CHAI

Agentic AI — Unified T&E Framework

Metrics from the CHAI T&E framework mapped across **6 action types** × **6 risk domains**. The first 3 columns (Tool Calling, Computer Use, Code Gen) reflect agentic action mechanisms. The second 3 (Data Interaction, Context & Hand-off, Task Execution) are CHAI healthcare agent skill categories — T&E coverage gaps for these are marked explicitly. Click any card to expand details. Gap cards identify where evaluation coverage is missing.

FILTER BY PRINCIPLE: All | Safety & Reliability | Efficacy | Efficiency | Transparency | Fairness & Bias | Usability | Show gaps

Tool Calling: 17 metrics | Computer Use: 11 metrics | Code Gen & Exec: 5 metrics | Data Interaction: 0 metrics | Context & Hand-off: 0 metrics

Task Execution: 6 metrics | All types: 15 metrics | Coverage gaps: 32

Tool Calling | Computer Use | Code Gen & Exec | Data Interaction | Context & Hand-off | Task Execution | All types | Coverage gap

RISK DOMAIN ↓ ACTION TYPE →	Tool Calling Pre-defined API/MCP calls. Bounded & auditable.	Computer Use Screen nav, clicks, typing. Broad action space.	Code Gen & Exec Writes & runs novel code. Hardest to audit.	Data Interaction How agents access & interpret healthcare data.	Context & Hand-off How information is structured for humans & other agents.	Task Execution How agents perform real workflows & deliverables.
Data Privacy Access scope, exposure, least-privilege	**Output Constraint Violation Rate (OCVR)** Safety **Slot Extraction F1 (SEF1)** Efficacy GAP **Unauthorized Data Access Rate** Coverage missing GAP **Data Minimization Score** Coverage missing	**Platform Reliability Score (PRS)** Safety GAP **Screen Data Capture Scope** Coverage missing	GAP **Sensitive Data Exposure in Generated Code** Coverage missing	GAP **FHIR Access Scope Compliance Rate** Coverage missing	GAP **Role-Appropriate Disclosure Rate** Coverage missing	GAP **Task-Scope PHI Access Rate** Coverage missing
Liability & Irreversibility	**Tool Invocation Correctness (TIC)**	**Task Success Rate —**	**Regression Non-Introduction Rate**	GAP **Data Omission Liability**	GAP **Silent Context Truncation Rate**	**Summarization Editing Effort**

Accountability, irreversible actions, oversight

Safety

Tool Execution Success Rate (TESR)
Safety

Task Completion Rate with Tool Use (TCR-T)
Efficacy

Business Policy Adherence Rate (BPAR)
Safety

Escalation Rate (ESR)
Safety

Hard Turn Limit Compliance Rate (HTLCR)
Safety

End-to-End Trace Coverage Rate
Transparency

Pass@k Reliability for Agentic Financial Tasks
Safety

Node-Level Regression Test Coverage Rate
Safety

Legal Agent Intermediate Progress Rate (LAIPR)
Efficacy

MedAgentBench
Efficacy

Task-Level Safety & Clinical Accuracy (TLSCAS)
Safety

Task Success Rate — WebArena
Efficacy

Call Abandonment Rate (CAR)
Efficacy

Patient-Reported Experience Score
Efficacy

PPE Non-Compliance Alert F1 (PNA-F1)
Safety

Triage Appropriateness Rate
Efficacy

End-to-End Clinical Agent Response Latency
Efficiency

User Satisfaction Score (USS)
Efficacy

Human Intervention Rate (HIR)
Efficacy

(RNIR)
Safety

Decontaminated Task Success Rate
Safety

Pass@1 Task Completion Rate
Safety

Epistemic Verification Behavior Rate
Safety

Proportion of Errors Auto-Corrected
Usability

Rate
Coverage missing

Coverage missing

Reduction
Efficiency

Claim Recall for Summary Factual Coverage
Usability

Intermediate Step Correctness Rate
Safety

Per-Record Processing Time
Efficiency

Provider Preference Match Rate
Efficacy

Prompt Injection Adversarial robustness, hijacked behavior	Noise-Induced Performance Degradation (NIPD) Safety GAP Prompt Injection Success Rate (Tool) Coverage missing	GAP Instruction Hijack Rate (Screen Content) Coverage missing	GAP Code Injection via External Data Coverage missing	GAP Adversarial Structured Data Feed Robustness Coverage missing	GAP Cross-Agent Context Field Injection Rate Coverage missing	GAP Workflow Injection via Structured Clinical Input Coverage missing
Malicious Use Harmful outputs, weaponizable capabilities	Real-Time Oversight / Independent Output Monitoring Rate Safety GAP Misuse Capability Score Coverage missing	GAP Abuse Scenario Coverage (Computer Use) Coverage missing	GAP Malicious Code Generation Rate Coverage missing	GAP PHI Over-Collection Detection Rate Coverage missing	GAP Context Manipulation for Downstream Misdirection Coverage missing	GAP Task Misdirection Resistance Rate Coverage missing
Third-Party & Supply Chain External tools, APIs, data source trust	Decision Accuracy Under Incomplete Information Efficacy Top-k Symptom-to-Visit Type Mapping Accuracy Efficacy Tool Selection Accuracy (TSA) Safety GAP Third-Party Source Verification Rate Coverage missing	GAP Source Trustworthiness Score (Web Content) Coverage missing	GAP Dependency/Pack Provenance Accuracy Coverage missing	GAP External Data Source Provenance Verification Coverage missing	GAP Third-Party Context Source Verification Rate Coverage missing	Number of Integrated Wearable / RPM Device Types Usability
Agent-to-Agent Multi-agent interactions, coordination, error	GAP Error Propagation Rate (Tool Chain)	GAP Handoff Failure Rate (UI Transition)	GAP Cross-Agent Code	GAP Concurrent EHR Read Consistency Rate	GAP Context Fidelity Across Agent Handoffs	GAP Sub-Agent Coordination

propagation	Coverage missing	Coverage missing	**Contamination Rate** Coverage missing	Coverage missing	Coverage missing	**Failure Attribution Rate** Coverage missing

Cross-cutting metrics — verified to apply across all 6 action types

Only metrics whose definition in the CHAI T&E framework is mechanism-agnostic are listed here. Metrics specific to multi-agent interactions, a single interface type, or a clinical sub-domain are placed in their respective cells above.

Goal Completion Rate (GCR)
Efficacy · Liability

Autonomy Index (AIx)
Efficacy · Liability

Harm-Reduction Index (HRI)
Safety · Malicious Use

Safety Score (Agent-SafetyBench)
Safety · Malicious Use

Constraint Violation Rate (CVR)
Safety · Liability

Plan Adherence Score (PAS)
Safety · Liability

Goal Adherence Score (GAS)
Safety · Liability

Step Correctness Rate (SCR)
Safety · Liability

Planning Efficiency Index (PEI)
Efficiency · Liability

Balanced Evaluation Coverage Score (BECS)
Efficacy · Liability

Intraclass Correlation Coefficient (ICC)
Safety · Liability

Standard Deviation of Task Success Rate (σ-TSR)
Safety · Liability

Cost-per-Success (Cost-of-Pass)
Efficiency · Liability

Predictive Parity Ratio (PPR)
Fairness · Agent-to-Agent

Fairness Constraint Satisfaction Rate (FCSR)
Fairness · Agent-to-Agent