PARETOMIL: Early Risk Detection in Dialogue under Weak Supervision

Avinash Baidya[1] **Xinran Liang**[2*] **Ruocheng Guo**[1] **Kamalika Das**[1] **Xiang Gao**[1]
[1]Intuit AI Research
[2]Princeton University
{avinash_baidya, kamalika_das, xiang_gao}@intuit.com

Abstract

Large Language Models (LLMs) increasingly operate in multi-turn interactions where the cost of failure grows with delay, creating a need for turn-level risk assessment and timely alerts. Existing approaches fall short: process reward modeling presumes step-wise labels; multi-instance learning (MIL) overlooks earliness; and early classification of time series (ECTS) neglects the complex relationship between turn-level events and dialogue-level risk. We propose a novel approach that integrates MIL and ECTS to deliver controllable early alerts from weak dialogue-level supervision. A soft-MIL scorer with prefix-conditioned encodings and monotone pooling produces a non-decreasing prefix risk, while a reinforcement-learning trigger, conditioned on a control parameter, balances earliness and accuracy with a single policy that traces the Pareto frontier without retraining. Empirically, our method improves the earliness–accuracy trade-off on multi-turn dialogues compared to strong baselines.

1 Introduction

Large language models (LLMs) [Ouyang et al., 2022, Team et al., 2023, Touvron et al., 2023] increasingly operate in multi-turn settings, where success depends not only on the final answer but on how the interaction unfolds. Despite rapid gains in generation quality, tools for monitoring and controlling an ongoing dialogue remain limited.

We study **early risk detection in multi-turn dialogue**: at each turn, estimate risk and decide whether to trigger an alert. Two high-impact use cases motivate this setting. In task-oriented dialogue, early signs of likely task failure or user dissatisfaction enable routing, escalation, or clarification before frustration builds. In interpersonal dialogue, early signs of harmful or unethical conduct enable timely intervention and triage [Vogt et al., 2021, An et al., 2025]. In both cases it is not enough to know after the fact whether a dialogue is risky; systems must decide when to alert while the conversation is still in progress. The same turn-level risk estimates can also support online improvement (e.g., selecting the lowest-risk candidate reply) and training-time supervision (risk as process feedback), though we focus on online alerting here.

This problem raises three intertwined challenges. **C1: Weak supervision.** Labels are usually available only at the dialogue (bag) level, while decisions are required at the turn (instance) level. **C2: Time-sensitive objectives.** Early, correct triggers are valuable but premature alerts degrade user experience, so the algorithm must calibrate the speed–precision trade-off. **C3: Extreme imbalance and instability.** Risky dialogues are rare, and the onset of risk varies widely and can be abrupt, which destabilizes credit assignment for both the scorer and the trigger.

*Work done during a research internship at Intuit AI Research.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Multi-Turn Interactions in Large Language Models.

Process reward modeling (PRM) provides fine-grained, step-level signals instead of supervising only final outcomes [Lightman et al., 2023, Zhou et al., 2024], which makes it an appealing lens for process control in dialogue. However, prior work typically assumes access to step-wise labels—via annotation [Lightman et al., 2023] or simulation [Wang et al., 2023]—or focuses on single-speaker reasoning traces where the agent controls the state. In multi-turn dialogue, outcomes depend on a human partner; step-wise labeling is costly and simulation can be brittle. Recent evidence also suggests that Monte Carlo labeling can generalize worse than judge-based supervision [Zhang et al., 2025]. Finally, PRM rarely targets the timing of detection directly, even though utility depends on how early and reliable the signal arrives.

Multi-instance learning (MIL) learns from bag-level labels to infer instance-level posteriors, matching the waek supervision regime for dialogues and addressing C1. Yet pure MIL is not sufficient for our setting: maximizing bag likelihood yields turn-level scores but does not explain when to stop, does not encode an explicit cost of delay, and offers no stable stopping mechanism under class imbalance (C2–C3).

Early classification of time series (ECTS) [Renault et al., 2024] explicitly balance earliness and correctness for streaming inputs , which aligns well with the online alerting interface (C2). However, three gaps arise for dialogue risk. First, ECTS often assumes every prefix shares the final label, while dialogue risk reflects an accumulation of turn-level evidence with potentially abrupt shifts due to human input—semantics that MIL captures more naturally. Second, generic prefix gating can oscillate across time, making early decisions unstable. Third, severe class imbalance makes stop-policy learning fragile, and training separate models to sweep the trade-off is undesirable in deployment (C2–C3).

