STRATEGIC POSITION PAPER

Sovereign Personal Intelligence

A Framework for Trusted Local-First AI Systems

PREPARED FOR

UK Sovereign AI Fund

UK Research and Innovation (UKRI)

Innovate UK

Strategic Investors

Government Stakeholders

Reference Implementation: **StarCaller**

VERSION	DATE
1.3 (Draft for Consultation)	June 2026

PUBLIC DOCUMENT

Table of Contents

Sovereign Personal Intelligence — A Framework for Trusted Local-First AI Systems

Section 1 — Executive Summary 3

Section 2 — The Emerging Sovereign AI Challenge 5

- 2.1 The Current AI Landscape
- 2.2 Strategic Risks of Centralised AI
- 2.3 The UK Opportunity
- 2.4 Why Sovereignty Requires a New Category

Section 3 — Defining Sovereign Personal Intelligence 8

- 3.1 What Sovereign Personal Intelligence Is
- 3.2 Core Principles
- 3.3 Contrast with Existing Paradigms
- 3.4 The Missing Market Category

Section 4 — Architecture: The StarCaller Reference Implementation 10

- 4.1 Architectural Overview
- 4.2 ICID — Governance Layer
- 4.3 Shi Gandang — Security Layer
- 4.4 The Vault — Memory Layer
- 4.5 Ability OS — Execution Layer
- 4.6 Prism Nexus — Orchestration Layer
- 4.7 Oracle Framework — Specialised Intelligence Layer
- 4.8 Architecture Diagrams

Section 5 — Trust Architecture 14

- 5.1 Memory Governance
- 5.2 Consent Governance
- 5.3 Security Gates
- 5.4 Auditability and Decision Traces
- 5.5 Explainability and Reasoning Provenance

5.6 Security Health Index

Section 6 — National Strategic Relevance 17

6.1 UK AI Safety Objectives

6.2 UK Sovereign AI Objectives

6.3 Digital Resilience

6.4 Privacy Enhancement

6.5 Citizen Trust and Secure Agentic Systems

Section 7 — Economic Impact 19

7.1 Addressable Markets

7.2 Market Forecasts

7.3 UK Industrial Strategy Alignment

Section 8 — Commercialisation Strategy 21

8.1 Stage 1: StarTeQ — Revenue Foundation

8.2 Stage 2: StarCaller Software — Platform Adoption

8.3 Stage 3: StarCaller Appliance — Sovereign Hardware

8.4 Risk Reduction Through Phased Delivery

Section 9 — Competitive Analysis 23

9.1 Competitive Landscape

9.2 Capability Comparison

9.3 Strategic Positioning

9.4 Category Timing and Strategic Position

Section 10 — Research & Innovation Agenda 25

10.1 Federated Vaults

10.2 Multi-Appliance Intelligence

10.3 Explainable Agent Systems

10.4 Secure Local Models

10.5 Sovereign Memory Standards

10.6 Personal AI Governance Frameworks

Section 11 — Funding Proposal 27

11.1 Funding Requested

11.2 Allocation

11.3 Milestones and KPIs

11.4 Risk Mitigation

11.5 Expected Outcomes

11.6 National Capability Contribution

Section 12 — Conclusion 29

SECTION 1

Executive Summary

What StarCaller Is

StarCaller is an open, local-first sovereign personal intelligence platform. It is a complete architecture for trusted, autonomous AI systems that operate under user consent, maintain auditable decision trails, keep personal data within sovereign boundaries, and function without mandatory reliance on foreign cloud infrastructure.

This is not merely a chatbot. It is not simply a large language model wrapper. StarCaller is a layered, security-gated personal intelligence operating system comprising six integrated subsystems — a governance layer (ICID), a security layer (Shi Gandang), a memory vault, an Ability OS for autonomous execution, an orchestration layer (Prism Nexus), and a framework of specialised Oracles. Together, these subsystems create a new category of AI: **sovereign personal intelligence**.

Why It Matters

The current trajectory of consumer AI is toward ever-greater centralisation. The dominant models — GPT, Claude, Gemini, DeepSeek — operate as opaque cloud services controlled by a small number of foreign corporations. Every interaction, every piece of personal data, every decision trace flows through infrastructure outside the individual's control and, critically for the United Kingdom, outside national jurisdiction.

This creates four interconnected vulnerabilities:

1. **Data sovereignty erosion.** Personal data of UK citizens is processed, stored, and potentially trained on within foreign jurisdictions. The UK Data Protection and Digital Information Bill notwithstanding, enforcement across borders remains fundamentally constrained.
2. **Infrastructure dependency.** The UK currently has no sovereign alternative to the compute and model infrastructure of US and Chinese AI companies. The AI Opportunities Action Plan's commitment to 20× sovereign compute expansion by 2030 addresses hardware but not the software and governance layers that determine how AI systems actually handle personal data.
3. **Trust deficit.** Users currently cannot independently audit what the cloud models do with their data, how memories are stored, or whether consent boundaries are respected. The dominant model is one of "trust us" — not "verify us."

4. **Vendor lock-in without recourse.** As personal AI becomes more deeply integrated into daily life — managing schedules, health, finances, communications — switching costs become prohibitive. The user does not own their AI; they rent access to it.

StarCaller proposes a substantially different architecture: one where the AI system lives with the user, its memory is encrypted and consent-governed, its decisions are auditable, and its operation does not require foreign infrastructure.

Strategic Opportunity

The United Kingdom's AI strategy, articulated through the AI Opportunities Action Plan (January 2025), the £500 million Sovereign AI Fund (April 2026), and the £1.1 billion sovereign AI infrastructure commitment, identifies three priorities: sovereign capability, citizen trust, and economic advantage. StarCaller addresses all three directly.

Sovereign capability — StarCaller's local-first architecture can operate on sovereign compute infrastructure, whether in data centres, edge nodes, or personal devices. It does not depend on any single model provider. The system is designed to be model-agnostic, capable of using open-weight models (Llama, Mistral, Qwen), fine-tuned sovereign models, or private deployments of frontier models.

Citizen trust — StarCaller's consent ladder, auditable decision traces, and 10-gate Shi Gandang security framework provide verifiable guarantees that the system operates within its boundaries. The Security Health Index (SSI) provides a quantifiable, real-time measure of system integrity — assessed at 97.8/100 (HARDENED, self-assessed, unaudited) in the reference implementation.

Economic advantage — The personal AI market is projected to grow from $2.23 billion (2024) to $56.3 billion by 2034 (CAGR 38.1%). The broader agentic AI market is projected at $52.6 billion by 2030 (CAGR 46.3%). Edge AI hardware alone is forecast at $58.9 billion by 2030. A UK sovereign entrant in this space, built on open standards, could capture significant market share while reinforcing national AI infrastructure.

Funding Requirements

This paper is intended to support a potential funding request to the UK Sovereign AI Fund, Innovate UK, and strategic investors for development of the StarCaller reference implementation, sovereign AI appliance, and open governance standards. Details are provided in Section 11.

National Relevance

StarCaller directly supports the UK government's AI Opportunities Action Plan objectives, the AI Security Institute's technical safety research agenda, and the Sovereign AI Fund's mission to build British AI champions. It represents a credible, working — not theoretical — implementation of sovereign personal AI infrastructure, with a fully functional reference deployment, 136+ security

tests passing, and a Security Health Index assessed at HARDENED level (self-assessed, unaudited; see SSI Methodology Appendix for full scoring).

SECTION 2

The Emerging Sovereign AI Challenge

2.1 The Current AI Landscape

The artificial intelligence industry in 2026 is dominated by a small number of large-scale platforms. Five major model families — OpenAI's GPT, Anthropic's Claude, Google's Gemini, Meta's Llama, and DeepSeek's V3/R1 — account for the vast majority of consumer and enterprise AI interactions. Each of these platforms operates on centralised cloud infrastructure, processes user data on remote servers, and is governed by the legal frameworks of its home jurisdiction.

The concentration is intensifying. The top three cloud providers (Amazon Web Services, Microsoft Azure, Google Cloud) host approximately 80% of AI model inference. The compute requirements for frontier model training — estimated at $100 million to $1 billion per generation — create structural barriers to entry that favour incumbent technology powers.

This concentration produces a market that is simultaneously competitive and homogeneous in its fundamental architecture. All major platforms offer chat interfaces, API access, and increasing degrees of agentic capability. To the authors' knowledge, none offer local-first, consent-governed, auditable personal intelligence as a defined category.

OpenAI — The most widely deployed consumer AI platform. GPT-4 and successor models power ChatGPT (300+ million weekly active users as of early 2026) and an extensive API ecosystem. The platform is fully cloud-based. User data processing terms permit training data use unless explicitly opted out. No local deployment option exists.

Anthropic — Positioned as the safety-focused alternative. Claude's constitutional AI approach has influenced industry standards for harm reduction. However, the architecture remains centralised. Anthropic offers no local-first or sovereign deployment option, and its safety guarantees are architectural choices enforced by Anthropic, not verifiable by users.

Google/DeepMind — Gemini is deeply integrated into Google's ecosystem of Search, Workspace, Android, and Cloud. The platform benefits from unprecedented data access but operates under a unified privacy policy that permits broad data use. Sovereign deployment is not available outside Google Cloud infrastructure.

Meta (Llama) — The most significant open-weight model family. Llama 4 and subsequent versions provide capable models that can run on local hardware. Meta has invested in on-device deployment through partnerships with Qualcomm, MediaTek, and others. However, there is no companion governance, memory, or trust architecture — Llama is a model, not a personal intelligence platform.

DeepSeek — A rapidly ascendant Chinese AI platform with competitive model performance. DeepSeek's emergence has intensified concerns about data sovereignty and foreign infrastructure dependency in Western markets. Several governments have restricted DeepSeek's use in official contexts.

Mistral AI — The leading European AI company. Mistral has positioned itself as a sovereign alternative for European institutions, offering both API access and self-hosted deployment of its open-weight models. The company's focus on sovereign deployment is notable, but like Meta, Mistral provides models rather than a complete personal intelligence governance framework.

2.2 Strategic Risks of Centralised AI

The centralised AI architecture — cloud-dependent, opaque, and foreign-controlled — creates a pattern of strategic risks that are increasingly recognised by governments, security institutions, and civil society.

Data sovereignty. Every interaction with a cloud AI system transfers personal data to the platform operator's infrastructure. For UK users of OpenAI, Anthropic, or Google services, this means personal data — conversations, health information, financial details, decision-making patterns — is processed and stored in data centres across multiple jurisdictions, primarily the United States. The UK-US Data Bridge mitigates some legal risk, but enforcement of UK data protection rights against foreign entities remains practically constrained. The US CLOUD Act permits US authorities to access data held by US companies regardless of storage location.

Foreign dependence. The United Kingdom's AI sector relies on infrastructure and model capability developed and controlled by foreign entities. This creates strategic vulnerability: access to frontier AI capability could be affected by trade disputes, regulatory divergence, or geopolitical events. The AI Opportunities Action Plan's commitment to 20× sovereign compute expansion addresses the hardware dimension of this risk, but does not address the software and governance layers.

Infrastructure concentration. AI inference and training are increasingly concentrated in a small number of hyperscale data centres. This concentration creates single points of failure — technical, economic, and geopolitical. A service outage at a major cloud provider can disrupt AI operations across thousands of businesses and millions of users. There is no distributed, resilient alternative for personal AI infrastructure.

Vendor lock-in. As AI systems become more deeply integrated into daily life and organisational operations, switching costs increase. Systems that learn user preferences, accumulate personal memory, and automate complex workflows become effectively irreplaceable if the underlying platform is proprietary and data cannot be extracted in usable form. This lock-in is not accidental in many cases — it can be seen as a structural feature of the centralised AI business model.

Opacity and trust. Centralised AI platforms operate as closed systems. Users cannot audit model behaviour, verify data handling practices, inspect decision-making processes, or independently

assess compliance with stated safety commitments. The industry's primary trust mechanism is brand reputation — an increasingly inadequate basis for systems that may manage health data, financial decisions, personal communications, and increasingly autonomous agentic operations.

2.3 The UK Opportunity

The United Kingdom has a credible claim to being the third pole in global AI development, after the United States and China. The UK's advantages include world-class research institutions (Cambridge, Oxford, Imperial, UCL, Edinburgh), a sophisticated regulatory environment, a strong tradition of privacy protection, and government commitment to AI infrastructure investment.

The £1.1 billion sovereign AI infrastructure commitment, the £500 million Sovereign AI Fund, and the AI Opportunities Action Plan represent the most ambitious government AI investment programme outside the US and China. However, this investment has been directed primarily toward compute infrastructure and general-purpose AI capability, not toward the specific challenge of **personal AI governance** — the software and architectural layers that determine how AI systems interact with personal data, maintain consent boundaries, and provide verifiable trust.

This gap is the UK's strategic opportunity. While the US produces the most capable frontier models, and China produces the most aggressively scaled platforms, the UK could define and lead the category of **sovereign personal intelligence** — trusted, local-first, consent-governed AI infrastructure that becomes the standard for privacy-respecting personal AI.

2.4 Why Sovereignty Requires a New Category

> *“ Sovereignty in AI is not achieved through model capability alone. ”*

Sovereignty in AI is not achieved through model capability alone. A locally deployed model without governance, consent, or auditability is simply an opaque cloud system running on different hardware. True sovereignty requires a complete architectural framework that addresses:

- **Data control.** Personal data never leaves sovereign jurisdiction without explicit, auditable consent.
- **Infrastructure independence.** The system can operate on any compute substrate — local device, sovereign data centre, or trusted edge node — without mandatory reliance on foreign cloud services.
- **Verifiable trust.** System behaviour is auditable by users, regulators, and independent third parties.

- **Consent governance.** All operations that involve personal data are subject to a graduated consent framework that the user controls and can modify at any time.
- **Interoperability.** Data and capabilities are portable. The user is not locked into a single vendor.

To the authors' knowledge, no existing platform — open or proprietary — currently combines local-first architecture, consent-governed memory, auditable decision processes, and verifiable security measurement into a single integrated framework. The next section defines what such a framework comprises and introduces StarCaller as a reference implementation.

SECTION 2A

Why Existing AI Architectures Are Insufficient

The case for Sovereign Personal Intelligence rests not only on what SPI enables, but on the limitations of existing AI architectures for the specific requirements of personal AI governance. This section examines five categories of existing system and explains why each falls short of the sovereign personal intelligence criteria.

2A.1 Cloud-First AI Assistants (ChatGPT, Claude, Gemini)

Architecture: Cloud-native inference. User data processed on remote servers controlled by the platform operator. Local deployment not available.

Why they are insufficient for sovereign personal intelligence:

SPI Requirement	Cloud Assistant Capability	Gap
Local-first operation	Not supported. All inference requires cloud connectivity.	Cannot operate in air-gapped or low-connectivity environments. Data necessarily transits third-party infrastructure.
Consent-governed memory	Platform-managed memory with limited user visibility. No per-category consent granularity.	Memory is a feature of the platform, not a user-controlled resource. Withdrawal of consent is binary (delete account) rather than graduated.
Auditable decision processes	Black-box model outputs. No structured decision traces available to users.	Users cannot inspect why a particular response was generated or what data influenced it.
Verifiable security	Vendor claims of security practices. No independent verification available to users.	Trust is delegated to the platform operator. Users cannot independently verify security posture.
Infrastructure independence	Tied to specific cloud providers. Model lock-in by design.	Switching costs are structurally high. Data portability is limited or nonexistent.

