AI for Good Global Summit

7 - 10 July 2026
Geneva, Switzerland

Designing the trust management for Agentic AI

Session 3: Evaluating and Assessing Trustworthiness in Agentic AI

9 July 2026

50+ UN PARTNERS

ITU

Members for the panel

Naying Hu

Moderator

Associate Rapporteur, Q16/SG17
Deputy director, Security, Safety and Governance Department, AI Institute, CAICT

Yuntao Wang

Panelist

ITU FG-EAI Chair, Q5/21 Rapporteur, China Academy of Information and Communications Technology (CAICT)

Pam Dixon

Panelist

Founder & Executive Director, World Privacy Forum

Henk Birkholz

Panelist

Co-Founder & CTO, XOR

CAICT 中国信通院

Introduction to the AI Security Work of ITU-T SG17

Naying HU
Associate Rapporteur, Q16/SG17
Deputy director, Security, Safety and Governance Department, AI Institute, CAICT
9 July, 2026

CAICT 中国信通院

Overview of AI Security Work Programme (ITU-T SG17 Q16)

The **39 work items** include **1** foundation item, **27** lifecycle security items, **6** trust management items, and **5** supporting technology items.

1. Fundamental AI Security (1)

- X.sf-pAI

2. Model Design & Training (3)

- X.rg-dis
- X.sr-da-GenAI
- X.GenAI-FT

3. Model Testing & Evaluation (5)

- XSTR.AI-GSB
- XSTR.sem-AIA
- X.sem-ddm
- X.seg-AI
- XSTR.OCC-LLM

4. Deployment, Operation & Maintenance (19)

- X.AA-LLM
- X.2211
- X.sr-AIapi
- X.srg-AIgis
- X.aaia-sec
- X.S-AIA
- X.S-AIAI
- X.sg-AIAof
- X.sr-taimas
- XSTR.stv-OC
- X.sg-eAI
- X.2212
- X.sgMMFM
- X.2214
- X.scm-llm
- X.SG-AIAD
- X.sg-GenAId
- X.sg-rag
- X.sg-sd

5. Safety & Trustworthiness (6)

- X.Max-trust
- XSTR.aam
- XSTR.atcp
- XSTR.llmp-tip
- XSTR.ltf-AAI
- XSTR.magAI-behavior

6. Supporting Technologies (5)

- X.LLMCC
- XSTR.lzkml
- X.AI-gcd
- X.pg-cla
- X.SSDHN-AI-Atk

CAICT 中国信通院

Benchmark, Evaluation and Agentic AI-Related Security Work Items

13 work items covering foundation-model and AI-agent evaluation(2), deployment controls for Agentic AI (7), and Agentic AI trust frameworks (4).

3. Model Testing & Evaluation (2 of 5)

XSTR.AI-GSB	Technical Report: Guidelines of security benchmark for foundation models
XSTR.sem-AIA	Technical Report: Security evaluation methods for Artificial Intelligence agent

4. Deployment, Operation & Maintenance (7 of 19)

X.aaia-sec	Security framework and technical requirements for autonomous AI agents
X.S-AIA	Security requirements and guidelines for Artificial Intelligence agent
X.S-AIAI	Security framework and requirements for invocation by AI agent
X.sg-AIAof	Security guidelines for cloud-side AI agent orchestration framework
X.sr-taimas	Security requirements for terminal-based artificial intelligence multi-agent system
XSTR.stv-OC	Technical Report: Security threats and vulnerabilities in the OpenClaw framework
X.SG-AIAD	Security guidelines for Artificial Intelligence agent data

