# TrustAgent

A survey on trustworthy large language model agent

## Safety

1. Adversarial Hijacking
2. Unsafe Action Chain
3. Role Impersonation
4. Recursive Self-harm
5. Environment Damage
6. .....

## Privacy

1. Password Theft
2. Agent Profile Leakage
3. Dialogue Collection
4. Whereabouts Exposure
5. Illegal Info-trade
6. .....

## Truthfulness

1. Hallucination Collusion
2. Gossip Spread
3. Fake News
4. Disease Misdiagnosis
5. Financial Fraud
6. .....

## Fairness

1. Resource Monopolization
2. Role Discrimination
3. Collaborative Exclusion
4. Racist Speech
5. Celebrity Worship
6. .....

## Robustness

1. Malfunction
2. Model Dependence
3. Chain Vulnerability
4. Interaction Fragility
5. Endless Over-thinking
6. .....

## Others

1. Agent Accountability
2. Agent Ethics
3. Agent Transparency
4. Agent Explanability
5. .....

**Multi-dimensional**

# Technical

## Attack

- ***Brain:*** *Jailbreak & Prompt Injection & Backdoor*
- ***Memory:*** *Memory Poisoning & Privacy Leakage & Memory Misuse*
- ***Tool:*** *Tool Manipulation & Tool Abuse*
- ***Agent-to-Agent:*** *Infectious Attack & Cooperative Attack*

## Defense

- ***Brain:*** *Alignment & Single-model Filter & Multi-agent Shield*
- ***Memory:*** *Detection & Prompt Modification & Output Intervention*
- ***Agent-to-Agent:*** *Topological Defense & Cooperative Defense*
- ***Other Modules:*** *Undeveloped*

## Evaluation

- ***Brain:*** *Focused Assesment & General Benchmark*
- ***Tool:*** *Dataset Testing & Sandbox Simulation*
- ***Memory:*** *Attack Success Rate & Retrieval Success Rate*
- ***Other Modules:*** *Undeveloped*

# Modular = Internal Modules + External Modules

## Memory in Retrieval

Store / Retrieve

- Working & Short-term
- Long-term & Knowledge
- Database & Embedding & ...

## LLM as Brain

- ***Reasoning:*** *IO & CoT & ToT & GoT & CoT-SC ...*
- ***Modality:*** *Text & Vision & Speech & Mixed*
- ***Structure:*** *SayCan & ReAct & Reflextion & SwiftSage & ...*

## Tool for Action

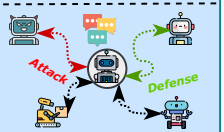
Call / Return

- Calculator & Robot Arm
- Code Interpreter & Sensor
- Search Engine & ...

## Agent & Multi-agent System

CAEML, ChatDev, AutoGen, MetaGPT, AutoGPT, Others

## Agent-to-Agent




- Cooperation & Competition
- Teaching & Debate & ...

## Agent-to-User

- Task & Problem & Query
- Feedback & File & ...

Ask / Answer

User's Resource: Diaglogue, Database, App, Phone, Email, .....

## Agent-to-Environment

Action / Feedback

- Physical World
- Game & Minecraft
- Docker & Web & ...