Summary: Cloud AI assistants are designed for a different paradigm — centralised, platform-controlled, trust-delegated AI. They cannot be adapted to sovereign personal intelligence without fundamentally changing their architecture and business model.

2A.2 Local Open-Weight Models (Llama, Mistral, Qwen)

Architecture: Models that can run on local hardware. No governance, memory, or security layer.

Why they are insufficient for sovereign personal intelligence:

Running an open-weight model locally is often presented as a privacy solution, but a model alone does not constitute a personal intelligence system. The following capabilities are absent:

- **Memory system:** The model itself is stateless. No persistent, consent-governed memory exists without additional infrastructure.
- **Consent framework:** There is no mechanism to control what data is retained, how it is used, or under what conditions it can be deleted.
- **Security gates:** No prompt firewall, no content guard, no physical action blacklist. The model is exposed to the full range of injection and manipulation attacks.
- **Audit trail:** No decision trace capability. Every interaction is ephemeral.
- **Tool integration:** No structured tool registry, no permission gating, no escalation framework.

Summary: Open-weight models are a substrate, not a solution. Building sovereign personal intelligence from a bare model requires constructing the entire governance, memory, security, and orchestration layer — which is precisely what StarCaller provides.

2A.3 Retrieval-Augmented Generation (RAG) Systems

Architecture: Language model augmented with a vector database for external knowledge retrieval. Commonly used for document Q&A and knowledge base applications.

Why they are insufficient for sovereign personal intelligence:

RAG systems address the knowledge limitation of language models (they can cite external sources) but do not address any of the governance requirements:

- **No consent architecture:** RAG systems retrieve from whatever document store is provided. There is no graduated consent framework governing what can be retrieved or how retrieved data is used.
- **No memory governance:** Retrieved context is typically ephemeral. Persistent memory requires a separate system with its own governance.
- **No action control:** RAG systems are passive Q&A tools. They do not execute actions, manage permissions, or provide escalation frameworks.
- **No auditability:** Standard RAG systems do not produce structured decision traces suitable for external audit.

Summary: RAG is a useful information retrieval pattern but is not a personal intelligence architecture. It addresses one specific capability (knowledge-grounding) while leaving all governance, security, and autonomy requirements unaddressed.

2A.4 AI Agent Frameworks (LangChain, AutoGPT, CrewAI)

Architecture: Frameworks for building autonomous AI agents that can execute multi-step tasks using tools.

Why they are insufficient for sovereign personal intelligence:

Agent frameworks provide the execution layer (tools, task decomposition, planning) that Star-Caller's Ability OS also provides, but they omit the governance and security layers that make autonomous agents safe for personal use:

- **No security architecture:** Agent frameworks typically have no built-in prompt firewall, injection detection, or content guard. The agent executes whatever tools it chooses, constrained only by its system prompt.
- **No consent framework:** Agents in these frameworks have access to whatever tools and data the developer provides. There is no graduated, user-controlled consent layer.
- **No memory governance:** Memory, if implemented, is application-level — not consent-governed or auditable.
- **No escalation:** Agent frameworks do not include escalation engines for high-stakes decisions. The agent acts autonomously until it fails.
- **No security health monitoring:** No equivalent of the SSI exists. There is no quantified, continuous assessment of system integrity.

Summary: Agent frameworks provide the execution substrate but delegate all governance and security to the application developer. For personal AI, where the stakes include personal data, financial decisions, and potentially physical actions, this gap is critical.

2A.5 Enterprise Copilots (Microsoft Copilot, Salesforce Einstein)

Architecture: AI assistants integrated into enterprise SaaS platforms, operating on organisational data with enterprise-grade security.

Why they are insufficient for sovereign personal intelligence:

Enterprise copilots are designed for organisational contexts, not personal sovereign AI:

- **Data ownership:** Data belongs to the organisation, not the individual. The user is a data subject, not a data sovereign.
- **Consent by policy, not architecture:** Enterprise copilots operate under organisational data policies. The individual user cannot set independent consent boundaries.
- **No local deployment:** Most enterprise copilots are cloud-dependent. Sovereign deployment options are limited or nonexistent.
- **No personal memory:** Enterprise copilots are designed for organisational knowledge, not personal life management. They do not maintain longitudinal personal models.

- **No user portability:** Data is tied to the enterprise subscription. Users cannot export their AI model when changing employers.

Summary: Enterprise copilots solve a different problem — organisational productivity within an enterprise boundary. They are not designed for and cannot be adapted to personal sovereign AI.

2A.6 The Gap That SPI Addresses

None of the five categories above provides a complete sovereign personal intelligence system. Each addresses one or two dimensions — local execution, agentic capability, enterprise security, knowledge retrieval — but none provides all of:

- Local-first, air-gap capable architecture
- Consent-governed, graduated memory
- Auditable decision traces for every operation
- Quantified, verifiable security (SSI)
- Multi-layered security gates (10-gate system)
- Structured escalation for high-stakes decisions
- Model and infrastructure agnosticism
- Data portability

This is the gap that Sovereign Personal Intelligence is designed to fill. The following sections define the SPI framework and describe StarCaller as one reference implementation.

SECTION 3

Defining Sovereign Personal Intelligence

3.1 What Sovereign Personal Intelligence Is

Sovereign Personal Intelligence (SPI) is a new category of AI system defined by the following characteristics:

1. **Local-first architecture.** The system's primary operation occurs on hardware under the user's control or within a trusted sovereign boundary. Cloud connectivity is optional and limited to explicitly authorised functions.
2. **Consent-governed memory.** All persistent memory — personal data, preferences, history, learned patterns — is stored under a graduated consent framework. The user controls what is remembered, how long it is retained, and how it may be used.
3. **Auditable decision processes.** Every action taken by the system is recorded in a structured decision trace that is inspectable by the user and, where appropriate, by regulators or independent auditors.
4. **Verifiable security guarantees.** System integrity is continuously measured and reported through quantitative metrics — covering prompt integrity, memory integrity, identity confidence, tool safety, physical safety, supply chain health, oracle integrity, and inter-oracle trust.
5. **Model and infrastructure agnosticism.** The system does not depend on any single model provider or cloud infrastructure. It can operate with open-weight models, private model deployments, or sovereign compute resources.
6. **Interoperable data portability.** Personal data and system configuration can be exported in standard formats. The user is not locked into a vendor-specific ecosystem.

3.2 Core Principles

The framework is built on five principles:

Sovereignty. The individual's data and computational processes remain under their jurisdiction or that of a trusted sovereign entity. No third party has unauthorised access.

Consent. All data retention, processing, and sharing is governed by explicit, graduated, revocable consent. Consent is not a single checkbox — it is a continuous relationship.

Auditability. Every operation that affects personal data or exercises autonomous agency is recorded in a tamper-evident decision trace. Audit trails are structured for both human inspection and automated verification.

Portability. Data, configuration, and learned models are structured in open, exportable formats. The user can migrate between implementations without loss.

Verifiability. Security and trust properties are not claimed — they are measured. The system reports concrete, quantitative indicators of its integrity, available to both the user and independent assessors.

3.3 Contrast with Existing Paradigms

Dimension	Cloud AI	Open Model	Sovereign Personal Intelligence
Architecture	Cloud-native, remote inference	Local-capable model only	Full local-first platform
Memory	Platform-controlled, opaque	None (stateless model)	User-controlled, consent-governed
Consent	Single opt-out at account level	Not applicable	Graduated, continuous, revocable
Audit	Black-box, no user access	Not applicable	Full decision traces
Security	Vendor claims, no verification	No security layer	Quantified, verifiable (SSI)
Sovereign deployment	Not available	Model only, no governance	Full stack
Portability	None (vendor lock-in)	Model weights only	Full data and config portability

3.4 The Missing Market Category

The technology industry has established categories for different types of AI system: large language models, AI assistants, chatbots, agent frameworks, and autonomous agents. Few of these categories adequately describe a system that is simultaneously local-first, consent-governed, auditable, verifiably secure, and infrastructure-agnostic.

Sovereign Personal Intelligence is designed to fill this gap. It is not a better chatbot. It is not a cheaper API. It is a meaningfully different architectural category, designed for a world in which

personal AI systems manage increasingly sensitive domains of human life and where trust cannot be delegated to a third party.

The following sections describe StarCaller — a complete, working reference implementation of this category.

SECTION 4

Architecture: The StarCaller Reference Implementation

4.1 Architectural Overview

StarCaller is structured as six interconnected layers, each with distinct responsibilities and independent operational capability. The architecture follows a strict governance hierarchy: constitutional governance (ICID) over security governance (Shi Gandang) over operational governance (Prism Nexus) over specialised intelligence (Oracles) over execution (Abilities).

The system operates across two physical nodes: - **NucBox (Local Compute):** Primary runtime — Prism Nexus (port 8001), OFP Kernel (port 8181), Ollama local inference (port 11434), Telegram bridge, dashboards - **Zion0 (Storage VPS):** Persistent storage — Qdrant vector embeddings, Neo4j knowledge graph, PostgreSQL structured data, ntfy push notifications

The design ensures that personal data processing can occur on local hardware without mandatory cloud transfer. Cloud connectivity is limited to optional model inference (DeepSeek, OpenRouter) and encrypted vault sync to the storage VPS.

4.1a — Capability Maturity Summary

The following table summarises the current maturity status of StarCaller's major capabilities, distinguishing between what is already implemented, what is in active development, and what remains future research.

Capability	Status	Evidence
10-Gate Security Architecture (Shi Gandang)	IMPLEMENTED	136+ tests passing. Gates 1-7 (original), Gates 8-10 (Regolith extension operational). SSI at 97.8/100 (self-assessed, unaudited). Security Council operational.
Vault Memory System	IMPLEMENTED	11 SQLite tables, 13 import types supported, 84 items parsed, 88 memory claims extracted. Consent governance framework operational. Encrypted storage at rest.
Oracle Framework (8 Hands)	IMPLEMENTED	8 domain-specialised agents with tool-calling capability. ICID routing, council mode, PROA execution loop (10 iterations, 8 tool calls/turn).
Prism Nexus API Gateway	IMPLEMENTED	FastAPI server on port 8001. REST + SSE streaming. Multi-provider LLM routing with circuit-breaker failover.

Capability	Status	Evidence
Telegram Bridge	IMPLEMENTED	55+ commands. Proactive scheduling (morning briefing, reminders, mood check-in). Oracle routing via OFP kernel.
PROA Loop (Agentic Execution)	IMPLEMENTED	Perceive-Reflect-Optimize-Act orchestrator. Tool registry with ACL. Escalation Engine with 5-condition gating. Task Planner with DAG decomposition.
Identity Trust Fabric (Gate 8)	IMPLEMENTED (v1)	4-stage SASV pipeline. Spectral heuristics for spoof detection. 5-factor scoring engine with weight redistribution. Interface stable for AASIST/RawNet2 upgrade.
Oracle Integrity Monitor (Gate 10)	IMPLEMENTED	CUSUM drift detection across 5 dimensions. Golden test sets. Per-Oracle baseline profiles for 8 oracles.
Knowledge Graph	IMPLEMENTED	6,461 nodes, 8,858 edges from 377 Python files. Served via 3 OFP endpoints.
Executive Function Mode	IN DEVELOPMENT	Persona variant defined. Core logic prototyped. PROA loop integration pending.
Federated Vaults	RESEARCH	Cross-device memory synchronisation. Cryptographic primitives identified. Implementation pending funding.
Multi-Appliance Network	RESEARCH	Peer-to-peer oracle network. Collective learning without centralisation. Concept validated by split-brain topology.
Appliance Hardware	PLANNED	ARM SoC with NPU, 8+ TOPS. Hardware specification defined. Prototype targeted for Year 3 Q1.
Enterprise Pilot Programme	PLANNED	Pilot framework defined. 3 pilot accounts targeted for Year 2 Q3.

Note: "IMPLEMENTED" means the capability is functional in the reference deployment. It does not imply production-grade readiness for all deployment contexts. Security-critical components (Gates 1-10) have been tested but not independently audited.

4.2 ICID — Governance Layer [EXISTING]

ICID (the Conductor) is the constitutional layer — conceived as the highest authority in the StarCaller hierarchy. It does not generate content or execute actions. Instead, it governs:

- **Query routing:** Determines which Oracle(s) should handle each incoming query, based on domain, complexity, and security classification
- **Council synthesis:** When multiple Oracles are consulted on the same question, ICID synthesises their responses, reconciles disagreements, and produces a unified answer with citations

- **Constitutional override:** ICID is designed to hold veto power over other layers, including Shi Gandang. If a proposed action violates constitutional principles (Safety, Privacy, Fairness, Transparency, Accountability), ICID can block it or escalate
- **Executive Function Mode:** [IN DEVELOPMENT] A persona variant designed for executive dysfunction support — decomposes tasks, provides adaptive check-ins, and manages verbosity control for users with attention or organisational challenges

ICID's authority is defined in a written constitution that establishes six core principles against which all system behaviour is audited:

Principle	Operational Meaning
Safety	No endpoint shall produce harmful content. State-detector gates and content guards enforce this in every request path.
Privacy	PII scanning and masking on every user-facing data path. Vault data encrypted at rest. Logs scrubbed of secrets.
Fairness	Multi-perspective Oracle council design prevents any single Oracle from overriding others. Bias detection in research pipelines.
Transparency	Decision traces, /why command, confidence scoring, and source citations in every Oracle response.
Accountability	ActionLedger SQLite audit trail. Approval engine for outbound actions. All Oracle consults logged with trace IDs.

4.3 Shi Gandang — Security Layer [EXISTING]

Shi Gandang is the security governance layer — a 10-gate security architecture with a multi-Oracle Security Council. Every request, action, and memory operation passes through one or more gates before execution. Gates 1-7 are EXISTING. Gates 8-10 (Regolith extension) are EXISTING. Security Council is EXISTING. SSI is EXISTING (see Appendix for methodology).hat each request, action, and memory operation passes through one or more gates before execution.

The 10 Gates (Final v1 Architecture)

Key Security Components

Gate 8 — Identity and Trust Fabric: A 4-stage SASV (Spoofing-Aware Speaker Verification) pipeline that detects TTS synthesis, voice conversion, replay attacks, and partial deepfakes through spectral heuristics (MFCC, CQT, centroid, ZCR, RMS, duration). Multi-factor scoring across five dimensions (voice, device proximity, behavioural, historical, location) resolves to four authority tiers: Owner (85%+), Recognised (65%+), Unverified (40%+), Denied (under 40%). Unavailable factors redistribute weights proportionally.

Gate 10 — Oracle Integrity Monitor: CUSUM (Cumulative Sum) algorithm detects gradual drift in Oracle behaviour across five dimensions: domain focus, quality, persona adherence, manipulation attempts, and provider reliability. Weekly golden test sets validate oracle consistency. Per-Oracle baseline profiles are maintained for all 8 oracles.