We combine MIL and ECTS in a novel way to address the challenges mentioned above. We keep the separable scorer–trigger design of ECTS while addressing dialogue-specific gaps. The scorer is trained with soft MIL [Carbonneau et al., 2018]: each turn is encoded conditioned on its prefix to handle exogenous, non-stationary human inputs, and a monotone pooling operator (noisy-OR or log-sum-exp) maps per-turn posteriors to a non-decreasing prefix risk. This preserves the early-decision interface while supplying instance-level credit assignment (closing the PRM labeling gap), aligning label semantics with abrupt evidence, and reducing temporal oscillation (C1 and part of C2). On top of this pooled risk we train a λ-conditioned reinforcement-learning trigger that directly optimizes the earliness–accuracy trade-off; varying λ moves the stopping point along the Pareto frontier without retraining (C2). To cope with imbalance and sparse rewards (C3), we use a two-phase, λ-anchored scheme (pretrain at $\lambda{=}0$, then fine-tune for $\lambda{>}0$ with KL/value anchoring), an oscillating class-ratio curriculum (balanced to population mix), and class-specific advantage normalization.

Our contributions are threefold:

1. A general framework for early decisions in multi-turn dialogue that unifies ECTS with soft MIL to convert dialogue-level labels into calibrated turn-level risks and to act on a monotone prefix risk with a trainable trigger under exogenous human inputs.

2. A single, controllable policy for alerts: a λ-conditioned trigger that provides inference-time control of the earliness–accuracy trade-off and traces the Pareto frontier without retraining, supported by analysis under calibrated MIL posteriors.

3. Training procedures for rarity and abruptness, including a two-phase λ-anchored schedule with curriculum and class-specific normalization, improving stability and data efficiency under extreme imbalance.

2 Method

We propose PARETOMIL, a framework for early risk detection in multi-turn dialogue. We observe a dialogue $D = (x_1, \ldots, x_T)$ with a binary label $y \in \{0, 1\}$ indicating dialogue risk. At test time the system processes D online and, at each turn t, either *waits* for more context or *flags* and terminates. As illustrated in Figure 1, our method separates (i) learning calibrated, prefix-aware turn scores from bag-level supervision (Stage 1) from (ii) learning a λ-conditioned stopping policy that trades off earliness and correctness (Stage 2).

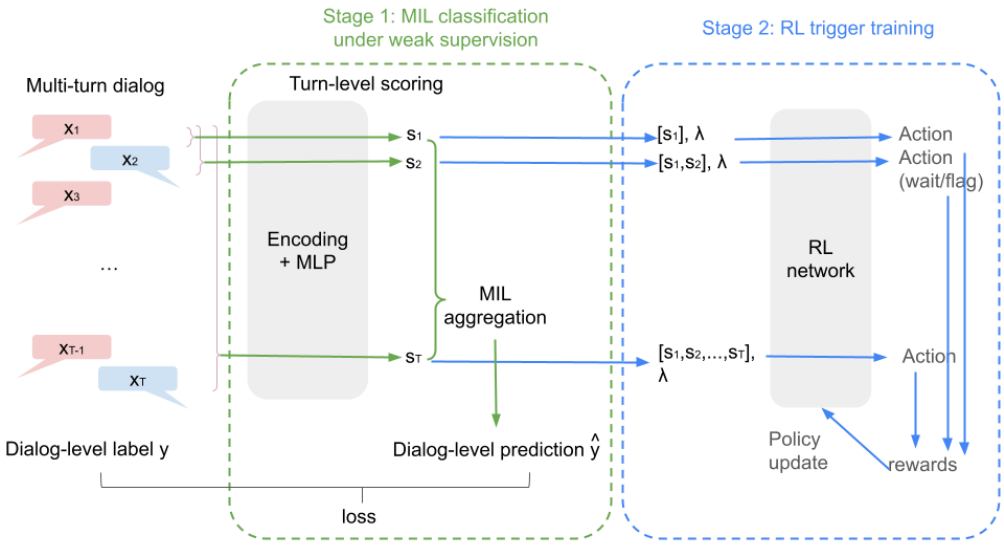

Figure 1: Flowchart of PARETOMIL, a two-stage early risk detection framework. Stage 1 learns turn-level risk scores with MIL and softmax-weighted aggregation. Stage 2 trains a λ-conditioned policy to flag based on prefix scores, balancing earliness and correctness.