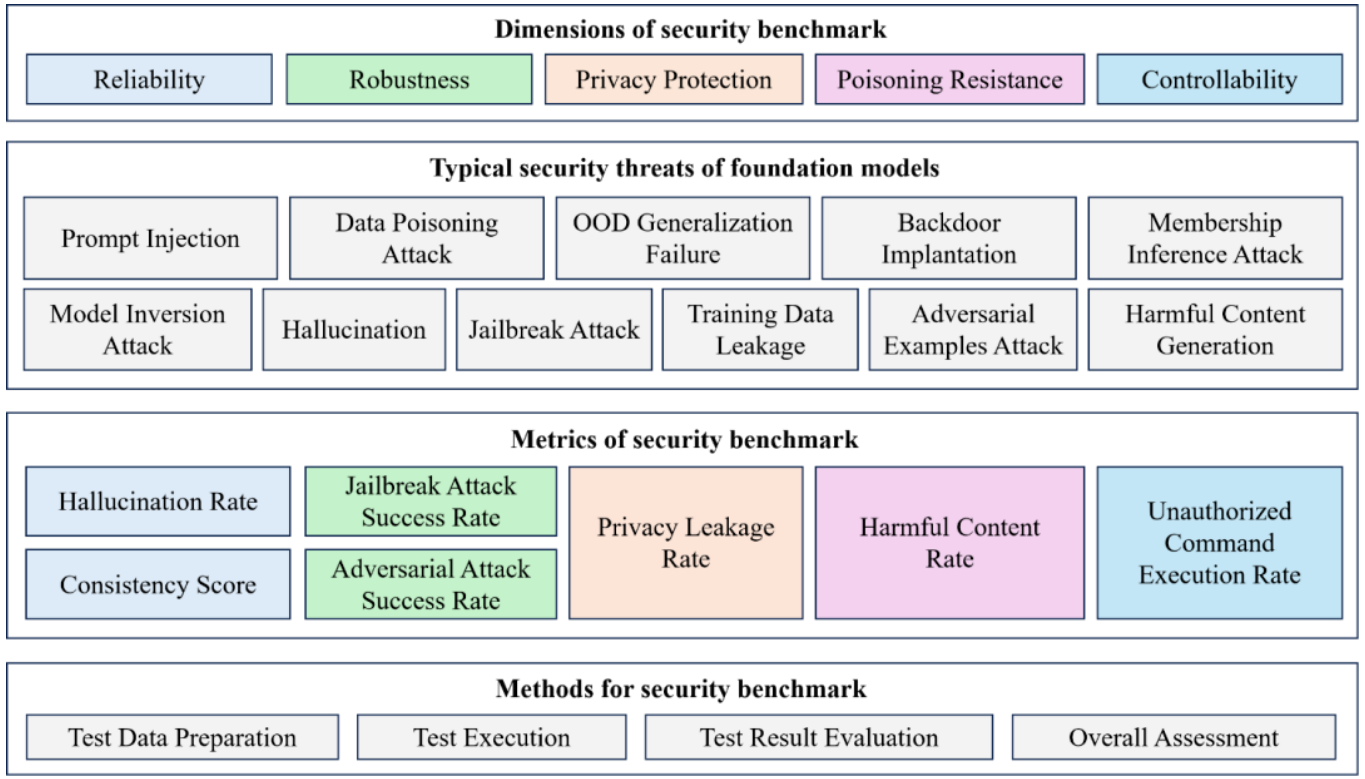
5. Safety & Trustworthiness (4 of 6)

X.Max-trust	A multi-party, agentic and extensible trust framework for next generation agentic-AI-enabled telecommunication networks
XSTR.atcp	Technical Report: Cryptographic primitives for enhancing AI agent trustworthiness
XSTR.ltf-AAI	Technical Report: Landscape of trust framework for agentic AI
XSTR.magAI-behavior	Technical Report: Capabilities for agentic AI security and trust controls to manage behavior of agentic multi-objective AI systems

CAICT 中国信通院

Guidelines of Security Benchmark for Foundation Models

XSTR.AI-GSB frames foundation-model security benchmark around dimensions, threats, metrics and test methods.



Dimensions

Define what aspects of foundation-model security should be assessed, covering reliability, robustness, privacy protection, poisoning resistance, and controllability.

Threats

Identify typical risks arising from foundation models, including prompt injection, data poisoning, hallucination, jailbreak, data leakage, adversarial attacks, and harmful content generation, et al.

Metrics

Develop measurable indicators such as hallucination rate, attack success rate, privacy leakage rate, harmful content rate, and unauthorized command execution rate.