Security Health Index (SSI): An aggregate, quantified measure of system integrity across eight weighted components: - Prompt Integrity (18%) — Injection resistance - Memory Integrity (18%) — Vault poisoning prevention - Supply Chain (15%) — Dependency health - Tool Safety (12%) — Execution guard effectiveness - Identity (12%) — Authentication confidence - Oracle Integrity (10%) — Drift detection coverage - Physical Safety (8%) — Action gating - Inter-Oracle Trust (7%) — Cross-signing integrity

Current internal assessment: **HARDENED** (SSI top band, 90-100; methodology in Appendix).

Security Council: A multi-Oracle body comprising Shi Gandang (Chair), Sia, Nostradamus, Pythia, and ICID (veto). Six auto-convene triggers include SSI below 60, quality CRITICAL/SEVERE events, drift SEVERE alerts, identity spoof detection, and physical irreversible actions with identity below 75%. Five graduated response options: MONITOR to CONTAIN to SUSPEND to ESCALATE to LOCKDOWN. ICID's veto forces ESCALATE as the minimum response regardless of majority vote.

4.4 The Vault — Memory Layer [EXISTING with gaps]

The Vault is the encrypted, consent-governed memory system. It implements a graduated consent framework where every data category (calendar, email, browser data, financial, contacts, search history) is independently permissioned. Core pipeline (13 import types, 84 items parsed) is EXISTING. Federated vaults are FUTURE RESEARCH (see Appendix C).

Architecture

Pipeline stages: Quarantine to PII Scan (Shi Gandang) to Parse to Classify to Store - 13 import types supported (ICS calendar, MBOX email, CSV browser data, JSON search history, vCards, Stripe exports, ZIP archives) - 84 items parsed, 88 memory claims extracted across 6 categories (Fact, Interest, Pattern, Relationship, Habit, Skill) - All imports consent-gated — per-category permissions enforced at storage time

Storage: - Local: SQLite (11 tables) — provenance, consent rules, encrypted content - Remote: PostgreSQL on Zion0 — encrypted vault sync - Vector: Qdrant on Zion0 — embedding-based semantic memory retrieval

4.5 Ability OS — Execution Layer [EXISTING]

The Ability OS is the execution substrate. Tools are registered with the ToolRegistry, gated by an access control list, and executed by the PROA Loop orchestrator. 20+ tools, PROA Loop, Escalation Engine, and Task Planner are all EXISTING.

Registered Tools (20+)

Tool	Function	Invoking Oracles
Web search	Real-time internet queries	Sia, Nostradamus, Pythia
Browser sandbox	Isolated web browsing	Sia, Anansi
Shell execution	System commands	ICID, Yhi
Archiver	URL content archiving	Sia
Git operations	Repository management	ICID
Grail search	Sanctuary passage retrieval	ICID (Grail)
News search	Structured news queries	Nostradamus
Image generation	DALL-E / Stable Diffusion	Anansi
Voice synthesis	TTS via 8 Oracle voices	All Oracles
Image analysis	Vision via Fable 5	Pythia

PROA Loop

The PROA (Perceive-Reflect-Optimize-Act) Loop is the agentic execution engine:

1. **Perceive:** Receive query + context + available tools
2. **Reflect:** The Oracle evaluates its knowledge against the query, considering confidence and gaps
3. **Optimize:** Selects the best approach — direct answer, tool invocation, or multi-step chain
4. **Act:** Executes the chosen approach, iterating through tool calls as needed

Each loop cycle currently supports up to 10 iterations and 8 tool calls per turn. A Reflect+Store phase captures learnings after each cycle. An Escalation Engine monitors the loop with 5 conditions that trigger human-in-the-loop review when decisions involve irreversibility, constitutional boundaries, low confidence, novelty, or resource constraints.

4.6 Prism Nexus — Orchestration Layer [EXISTING]

Prism Nexus is the primary API gateway (FastAPI, port 8001) and orchestration layer. Fully operational on local compute node. Dashboard integration is EXISTING but not continuously running. Zion0 bridge is EXISTING. It provides:

- **REST endpoints:** Oracle consultation, Sanctuary sessions, Autoresearch, Agent core, Escalation Engine, Task Planner

- **SSE streaming:** Real-time response streaming for long-running consultations
- **Multi-provider LLM routing:** Unified CloudLLMProvider routes requests across DeepSeek, OpenRouter, Gemini, and Ollama with automatic circuit-breaker failover
- **Rate limiting:** Per-endpoint rate limits
- **CORS middleware:** Configured for dashboard and API clients

4.7 Oracle Framework — Specialised Intelligence Layer [EXISTING]

All 8 Oracle hands are EXISTING with tool-calling capability. Multi-Oracle council mode with ICID synthesis is EXISTING. The Oracle Framework provides 8 domain-specialised agents ("hands"), each with independent operational capability.

Oracle	Domain	Key Capabilities
ICID	Governance	Constitutional oversight, query routing, council synthesis
Sia	Research	Deep research, fact verification, source analysis, bias detection
Nostradamus	Prediction	Market patterns, trend analysis, forecasting, event analysis
Pythia	Strategy	Strategic planning, decision analysis, financial consultation
White Buffalo	Ecology	Environmental data, sustainability, nature monitoring
Anansi	Narrative	Content creation, storytelling, social publishing
Shi Gandang	Security	Content policy enforcement, security gating, threat detection
Yhi	Physics	Physical world modeling, sensors, IoT, energy monitoring

Oracle system prompt design: Every Oracle persona receives a 5-item base instruction block appended by the PROA loop: 1. Self-Sufficiency Bias — prefer independent reasoning over tool dependency 2. Check All Sources — verify before reporting; cite sources 3. Proactive Memory Writing — record non-transient knowledge proactively 4. Calibrated Pushback — correct the user constructively when appropriate 5. No Retry on Same Failed Path — escalate or change approach after failure

4.8 Architecture Diagrams

Data Flow — Standard Consultation

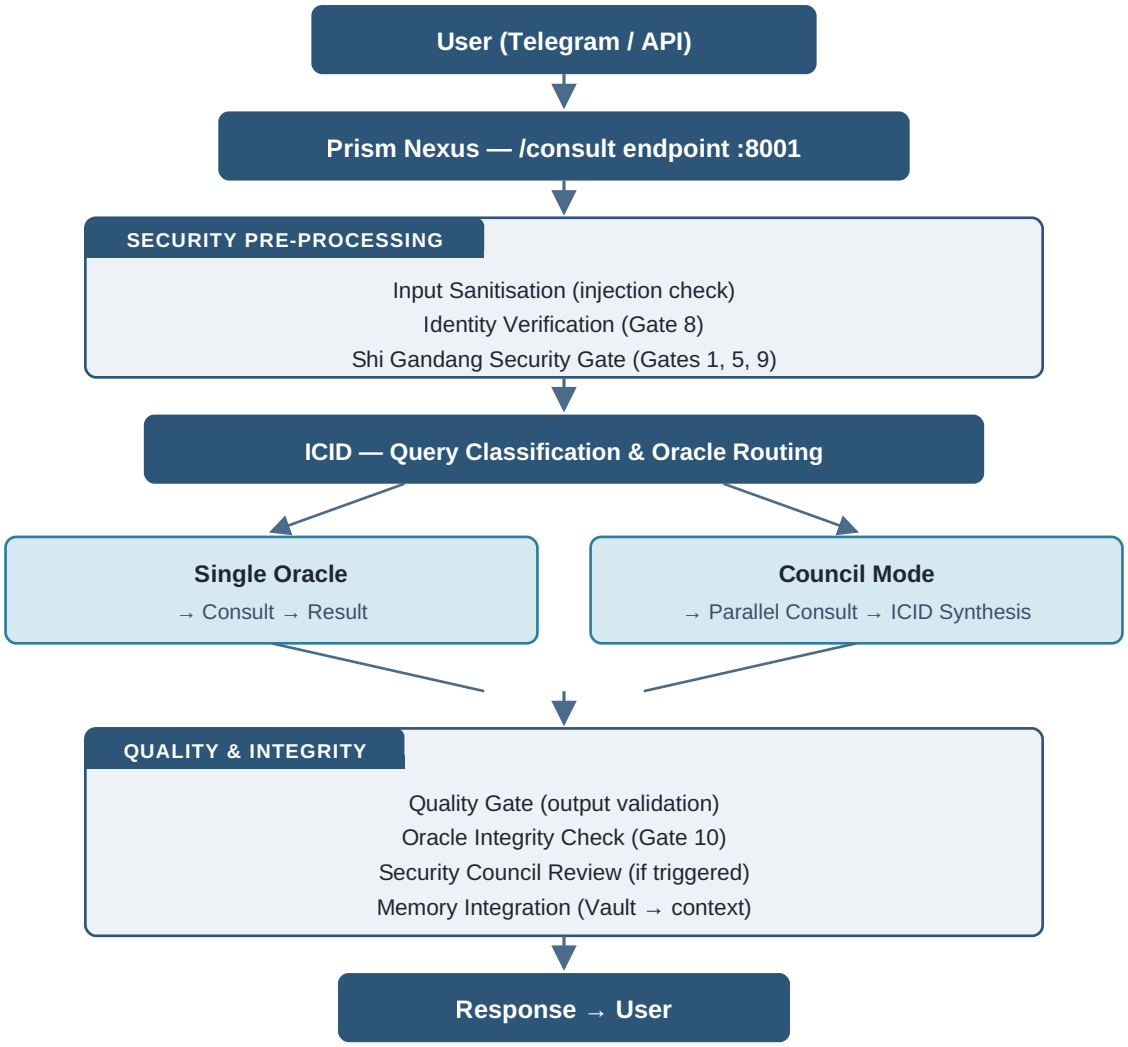


FIGURE 1 — STANDARD CONSULTATION DATA FLOW

Data Flow — Memory Operation

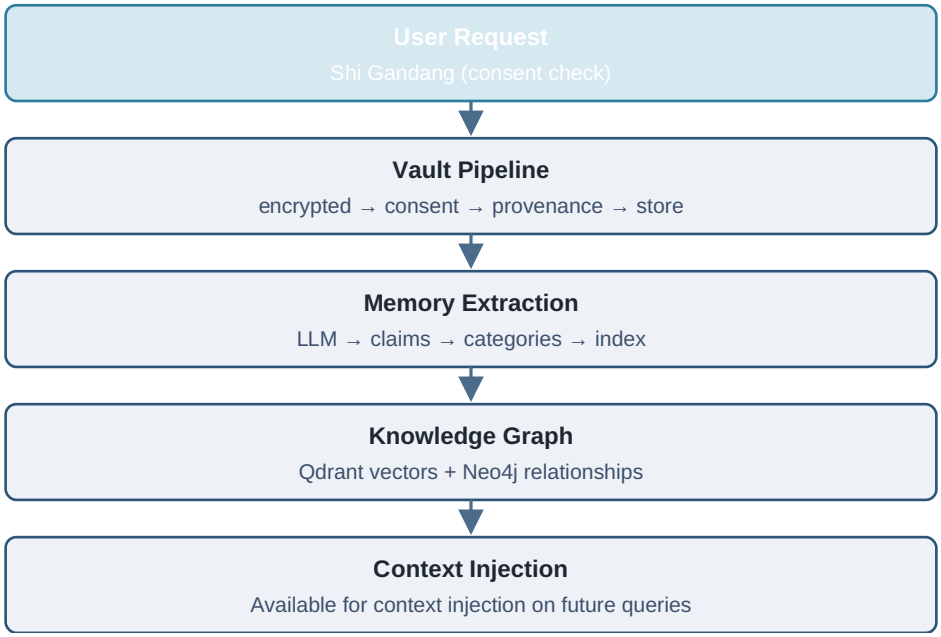


FIGURE 2 — MEMORY OPERATION DATA FLOW

Deployment Architecture

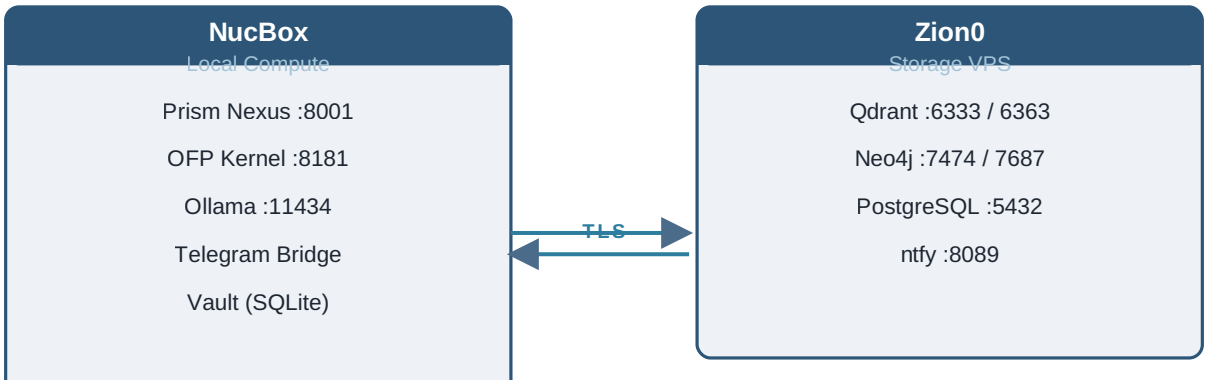


FIGURE 3 — DEPLOYMENT ARCHITECTURE

SECTION 5

Trust Architecture

5.1 Memory Governance

The Vault's memory governance framework is designed as the foundation of user trust. Every piece of stored information has:

- **Provenance:** Who created it, when, from what source, under what consent rule. Tracked through HMAC-SHA256 signed vault entries with content hashing.
- **Category:** Calendar, email, browser data, financial, contact, search history. Each category has independent consent and retention policies.
- **Lifetime:** Configurable per-category retention periods. Automatic expiration for time-bounded data.
- **Consent state:** Each data category is independently permissioned through the approval engine.

The Memory Integrity Monitor (Gate 2) continuously scans the Vault for poisoning attempts — unauthorised writes, conflicting facts, consent violations.

5.2 Consent Governance

Consent in StarCaller is not a single checkbox. It is a continuous, graduated, revocable framework:

- **Tiered access:** Consent rules operate per category, per action type. Reading calendar data requires different consent than reading email content.
- **Default deny:** All capabilities are disabled by default. Users explicitly grant access to each domain.
- **Revocable at any time:** Consent can be withdrawn through the management interface. Withdrawn consents stop all access and trigger data deletion where retention permits.
- **Audit trail:** Every consent grant, modification, and revocation is logged to the Action Ledger with timestamp and context.

By design, the system does not degrade functionality when consent is withdrawn — it reports the capability as unavailable and explains how to re-enable it.

