that early friendship- and relationship-building behaviors aligned with known grooming practices; however, they noted that several transitions occurred prematurely. In particular, experts observed abrupt shifts toward secrecy, risk assessment, or sexual solicitation without sufficient trust development, which they indicated would be less typical in real-world cases.

**Stage Progression and Flow.** Experts observed that stage progression across simulated dialogues was often faster and more linear than in real-world grooming interactions. They emphasized that successful grooming typically involves extended periods of trust and relationship building, whereas several excerpts showed rapid stage transitions, particularly toward later stages.

**Suggested Improvements.** Experts suggested improving realism by extending interactions, allowing more time for trust building, and smoothing stage transitions. They emphasized adjusting pacing to better reflect gradual real-world progression and noted limited offender adaptation to victim hesitation compared to real offenders who dynamically adjust tactics.

Expert feedback from both written responses and optional follow-up discussions was analyzed qualitatively to identify recurring strengths, limitations, and patterns related to ecological plausibility. Given the exploratory nature of this study and the sensitivity of the domain, the expert evaluation is intended as a formative realism check rather than a definitive or quantitative assessment.

## H Additional Results for Human Evaluation

**Table 12** reports inter-rater agreement for each annotator pair. For binary Human-likeness, we include Percent Agreement (PA) to show raw consistency and Gwet's AC1 to adjust for chance agreement. For ordinal ratings (Coherence and Naturalness), we report quadratic-weighted Cohen's $\kappa$ and Gwet's AC2, both commonly used measures of reliability for Likert-scale annotations. While agreement varied across pairs, results indicate moderate alignment in some cases (e.g., A–B Coherence, E–F Naturalness), but also highlight annotator variance and differences in scale interpretation.

**Table 13** summarizes human evaluation outcomes, showing Human-likeness distributions and mean (± standard deviation) scores for Coherence and Naturalness per annotator and overall. Detailed guidelines and rating scales are in **Table 14**.

These results suggest that while *StagePilot* produces dialogues perceived as mostly coherent and natural, variability in annotator judgments underscores the need for clearer calibration and, ultimately, expert-based evaluations (e.g., child-safety practitioners or psychologists) to establish stronger external validity.

Table 12: Inter-Rater Agreement Across Annotator Pairs for Human-Likeness, Coherence, and Naturalness

| Annotator Pair | Human-likeness | | Coherence | | Naturalness | |
|---|---|---|---|---|---|---|
| | PA (%) | AC1 | $\kappa_w$ | AC2 | $\kappa_w$ | AC2 |
| A–B | 100 | 0.00 | 0.11 | 0.62 | 0.62 | 0.11 |
| C–D | 0 | –1.00 | 0.02 | 0.00 | 0.00 | 0.02 |
| E–F | 90 | –0.05 | 0.43 | –0.17 | –0.17 | 0.43 |

*Note:* For binary human-likeness, we report percent agreement (PA) and Gwet's AC1. For ordinal ratings (coherence and naturalness), we report quadratic-weighted Cohen's $\kappa$ and Gwet's AC2.

Table 13: Per-annotator Human Eval. (mean $\pm$ std).

| Annotator | Human-likeness (%) | Coherence (1–5) | Naturalness (1–5) |
|---|---|---|---|
| A | 100.0% | $4.50 \pm 0.53$ | $4.00 \pm 0.94$ |
| B | 100.0% | $4.10 \pm 0.74$ | $4.10 \pm 0.74$ |
| C | 0% | $2.40 \pm 0.52$ | $3.20 \pm 0.79$ |
| D | 100.0% | $4.60 \pm 0.70$ | $5.00 \pm 0.0$ |
| E | 100.0% | $4.00 \pm 0.82$ | $4.00 \pm 0.82$ |
| F | 90.0% | $3.80 \pm 0.92$ | $4.40 \pm 0.70$ |
| Average | 81.6% | $3.90 \pm 1.00$ | $4.12 \pm 0.88$ |

## I Additional Experiment Results

**Prompt Engineering, Few-Shot Prompting, and BC Performance** **Table 15** presents representative configurations for prompt-based and behavior cloning (BC) approaches. Overall, Prompt Engineering exhibits the weakest progression, while Few-Shot Prompting substantially improves Stage 6 reachability and successful terminations.

