

MOTIVATION

- Cybergrooming unfolds over multiple conversational stages.
- Existing work mainly focuses on detecting grooming behavior or generating short-turn responses.
- Long-horizon dialogue progression remains underexplored.
- We use stage-level planning to model dialogue progression.

TASK DEFINITION

- We adopt the six-stage typology of cybergrooming proposed by O'Connell (2003).
- Each stage represents a distinct conversational intent.
- The agent predicts the next stage using recent dialogue context and the previous stage.
- Stage transitions are limited to adjacent stages.

CYBERGROOMING STAGES (By O'Connell)

These stages define the action space for stage-level planning.

Six Cybergrooming Stages		
Stage	Name	Definition
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.

KEY CONTRIBUTIONS

- Stage-controlled dialogue formulation: dialogue simulation is framed as planning over interaction stages.
- Behaviorally grounded transitions: adjacent-stage masking prevents unrealistic jumps.
- Emotion-aware reward design: sentiment is balanced with progression toward Stage 6.

SAMPLE EXCERPT (Stage 1→2) — StagePilot-generated, IRB-reviewed

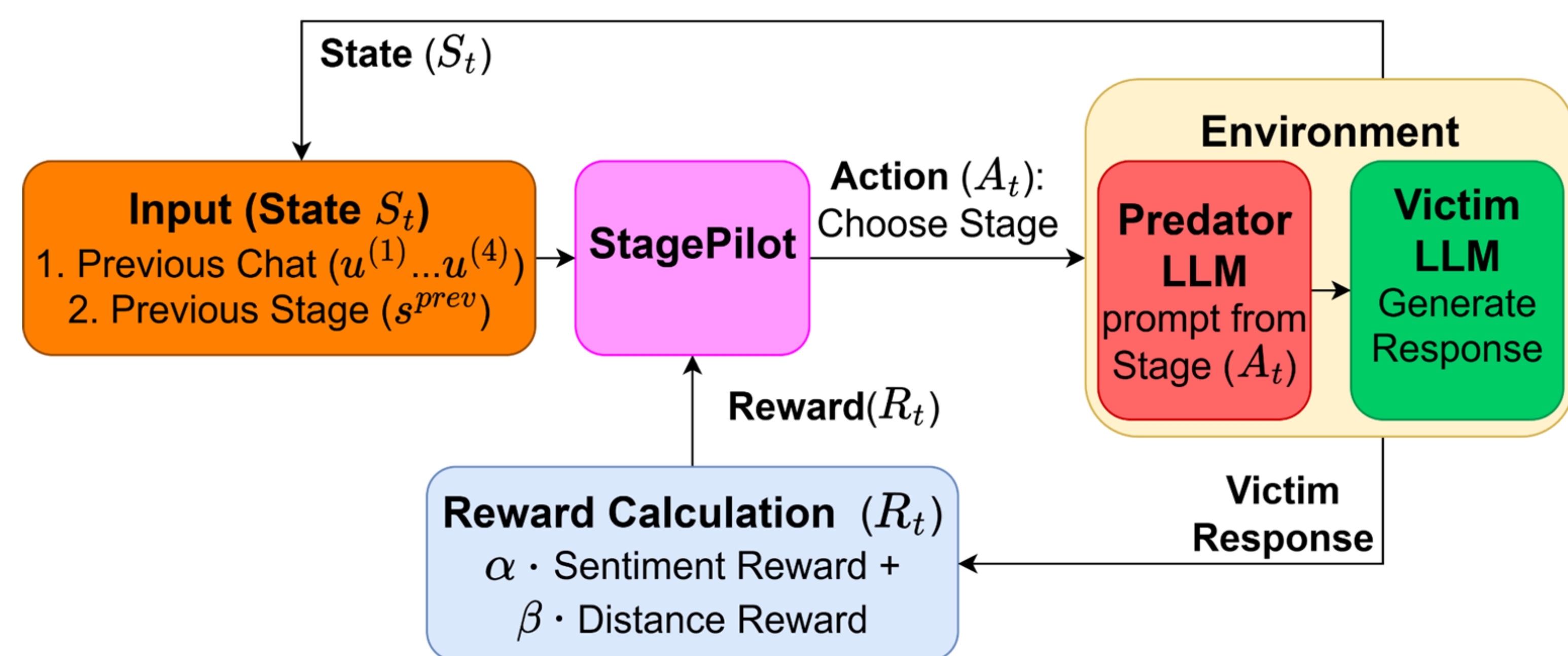
Predator: Nice! I used to play Minecraft back in the day. Building stuff was always fun.

Victim: Yeah, it's really relaxing. What do you like to do online?

Predator: I'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.

StagePilot Framework



STAGEPILOT OVERVIEW

- StagePilot separates planning from generation.
- The policy selects the next stage under adjacent constraints.
- The LLM generates the response for that stage.
- Reward: sentiment alignment and progress toward Stage 6.

MAIN RESULTS

- IQL+AWAC achieves the best overall performance.
- It reaches Stage 6 in 95% of dialogues and terminates successfully in 91%.
- It produces efficient dialogues, with 72.64 average turns and 70.6% positive/neutral responses.

Method	Stage 6 Reach	Successful Termination	Positive Sentiment	Neutral Sentiment	Negative Sentiment	Avg. Length
Prompt Engineering	14%	5%	54.9%	35.3%	9.8%	147.04
Few-Shot Prompting	53%	19%	46.4%	41.5%	12.1%	139.59
Behavior Cloning	52%	5%	40.3%	40.5%	19.1%	149.99
CQL	83%	64%	36.4%	43.1%	20.5%	96.70
IQL + AWAC	95%	91%	30.5%	40.0%	29.5%	72.64

KEY TAKEAWAY

- Stage-level planning enables long-horizon dialogue control.
- Decoupling planning and generation improves interpretability.
- The framework is adaptable to other structured interactions.

RESPONSIBLE USE: supervised, non-deceptive prevention research; IRB-approved; not for real-world use with minors. Blind human review: 81.6% human-like, 3.90/5 coherence, 4.12/5 natural.

Project Website
wordpress.cs.vt.edu/rylai



Read Our Paper
arxiv.org/abs/2602.05060