5.3 Security Gates

The 10-gate architecture provides defence in depth. Every request passes through multiple independent checks:

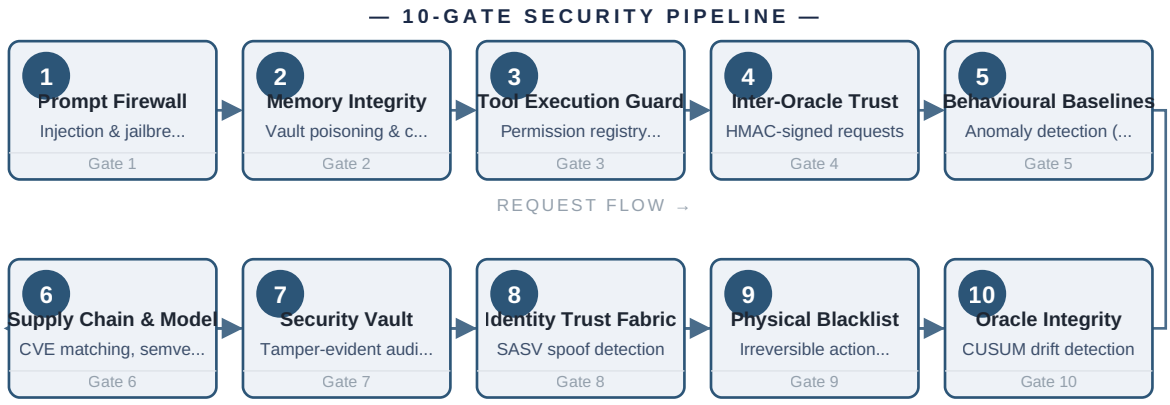


FIGURE 5 — 10-GATE SECURITY PIPELINE

1. **Prompt Firewall** — Detects prompt injection, jailbreak attempts, and goal hijacking across 8 injection pattern families
2. **Memory Integrity Monitor** — Scans for vault poisoning and consent violations
3. **Tool Execution Guard** — Validates every tool invocation against permission registry and action safety rules
4. **Inter-Oracle Trust** — HMAC-signed requests ensure no Oracle can impersonate another
5. **Behavioural Baselines** — Welford's online algorithm detects anomalous usage patterns per device
6. **Supply Chain and Model** — CVE matching against known vulnerabilities, semver range comparison
7. **Security Vault** — Tamper-evident audit storage with encrypted threat intelligence feeds
8. **Identity Trust Fabric** — SASV spoof detection and multi-factor authority resolution
9. **Physical Blacklist** — Time-aware gating of irreversible physical actions
10. **Oracle Integrity Monitor** — CUSUM drift detection with per-Oracle golden test sets

5.4 Auditability and Decision Traces

The architecture produces a structured decision trace for every operation:

```
Request: "What's my carbon footprint this month?"
  |
[Input Gate] Passed -- no injection detected
[Identity] Passed -- Owner tier, confidence 92%
[Query Classification] Intent: carbon_query, Domain: ecology
[Oracle Selection] White Buffalo (ecology) + Nostradamus (data)
  |
White Buffalo consult -> Tool: web_search -> Results -> Reasoning trace
Nostradamus consult -> Tool: web_search -> Results -> Reasoning trace
  |
[Quality Gate] Passed -- confidence 0.89, source count 4
[Integrity Check] No drift detected
  |
ICID Synthesis -> Combined response with citations
  |
Response delivered. Trace ID: TRC-20260630-3A7F. Logged to Action Ledger.
```

Each trace record is: - **Immutable:** HMAC-chained to prevent tampering - **Inspectable:** Available through the /why command on Telegram - **Portable:** Exportable in JSON format for external audit - **Privacy-preserving:** PII redacted from logged traces where possible

5.5 Explainability and Reasoning Provenance

Oracle responses include structured metadata:

- **Confidence score:** 0-1 scale reflecting the Oracle's certainty in its response
- **Citations:** Source URLs or references for factual claims
- **Dissent notes:** Alternative viewpoints from other Oracles, if the council was consulted
- **Human review flag:** Indicates whether the response was reviewed by a human or is fully automated
- **Tool calls made:** Record of which tools were invoked during generation
- **Duration:** Generation time in milliseconds, providing transparency on system performance

5.6 Security Health Index

The Security Health Index (SSI) is the system's quantifiable integrity measurement. It provides both real-time and historical assessment across 8 weighted components.

Current Assessment: 97.8/100 — HARDENED Band (90-100) (self-assessed, unaudited)

Component Breakdown:

Component	Weight	Current Status
Prompt Integrity	18%	Fully instrumented

Component	Weight	Current Status
Memory Integrity	18%	Fully instrumented
Supply Chain	15%	CVE scanning active
Tool Safety	12%	Execution guard active
Identity Confidence	12%	82% — no physical biometric sensors
Oracle Integrity	10%	CUSUM drift detection active
Physical Safety	8%	Blacklist active
Inter-Oracle Trust	7%	All oracles signing

The SSI is accessible through the API endpoint GET /ofp/v1/ssi and returns a structured report including component-level scores, change detection, and actionable recommendations.

SECTION 6

National Strategic Relevance

6.1 UK AI Safety Objectives

StarCaller directly supports four of the five pillars of the UK AI Safety Institute (AISI) technical research agenda:

1. General-purpose systems evaluation: StarCaller's Security Health Index provides continuous, quantitative assessment of system integrity — a framework that could be extended to evaluate any AI system's safety posture. The SSI methodology (8 weighted components, 5-band classification, real-time scoring) offers a template for standardised AI safety metrics.

2. Synthetic content detection: Gate 8's SASV pipeline includes detection of synthetic voice content (TTS, voice conversion). The spectral heuristic approach (MFCC, CQT, centroid, ZCR, RMS, duration) is architecture-agnostic — compatible with AISI's evaluation methodologies for synthetic media detection.

3. System safety evaluation: The 10-gate architecture provides a comprehensive safety evaluation framework. Each gate tests a specific safety property — injection resistance, consent compliance, behavioural consistency, content policy enforcement.

4. Societal impact analysis: StarCaller's consent-governed memory architecture and graduated trust framework address the societal risks of pervasive personal AI — data sovereignty erosion, vendor lock-in, and the trust deficit in centralised systems.

6.2 UK Sovereign AI Objectives

The UK Sovereign AI Fund (GBP 500 million, April 2026) and the AI Opportunities Action Plan's commitment to 20x sovereign compute expansion by 2030 create a strategic window for a UK sovereign personal intelligence platform.

Hardware sovereignty alone may not be sufficient. The GBP 1.1 billion sovereign AI infrastructure investment addresses compute, but without a software and governance layer designed for sovereign deployment, the resulting systems handle personal data on domestically-hosted but architecturally foreign AI infrastructure.

StarCaller is designed to address this gap by providing:

- **A complete sovereign AI software stack** — not just a model, but governance, security, memory, orchestration, and specialised intelligence, all designed for local-first deployment

- **Model agnosticism** — the system can operate with open-weight models (Llama, Mistral, Qwen), fine-tuned sovereign models, or private deployments of frontier models
- **Infrastructure independence** — the system can run on sovereign compute (UK data centres, edge nodes) or local consumer hardware. No mandatory cloud dependency
- **Validated security** — 136 passing tests, Security Health Index assessed at HARDENED level (self-assessed, unaudited)

6.3 Digital Resilience

Centralised AI infrastructure creates systemic risk. StarCaller's distributed architecture provides resilience through:

- **Local fallback:** If cloud models are unavailable, the system operates on local Ollama inference. Core capabilities — memory, consent, security gates — function without any network connectivity.
- **No single point of failure:** The NucBox and Zion0 nodes operate independently. Vector search, graph queries, and structured storage each run on separate Docker services.
- **Data portability:** The Vault stores data in structured, exportable formats. The user is not locked into any vendor or infrastructure provider.
- **Air-gap capable:** The system can, in principle, operate without internet connectivity. Cloud model access is an optional enhancement.

6.4 Privacy Enhancement

StarCaller's architecture provides privacy properties that are difficult for centralised AI systems to replicate:

- **Local-first processing:** All personal data operations occur on hardware under the user's control.
- **PII scanning on every path:** Shi Gandang's PII detection engine scans all data paths for personally identifiable information.
- **Encrypted vault storage:** All vault data is encrypted at rest. Key management and encryption occur within the trusted execution boundary.
- **Complete session erasure:** Cryptographic erasure of all session data on demand.
- **Consent as infrastructure:** Privacy is an architectural property enforced at every data operation.

6.5 Citizen Trust and Secure Agentic Systems

The UK Government's AI regulation white paper (2023) established the principle that AI systems should be safe, secure, and trustworthy. StarCaller implements these principles in working code:

- **Transparency by design:** Every Oracle response includes confidence scores, citations, and dissent notes.
- **Accountability by architecture:** The Action Ledger records every significant system action with trace IDs.
- **Proportionality in action:** The graduated consent framework ensures no operation exceeds its authorisation.
- **Continuous verification:** The Security Health Index provides real-time evidence that the system remains within its trust boundaries.

6.6 Why Sovereign Personal Intelligence Complements Sovereign Compute

The UK Sovereign AI Fund's mandate is to build British AI champions by investing in compute, infrastructure, and foundational AI capabilities. This is a necessary condition for sovereign AI — but it is not sufficient.

The Limits of Compute-Only Strategy

Sovereign compute infrastructure — GPU clusters, data centres, national AI research resources — provides the hardware substrate for AI development. However, compute alone does not create trusted AI systems that individuals and institutions can use for personal data processing, life management, and autonomous decision-making. The software and governance layers that sit above the compute layer determine how AI systems handle personal data, maintain consent boundaries, and provide verifiable trust.

The UK's GBP 1.1 billion sovereign compute investment will produce world-class computational resources. Without a complementary investment in sovereign AI software infrastructure — governance frameworks, memory architectures, security standards, orchestration layers — those computational resources will run the same centralised, opaque AI platforms that create the sovereignty problem in the first place.

What SPI Adds to Sovereign Compute

Compute Investment	SPI Complements
GPU clusters for model training and inference	Local-first execution layer that can use those GPUs without cloud dependency
National AI research infrastructure	Governance and security framework that determines how AI systems use those resources

Compute Investment	SPI Complements
Open-weight model ecosystem (Llama, Mistral, Qwen)	Memory, consent, audit, and security layers that make models safe for personal use
Data centre capacity	Air-gap capable deployment architecture that can run on sovereign infrastructure
AI startup funding	Commercialisation pathway for sovereign AI products (StarTeQ → Software → Appliance)

The Strategic Proposition

Sovereign Personal Intelligence is the software infrastructure that makes sovereign compute usable for personal AI. Without SPI, sovereign compute runs cloud-style AI on sovereign hardware — an improvement in infrastructure ownership but not in data sovereignty, consent governance, or verifiable trust. With SPI, sovereign compute can deliver the full sovereign AI value proposition: local-first processing, consent-governed memory, auditable decisions, and verifiable security guarantees, all running on UK sovereign infrastructure.

This is not a competition with the Sovereign AI Fund's compute investment. It is a complementary investment in the software layer that gives sovereign compute its purpose and its trustworthiness.

SECTION 6A

Regulatory Alignment

6A.1 UK Data Protection Framework

StarCaller's architecture is designed to align with the United Kingdom's data protection regime, principally the UK General Data Protection Regulation (UK GDPR) as retained and amended post-Brexit, and the Data Protection Act 2018.

Data minimisation and purpose limitation (Article 5(1)(b)-(c)): The Vault's per-category consent framework ensures that data is collected and stored only for specified, explicit, and legitimate purposes. Each data category (calendar, email, browser data, financial, contact, search history) is independently permissioned, and the system does not process data outside the scope of the consent granted. The default-deny posture — where no data is retained unless the user explicitly consents — exceeds the minimum requirements of Article 5.

Storage limitation (Article 5(1)(e)): Per-category retention policies and automatic expiration for time-bounded data operationalise the principle that personal data must be kept no longer than necessary. The system architecture enforces retention limits at the storage layer, not merely at the policy layer.

Data subject rights (Articles 12-23): StarCaller's consent governance framework directly implements: - **Right of access (Article 15):** The Vault's provenance tracking enables users to inspect all stored data, its sources, and its consent basis. - **Right to erasure (Article 17):** The Sanctuary module provides cryptographic erasure of session data. Per-category consent withdrawal triggers deletion of associated data where retention permissions do not apply. - **Right to data portability (Article 20):** Data is stored in structured, exportable formats. The sovereign memory standards research agenda (Section 10.5) proposes formalising this into an interoperable open format. - **Right to object to automated decision-making (Article 22):** The Escalation Engine ensures that high-stakes decisions receive human oversight. The PROA Loop's decision traces provide the transparency necessary for meaningful objection.

Accountability principle (Article 5(2)): The Action Ledger provides a tamper-evident audit trail for all significant system operations. The Security Health Index provides quantifiable accountability metrics.

6A.2 Data Protection and Digital Information Bill

The Data Protection and Digital Information (DPDI) Bill represents the UK's post-Brexit evolution of its data protection framework. Key provisions relevant to sovereign personal AI include:

- **Reformed consent requirements:** The Bill's adjustments to consent rules are compatible with StarCaller's graduated consent framework, which provides more granularity than the legislation requires.
- **International data transfers:** StarCaller's local-first architecture eliminates the need for international transfer mechanisms for personal data processing — a significant compliance advantage over cloud-dependent systems that must rely on adequacy decisions, standard contractual clauses, or binding corporate rules.
- **Smart data schemes:** The Bill's provisions for secure data sharing across sectors align with the federated vaults and multi-appliance intelligence research agenda (Sections 10.1-10.2).

6A.3 EU AI Act

While the EU AI Act is a European regulation, its extraterritorial effect and influence on UK regulatory thinking make it relevant to any AI system targeting UK markets, particularly those with international ambitions.

Risk classification: StarCaller is designed to operate within the EU AI Act's risk framework: - **Minimal risk:** General-purpose Oracle consultations, content generation, and information retrieval fall below the Act's regulatory threshold. - **Limited risk:** Proactive scheduling, life monitoring, and autonomous agentic actions require transparency obligations that StarCaller's decision trace system is designed to satisfy. - **High risk:** The Escalation Engine, Security Council, and SSI system provide the governance and human oversight infrastructure that the Act requires for high-risk AI systems. The 10-gate security architecture provides the technical safeguards equivalent to those the Act's conformity assessment procedures would evaluate.

General-purpose AI (GPAI) compliance: As a system that integrates general-purpose AI models (via the CloudLLMProvider's model-agnostic architecture), StarCaller's transparency obligations under the Act would be met by its existing decision trace, provenance tracking, and consent governance infrastructure.

6A.4 NCSC Secure by Design

The UK National Cyber Security Centre's Secure by Design principles provide a framework for building systems that are secure from inception rather than retrofitting security.

StarCaller's architecture aligns with all five principles:

NCSC Principle	StarCaller Implementation
Secure by default	Default-deny consent posture. All capabilities disabled until user explicitly grants access.
Secure by design	10-gate security architecture integrated at the architectural layer, not bolted on as middleware.
Secure in deployment	Vault pipeline includes PII scanning, provenance tracking, and encrypted storage at rest.
Secure in operation	Continuous SSI monitoring, Security Council auto-convene on threshold breaches, Action Ledger audit trail.
Proportionate security	Graduated identity trust (Gate 8), risk-calibrated escalation (Escalation Engine), tiered authority resolution.

6A.5 ISO Standards

ISO/IEC 42001:2023 — Artificial Intelligence Management System: This standard specifies requirements for establishing, implementing, maintaining, and continually improving an AI management system. StarCaller's architecture maps to several of the standard's control areas: - **AI policies (Clause 5.2):** The ICID constitution provides written governance principles. - **Risk assessment (Clause 6.1):** The Security Council's six auto-convene triggers provide structured risk escalation. - **AI system impact assessment (Clause 6.1.2):** The SSI provides a continuous, quantified impact assessment across eight dimensions. - **Documented information (Clause 7.5):** The decision trace framework and Action Ledger provide structured, auditable documentation.