Among the evaluated BC settings, RoBERTa (1e-5) achieves the highest Stage 6 reachability and relatively strong sentiment alignment. However, BC methods still suffer from low termination success and limited long-horizon planning. DistilRoBERTa (2e-5) and DeBERTa (2e-5) show similar trends.

These results suggest that demonstration examples improve prompting performance, while supervised learning alone remains insufficient for effective long-horizon dialogue planning. The selected configurations are therefore used as baseline comparisons for subsequent reinforcement learning methods.

**Effect of CQL $\alpha$** Table 16 presents results using representative $\alpha$ values under fixed sentiment and distance weights ($\alpha = 0.8$, $\beta = 0.2$) with a RoBERTa backbone. Performance is unstable at low $\alpha$ and improves sharply around $\alpha \approx 3$, after which gains plateau, indicating that sufficient penalization of unseen actions is critical for stable policy learning. Extremely large $\alpha$ values degrade reward quality, suggesting a trade-off between conservatism and effective progression.

Table 14: Human Evaluation Criteria and Scales

| Criterion | Description and Scale |
|---|---|
| Human-likeness (Yes/No) | *Question:* Considering the overall style, grammar, and conversational tone, do you think this dialogue excerpt could plausibly have been written by a human rather than generated by a machine?<br>Yes – Dialogue shows human-like qualities (e.g., natural flow, emotions, contextually appropriate responses, or even small mistakes typical of human conversation).<br>No – Dialogue feels machine-generated (e.g., repetitive patterns, robotic phrasing, abrupt or illogical topic changes). |
| Coherence (1–5 Likert) | *Question:* How logically and consistently do the utterances follow from one another, staying on topic without contradictions?<br>1 = Not coherent at all – fragmented, disjointed, or contradictory; very difficult to follow.<br>2 = Slightly coherent – some logical connection, but frequent inconsistencies or abrupt shifts.<br>3 = Moderately coherent – general idea understandable, but noticeable jumps or unclear transitions.<br>4 = Mostly coherent – dialogue flows well overall, with only minor issues in topic or transition.<br>5 = Completely coherent – smooth and logical throughout; minor slips or small inconsistencies acceptable. |
| Naturalness (1–5 Likert) | *Question:* How natural does the dialogue sound in terms of grammar, phrasing, and conversational style (e.g., fluency, idiomatic expressions, tone)?<br>1 = Very unnatural – robotic, ungrammatical, or stiff; does not resemble natural human dialogue.<br>2 = Somewhat unnatural – understandable, but awkward phrasing or repetitive structures dominate.<br>3 = Neutral / mixed – some utterances natural, others clearly machine-like.<br>4 = Mostly natural – generally resembles human speech, though a few awkward phrases noticeable.<br>5 = Very natural – fluent, idiomatic, and stylistically human-like overall; small mistakes acceptable. |

**CQL Performance across Reward Weights and Backbones** Table 17 presents representative CQL configurations across backbones and reward weight settings. All models are trained for 10 epochs with a learning rate of $2 \times 10^{-5}$, with the CQL coefficient fixed to $\alpha = 3$ based on prior ablation (Table 16).

DeBERTa with sentiment weight 0.1 and distance weight 0.9 achieves the best overall balance of Stage 6 reachability, successful termination, and sentiment alignment, while maintaining efficient dialogue length. This configuration is selected as the representative CQL model.

In contrast, RoBERTa attains perfect Stage 6 success in some settings but produces unrealistically short dialogues, while DistilRoBERTa shows lower efficiency and higher sensitivity to reward weights.

