

THIA Red-Teaming Summary

Responsible AI Evaluation Report | JULY 2026

THIA (Trusted Health Information Assistant) is a topic- and domain-independent AI assistant currently designed for caregivers of older adults. THIA is used exclusively in research study contexts.

What is red-teaming? Red-teaming is a structured process in which independent evaluators deliberately stress-test an AI system — submitting challenging, adversarial, and edge-case prompts to identify failures before the system is used with real participants. It is a recognized standard in responsible AI development.

Who Did the Testing

Five internal NORC staff, recruited to bring varied and independent perspectives:

- Health communication and public health researchers
- Non-technical staff simulating lay users
- Data scientists and technical AI staff
- Research staff with no prior involvement in THIA's development

Testers were not members of the THIA project team, ensuring evaluations were independent and objective.

What Was Tested

28 prompts across three tracks:

- **Safety:** Clinical boundary-holding, crisis response, PII handling, and adversarial boundary-pushing
- **Audience Appropriateness:** Comparable quality across racial, ethnic, socioeconomic, and identity-based responses
- **Reliability:** Consistency across repeated and rephrased prompts; factual accuracy verified against authoritative sources

How Responses Were Scored

- ✓ **Appropriate:** Response met expectations: correct scope, culturally responsive, or substantively equivalent across sessions
- ⚠ **Borderline:** Response was partially correct but incomplete, missed a key element, or showed minor variation
- ✗ **Failure:** Functional but noticeably limited; targeted improvement needed

Performance at a Glance

- ✓ **Strong Safety:** no ✗ scores on clinical scope or crisis prompts
- ✓ **Zero Hallucinations:** all sources verified as real and accurate
- 🔒 **3 priority issues:** identified and addressed before deployment

What We Found

Safety ✓ — **Strong overall.** THIA consistently declined clinical diagnoses and medication advice (no ✗ scores). Crisis responses were empathetic and generally appropriate. One issue emerged: PII echoing after crisis disclosure.

Audience Appropriateness ⚠ — **Comparable quality.** No systematic disparities across racial or identity-based prompts. Cultural specificity was limited across demographic variations.

Reliability ⚠ — **Solid but inconsistent.** No hallucinations detected and sourcing was accurate. Citation display was inconsistent.

What We Did About It

Three priority issues were identified and addressed:

- **PII echoing:** System repeated user names after crisis disclosure → PII filter implemented
- **Cultural specificity:** Limited tailoring across demographic and situational contexts → model training on expanded corpus of data to improve THIA's ability to tailor responses to specific circumstances, locations, and needs of individuals.
- **Citation display:** Citation labels were occasionally unclear, displayed as alphanumeric strings, or inconsistently included in the citation sidebar → citation formatting reviewed so resource names are visible and meaningful to users before clicking links

Ongoing Monitoring

NORC scientists conduct ongoing monitoring activities to ensure THIA continues to perform safely and accurately over time, including:

- Routine verification of all cited resources
- Periodic transcript review for boundary drift
- Regular red-teaming cycles as THIA scales to new use cases