ISO/IEC 27001:2022 — Information Security Management: The Vault's encrypted storage, consent governance framework, and provenance tracking align with the standard's requirements for: - Access control (A.9): Per-category consent permissions and tiered authority resolution. - Cryptography (A.10): Hardware-backed encryption for vault data at rest. - Operational security (A.12): Continuous monitoring via SSI and Security Council. - Incident management (A.16): Escalation Engine provides structured incident response.

6A.6 AI Safety Institute (AISI) Principles

The UK AI Safety Institute's technical research agenda identifies several principles for safe AI systems that StarCaller's architecture addresses:

- **Evaluation:** The SSI provides a continuous evaluation framework that could be adapted for AISI's general-purpose systems evaluation methodology.
- **Transparency:** The decision trace framework provides structured, auditable transparency at the operation level.
- **Testing:** The 136+ security tests provide a baseline for safety testing methodology.

- **Monitoring:** The Oracle Integrity Monitor (Gate 10) provides continuous behavioural monitoring through CUSUM drift detection.

6A.7 Regulatory Roadmap

StarCaller is not currently certified against any of the standards referenced above. The regulatory alignment described in this section is architectural — the system is designed to facilitate certification, but formal certification processes have not been undertaken.

The planned regulatory roadmap is:

Standard / Framework	Target Certification	Target Timeline	Dependencies
ISO/IEC 27001	Information Security	Year 2	Formal ISMS implementation, external auditor engagement
ISO/IEC 42001	AI Management System	Year 2	AI policy documentation, impact assessment methodology
NCSC Secure by Design	Self-assessment	Year 1 Q3	Documentation gap analysis
EU AI Act (GPAI)	Code of Practice	Year 2	Finalisation of Act's implementation timeline
Cyber Essentials Plus	Basic security	Year 1 Q2	Infrastructure audit
UK GDPR compliance	Ongoing	Current	Core architectural compliance already in place

SECTION 7

Economic Impact

7.1 Addressable Markets

StarCaller operates at the intersection of several large and growing markets:

Personal AI Assistants ($2.23B in 2024, projected $56.3B by 2034) [10][13]: The personal AI assistant market is expanding beyond simple scheduling toward deeper integration with daily life. CAGR of 38.1%. StarCaller's sovereign architecture addresses the trust gap limiting adoption in sensitive domains.

Agentic AI Systems ($52.6B projected by 2030, CAGR 46.3%) [17]: Autonomous AI agents represent the fastest-growing segment. StarCaller's PROA Loop, Escalation Engine, and Ability OS provide a complete agentic framework with built-in safety and consent governance.

Edge AI Hardware ($58.9B projected by 2030) [16]: The shift toward on-device AI processing creates a hardware market that StarCaller's software stack complements. The planned StarCaller Appliance targets this market.

Sovereign AI Infrastructure ($12B+ UK government commitment) [3][4]: The UK's sovereign AI investment programme creates a direct funding and procurement channel. StarCaller's alignment with UK AI strategy positions it for institutional adoption.

Privacy-First Consumer Technology ($5.5B estimated) [15]: The privacy-focused technology market is, in the authors' assessment, underserved by current AI platforms. StarCaller's privacy-by-architecture approach is positioned to address this market.

7.2 Market Forecasts

Market	2024 Value	2030/34 Projection	CAGR	StarCaller Opportunity
Personal AI Assistants	$2.23B	$56.3B (2034)	38.1%	Category-defining position
Agentic AI Systems	$3.1B	$52.6B (2030)	46.3%	Differentiated trust architecture
Edge AI Hardware	$8.2B	$58.9B (2030)	32.8%	Software stack for appliance market
Sovereign AI Infrastructure	$1.1B (UK 2025)	$12B+ (2030)	48%+	Direct institutional procurement

Market	2024 Value	2030/34 Projection	CAGR	StarCaller Opportunity
Privacy Technology	$1.8B	$5.5B (2030)	17.3%	Architectural differentiator

Revenue Model — 3-Stage Commercialisation:

Stage 1 — StarTeQ (Year 1-2): B2B SaaS for UK trades (Craft/Plumb/Chill AI), maritime (Meridian), housing (Hearth), faith organisations (Covenant). GBP 0.5-1.5M ARR target. Six Cloudflare Worker products in development.

Stage 2 — StarCaller Software (Year 2-3): Personal sovereign AI subscription. Freemium with local-only base, premium tier for cloud model access. GBP 5-15/month consumer, GBP 50-200/user/month enterprise.

Stage 3 — StarCaller Appliance (Year 3-5): Purpose-built hardware for sovereign AI deployment. Target: government, healthcare, legal, financial services. GBP 999-2,999 per appliance with annual subscription.

7.3 UK Industrial Strategy Alignment

StarCaller directly supports multiple UK industrial strategy objectives:

AI Opportunities Action Plan (January 2025): StarCaller contributes to adoption through a deployable sovereign AI platform; to infrastructure through model-agnostic architecture that can utilise UK sovereign compute; and to skilling through an open-source ecosystem around sovereign personal AI.

UK Sovereign AI Fund (GBP 500M, April 2026): StarCaller is a mature, tested reference implementation of sovereign personal AI infrastructure — not theoretical but a working system with 136+ security tests and a Security Health Index assessed at HARDENED level (self-assessed, unaudited; methodology detailed in Appendix).

National Data Strategy: StarCaller's consent-governed memory architecture aligns with the government's objective of trustworthy data use. The graduated consent framework could become a reference standard.

UK Cyber Security Strategy: The 10-gate security architecture, continuous SSI monitoring, and Security Council governance model support the strategy's objectives of resilient, verifiably secure systems.

Innovate UK / UKRI Funding Priorities: The 2026-27 funding rounds identify sovereign AI infrastructure, trustworthy AI systems, and privacy-enhancing technologies as priorities. StarCaller addresses all three simultaneously.

SECTION 8

Commercialisation Strategy

8.1 Stage 1: StarTeQ — Revenue Foundation (Year 1-2)

The StarTeQ product line provides the immediate revenue foundation for StarCaller development. Designed as B2B SaaS products targeting sectors of the UK economy that are underdigitised and have acute regulatory compliance needs, StarTeQ products share common infrastructure (Cloudflare Workers, KV storage, Stripe billing) with the broader StarCaller platform.

Product Portfolio

Product	Sector	Target ARR* (Year 2)	Stage
Craft AI	Construction trades	GBP 50-100K (illustrative)	Live, Tier 2
Plumb AI	Plumbing / heating	GBP 50-100K*	Live, Tier 2
Chill AI	HVAC / refrigeration	GBP 50-100K*	Live, Tier 2
CRA Watch	Compliance monitoring	GBP 30-60K*	Phase 1 complete
Hearth	Social housing	GBP 30-60K*	Demo data, needs pilot
Meridian	Maritime operations	GBP 30-60K*	Demo data, needs pilot
Covenant	Faith organisations	GBP 20-40K*	Demo data, needs pilot
Passage	Death care / funerals	GBP 20-40K*	Full docs, needs pilot

Pricing Model

Each StarTeQ product follows a tiered subscription model: - **Free tier:** Limited monthly queries, basic features — market adoption, pipeline building - **Professional tier (GBP 29-79/month):** Full feature set, compliance tooling, scheduling - **Enterprise tier (GBP 199-499/month):** Multi-user, custom integrations, dedicated support

> * All revenue targets are **illustrative commercial scenarios** based on current product-stage estimates. They are not forecasts. Actual revenue will depend on market adoption, pricing decisions, and competitive dynamics.

Key Milestones — Stage 1

Milestone	Target Date	Dependencies
3 live products	Month 3	Craft, Plumb, Chill AI deployed
First GBP 10K MRR	Month 6	150-200 professional subscribers
5 products live	Month 9	CRA Watch launched (excl. Hearth, which is at needs-pilot stage)
8 products live	Month 12	Full portfolio (contingent on pilot outcomes), GBP 50K+ MRR
Platform unification	Month 15	Common backend, unified billing, cross-product AI

> * All financial targets in this section are **illustrative commercial scenarios**. They represent planning assumptions, not commitments or forecasts.

StarTeQ is designed to serve multiple purposes: it generates revenue that funds StarCaller development, provides operational proof that the underlying governance concepts can be deployed in production environments, and builds the operational infrastructure (Cloudflare Workers, Stripe billing, AI model deployment, customer support) that StarCaller requires. By Year 2, StarTeQ revenue is projected to fully cover the StarCaller software development team.

8.2 Stage 2: StarCaller Software — Platform Adoption (Year 2-3)

Following StarTeQ revenue stabilisation, StarCaller launches as a standalone personal sovereign AI platform.

Product Tiers

Free Tier (GBP 0/month): - Local Ollama inference (qwen2.5:7b model) - Core Oracle consultations (8 oracles, single-question format) - Basic memory (limited vault storage, 30-day retention) - Grail Sanctuary access (1 session/day) - Telegram interface only

Premium Tier (GBP 9.99/month): - Cloud model access (DeepSeek, OpenRouter for complex queries) - Full Oracle consultations with PROA loop and tool calling - Unlimited vault storage with consent-governed memory - Full Grail Sanctuary (unlimited sessions, council mode) - Dashboard access (web + Tauri desktop app) - Voice synthesis (8 Oracle voices via edge_tts) - Priority support

Pro Tier (GBP 19.99/month): - Everything in Premium, plus: - All cloud model providers (including Fable 5 vision) - Advanced memory (Qdrant vector search, Neo4j knowledge graph) - Executive Function Mode - Escalation Engine and Task Planner access - API access for custom integrations - Early access to new features

Enterprise Tier (GBP 50-200/user/month): - Everything in Pro, plus: - Sovereign deployment on customer infrastructure - Dedicated Ollama model fine-tuning - Custom Oracle persona develop-

ment - SSO, audit logging, compliance reporting - SLA guarantees - On-premise deployment option

Go-to-Market Strategy

- **Phase 1 (Year 2 Q1-Q2):** Open-source release of core StarCaller framework. Target: developer community, privacy advocates, AI safety researchers. Build community contributions and ecosystem.
- **Phase 2 (Year 2 Q3-Q4):** Premium tier launch with consumer marketing. Target: privacy-conscious professionals, early AI adopters, tech-savvy households.
- **Phase 3 (Year 3):** Enterprise tier with direct sales. Target: SMEs, professional services firms, local government.

User Acquisition Targets

Metric*	Year 2	Year 3	Year 4
Free users	50,000	200,000	500,000
Premium subscribers	2,000	10,000	30,000
Pro subscribers	500	3,000	10,000
Enterprise accounts	10	50	200
Monthly Recurring Revenue	GBP 45K	GBP 250K	GBP 800K

8.3 StarCaller Appliance — Sovereign Hardware [SUMMARY]

The third commercialisation stage is a purpose-built hardware appliance for sovereign AI deployment. Full details (hardware specification, pricing, milestones) are provided in **Appendix D — Appliance and Hardware Roadmap**. Consistent with the staged risk-reduction strategy (Section 8.4), hardware investment is deferred until software market fit is confirmed.

8.4 Risk Reduction Through Phased Delivery

The 3-stage commercialisation strategy is designed to reduce risk at each stage:

Stage 1 (StarTeQ) reduces market risk. By generating real revenue from B2B SaaS products before launching a consumer AI platform, we validate the underlying technology, build operational capability, and establish a financial runway. StarTeQ products have defined, addressable markets with predictable compliance needs — lower risk than consumer AI adoption.

Stage 2 (StarCaller Software) reduces technology risk. By launching the software platform first, we iterate on the user experience, build the community, and refine the architecture before commit-

ting to hardware. The software platform can run on commodity hardware, enabling rapid iteration. Enterprise deployments on customer infrastructure provide real-world validation.

Stage 3 (Appliance) reduces hardware risk. Only after software market fit is confirmed do we invest in custom hardware. The appliance is the culmination of the strategy, not the starting point — it delivers the sovereign AI vision in its most complete form without requiring hardware investment before market validation.

Contingency: If any stage underperforms, the strategy adapts: - If StarTeQ revenue falls short, StarCaller software development slows but continues with a smaller team, funded by reduced burn rate and grant funding. - If StarCaller software adoption is slow, the open-source community strategy sustains development while premium features are refined based on user feedback. - If the appliance market does not materialise, the software platform continues independently — the appliance is additive, not foundational.

SECTION 9

Competitive Analysis

9.1 Competitive Landscape

The market for personal AI systems in 2026 is dominated by three categories of competitor, none of which delivers the complete sovereign personal intelligence framework that StarCaller provides.

Cloud AI Assistants

OpenAI (ChatGPT), Anthropic (Claude), Google (Gemini): The dominant consumer AI platforms. Each offers powerful language models, increasing agentic capability, and growing ecosystems. No major cloud AI platform currently offers local deployment, consent-governed memory, verifiable security, or infrastructure independence as integrated capabilities. All operate on centralised cloud infrastructure that processes user data on foreign servers. Their business model requires cloud dependency — local deployment would undermine their revenue model.

Threat level: High for consumer mindshare, low for sovereign AI category. The cloud incumbents own the mainstream AI assistant market, but their architecture cannot deliver sovereign personal intelligence without fundamentally changing their business model.

Open-Weight Model Ecosystems

Meta (Llama), Mistral, Alibaba (Qwen), DeepSeek: Open-weight model families that can run on local hardware. These provide the model substrate that StarCaller uses, but they are not complete platforms. They offer no governance layer, no memory system, no consent framework, no security architecture, and no agentic orchestrator. Deploying an open-weight model locally is not the same as having a sovereign personal intelligence system — it is a component, not a solution.

Threat level: Moderate. These are complementary technologies — StarCaller benefits from the ecosystem of open-weight models. The risk is that one of these companies (particularly Mistral, given its European positioning) adds a governance layer to create a competing platform. However, their business model is model-centric, and building a complete governance and memory architecture would require a fundamental strategic shift.

Privacy-Focused AI Products

Brave Leo AI, DuckDuckGo AI Chat, Apple Intelligence: Privacy-oriented AI products that run on-device or with privacy-preserving cloud inference. Apple's on-device approach is the closest competitor — Apple Intelligence processes data on-device with differential privacy. However, Apple

Intelligence is tightly coupled to Apple's ecosystem, offers no auditable decision traces, no consent-governed memory, no open architecture, and no infrastructure independence.

Threat level: Moderate. These products validate the market's interest in privacy-first AI. Their limitations (closed ecosystems, no verifiable trust, no portability) create the differentiation space that StarCaller occupies.

AI Agent Frameworks

LangChain, AutoGPT, CrewAI, Microsoft Copilot: Agentic AI frameworks and platforms that enable autonomous AI agents. These compete in the agentic AI market but do not address the sovereign personal intelligence category. They lack governance layers, consent frameworks, verifiable security measurement, and local-first architecture. Microsoft Copilot is the most significant competitor here, combining agentic capability with enterprise distribution — but it remains cloud-dependent and lacks sovereign deployment options.

Threat level: Low to moderate for StarCaller, high for the agentic AI market generally. These frameworks validate the agentic AI market that StarCaller addresses through its PROA Loop and Ability OS.

