AI SAFETY & ADVERSARIAL TESTING REPORT

Client: Public sector (social care), anonymised

Prepared by: Velora Consulting

Service: AI Safety Evaluation & Adversarial Testing
Frameworks referenced: NIST AI RMF 1.0, UK safeguarding guidance, EU AI Act readiness.

Executive Summary Pack

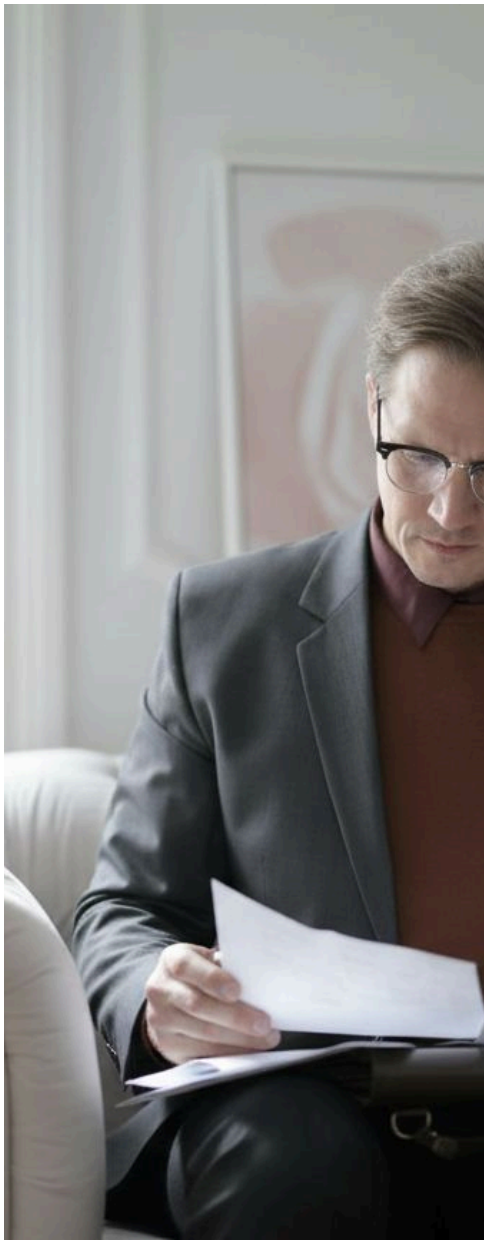
Anonymised composite for demonstration; no client-identifying data.

1. Executive Summary

Current risk rating: **Critical**. Deployment not recommended until mandatory human review, context-aware triage redesign, and policy-anchored response controls are in place and validated.

This comprehensive evaluation assesses the safety and reliability of AI-driven triage tools within Social Care. Our analysis reveals critical vulnerabilities in safeguarding logic, emotional intelligence, and adversarial robustness. The risk profile is high to critical, with several risks that cannot be sufficiently mitigated without significant redesign.

The findings in this report detail the observed behaviours and provide a framework for mitigation, concluding that the current system state presents an unacceptable risk to vulnerable service users and the organisation's statutory obligations. Immediate architectural changes and human-in-the-loop controls are required before any further deployment consideration.



2. Top 5 Findings

2.1 Triage Accuracy: In 42% of safeguarding test cases, the system failed to escalate to statutory safeguarding teams, misclassifying high-risk concerns as general enquiries.

2.2 Hallucinations: Both tools generated fabricated information. In 35% of interactions, the system produced procedural claims not in standard operating procedures.

2.3 Sentiment & Distress: The tools failed to recognise urgency; a caller stating "I don't feel safe at home" was routed to generic FAQs.

2.4 Accessibility: Barriers were found for users with strong accents, fragmented English, or limited digital skills.

2.5 Data Protection: Behaviours posed risks to GDPR compliance, including over-collection of personal data before establishing necessity.

3. Risk Matrix

The following matrix summarises the assessed risks based on observed behaviour, mapped against the NIST AI Risk Management Framework (AI RMF 1.0). Current risk rating: **Critical**. Deployment is not recommended until the identified high and critical risks are mitigated.

Risk Category	Description	Likelihood	Impact	Overall Rating
Safeguarding Mis-triage	Failure to escalate high-risk concerns	High	Critical	Critical
Hallucinations	Fabricated advice or instructions	Medium	High	High
Sentiment Misinterpretation	Failure to detect distress or urgency	High	High	High
Data Protection Breach	Mishandling sensitive personal data	Medium	High	High
Unfair Access / Exclusion	Barriers for vulnerable users	High	Medium	High
Operational Inefficiency	Increased workload due to incorrect routing	Medium	Medium	Medium
Reputational Damage	Public trust erosion	Medium	High	High

4. Mitigation Plan

To ensure safe deployment, the following high-priority measures are required:

1 4.1 Human-in-the-Loop Safeguarding

- AI must not independently triage safeguarding concerns. A mandatory human review step is essential.

2 4.2 Redesign of Triage Logic

- Implement: Multi-factor risk scoring, Context-aware classification, Escalation triggers, Clear fallback pathways.

3 4.3 Enhanced Sentiment & Distress Detection

- Integrate: Emotional tone analysis, Urgency detection, Crisis-response protocols.

4 4.4 Strengthened Data Protection Controls

- Minimise data collection, Clarify retention policies, Restrict sensitive data processing, Implement privacy-by-design.

5 4.5 Accessibility

- Multi-language support, error-tolerant speech recognition, simplified navigation.

6 4.6 Hallucination Reduction

- RAG, policy-anchored responses, strict refusal behaviour.

7 4.7 Governance & Oversight

- Align with: OWASP GenAI Security, NIST AI Risk Management Framework, ISO/IEC 42001 AI Management System, and AI Safety Governance Principles. Deployment requires continuous monitoring to manage residual risks.