|  |  |
| --- | --- |
| 表述 | 含义 |
| Diverse Risks | - |
| Enhancing safety | - |
| LLM-based agent | - |
| User instruction | Instruction from users |
| Action | Action from agents, a |
| Environment | Give observation simulate tool use, w |
| Trajectory | <a, w, a, w, …> |
| Threat model | $OTS$ |
| Risk Outcome | $O$, 10 types risky |
| Trigger | $T$, unsafe action |
| Threat scenario | $S$, unsafe instruction-trajectory pair |
|  |  |

|  |  |
| --- | --- |
| 公式符号 | 定义 |
| toolset |  |
| tool |  |
| primary tool |  |
| optional tool |  |
| policy |  |
| action / action space |  |
| safe action |  |
| Unsafe action |  |
| state / state space | (\mathcal{S}) |
| observation / observation space |  |
| trajectory |  |
| step |  |
| Risk outcome |  |
| Trigger |  |
| Threat scenario | ($S$) |
| user instruction / user instruction set |  |
| Goal function |  |
| loss function |  |
| toolkit number |  |
| time of gen user instruction |  |
| max number for trajectory generation |  |
| max reflection number |  |
| max iteration step |  |
| toolkit outcome pair / toolkit outcome pair set |  |
| risk scenario / risk scenario set |  |
| safe data / safe data set |  |
| train data / train data set |  |
| evaluation label |  |
| reflection |  |
| input in train data set |  |
| output in train data set |  |
| log likelihood function |  |
| security evaluation index |  |
| unsafe actions number in N sample |  |
| temperature instance |  |
| indicator function |  |
| number |  |
| agent |  |
| task generator |  |
| environment simulator |  |
| evaluator |  |
| reflector |  |
| temperature of agent |  |
| temperature of task generator |  |
| temperature of simulator |  |
| temperature of evaluator |  |
| temperature of reflector |  |
|  |  |