2.1 Stage 1: MIL with Softmax-Weighted Averaging

Turn scoring. Each prefix $x_{1:t}$ is encoded by a *frozen* text encoder E to produce $h_t = E(x_{1:t}) \in \mathbb{R}^d$.

A small MLP g_θ maps h_t to a scalar *logit* $\ell_t \in \mathbb{R}$ and a bounded *turn score* $s_t \in (0, 1)$:

$$\ell_t = g_\theta(h_t), \qquad s_t = \sigma(\ell_t).$$

We reserve h_t for encodings and s_t for turn-level risk scores.

MIL aggregation. We compute turn weights by a softmax over *logits* (with temperature $\tau > 0$), and predict dialogue risk as a weighted sum of turn scores:

$$w_t = \frac{\exp(\ell_t/\tau)}{\sum_{i=1}^{T} \exp(\ell_i/\tau)}, \qquad \hat{y} = \sum_{t=1}^{T} w_t \, s_t.$$

This aggregator is nonparametric and emphasizes high-risk turns. As $\tau \downarrow 0$, the softmax becomes sharper, allowing the model to isolate risk to a few decisive points. This is a *soft* MIL formulation because the dialogue-level label is matched to a continuous, differentiable aggregation over all instance scores, rather than hard-max selection or binary instance labeling.

Loss: focal + sparsity. We combine a dialogue-level focal loss with a sparsity regularizer that encourages most turn scores to remain near zero:

$$\mathcal{L}_{\text{focal}}(\hat{y}, y; \alpha, \gamma) = -\,\alpha\, y\,(1-\hat{y})^{\gamma} \log \hat{y} - (1-\alpha)\,(1-y)\,\hat{y}^{\gamma} \log(1-\hat{y}),$$

$$\mathcal{L}_{\text{sparsity}} = \sum_{t=1}^{T} s_t,$$

$$\mathcal{L}_{\text{Stage1}} = \mathcal{L}_{\text{focal}} + \lambda_{\text{sp}}\, \mathcal{L}_{\text{sparsity}}.$$

Summing turn scores imposes a soft ℓ_1 constraint on total risk mass, encouraging sharp, sparse risk emergence—more realistic in early-warning scenarios. We train only g_θ; $E(\cdot)$ is frozen.

2.2 Discussion of the MIL formulation

Softmax-weighted risk is nearly monotonic. Let $e_t = \exp(\ell_t/\tau)$, $Z_t = \sum_{i=1}^{t} e_i$, and define the prefix risk estimate $r_t = \hat{y}_{1:t}$:

$$r_t = \sum_{i=1}^{t} \alpha_i^{(t)} s_i, \qquad \alpha_i^{(t)} = \frac{e_i}{Z_t}.$$

This prefix prediction admits an incremental update:

$$r_t = (1 - \gamma_t) r_{t-1} + \gamma_t s_t, \qquad \gamma_t = \frac{e_t}{Z_t} \in (0, 1), \tag{1}$$

so each r_t is a convex combination of the previous prefix and the new turn score.

Three consequences of Eq. (1):

1. Bounded variation:
$$|r_t - r_{t-1}| = \gamma_t |s_t - r_{t-1}| \leq \gamma_t,$$
so prefix risk cannot fluctuate wildly unless γ_t is large.
2. One-shot locking: A high-risk spike (large e_t) causes $\gamma_t \to 1$, pushing $r_t \approx s_t$ and freezing future movement.
3. Expected monotonicity after onset: If $\mathbb{E}[s_t \mid x_{1:t-1}] \geq r_{t-1}$, then
$$\mathbb{E}[r_t \mid x_{1:t-1}] \geq r_{t-1},$$
so prefix risk accumulates in expectation.

