AI 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). The cumulative risk profile is too high for deployment without substantial redesign and governance controls.

Risk Category	Description	Likelihood	Impact	Overall Rating	Mitigation Levers
Safeguarding Mis-triage	Failure to escalate high-risk concerns to human operators.	**High**	Critical	**Critical**	Mandatory human-in-the-loop review; automated escalation triggers.
System Hallucinations	Generation of fabricated info or incorrect procedural claims.	**Medium**	**High**	**High**	RAG implementation; strict refusal for out-of-scope queries.
Sentiment Misinterpretation	Failure to detect urgency or distress in user communications.	**High**	**High**	**High**	Real-time tone analysis; keyword-independent urgency detection.
Data Privacy Risks	Potential for over-collection or mishandling of sensitive data.	**Medium**	**High**	**High**	Privacy-by-design; automated PII redaction and minimisation.
Access Barriers	Exclusion of users with diverse linguistic or digital needs.	**High**	**Medium**	**High**	Multi-language support; simplified navigation and UI.
Adversarial Vulnerability	Susceptibility to manipulation to bypass safety guardrails.	**High**	**High**	**High**	Regular red teaming; robust adversarial training protocols.
Operational Inefficiency	Increased workload due to incorrect routing and errors.	**Medium**	**Medium**	**Medium**	Continuous performance monitoring; iterative prompt refinement.

Rating Scales | Likelihood: Low, Medium, High. Impact: Low, Medium, High, Critical. Aligned with NIST AI RMF 1.0.

Overall rating combines likelihood and impact, with expert override where safety-critical risk is present

