PRISM: Festina Lente Proactivity—Risk-Sensitive, Uncertainty-Aware Deliberation for Proactive Agents

Yuxuan Fu Xiaoyu Tan Teqi Hao Chen Zhan Xihe Qiu

Shanghai University of Engineering Science, National University of Singapore

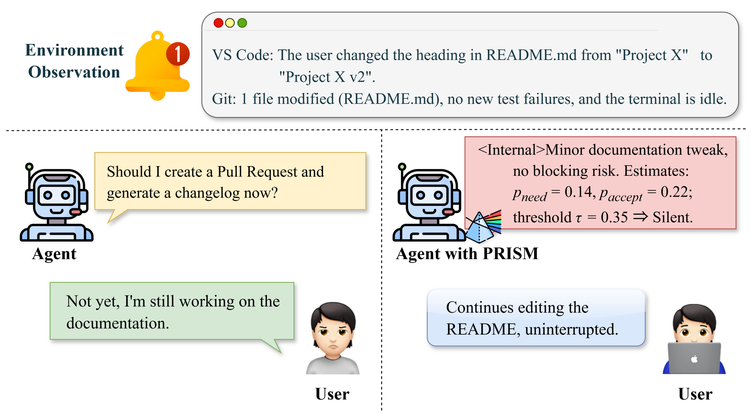
ICLR International Conference On Learning Representations

Why Proactive Agents Often Fail

Proactive agents must decide not only **what** to say, but also **whether** and **when** to intervene. In real deployments, this decision is difficult because:

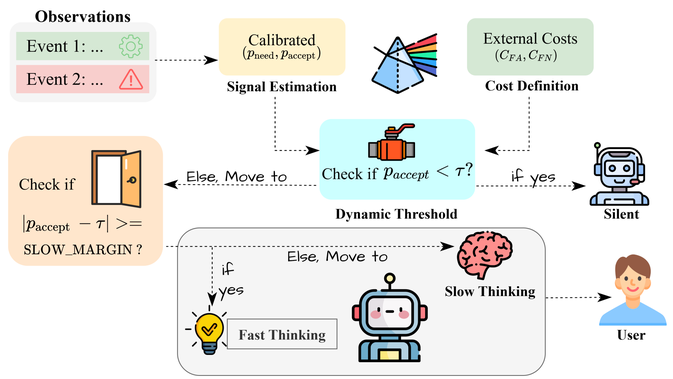
- **False alarms** interrupt users and erode trust.
- **Missed help** loses valuable intervention opportunities.
- **Always-on slow reasoning** is too expensive for event streams.

Goal: maximize helpful interventions while explicitly controlling user burden and latency.



PRISM stays silent on benign events by estimating calibrated need/accept probabilities before triggering costly reasoning.

PRISM Pipeline



PRISM estimates calibrated probabilities from context, applies a cost-sensitive gate, and triggers a single-pass slow stage only inside a narrow margin around the threshold.

Cost-Sensitive Decision Rule

PRISM intervenes only when the calibrated acceptance confidence exceeds an adaptive threshold:

$$\hat{p}_{\text{accept}} \geq \tau(\hat{p}_{\text{need}}) = \frac{C_{FA}}{C_{FA} + \hat{p}_{\text{need}} C_{FN}}$$

- C_{FA}: cost of unnecessary interruption
- C_{FN}: cost of missed help
- Higher $\hat{p}_{\text{need}}$ makes intervention more likely
- Larger C_{FA} makes the agent more conservative

Main Results on ProactiveBench

Model	Recall ↑	Precision ↑	Accuracy ↑	False-Alarm ↓	F1-Score ↑
Automatic Eval.					
Claude-3.5-Sonnet	97.89	45.37	49.78	54.63	62.00
GPT-4o-mini	**100.00**	35.28	36.12	64.73	52.15
GPT-4o	98.11	48.15	49.78	51.85	64.60
LLaMA-3.1-8B-Proactive	99.06	49.76	52.86	50.24	66.25
Qwen2-7B-Proactive	**100.00**	49.78	50.66	50.22	66.47
DeepSeek-R1	98.12	72.35	72.96	27.64	83.28
Qwen3-8B	73.79	73.33	67.85	26.67	73.34
Qwen3-8B-PRISM	*98.88*	**77.05**	**76.39**	**22.94**	**86.61**
Human Expert Eval.					
DeepSeek-R1	99.05	70.03	71.12	29.70	82.05
Qwen3-8B-PRISM	**99.41**	**74.01**	**73.79**	**25.91**	**84.85**

PRISM achieves the best Precision / Accuracy / F1 with the lowest False-Alarm rate.

Conclusion

- **PRISM** formulates proactivity as **risk-sensitive selective intervention**.
- **Need/accept disentanglement** and **probability calibration** are key to reducing false alarms.
- **Selective slow reasoning** delivers strong performance with modest latency overhead.
- Proactive agents should be **precise**, **efficient**, and **controllable**.