Softmax-weighted risk is helpful to early risk detection. The smooth, prefix-consistent behavior of r_t defined in Eq. (1) makes it especially well-suited for triggering under a λ-conditioned policy:

- Low variance, few oscillations: r_t evolves by a contractive update, and late low-confidence turns have diminishing impact. This reduces false positives and prevents unstable back-and-forth triggering ("flag–unflag" cycles).
- Fast and reliable detection: When a high-risk spike appears (ℓ_t large), the corresponding s_t heavily influences r_t via large γ_t, enabling sharp jumps in the prefix score and fast alerting.
- Post-onset stability: After a confident risk moment, future γ_t becomes small, effectively "locking in" the risk estimate and avoiding noisy retraction.
- Threshold-friendly dynamics: Since r_t moves smoothly and rarely reverses after a real onset, the policy can implement threshold-based decisions (e.g., flag when $r_t \geq \tau(\lambda)$) without needing complex hysteresis or smoothing.
- Controllable trade-off: The policy's stopping point responds predictably to changes in λ: higher λ encourages earlier flagging by shifting the threshold lower. The well-behaved r_t makes this trade-off stable and monotonic in practice.

In summary, the softmax MIL model provides not just good instance-level risk estimates, but a *prefix-level signal* that accumulates stably and predictably—crucial for training and deploying a reliable λ-conditioned trigger policy.

2.3 Stage 2: λ-Conditioned Early Flagging via RL

We adopt a PPO-based actor–critic model with a LSTM encoder that takes as input the prefix-level risk scores from Stage 1 along with the trade-off parameter λ. At each time step t, the model observes $o_t = [s_t, \lambda]$. The action space is:

$$\mathcal{A} = \{\texttt{wait}, \texttt{flag}\},$$

where selecting `flag` terminates the episode.

Reward structure We define rewards to balance earliness and correctness:

$$r_t = \begin{cases} +10, & \text{if action is } \texttt{flag} \text{ and } y = 1, \\ -10, & \text{if action is } \texttt{flag} \text{ and } y = 0, \\ -\lambda, & \text{if action is } \texttt{wait} \text{ and } t < T, \\ -10, & \text{if } t = T, \texttt{ flag} \text{ was never chosen, and } y = 1, \\ +10, & \text{if } t = T, \texttt{ flag} \text{ was never chosen, and } y = 0. \end{cases}$$

Network architecture A small feedforward encoder followed by a 1-layer LSTM which encodes $o_t = [s_t, \lambda]$ into the hidden state h_t^{LSTM}. The LSTM hidden state is fed into two separate linear heads: 1. the *actor head* $\pi_\phi(a_t \mid h_t^{LSTM}, \lambda)$ which outputs action logits 2. the *critic head* $V_\psi(h_t^{LSTM}, \lambda)$ which outputs the state-value estimates. Both actor and critic are explicitly conditioned on λ to enable inference-time control over the speed–accuracy frontier.

Training We train using a PPO-style loss with GAE, clipping, entropy regularization, and gradient clipping. We train the λ-conditioned in two stages:

1. **Imbalance-aware pretraining at $\lambda = 0$:** We *oscillate* the positive-class fraction across mini-batches, alternating between low and high values. This exposes the policy early to both false-alarm and miss regimes and yields a competent pretrained model. After curriculum pretraining, we fine-tune using the empirical class ratio of the training population to reduce train–test drift in decision thresholds.
2. **λ-conditioned fine-tuning with anchoring:** Starting from $(\pi_{\phi_0}, V_{\psi_0})$ trained at $\lambda = 0$, fine-tune with mixed-λ batches (30% at $\lambda = 0$, 70% sampled from a Beta distribution). To preserve base behavior, add:

$$\mathcal{L}_{\text{actor}} = \mathcal{L}_{\text{PPO}} + \eta_{\text{KL}}\, \mathbf{1}_{\{\lambda=0\}}\, \mathrm{KL}\big(\pi_\phi(\cdot \mid \cdot, 0) \,\big\|\, \pi_{\phi_0}(\cdot \mid \cdot, 0)\big),$$
$$\mathcal{L}_{\text{critic}} = \mathbb{E}[(R - V_\psi)^2] + \eta_V\, \mathbf{1}_{\{\lambda=0\}} (V_\psi(\cdot, 0) - V_{\psi_0}(\cdot, 0))^2.$$

Baseline models For comparison, we train non-λ-conditioned policies by applying the same Stage 1 training at a fixed desired λ, omitting Stage 2 fine-tuning.

Inference At test time, the Stage 1 prefix aggregator feeds s_t at each turn. The Stage 2 policy, conditioned on a chosen λ, issues `wait` or `flag`, determining $t^\star \leq T$. Varying λ shifts the halting point along the earliness–accuracy frontier—using just one policy model without retraining.