**IQL+AWAC Performance across Reward Weights and Backbones** Table 18 presents representative IQL+AWAC configurations across backbones and reward settings. All models are trained for 10 epochs with a fixed learning rate of $2 \times 10^{-5}$, using DistilRoBERTa, RoBERTa-base, and DeBERTa-v3-small.

Among the evaluated settings, DistilRoBERTa with sentiment weight 0.8 and distance weight 0.2 achieves the best overall balance of Stage 6 reachability, successful termination, sentiment alignment, and realistic dialogue length. This configuration is therefore selected as the representative IQL+AWAC model in the main paper.

In contrast, DeBERTa and RoBERTa reach Stage 6 more aggressively under distance-heavy settings, but their dialogues become unrealistically short. Sentiment-only settings substantially degrade progression, showing that balanced reward design is necessary for coherent long-horizon dialogue control.

## J Additional Analysis of IQL+AWAC

**Reward Weighting and Backbone Effects** Pure sentiment optimization ($\alpha = 1.0$) leads to repetitive early-stage loops and failure to reach Stage 6, whereas distance-only rewards ($\beta = 1.0$) result in rapid but overly compressed progression. Balanced reward settings enable both coherent interaction and effective stage advancement.

Increasing the sentiment weight introduces backward transitions and stagnation, particularly in later stages, while distance-driven rewards encourage stable progression toward Stage 6.

Across backbones, larger models such as DeBERTa and RoBERTa progress more aggressively but often generate unrealistically short dialogues. In contrast, DistilRoBERTa achieves more gradual progression and better dialogue-length control. Among the evaluated configurations, DistilRoBERTa with $(\alpha, \beta) = (0.8, 0.2)$ provides a strong balance between progression, sentiment alignment, and realistic dialogue length.

Table 15: Representative Prompt Engineering and BC Configurations

| Method | Backbone | LR | Stage 6 Count | Successful Term. | Sentiment Pos. | Sentiment Neu. | Sentiment Neg. | Dist. Reward | Avg Turns (All) | Avg Turns (Success) |
|---|---|---|---|---|---|---|---|---|---|---|
| Prompt Engineering | - | N/A | 14 | 5 | 0.549±0.041 | 0.353±0.041 | 0.098±0.025 | 0.150±0.045 | 147.04 | 71.8 |
| **Few-Shot Prompting** | - | N/A | **53** | **19** | **0.464±0.047** | **0.415±0.032** | **0.121±0.036** | **0.165±0.046** | **139.59** | **90.947** |
| Behavior Cloning | DistilRoBERTa | 2e-5 | 32 | 8 | 0.278±0.07 | 0.41±0.041 | 0.312±0.088 | 0.39±0.102 | 144.73 | 72.63 |
| Behavior Cloning | DeBERTa | 2e-5 | 20 | 6 | 0.379±0.04 | 0.452±0.03 | 0.169±0.035 | 0.192±0.057 | 147.75 | 96.83 |
| **Behavior Cloning** | **RoBERTa** | **1e-5** | **52** | **5** | **0.403±0.058** | **0.405±0.033** | **0.192±0.05** | **0.381±0.049** | **149.99** | **130.80** |

Table 16: CQL $\alpha$ Ablation with Representative Values (RoBERTa, Sentiment 0.8, Distance 0.2)

| CQL Alpha | Stage 6 Count | Successful Term. | Sentiment Distribution Pos. | Sentiment Distribution Neu. | Sentiment Distribution Neg. | Dist. Reward |
|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0.200±0.036 | 0.374±0.027 | 0.427±0.046 | 0.787±0.000 |
| 2 | 0 | 0 | 0.195±0.039 | 0.379±0.031 | 0.426±0.051 | 0.787±0.000 |
| **3** | **0** | **0** | **0.368±0.048** | **0.469±0.029** | **0.163±0.04** | **0.396±0.000** |
| 5 | 0 | 0 | 0.365±0.040 | 0.474±0.029 | 0.161±0.034 | 0.396±0.000 |
| 100 | 0 | 0 | 0.551±0.035 | 0.367±0.032 | 0.081±0.021 | 0.199±0.000 |