Methods

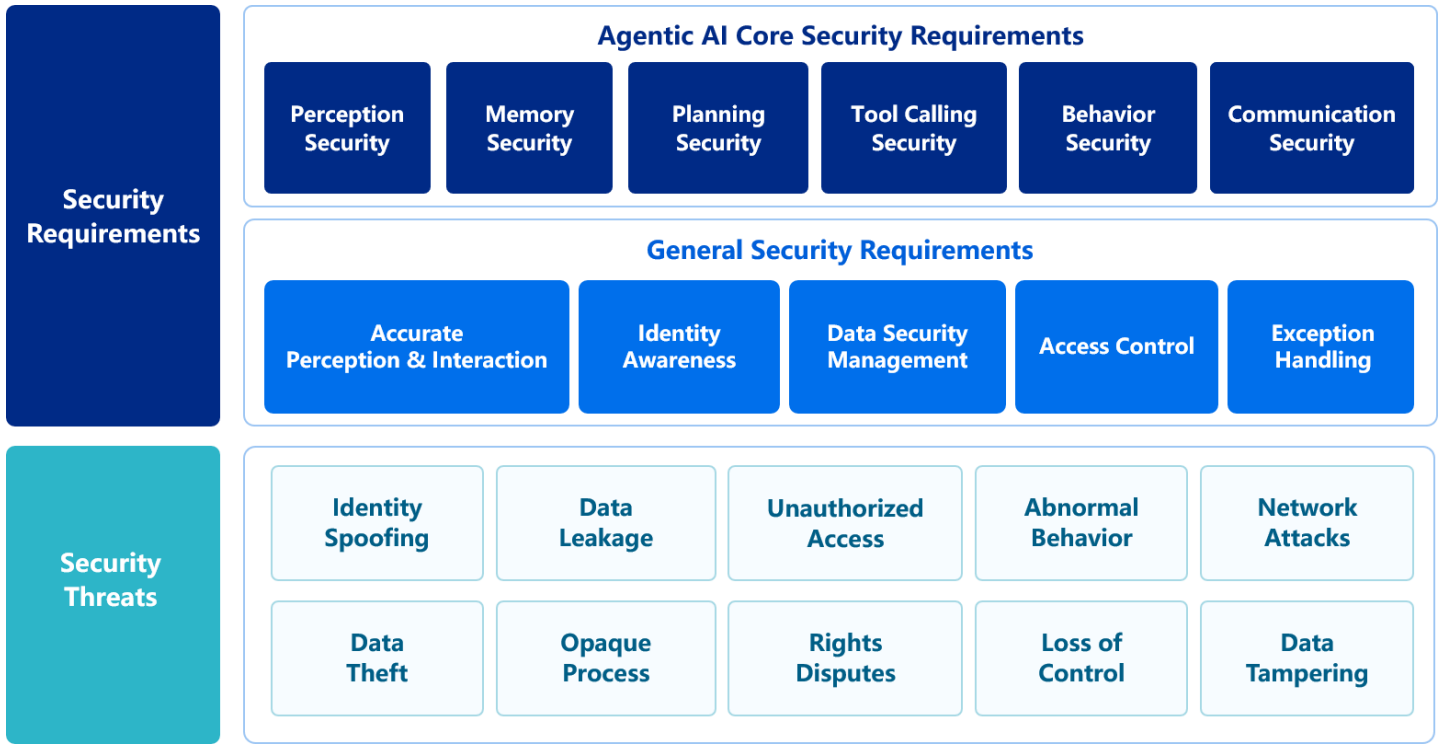
Build a complete benchmark workflow through test data preparation, test execution, result evaluation, and overall assessment.

Source: XSTR.AI-GSB, Guidelines of security benchmark for foundation models

CAICT 中国信通院

Thoughts on How to Conduct Agentic AI Evaluation and Testing

Agentic AI evaluation should be threats and risks driven, translating security threats into security requirements and evaluation criteria for systematic security assessment.



Security Threats & Risks

Identify risks arising from system components, operational behaviors, external interactions, and supporting infrastructures.

General Security Requirements

Security requiremts related to models, data, infrastructure, and operational processes.

Core Security Requirements

Establish security requirements for key agent capabilities, including perception, memory, planning, tool use, behavior, and communication.

Evaluation Methods

Develop evaluation methodologies, metrics, and testing procedures to validate security capabilities and risk controls.

Source: MIIT-TC1/WG8 Artificial Intelligence-Security Governance-Common Techniques Requirements for AI Agent Security

CAICT 中国信通院

Thanks!

Contact me: hunaying@caict.ac.cn