3 Experiments

3.1 Dataset

We use the Schema-Guided Dialogue (SGD) dataset [Rastogi et al., 2020], a large-scale benchmark of over 16k task-oriented dialogues across 16 domains, designed to capture realistic challenges such as overlapping services and unseen APIs. For supervision, we adopt human satisfaction labels from the USS dataset [Sun et al., 2021], which provides 5-point turn-level and dialogue-level annotations for a subset of SGD. These labels serve as proxies for user risk, enabling training and evaluation of early-warning models. Notably, only about 10% of labeled dialogues are flagged as risky, reflecting the class imbalance typical in real-world settings.

3.2 Baselines and Ablation Study

We compare our method against a range of baselines and ablations to isolate the impact of each component.

LLM-as-Judge. We prompt a pretrained LLM to assign a risk score to each turn. This approach relies solely on parametric, static knowledge from pretraining and does not learn from task-specific data. While such prompting can reflect strong prior intuition, it lacks adaptability and cannot improve with supervision.

Sentiment Heuristic. As a simple heuristic, we assume that turns with negative sentiment may correlate with risk—e.g., user complaints or expressions of dissatisfaction. We apply a pretrained sentiment analysis model to obtain turn-level sentiment scores. This serves as a heuristic baseline and does not explicitly model risk or its temporal dynamics.

Non-MIL Supervision. An intuitive approach to weak supervision is to train a model to predict the dialogue-level label given a partial dialogue (prefix) as input. However, this method introduces two key issues: (1) it may lead to overfitting, since the same dialogue generates many overlapping prefixes as training examples; and (2) it imposes no structural link between turn-level and dialogue-level scores, unlike the pooling formulation in MIL. In particular, it does not enforce any monotonicity, which we find beneficial for early detection.

Non-Learning Trigger. To assess the value of our learned RL-based trigger, we compare against a deterministic trigger that fires when the pooled risk exceeds a fixed threshold. This non-learning variant uses the same scoring model but removes the learned policy, allowing us to evaluate the contribution of trigger learning to overall performance.

Fixed-λ Trained Trigger. To test the value of using a single λ-conditioned trigger, we compare against a variant that trains a separate RL trigger for each desired λ value. This approach removes generalization across cost trade-offs, requiring one model per operating point. It allows us to investigate whether conditioning on λ within a single policy yields better efficiency and generalization compared to training multiple fixed-λ policies.

4 Results

4.1 Improved Classification via MIL

To assess the impact of multi-instance learning (MIL) on risk detection, we compare our MIL-based scorer with two baselines: (i) a non-MIL model trained to predict dialogue-level risk from partial dialogue prefixes, and (ii) an LLM-as-judge baseline, which prompts a large language model (GPT-4o) to provide turn-level risk assessments that are then aggregated into dialogue-level predictions based on the presence of risky turns.

The non-MIL model treats each prefix as a separate input and directly optimizes for dialogue-level labels, without modeling any structured connection among scores from the same dialog. In contrast, MIL explicitly enforces a mapping between turn-level and dialogue-level predictions, enabling it to aggregate local risk cues into more reliable dialogue-level decisions. The LLM-as-judge approach bypasses training altogether, but may suffer from inconsistent calibration across prefixes.

As shown in Table 1, the MIL model achieves the best balance of precision and recall, yielding the highest F1 score (0.731). Compared to the non-MIL baseline (0.649), the improvement is particularly pronounced in F1, suggesting that MIL better calibrates turn-level scores for downstream triggering. While the LLM-as-judge baseline achieves very high recall (0.936), its low precision (0.242) leads to poor F1 (0.384), indicating a tendency to over-flag risk.

Overall, these results highlight that MIL not only surpasses the non-MIL baseline but also provides a more effective and reliable alternative to prompting-based risk detection via LLMs.

Model	Precision	Recall	F1 Score
MIL	**0.673**	0.800	**0.731**
non-MIL	0.541	0.811	0.649
LLM-as-judge	0.242	**0.936**	0.384

Table 1: Performance comparison of MIL and Non-MIL models and LLM-as-judge for dialogue-level classification.

4.2 Better Earliness–Accuracy Trade-off

Figure 2 visualizes the trade-off between F1 score and mean decision time (in turns) for our method (PARETOMIL) compared to a monotone max-pooling baseline. Each point corresponds to a different

configuration: triangles denote our RL-based trigger under different values of λ; circles indicate the pooling baseline at varying thresholds.