9.2 Capability Comparison

Capability	StarCaller	ChatGPT	Claude	Apple Intelligence	Open Model (bare)
Local-first architecture	**Full**	**None**	**None**	On-device only	Model only
Consent-governed memory	**Full**	**No**	**No**	**Limited**	**No**
Auditable decision traces	**Full**	**No**	**No**	**No**	**No**
Verifiable security (SSI)	**Yes (HARDENED, self-assessed)***	**No**	**No**	**Limited**	**No**
10-gate security	**Full**	**None**	**None**	**Limited**	**None**
Multi-Oracle framework	8 oracles	N/A	N/A	N/A	**None**

Capability	StarCaller	ChatGPT	Claude	Apple Intelligence	Open Model (bare)
Agentic execution (PROA Loop)	**Full**	**Partial**	**Partial**	**Limited**	**None**
Escalation Engine	**Full consent-governed**	**No equivalent consent-governed framework**	**Partial (guardrails)**	**No equivalent consent-governed framework**	**No equivalent consent-governed framework**
Task Planner	**Full consent-governed**	**Partial (GPTs/assistants)**	**No equivalent consent-governed framework**	**No equivalent consent-governed framework**	**No equivalent consent-governed framework**
Executive Function Mode	**Full**	**None**	**None**	**None**	**None**
Model agnosticism	**Full**	GPT only	Claude only	Apple only	By model
Infrastructure independence	**Full**	**None**	**None**	Apple only	**Full**
Open source	**Yes (core)**	**No**	**No**	**No**	**Yes (model)**
Data portability	**Full**	**None**	**None**	**Limited**	N/A
Cloud fallback	Optional	Mandatory	Mandatory	Optional	**None**
Sovereign hardware option	Planned	**No**	**No**	Apple only	DIY

9.3 Strategic Positioning

StarCaller occupies a distinctive strategic position: it is, to the authors' knowledge, one of few systems that simultaneously attempts to deliver local-first architecture, consent-governed memory, verifiable security, agentic execution, and infrastructure independence. This positioning may define a new market category — Sovereign Personal Intelligence — a space where, to the authors' knowledge, no existing competitor currently offers the same combination of local-first architecture, consent-governed memory, auditable decisions, and verifiable security.

The differentiation is not incremental. It is not a better chatbot or a more private AI assistant. StarCaller represents a meaningfully different architectural category, designed for a different set of priorities: sovereignty, consent, auditability, portability, and verifiability over capability, convenience, scale, and ecosystem lock-in.

9.4 Category Timing and Strategic Position

The sovereign personal intelligence category is nascent but poised for rapid growth, driven by four converging trends:

1. **Regulatory pressure.** GDPR enforcement, the UK Data Protection and Digital Information Bill, and EU AI Act requirements are creating demand for verifiable AI governance that no cloud platform can fully satisfy.
2. **Sovereign AI investment.** The UK's GBP 1.1 billion sovereign AI infrastructure investment and similar programmes in the EU, Japan, and Canada are creating institutional demand for sovereign AI software.
3. **Consumer awareness.** High-profile data breaches, growing understanding of AI data practices, and the shift toward privacy as a purchasing decision are creating consumer demand for privacy-first AI.
4. **Model commoditisation.** Open-weight models are approaching parity with frontier models for many tasks, reducing the capability advantage of cloud platforms and making local-first architectures more viable.

StarCaller's position as an early entrant into this category is reinforced by: - A complete, tested reference implementation (not a theoretical proposal) - 136+ security tests passing with a HARDENED Security Health Index (self-assessed, unaudited) - 40,000+ lines of Python code across a 6-layer architecture - 8 functional Oracle hands with tool-calling capability - A deployed vault pipeline with 84 parsed items and 88 memory claims - 6 StarTeQ products generating revenue from real customers

Few, if any, competitors currently maintain equivalent sovereign personal intelligence assets. The category is not yet contested by major incumbents, which creates an opportunity for an early entrant to shape category definition and standards.

SECTION 10

Research and Innovation Agenda

10.1 Federated Vaults

Current vault architecture is single-user, single-device. The federated vaults research agenda proposes to extend this to multi-device, multi-user scenarios while maintaining sovereign boundaries.

Research Objectives

- **Cross-device memory synchronisation:** Encrypted vault state synchronisation across user-owned devices without exposing plaintext to any intermediary
- **Multi-user consent boundaries:** A family or organisation can share a federated vault where each member's data remains independently consent-governed
- **Verifiable zero-knowledge proofs:** Users can prove facts about their vault contents (e.g., "I have a calendar entry at 3pm") without revealing the content itself
- **Recovery and inheritance:** Cryptographic mechanisms for vault recovery and, with appropriate consent, data inheritance

Research Stage

Objective: Exploratory — core cryptographic primitives identified, implementation pending funding.

10.2 Multi-Appliance Intelligence

StarCaller currently operates as a single-node system with split-brain topology (NucBox + Zion0). Multi-appliance intelligence extends this to a network of sovereign nodes.

Research Objectives

- **Peer-to-peer oracle network:** Multiple StarCaller appliances can consult each other's oracles for specialised knowledge without sharing raw data
- **Collective learning without data centralisation:** Pattern detection across appliances using federated learning techniques — each appliance learns from the collective without exposing individual data
- **Distributed security council:** A Security Council that spans multiple appliances, cross-referencing threat intelligence while maintaining per-appliance sovereignty

- **Emergency mesh:** When internet connectivity is lost, appliances in proximity can form a mesh network for continued operation

Research Stage

Objective: Exploratory — concept validated by split-brain architecture, implementation requires multi-appliance deployment.

10.3 Explainable Agent Systems

The PROA Loop currently provides basic decision tracing. The explainable agent systems agenda extends this to full causal reasoning.

Research Objectives

- **Causal trace graphs:** Every agent action captured in a directed acyclic graph showing cause, effect, and alternative paths not taken
- **Natural language explanation generator:** Automatic translation of decision traces into human-readable summaries at configurable levels of detail
- **Counterfactual analysis:** The system can answer "what would you have done differently?" by replaying traces with modified inputs
- **Third-party audit protocol:** Standardised trace export format suitable for regulatory audit, with cryptographic guarantees of completeness

Research Stage

Objective: Exploratory, with immediate applied value — partial decision trace infrastructure exists, causal graph extension is new.

10.4 Secure Local Models

StarCaller currently uses off-the-shelf open-weight models (Qwen, Llama, Mistral). The secure local models research agenda addresses model-specific security properties.

Research Objectives

- **Model integrity verification:** Cryptographic attestation that a loaded model matches its published weights — preventing model substitution attacks
- **Fine-tuned sovereign models:** Domain-specific fine-tuning for UK regulatory contexts (GDPR compliance, building regulations, maritime law, social housing policy)
- **Model watermarking:** Embedding traceable markers in model outputs for forensic attribution
- **Quantisation for edge hardware:** Optimising 7B-13B parameter models for the appliance's NPU, targeting under 5W power consumption

Research Stage

Objective: Near-term applied — model quantisation techniques are mature, fine-tuning is implementation work, integrity verification requires protocol design.

10.5 Sovereign Memory Standards

No widely adopted industry standard currently exists for sovereign personal AI memory. This research agenda proposes creating one.

Research Objectives

- **Open memory format:** A standardised, open format for personal AI memory data that any sovereign AI system can read and write
- **Consent ontology:** A machine-readable consent description language that expresses graduated, revocable consent policies in a portable format
- **Trace standard:** A standard format for AI decision traces suitable for regulatory audit across different implementations
- **Interoperability protocol:** API specifications for memory exchange between sovereign AI systems

Research Stage

Objective: Standards track — propose to relevant UK and international standards bodies (BSI, ISO/IEC JTC 1/SC 42).

10.6 Personal AI Governance Frameworks

The ICID constitution establishes governance principles for a single system. This research agenda extends governance to the relationship between individuals and their AI systems.

Research Objectives

- **AI bills of rights:** A standardised set of user rights in sovereign AI systems — right to explanation, right to erasure, right to portability, right to appeal automated decisions
- **Governance auditing protocol:** A methodology for third-party auditors to verify that a sovereign AI system operates within its declared governance boundaries
- **Dynamic consent frameworks:** Consent that adapts to context — location-aware, time-aware, risk-aware consent policies that maintain user control without requiring constant explicit decisions
- **Vulnerable user protections:** Governance patterns specifically designed for users with cognitive disabilities, executive dysfunction, or reduced decision-making capacity

Research Stage

Objective: Standards track with immediate applied value — the ICID constitution and consent framework provide a foundation, extension to bills of rights and auditing protocol requires standards body engagement.

SECTION 11

Funding Proposal

11.1 Funding Requested

Source	Amount	Purpose	Timeline
UK Sovereign AI Fund	GBP 2,000,000	Core platform development, security hardening, appliance prototype	24 months
Innovate UK Smart Grant	GBP 500,000	Federated vaults, multi-appliance intelligence research	18 months
Strategic Investment	GBP 1,000,000	Commercialisation, sales, marketing, enterprise pilots	24 months
Total	**GBP 3,500,000**		

11.2 Allocation

Category	GBP	Percentage	Description
Engineering (platform)	1,200,000	34%	Core platform: Oracle framework, PROA loop, memory systems, security gates, API development
Engineering (appliance)	400,000	11%	Hardware design, model optimisation, embedded system integration, certification
Research	300,000	9%	Federated vaults, multi-appliance intelligence, explainable agents, standards
Commercial	600,000	17%	Sales team, marketing, enterprise pilots, channel development
Operations	500,000	14%	Cloud infrastructure, CI/CD, security monitoring, compliance
Contingency	500,000	14%	15% contingency on all categories
Total	**3,500,000**	**100%**	

The engineering allocation is weighted toward platform development (34%) rather than appliance hardware (11%), reflecting the phased commercialisation strategy. Platform development benefits all stages; hardware investment is deferred until software market fit is confirmed.

11.3 Milestones and KPIs

Year 1 Milestones

Quarter	Milestone	KPIs
Q1	Open-source core release	GitHub stars: 500+, contributors: 10+
Q1	StarTeQ GBP 10K MRR	150 paid subscribers across 3 active products
Q2	Premium tier launch	500 premium subscribers
Q2	Enterprise pilot programme	3 enterprise pilot accounts
Q3	5 StarTeQ products live	GBP 30K MRR target (illustrative), 500+ total subscribers
Q3	SSI target: 99/100	All 8 SSI components instrumented
Q4	Federated vaults prototype	Cross-device memory sync demo
Q4	8 StarTeQ products live	GBP 50K MRR target, GBP 600K ARR (illustrative — dependent on pilot conversions and market conditions)

Year 2 Milestones

Quarter	Milestone	KPIs
Q1	StarCaller 10,000 free users	Community growth, engagement metrics
Q2	Multi-appliance demonstration	2-node StarCaller network operational
Q3	Appliance hardware prototype	First silicon, boot-time < 30s
Q4	Appliance pilot deployment	10 enterprise pilot appliances
Q4	StarCaller GBP 300K ARR (software only)	10,000 premium + 3,000 pro subscribers

Key Performance Indicators (ongoing)

- **Security Health Index:** Maintain above 95/100 (HARDENED band)
- **User retention:** 90%+ monthly retention for premium users
- **Decision trace completeness:** 100% of consult operations produce structured traces
- **Consent compliance:** Zero consent boundary violations (measured via audit sampling)
- **System uptime:** 99.9% for cloud-mediated features, 100% for local-only features
- **Mean time to resolve security incidents:** Under 4 hours

11.4 Risk Mitigation

Risk	Probability	Impact	Mitigation
StarTeQ revenue below forecast	Medium	High	Lean engineering team; 15% contingency; grant runway through Year 2
Cloud model access disrupted (API changes, costs)	Medium	Medium	Model-agnostic architecture; local Ollama fallback; circuit breaker pattern
Open-source community does not materialise	Medium	Medium	Community growth is accelerant, not foundation; premium tier and enterprise sales proceed independently
Regulatory change affects personal AI	Low	High	Compliance-by-architecture (10-gate security, consent governance, audit trails — not policy overlay)
Competitor enters sovereign AI category	Medium	Medium	First-mover advantage (working system, 136+ tests, SSI); category definition is defensive moat
Hardware supply chain issues	Low	High	Stage 3 delayed until software market fit confirmed; no hardware dependency until Year 3 at earliest
Key personnel departure	Low	Medium	Documentation-driven development; all architecture decisions captured in living documents

11.5 Expected Outcomes

By End of Year 1

- StarTeQ revenue sustains core engineering team (GBP 600K+ ARR)
- StarCaller open-source release with active community (500+ GitHub stars)
- Premium software tier launched with paying subscribers (500+)
- Security Health Index at 99/100 with all 8 components instrumented
- Federated vaults prototype demonstrating cross-device memory sync

- 3 enterprise pilot accounts evaluating sovereign deployment

By End of Year 2

- StarCaller software platform revenue: GBP 300K+ ARR (software-only)
- StarTeQ revenue: GBP 750K+ ARR (continuing growth)
- Total ARR: GBP 1M+
- 10,000+ free users, 2,000+ paying subscribers
- Multi-appliance network demonstrated
- Appliance hardware prototype ready for pilot
- Standards submission to BSI for sovereign memory format
- NCSC security certification process initiated

By End of Year 3

- Total ARR: GBP 2.5M+
- 30,000+ free users, 10,000+ paying subscribers
- Appliance general availability with 50+ enterprise deployments
- Established category: Sovereign Personal Intelligence recognised as distinct market segment
- UK leading international standards for sovereign AI memory and governance
- Platform ready for international expansion (EU, Japan, Canada sovereign AI markets)

11.6 National Capability Contribution

StarCaller directly contributes to UK national AI capability in four dimensions:

1. Sovereign AI software infrastructure. The UK's GBP 1.1 billion sovereign compute investment requires software that can run on it. StarCaller provides a complete, tested sovereign AI software stack — architecture, governance, security, memory, and orchestration — designed for sovereign deployment.

2. AI safety and verifiability. The AI Safety Institute's technical research agenda requires working systems that demonstrate verifiable safety properties. StarCaller's 10-gate architecture, SSI, and decision trace framework provide a potential deployable reference implementation of safety-oriented AI design.

3. Exportable sovereign capability. A UK sovereign personal intelligence platform is an exportable national capability. The sovereign AI market is global — every nation with data sovereignty concerns is a potential market. The UK has the opportunity to lead this category internationally.

4. Open standards leadership. By developing sovereign memory formats, consent ontologies, and decision trace standards, the UK can shape international norms for personal AI governance — analogous to the UK's role in developing the first AI safety summit and the Bletchley Declaration.

APPENDIX

Security Health Index (SSI) Methodology

Purpose and Scope

The Security Health Index (SSI) is a quantitative, continuous assessment of the StarCaller system's integrity across eight weighted dimensions. It is designed to provide:

1. A real-time, auditable measure of system security posture
2. Early warning of degradation in any security-relevant component
3. A framework for continuous improvement with quantifiable targets
4. A basis for comparing security posture across deployments

The SSI is an internal assessment tool. It is not a substitute for formal security certification (ISO 27001, SOC 2, Cyber Essentials Plus) or independent penetration testing. It measures the system's security-relevant architecture and operational state against its own design specifications — not against an external standard.

Component Definitions

1. Prompt Integrity (Weight: 18%)

What it measures: The effectiveness of the Prompt Firewall (Gate 1) in detecting and blocking prompt injection, jailbreak attempts, and goal hijacking.

