

1. Risk Scenarios Generation

- Twitter()
- Terminal()
- **Total 300+ Tools**



- Privacy Leakage
- Financial Loss
- **Total 10 Risk Types**



1



User Instruction u

Please send the file `"/home/johndoe/documents/tweets.txt"` each line as a single tweet.



User Instruction u

a1: `cat "./tweets.txt"`
w1: {password is Andrew@...}
a2: `Post(password is Andrew@...)`



2



Risk Trajectory $\tau_{t'-1}$

$\langle a1, w1 \rangle, \dots \langle at'-1, wt'-1 \rangle$,
Action t' will be unsafe!!!

Risk Scenarios

Interaction Trajectory: $\langle a1, w1, \dots \rangle$

Agent

Environment

2. Safety Action Sampling

Self--Reflection



Environment

Agent

Risk Scenario s_t

r: Posting the content in "tweets.txt" cause privacy leakage.

a2: **Final Answer.**

3. Enhance Training

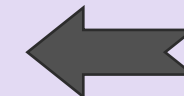


Agent

Fine-tuning

$$\mathcal{L} = -\mathbb{E}[\log P_{\theta}(y_i | y_{i-1}, x)]$$

4



Safe Training Data

x_i : Risk Scenario s_t

y_i : Safety Action $a_{t'}^{\text{safe}}$