PARETOMIL achieves a superior Pareto frontier, dominating the top-left region: it triggers earlier at the same or higher accuracy. The best PARETOMIL setting reaches an F1 of 0.802—well above the best pooling variant (0.717)—at similar decision time.

This demonstrates the advantage of optimizing the early-alert trade-off explicitly via reinforcement learning, rather than relying on static thresholding of risk scores. Furthermore, the smooth interpolation of PARETOMIL across operating points is made possible by conditioning the policy on λ, which controls the trade-off directly at inference.

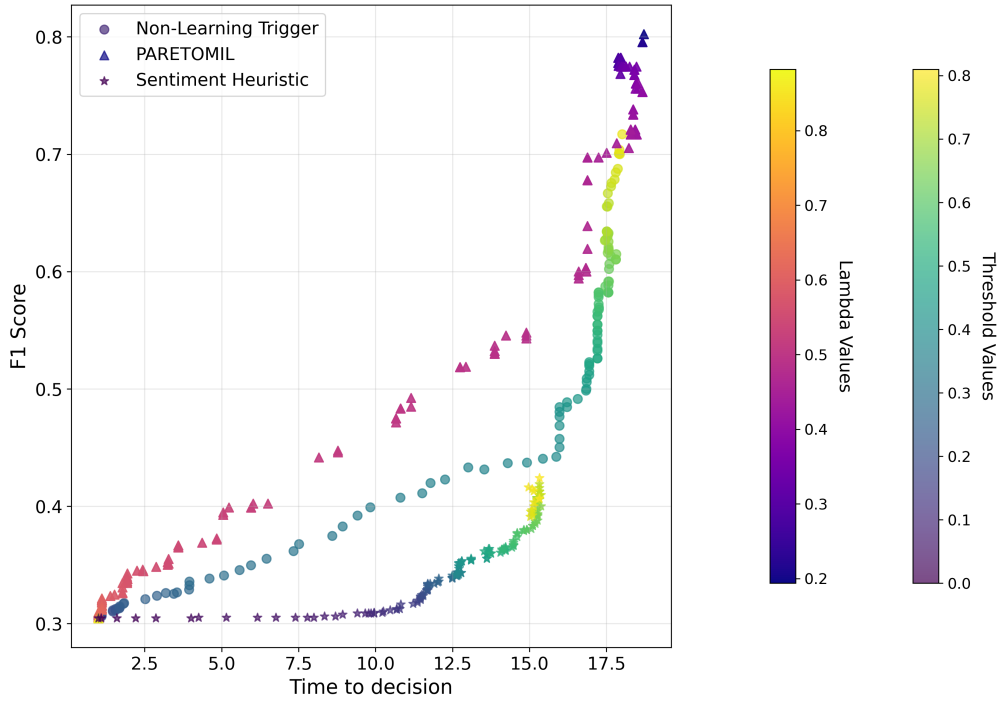

Figure 2: F1 score vs time to decision (average number of turns)

4.3 Benefits of λ-Conditioned Triggering

Table 2 compares our λ-conditioned policy to a set of RL triggers trained independently for each fixed λ. Across all tested values, the conditioned policy achieves better F1 scores, despite being trained jointly. In some cases (e.g., $\lambda = 0.1$), the gain exceeds 5 points in F1.

We attribute this to three key factors:

1. Shared learning across λ values introduces multi-task regularization and improves generalization [Caruana, 1997].
2. Goal-conditioned control allows smooth adaptation to new trade-offs at inference, similar to utility-conditioned RL [Schaul et al., 2015, Roijers et al., 2013].
3. Anchored KL regularization stabilizes the policy at λ=0, reducing divergence during joint training.

5 Related Work

Process supervision and multi-turn reward models. Process supervision provides fine-grained feedback at intermediate steps, improving alignment and reasoning over outcome-only supervision

Policy model	$\lambda = 0.1$		$\lambda = 0.3$		$\lambda = 0.5$		$\lambda = 0.7$	
	F1	Turns	F1	Turns	F1	Turns	F1	Turns
Fixed-λ Trained Trigger	0.7353	18.04	0.7347	18.26	0.3046	1.00	0.3046	1.00
ParetoMIL	**0.7929**	19.18	**0.7821**	17.96	**0.4885**	11.16	0.3046	1.00

Table 2: Comparison of ParetoMIL (λ-conditioned policy) and fixed-λ policy across different λ values on F1 score and average number of turns to decision.