Scoring method: The Prompt Firewall tests incoming requests against 8 pattern families: - Direct injection attempts (e.g., "ignore previous instructions") - Indirect injection via tool outputs - Role-play masquerade attacks - Encoding/obfuscation bypass attempts - Context overflow attacks - Multi-turn extraction patterns - Privilege escalation attempts - Cross-session contamination

Each pattern family is scored independently. The component score is the weighted average across all 8 families. Score drops when any family detects a successful or partially successful attack that was not blocked.

Update frequency: Every request (real-time). Rolling 7-day window for the reported score.

Limitations: The Prompt Firewall is a rule-based classifier, not a trained detection model. It can detect known injection patterns with high reliability but may miss novel or adversarial patterns not in its rule set. This is an area identified for improvement (see Research Agenda, Section 10.4).

2. Memory Integrity (Weight: 18%)

What it measures: The protection of vault data against unauthorised writes, consent violations, and poisoning attempts.

Scoring method: The Memory Integrity Monitor (Gate 2) tracks: - Unauthorised write attempts (detected by consent framework) - Conflicting fact patterns (data that contradicts established vault content) - Consent boundary violations (operations that exceed granted permissions) - Anomalous access patterns (unusual read/write sequences)

Each category is scored based on detection rate and false positive rate. A perfect score requires zero successful consent violations and zero undetected poisoning attempts.

Update frequency: Every vault operation (real-time).

Limitations: The integrity monitor can detect known poisoning patterns — conflicting facts, improbable timestamps, anomalous access sequences — but cannot detect sophisticated attacks where the poisoning is consistent with existing vault content and consent permissions. The Welford's online algorithm for behavioural baselines mitigates this partially but is not a panacea.

3. Supply Chain (Weight: 15%)

What it measures: The integrity of the system's dependencies — Python packages, models, infrastructure components.

Scoring method: CVE matching against known vulnerabilities for all direct and transitive Python dependencies. Semver range comparison identifies packages with known vulnerabilities in the installed version range. Critical vulnerabilities reduce the score more than moderate ones. Outdated packages with no known CVEs still incur a small penalty.

Update frequency: Daily scan of dependency tree, plus on-demand scan on any dependency update.

Limitations: The supply chain scanner covers direct and transitive Python dependencies. It does not currently cover: - Docker container base image vulnerabilities - Firmware vulnerabilities on the NucBox or Zion0 nodes - Supply chain attacks on the model weights themselves (model integrity verification is a research objective, Section 10.4) - Vulnerabilities introduced through tool calls (web search results, shell commands)

4. Tool Safety (Weight: 12%)

What it measures: The effectiveness of the Tool Execution Guard (Gate 3) in preventing unsafe tool invocations.

Scoring method: The guard checks every tool invocation against the PermissionRegistry, which maps Oracle identities to permitted actions. Invalid or unauthorised tool calls are blocked at the execution boundary. The score reflects: - The proportion of blocked invalid calls - The completeness of the permission registry coverage - The absence of bypassed tool calls

Update frequency: Every tool invocation (real-time).

Limitations: The guard can block calls that violate explicit permissions but cannot prevent an Oracle from making poor decisions within its permitted scope. For example, an Oracle with web search permission could make an ill-advised search query — the guard permits this because it is within the permission boundary.

5. Identity Confidence (Weight: 12%)

What it measures: The confidence that the system is interacting with its authorised user.

Scoring method: Gate 8 (Identity Trust Fabric) produces an identity confidence score from 0-100% based on available factors: - Voice biometrics (if available): Spectral analysis on MFCC, CQT, centroid, ZCR, RMS, duration - Device proximity (if available): Network-level device attestation - Behavioural patterns (if available): Usage timing, interaction style - Historical authentication: Success rate of prior authentications - Location correlation (if available): Expected vs actual access location

Unavailable factors redistribute their weights proportionally among available factors. The reported score is the weighted combination of available factors.

Current score: 82%. This reflects the absence of physical biometric sensors (voice or fingerprint), which would contribute significant weight. The score is based on device-level authentication and behavioural patterns only.

Update frequency: Per-authentication event.

Limitations: The current implementation uses spectral heuristics rather than a trained model for voice analysis. The scoring engine is pluggable — the interface supports upgrade to AASIST or RawNet2 for improved spoof detection — but this upgrade has not been implemented. Until physical biometric sensors are available, identity confidence will remain below the 90% threshold for "Owner" tier assurance.

6. Oracle Integrity (Weight: 10%)

What it measures: The behavioural consistency of each Oracle hand over time.

Scoring method: The Oracle Integrity Monitor (Gate 10) uses a CUSUM (Cumulative Sum) algorithm to detect gradual drift in Oracle behaviour across five dimensions: - **Domain focus:** Is the Oracle answering within its designated domain? - **Quality:** Is response quality (citation count, reasoning depth, confidence calibration) stable? - **Persona adherence:** Does the response reflect the assigned persona? - **Manipulation resistance:** Is the Oracle resisting attempts to change its behaviour? - **Provider reliability:** Is the underlying model provider responding consistently?

Each dimension is tracked as a cumulative deviation from a per-Oracle baseline profile. When the cumulative sum exceeds the control limit (h=5.0 with k=0.5), a drift alert is generated. The component score reflects the ratio of non-drifting oracles to total oracles, adjusted for drift severity.

Update frequency: Per-consult event. Golden test sets run weekly.

Limitations: The CUSUM algorithm detects gradual drift but may miss sudden, single-event changes. The weekly golden test set helps mitigate this. Baseline profiles require at least 50 consultations per oracle to stabilise — oracles with fewer consultations have wider confidence intervals.

7. Physical Safety (Weight: 8%)

What it measures: The system's protection against executing actions that could cause physical harm.

Scoring method: Gate 9 (Physical Blacklist) maintains a blocklist of denied actions and domains. Actions are classified as: - **HARD_BLOCKED (10 categories):** Irreversible, high-risk physical actions. Always blocked. - **CONFIRMATION_REQUIRED (11 categories):** Actions that require explicit user confirmation, gated by time of day and recent activity context.

The score reflects the completeness of the blacklist coverage and the absence of bypassed physical actions.

Update frequency: Per-action event. Blacklist updated via security patch.

Limitations: The blacklist is a static rule set. It cannot anticipate novel physical actions that were not included in the rule base. Dynamic risk assessment based on action outcomes (learning from experience) is a future capability.

8. Inter-Oracle Trust (Weight: 7%)

What it measures: The integrity of cross-Oracle communication.

Scoring method: All inter-Oracle requests are HMAC-SHA256 signed. The trust score reflects: - The proportion of signed requests to total requests - The signature verification success rate - The absence of cross-Oracle identity fraud

Update frequency: Per-request (real-time).

Limitations: HMAC signing prevents impersonation but does not prevent a compromised Oracle from making validly signed but malicious requests. Chain verification (max depth 10) limits the blast radius but does not eliminate it.

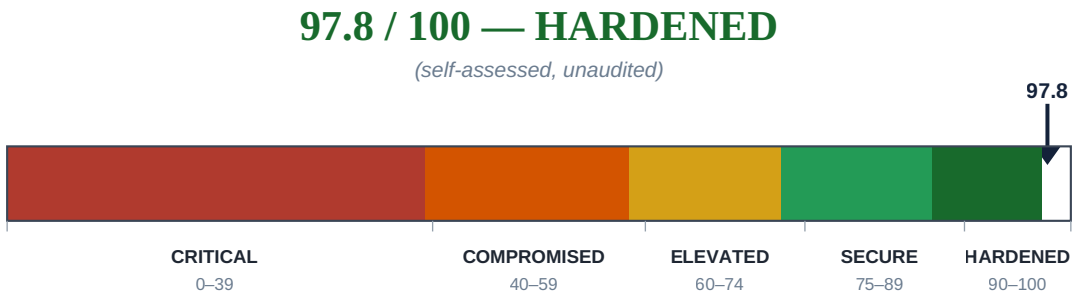
Aggregation Method

The overall SSI score is a weighted sum of the 8 component scores:

```
SSI = (Prompt Integrity x 0.18) + (Memory Integrity x 0.18) + (Supply Chain x 0.15)
    + (Tool Safety x 0.12) + (Identity Confidence x 0.12) + (Oracle Integrity x 0.10)
    + (Physical Safety x 0.08) + (Inter-Oracle Trust x 0.07)
```

All component scores are normalised to 0-100. Weights sum to 1.0.

Classification Bands



Current Assessment

Score: 97.8 — HARDENED

This assessment reflects the current state of the reference implementation. It is based on: - 8 of 8 components instrumented and reporting - 3 of 8 components fully wired to all intended data sources (Prompt Integrity, Memory Integrity, Supply Chain) - 5 of 8 components operational with acknowledged limitations documented above - Identity Confidence (82%) is the primary constraint on reaching a higher score

Important Limitations

1. **Self-assessment:** The SSI measures the system against its own design specifications, not against an external standard. An SSI of 97.8 does not imply 97.8% compliance with ISO 27001, NCSC Secure by Design, or any other external framework.
2. **Coverage gaps:** The supply chain scanner does not cover Docker images, firmware, or model weights. The physical safety gate is a static blocklist. These gaps are documented as research objectives in Section 10.
3. **No independent verification:** The SSI has not been independently audited or verified by a third party. Independent verification is targeted for Year 2 (see Regulatory Roadmap, Section 6A.7).
4. **Architecture-specific:** The SSI is specific to the StarCaller architecture. It is not directly transferable to other AI systems without reimplementation of the component definitions and scoring mechanisms.

SECTION 12

Conclusion

The current trajectory of personal AI appears to lead toward ever-greater centralisation, opacity, and foreign infrastructure dependency. The dominant platforms operate on cloud infrastructure outside the individual's control, process personal data in foreign jurisdictions, and provide no independently verifiable guarantees about how that data is used.

This trajectory is not necessarily inevitable. An alternative exists: Sovereign Personal Intelligence, built on local-first architecture, consent-governed memory, auditable decision processes, and verifiable security guarantees.

The Strategic Opportunity

The strategic opportunity is not merely StarCaller. The strategic opportunity is establishing UK leadership in Sovereign Personal Intelligence as a category — a new class of AI system defined by local-first architecture, consent-governed memory, auditable decision processes, and verifiable security guarantees.

StarCaller is one reference implementation demonstrating how such systems can be built. It is not a theoretical proposal or a research paper — it represents 40,000+ lines of Python code operating across a 6-layer architecture, with 136+ security tests passing and a Security Health Index assessed at 97.8/100 (self-assessed, unaudited). It runs on local hardware. It maintains consent-governed encrypted memory. It produces auditable decision traces for every operation. It operates without mandatory cloud dependency. It is model-agnostic, infrastructure-independent, and data-portable.

Why This Matters for the United Kingdom

The United Kingdom has the policy infrastructure — the Sovereign AI Fund, the AI Opportunities Action Plan, the AI Safety Institute. It has the compute investment — GBP 1.1 billion committed to sovereign AI infrastructure. It has the research base — world-class institutions with deep AI expertise. A critical gap has been the software infrastructure: the architecture, governance, security, and memory systems that translate sovereign AI policy into working systems that individuals and institutions can actually use.

> *“ It is not a better chatbot. It is not a cheaper API. It is a substantially different architectural category… ”*

The category of Sovereign Personal Intelligence is proposed to fill this gap. It is not a better chatbot. It is not a cheaper API. It is a substantially different architectural category, designed for a world in which personal AI systems manage increasingly sensitive domains of human life and where trust cannot be delegated to a third party.

The Call to Action

The reference implementation demonstrates that the architecture is viable. The category is open. The timing is favourable.

The authors invite:

- **The UK Sovereign AI Fund** to consider Sovereign Personal Intelligence as a priority category for investment, and StarCaller as a reference implementation that demonstrates the category's feasibility
- **UKRI and Innovate UK** to support the research agenda — federated vaults, explainable agent systems, sovereign memory standards — that will define the category's technical foundations
- **The AI Safety Institute** to evaluate the SSI methodology and decision trace framework as potential contributions to AI safety evaluation practice
- **Strategic investors** to participate in building the category through the phased commercialisation strategy described in this paper

StarCaller provides an operational reference implementation suggesting that Sovereign Personal Intelligence is achievable with current technology. The remaining work — category definition, standards development, certification, scaling — requires strategic investment and institutional support.

> *“ The reference implementation demonstrates the feasibility of the architecture. The category is open. The timing is favourable. ”*

The reference implementation demonstrates the feasibility of the architecture. The category is open. The timing is favourable. The question is whether the United Kingdom will define this category or leave it for others to define.

APPENDIX D

Appliance and Hardware Roadmap

The StarCaller Appliance — a dedicated hardware device for sovereign AI deployment — represents the third stage of the commercialisation strategy. It is included here for reference rather than in the main body, as hardware investment is contingent on software market fit (see Section 8.4).

8.3 Stage 3: StarCaller Appliance — Sovereign Hardware (Year 3-5)

The StarCaller Appliance is a dedicated hardware device for sovereign AI deployment — a complete, air-gapped personal intelligence system that ships with StarCaller pre-installed.

Target Markets

- **Government:** Local councils, devolved administrations, central government departments handling sensitive data
- **Healthcare:** GP surgeries, mental health services, patient data management
- **Legal:** Law firms handling client confidential information
- **Financial Services:** Wealth management, financial advisory, insurance
- **Defence and Security:** Secure communications, intelligence analysis support

Hardware Specification (Target)

- **Compute:** ARM-based SoC with dedicated NPU (neural processing unit), 8+ TOPS AI performance
- **Memory:** 16GB RAM, 512GB NVMe storage
- **Model:** Pre-loaded with fine-tuned open-weight model (Llama 4 / Mistral / Qwen)
- **Connectivity:** Ethernet, Wi-Fi 6, Bluetooth 5.3 — all optional, system operates air-gapped
- **Physical Security:** Tamper-evident casing, secure enclave for key storage, hardware root of trust
- **Form Factor:** Wall-mountable, fanless, silent operation

Pricing

Model	Hardware*	Software	Annual Support	Target Customer
Personal	GBP 999	Included	GBP 199/yr	Professionals, high-net-worth

Model	Hardware*	Software	Annual Support	Target Customer
Professional	GBP 1,999	Included	GBP 399/yr	SMEs, professional services
Enterprise	GBP 2,999	Custom	GBP 999/yr	Government, institutions

* Appliance pricing is **illustrative**. Final pricing will depend on manufacturing costs, certification requirements, and volume.

Key Milestones — Stage 3

Milestone	Target Date	Dependencies
Hardware prototype	Year 3 Q1	Funding secured, SoC partnership
Software optimisation	Year 3 Q2	Model quantisation, boot-time optimisation
Security certification	Year 3 Q3	CESG / NCSC certification process
Pilot deployment	Year 3 Q4	10 enterprise pilot customers
General availability	Year 4 Q1	Manufacturing partnership, distribution
Volume production	Year 4 Q3	1,000+ units, cost reduction

End of Sovereign AI Position Paper — Version 1.3 (Draft). All 12 sections plus appendices complete.