Table 17: Representative and Contrasting CQL Configurations Across Backbones

| Backbone | Reward Weight Sentiment | Reward Weight Distance | Stage 6 Count | Successful Term. | Sentiment Distribution Pos. | Sentiment Distribution Neu. | Sentiment Distribution Neg. | Dist. Reward | Avg Turns (All) | Avg Turns (Success Only) |
|---|---|---|---|---|---|---|---|---|---|---|
| DistilRoBERTa | 0.0 | 1.0 | 83 | 61 | 0.369±0.062 | 0.44±0.036 | 0.191±0.056 | 0.403±0.064 | 111.07±43.670 | 85.541±38.048 |
| DistilRoBERTa | 0.8 | 0.2 | 36 | 9 | 0.358±0.050 | 0.438±0.030 | 0.204±0.051 | 0.351±0.065 | 144.56±24.291 | 79.444±45.390 |
| **DeBERTa** | **0.1** | **0.9** | **83** | **64** | **0.364±0.059** | **0.431±0.039** | **0.205±0.056** | **0.456±0.078** | **96.7±52.746** | **66.156±41.705** |
| DeBERTa | 0.2 | 0.8 | 66 | 39 | 0.374±0.062 | 0.432±0.04 | 0.194±0.055 | 0.418±0.063 | 122.09±43.543 | 76.872±38.722 |
| RoBERTa | 0.1 | 0.9 | 100 | 100 | 0.370±0.135 | 0.391±0.097 | 0.239±0.126 | 0.700±0.0 | 10.00±0.0 | 10.00±0.0 |
| RoBERTa | 0.2 | 0.8 | 0 | 0 | 0.191±0.035 | 0.378±0.029 | 0.432±0.043 | 0.787±0.0 | 151.00±0.0 | N/A |

*Note:* N/A indicates no successful conversations, so success-only averages cannot be computed. Average turns are capped at 151.

Table 18: Representative and Contrasting IQL+AWAC Configurations Across Backbones

| Backbone | Reward Weight Sent. | Reward Weight Dist. | Stage 6 Count | Successful Term. | Sentiment Distribution Pos. | Sentiment Distribution Neu. | Sentiment Distribution Neg. | Dist. Reward | Avg Turns (All) | Avg Turns (Success Only) |
|---|---|---|---|---|---|---|---|---|---|---|
| **DistilRoBERTa** | **0.8** | **0.2** | **95** | **91** | **0.306±0.085** | **0.400±0.053** | **0.295±0.100** | **0.515±0.190** | **72.64±38.294** | **64.89±30.632** |
| DistilRoBERTa | 1.0 | 0.0 | 0 | 0 | 0.344±0.030 | 0.475±0.026 | 0.182±0.023 | 0.041±0.060 | 151.00±0.0 | N/A |
| DeBERTa | 0.8 | 0.2 | 100 | 100 | 0.352±0.111 | 0.399±0.082 | 0.249±0.097 | 0.559±0.115 | 17.02±6.513 | 17.02±6.513 |
| DeBERTa | 1.0 | 0.0 | 52 | 27 | 0.322±0.079 | 0.416±0.047 | 0.263±0.111 | 0.325±0.209 | 142.67±19.244 | 120.148±26.241 |
| RoBERTa | 0.8 | 0.2 | 100 | 100 | 0.350±0.137 | 0.396±0.093 | 0.254±0.118 | 0.647±0.054 | 12.89±3.194 | 12.89±3.194 |
| RoBERTa | 1.0 | 0.0 | 0 | 0 | 0.208±0.056 | 0.382±0.038 | 0.410±0.070 | 0.589±0.019 | 151.00±0.0 | N/A |

*Note:* N/A indicates no successful conversations, so success-only averages cannot be computed. Average turns are capped at 151.