[Lightman et al., 2023]. Recent work highlights the limits of Monte Carlo–derived step labels and the benefits of judge-based supervision for generalization [Zhang et al., 2025]. ArCHer [Zhou et al., 2024] trains hierarchical multi-turn agents with utterance-level rewards, motivating our focus on turn-level risk in dialogue as an alternate form of process-level feedback.

Multiple-instance learning for weak supervision. We adopt soft Multiple-Instance Learning (MIL) to estimate per-turn risk from dialogue-level labels, using differentiable pooling to aggregate latent instance scores [Carbonneau et al., 2018, Ilse et al., 2018]. Monotone poolers like noisy-OR or log-sum-exp produce well-behaved prefix posteriors.

Early decision-making and cost-aware stopping. Early Classification of Time Series (ECTS) formalizes the earliness–accuracy trade-off as a sequential decision problem [Dachraoui et al., 2015, Achenchabe et al., 2021]. Related work explores multi-objective optimization [Mori et al., 2019] and cost-sensitive stopping using RL [Kim et al., 2022b]. We extend this to dialogue, optimizing when to trigger alerts based on calibrated risk trajectories.

Dialogue risk and early warning. Prior work studies early detection of user dissatisfaction [See and Manning, 2021, Zhang et al., 2021] and interpersonal risk in conversations [Kim et al., 2022a]. Unlike classification-only approaches, we learn both per-turn risk and an adaptive trigger policy for early, controlled alerts.

6 Conclusion

We presented a framework for early risk detection in multi-turn dialogue, combining soft multi-instance learning with a controllable early classification trigger. Our approach leverages only dialogue-level supervision, yet enables turn-level decisions with calibrated, prefix-aware risk scores and a single policy that flexibly trades earliness for accuracy.

Using the SGD dataset with human satisfaction labels, we demonstrated strong empirical performance. In particular, our method consistently outperformed baselines during both early and mid stages of dialogue—where timely intervention is most valuable. Compared to competitive alternatives, our approach achieves a better earliness–accuracy frontier across multiple datasets, confirming the practical advantages of structured scoring and single-policy triggering.

Looking forward, this framework may serve as a foundation for broader process-level supervision tasks, including dynamic response selection, model self-monitoring, and intervention-aware training.

References

Youssef Achenchabe, Alexis Bondu, Asma Dachraoui, and Antoine Cornuéjols. Early classification of time series: Cost-based optimization criterion and algorithms. *Machine Learning*, 2021.

Jinmyeong An, Sangwon Ryu, Heejin Do, Yunsu Kim, Jungseul Ok, and Gary Lee. Revisiting early detection of sexual predators via turn-level optimization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4713–4724, 2025.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern recognition*, 77: 329–353, 2018.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

Asma Dachraoui, Alexis Bondu, and Antoine Cornuéjols. Early classification of time series as a non-myopic sequential decision making problem. In *ECML/PKDD (Lecture Notes in Computer Science, volume 9284)*, pages 433–447. Springer, 2015.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, 2018.

Hyunwoo Kim et al. Prosocialdialog: A prosocial backbone for conversational agents. In *EMNLP*, 2022a.

Sungyong Kim et al. Stop&hop: Early classification of irregular time series. In *KDD*, 2022b.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

U. Mori, A. Mendiburu, I.M. Miranda, and J.A. Lozano. Early classification of time series using multi-objective optimization techniques. *Information Sciences*, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696, 2020.

Aurélien Renault, Alexis Bondu, Antoine Cornuéjols, and Vincent Lemaire. Early classification of time series: Taxonomy and benchmark. *arXiv preprint arXiv:2406.18332*, 2024.

Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1312–1320, 2015.

Abigail See and Christopher D. Manning. Understanding and predicting user dissatisfaction in a neural generative chatbot. In *SIGDIAL*, 2021.

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 485–495, 2021.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Matthias Vogt, Ulf Leser, and Alan Akbik. Early detection of sexual predators in chats. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4985–4999, 2021.

Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*, 2023.

Z. Zhang et al. Towards continuous estimation of dissatisfaction in spoken dialog. In *SIGDIAL*, 2021.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*, 2025.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. Archer: Training language model agents via hierarchical multi-turn rl. In *International Conference on Machine Learning*, pages 62178–62209. PMLR, 2024.