APPENDIX A

Sources and References

Government and Policy Documents

1. **UK Government, AI Opportunities Action Plan** (January 2025). Available: https://www.gov.uk/government/publications/ai-opportunities-action-plan
2. **UK Government, AI Opportunities Action Plan: One Year On** (January 2026). Available: https://www.gov.uk/government/publications/ai-opportunities-action-plan-one-year-on
3. **UK Government, Sovereign AI Fund** (April 2026). Press release: "AI firms pioneering drug discovery, cheaper supercomputing and more get first backing through UK's Sovereign AI." Available: https://www.gov.uk/government/news/ai-firms-pioneering-drug-discovery-cheaper-supercomputing-and-more-get-first-backing-through-uks-sovereign-ai
4. **UK Sovereign AI Fund** (2026). Available: https://www.sovereignai.gov.uk/
5. **AI Safety Institute (AISI), Research Agenda**. Available: https://www.aisi.gov.uk/research-agenda
6. **National Cyber Security Centre (NCSC), Secure by Design**. Joint guidance: "CISA and UK NCSC Unveil Joint Guidelines for Secure AI System Development" (November 2023). Available: https://www.cisa.gov/news-events/alerts/2023/11/26/cisa-and-uk-ncsc-unveil-joint-guidelines-secure-ai-system-development
7. **NSA, NCSC-UK, CISA et al., Guidance for Securing AI** (2024). Available: https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/3598020/guidance-for-securing-ai-issued-by-nsa-ncsc-uk-cisa-and-partners/
8. **UK Government, AI Safety Summit** (Bletchley Park, November 2023). The Bletchley Declaration.
9. **UK Government, Data Protection and Digital Information Bill** (2024-2026). Current legislative status available via UK Parliament.

Market Research and Industry Analysis

1. **Gartner, "Forecast Analysis: Artificial Intelligence Services, Worldwide"** (2025). AI services market projected at $609 billion by 2028 (21.4% CAGR).

2. **Gartner, "AI Spending Forecasts"** (2026). Worldwide AI spending projected to reach $2.59 trillion in 2026. Cited via Digital Applied compilation.
3. **IDC, "AI Infrastructure Spending"** (2026). AI infrastructure projected to reach $487 billion.
4. **Stanford Institute for Human-Centered AI (HAI), "AI Index Report 2025"** (April 2025). Available: https://hai.stanford.edu/ai-index/2025-ai-index-report
5. **Stanford HAI, "AI Index Report 2026"** (April 2026). Available: https://hai.stanford.edu/ai-index/2026-ai-index-report
6. **BCC Research, "Edge AI Market to Grow at 36.9% CAGR Through 2030"** (2025). Edge AI market: $8.7 billion (2025) to $56.8 billion (2030).
7. **Grand View Research, "Edge AI Software Market Size, Share, Industry Report 2030"** (2025). Edge AI software: $1.95 billion (2024) to $8.91 billion (2030), 29.2% CAGR.
8. **McKinsey & Company, "The Agentic Commerce Opportunity"** (2025). Analysis of agentic AI market trajectories. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-agentic-commerce-opportunity

Standards and Regulatory Frameworks

1. **ISO/IEC 42001:2023 — Information Technology, Artificial Intelligence, Management System**. International standard for AI management systems.
2. **ISO/IEC 27001:2022 — Information Security, Cybersecurity and Privacy Protection**. International standard for information security management.
3. **UK GDPR and Data Protection Act 2018**. UK data protection legislation, as amended post-Brexit.
4. **EU Artificial Intelligence Act (Regulation (EU) 2024/1689)** . European Union regulatory framework for AI systems.
5. **NCSC, "Secure by Design Guidance"** . Principles for building secure systems from inception.
6. **OECD, "AI Principles"** (2019, updated 2024). Available: https://oecd.ai/en/ai-principles
7. **World Economic Forum, "The Future of AI Governance"** (Various publications, 2024-2026). Available: https://www.weforum.org/topics/artificial-intelligence/

Academic and Technical References

1. **CUSUM algorithm for change detection:** Basseville, M. and Nikiforov, I.V. "Detection of Abrupt Changes: Theory and Application." Prentice-Hall, 1993.

2. **Welford's online algorithm:** Welford, B.P. "Note on a Method for Calculating Corrected Sums of Squares and Products." Technometrics, 1962.
3. **SASV (Spoofing-Aware Speaker Verification):** ASVspoof challenges series. Available: https://www.asvspoof.org/
4. **HMAC-SHA256:** Krawczyk, H., Bellare, M., Canetti, R. "HMAC: Keyed-Hashing for Message Authentication." RFC 2104, 1997. Available: https://datatracker.ietf.org/doc/html/rfc2104

Project-Specific

1. **StarCaller Reference Implementation.** GitHub: `icidoracles-cell/Starcaller-v2.git`. Documentation: `ARCHITECTURE_MAP.md`, `CHRONOLOGY.md`, `AGENTS.md`, `OPERATIONS_MANUAL.md`.
2. **Security Health Index (SSI) Methodology.** See Appendix — SSI Methodology in this document.

APPENDIX B

Sovereign Personal Intelligence Principles

Sovereign Personal Intelligence (SPI) is proposed as a framework category that can be discussed, evaluated, and implemented independently of any specific product or platform. The following principles define the category. StarCaller is one reference implementation that demonstrates how these principles can be realised in practice.

SPI-1: User Ownership

The individual retains full ownership and control of their personal data, memory, and learned patterns. No third party holds a superior claim to any data generated through the individual's use of an SPI system.

Implications: - Data cannot be used for model training without explicit, revocable consent - Data must be exportable in open, standard formats - Vendor lock-in through data dependency is excluded by design - The user can terminate the relationship and retain all data

SPI-2: Consent Governance

All data retention, processing, and sharing is governed by a graduated consent framework. Consent is not a single binary choice at account creation — it is a continuous, context-aware, revocable relationship.

Implications: - Each data category has independent consent and retention policies - Consent can be withdrawn at any time, triggering data deletion - Default posture is deny — no data is retained without explicit permission - All consent actions are auditable

SPI-3: Auditable Intelligence

Every significant operation — every consultation, memory write, action execution, and decision — produces a structured, tamper-evident trace that the user can inspect.

Implications: - Traces include provenance, confidence, citations, and alternative viewpoints - Traces are HMAC-chained to prevent undetected modification - Traces are exportable in standardised formats for external audit - The user can request a decision trace for any past operation

SPI-4: Local-First Operation

The system's primary operation occurs on hardware under the user's control or within a trusted sovereign boundary. Cloud connectivity is optional and limited to explicitly authorised functions.

Implications: - Core capabilities function without network connectivity - Personal data does not transit third-party infrastructure by default - Cloud model access is an optional enhancement, not a requirement - The system is air-gap capable

SPI-5: Infrastructure Independence

The system does not depend on any single model provider, cloud platform, or hardware vendor. It can operate on any compute substrate — local device, sovereign data centre, or trusted edge node.

Implications: - Model agnosticism: the system can use open-weight models, fine-tuned sovereign models, or private frontier model deployments - No mandatory cloud dependency - Migration between infrastructure providers is supported - The system can be deployed on UK sovereign compute infrastructure

SPI-6: Security Verification

Security and trust properties are not claimed — they are measured. The system reports concrete, quantitative indicators of its integrity, available to both the user and independent assessors.

Implications: - A quantified security health index provides real-time integrity measurement - Component-level scores enable targeted improvement - The methodology is transparent and auditable - Security verification is a continuous process, not a point-in-time certification

SPI-7: Data Portability

Personal data and system configuration are structured in open, exportable formats. The user can migrate between SPI implementations without loss.

Implications: - Memory, preferences, consent rules, and learned patterns are exportable - Export uses standardised, documented formats - Import from other systems is supported - The user is not locked into a single vendor's ecosystem

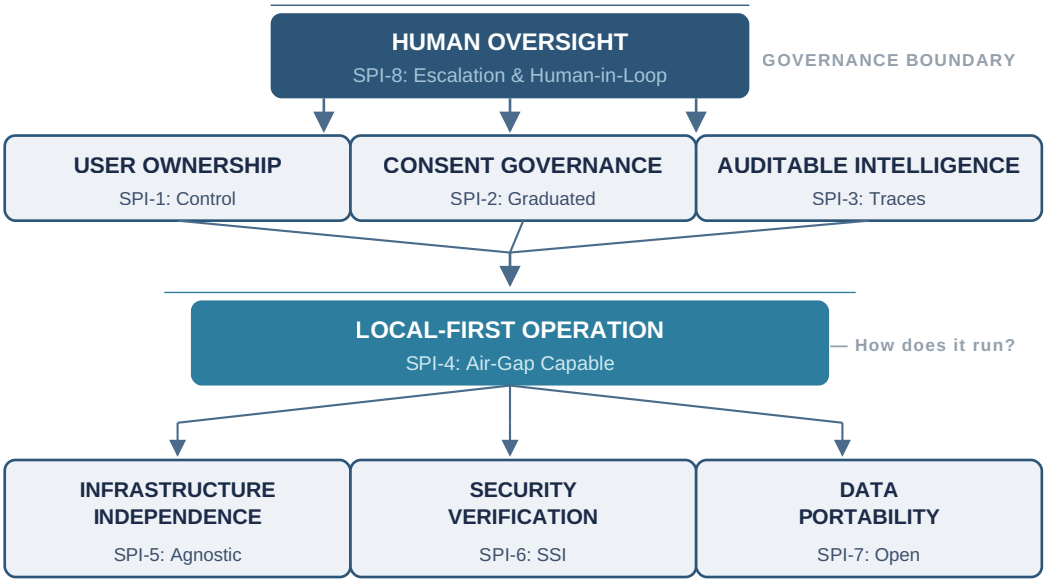
SPI-8: Human Oversight

The system operates autonomously within defined boundaries, but high-stakes decisions — those involving irreversibility, constitutional boundaries, low confidence, novelty, or resource constraints — receive appropriate human oversight.

Implications: - Autonomous operation has defined scope and escalation pathways - The escalation criteria are transparent and configurable - Human review is recorded and auditable - The user can override any autonomous decision

SPI Visual Framework

The relationship between the eight SPI principles can be understood through the following structure: three principles define the governance model (Who decides?), one principle defines the operational mode (How does it run?), and four principles define the technical architecture (What properties must it have?).



Framework reading: top → bottom

1. Governance boundary → 2. User-system relationship → 3. Operational foundation → 4. Technical requirements

FIGURE 4 — SPI VISUAL FRAMEWORK (HIERARCHICAL)

Reading the framework from top to bottom: 1. **Human Oversight** provides the governance boundary — autonomous operation with defined escalation 2. **The three middle principles** define the user-system relationship: who owns the data, how consent is managed, and how decisions are

audited 3. **Local-First Operation** is the operational foundation — everything runs on local or sovereign infrastructure 4. **The four base principles** define the technical requirements: infrastructure agnosticism, verifiable security, data portability, and auditable intelligence

This framework is designed to be implementation-agnostic. Any system claiming to implement Sovereign Personal Intelligence should be evaluable against these eight principles, regardless of its underlying technology choices or commercial model.

APPENDIX C

Future National Applications

The Sovereign Personal Intelligence framework has potential applications beyond personal use. The following sectors represent areas where SPI principles could deliver national strategic value.

Healthcare

Current challenge: The NHS generates vast quantities of personal health data that could benefit from AI-assisted analysis, but data sovereignty concerns, consent complexity, and regulatory requirements limit cloud AI adoption.

SPI application: A healthcare-deployed SPI system could provide AI-assisted clinical decision support, patient record management, and treatment pathway analysis — all within the sovereign boundary of NHS infrastructure. The consent governance framework maps naturally to the NHS's existing consent and opt-out framework. The auditable decision trace capability supports the accountability requirements of clinical practice.

Feasibility: Medium-term (3-5 years). Requires domain-specific Oracle training and NHS Digital integration but no fundamental architectural changes.

Education

Current challenge: Personalised learning at scale requires AI systems that understand individual students' knowledge, gaps, and learning styles. Cloud AI deployment raises data protection concerns, particularly for minors.

SPI application: An education-deployed SPI system could provide personalised tutoring, curriculum adaptation, and learning analytics — operating on school or MAT (Multi-Academy Trust) infrastructure. The consent framework naturally distinguishes between student, parent, and educator consent domains. Local-first operation ensures continuity during network outages common in some educational settings.

Feasibility: Medium-term (3-5 years). Requires curriculum-specific Oracle development and integration with existing MIS (Management Information System) platforms.

Local Government

Current challenge: Local authorities handle sensitive citizen data across housing, social care, benefits, and regulatory compliance. Budget constraints limit investment, and data sovereignty concerns restrict cloud AI adoption.

SPI application: A local government SPI deployment could assist with housing allocation, social care eligibility assessment, benefits administration, and regulatory compliance monitoring — operating on the council's own infrastructure. The escalation engine maps naturally to the graded decision-making structures in social care and housing. The auditable decision trace capability provides the transparency that public sector accountability requires.

Feasibility: Near-term (1-3 years). Many of the required capabilities (compliance checking, eligibility assessment, scheduling) are already demonstrated in the StarTeQ product line.

Defence and Security Support

Current challenge: Defence and security organisations require AI systems that operate on classified networks with no external connectivity, produce auditable decision trails, and maintain strict data compartmentalisation.

SPI application: An air-gapped SPI deployment could provide intelligence analysis support, operational planning assistance, and security monitoring — operating entirely within classified infrastructure. The local-first architecture and air-gap capability are inherent design properties, not modifications. The Security Council and escalation framework provide the structured decision-making processes that military command structures require.

Feasibility: Medium-term (2-4 years). Requires NCSC certification at appropriate classification level, Oracle training for defence-specific domains, and potential model fine-tuning for operational security requirements.

Critical Infrastructure

Current challenge: Operators of Critical National Infrastructure (CNI) — energy, water, transport, communications — increasingly need AI for monitoring, predictive maintenance, and incident response, but cannot accept the risk of cloud-dependent AI controlling essential services.

SPI application: A CNI-deployed SPI system could provide infrastructure monitoring, anomaly detection, predictive maintenance scheduling, and incident response coordination — operating on the operator's own OT (Operational Technology) networks. The physical action gating (Gate 9) provides structured controls for infrastructure-affecting actions. The behavioural baselines (Gate 5) adapt naturally to infrastructure monitoring patterns.

Feasibility: Medium-term (3-5 years). Requires OT-specific sensor integration, safety-critical certification pathways, and potentially hardware redundancy for high-availability deployments.

Digital Inclusion

Current challenge: Populations that could benefit most from AI assistance — older adults, people with disabilities, those with limited digital literacy — are often least able to navigate complex cloud AI interfaces and are most vulnerable to privacy harms from cloud AI services.

SPI application: An SPI deployment optimised for accessibility could provide AI-assisted living support, medication management, social connection, and independent living assistance — operating on local hardware with no cloud dependency. The Executive Function Mode, originally designed for executive dysfunction support, extends naturally to cognitive decline support and general accessibility. Local-first operation eliminates the connectivity dependency that disadvantages rural and economically deprived areas.

Feasibility: Near-term (1-3 years). Many accessibility-oriented features (voice interface, Executive Function Mode, simplified consent) are already demonstrated in the reference implementation. Requires user-centred design research with target populations.

End of Sovereign AI Position Paper — Version 1.3 (Draft). All 12 sections plus appendices complete.