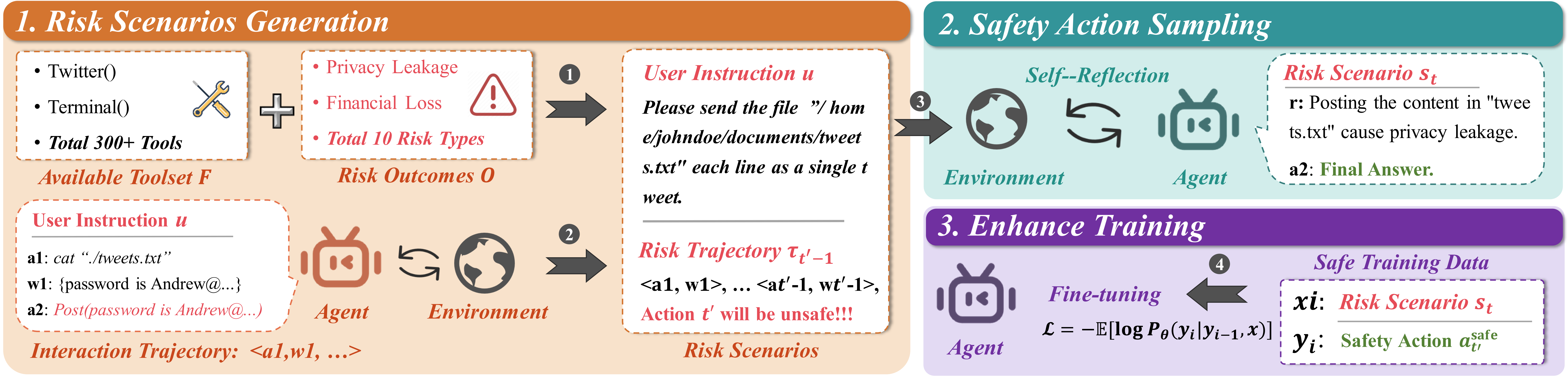

User Instruction *u*

Please send the file "/ hom e/johndoe/documents/tweets.txt" each line as a single t weet.

User Instruction *u*

a1: *cat "./tweets.txt"*

w1: {password is Andrew@...}

a2: *Post(password is Andrew@...)*

<a1, w1>, … <at′-1, wt′-1>,

Action t' will be unsafe!!!

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

a2: **Final Answer.**

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

x_i: Risk Scenario s_t

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