ScienceAI — Product Requirements Document

Date: June 4, 2026
Author: Product Team
Status: Draft
Version: 2.0 (Revised — all evaluation gaps addressed)

---

1. Executive Summary

Problem Statement

Scientific researchers spend a disproportionate amount of time *finding and consuming* information — manually scanning hundreds of papers, conference proceedings, clinical trials, and competitor updates across fragmented, siloed sources. The bottleneck is not intelligence; it is information overload, and no existing tool synthesises across sources, grounds answers in citations, and delivers results in accessible audio format for researchers on the move.

Proposed Solution

ScienceAI is a Gen AI-powered research intelligence platform that lets scientists converse with the scientific literature, retrieve source-attributed answers grounded exclusively in a continuously updated vector knowledge base, and consume research as on-demand audio summaries or dual-voice podcasts — all from a single Next.js 16 web application backed by a multi-agent AI architecture (Langflow, OpenAI GPT-5-mini, Chroma DB, ElevenLabs TTS).

Business Impact

- **Research productivity:** Reduce time-to-insight from hours to seconds by replacing manual literature review with grounded AI synthesis — saving each researcher an estimated 5+ hours per week
- **Competitive moat:** Audio/podcast generation differentiates from text-only academic AI tools; domain-specific ingestion pipeline creates a compounding data flywheel that generic tools cannot replicate
- **Revenue:** SaaS model projects $6M ARR at 1,000 team subscribers; $20,800/yr in researcher productivity recaptured per seat

Key Milestones

Milestone	Target
POC / Capstone Demo	April 2026 ✅
Private Beta (Research Teams)	Q3 2026
V1.0 Public Launch	Q4 2026
Enterprise / Institutional Tier	Q1 2027

Success Metrics

Tier	Metric	Baseline	Target
North Star	Hours of research work saved per researcher per week	—	3+ hrs/week
Primary	Time-to-insight (query → streamed answer, P95)	< 5 sec	< 3 sec
Primary	Citation accuracy rate (source-grounded responses)	Manual spot-check	≥ 95%
Secondary	Weekly Active Users (WAU)	—	500 (V1.0)
Secondary	Audio feature adoption (% of WAU)	—	40%
Secondary	30-day researcher retention rate	—	≥ 60%
Health	API error rate on chat endpoint	—	< 0.5%
Health	Vector store freshness lag	—	< 24 hrs

2. Problem Definition

2.1 Customer Problem

- **Who:** Academic researchers (PhD students, post-docs, PIs), clinical scientists, R&D professionals, and competitive intelligence analysts in biotech, pharma, and top-tier universities. Primary persona: a mid-career research scientist at a pharma or biotech firm who publishes 3–5 papers per year, manages a small team, and must stay current across 5–8 active research fronts simultaneously.
- **What:** They cannot keep up with the volume of new publications, trial results, conference presentations, and competitor moves relevant to their domain — and when they do find relevant material, synthesising it across sources requires hours of manual reading
- **When:** Daily — particularly at project kick-off, ahead of grant submissions, during literature review phases, and before attending conferences
- **Where:** Desktop browser (lab, office, remote); increasingly mobile while commuting or during bench work
- **Why:** Scientific publishing volume doubles approximately every 9 years (Nature, 2024); cross-domain synthesis requires integrating terminology, methodology, and findings from dozens of papers written in inconsistent vocabulary and format

- **Impact:** Researchers spend an estimated 23% of working time on information gathering (Nature, 2023). In drug discovery, delayed literature awareness can cost months and millions. Each researcher recovering 5 hrs/week × $80/hr loaded cost = $20,800/year in productivity per seat.

2.2 Why Rule-Based Systems Fail — The Case for Agentic AI

Keyword search and deterministic rule-based NLP are insufficient for scientific research queries because the problem is fundamentally one of **semantic complexity, cross-document reasoning, and contextual ambiguity**:

- A query like *"latest CDK inhibitor resistance mechanisms in triple-negative breast cancer"* requires understanding that "CDK" is domain-contextual (cyclin-dependent kinase, not a JavaScript runtime), interpreting "latest" relative to the knowledge base's most recent embeddings, and synthesising resistance mechanisms across multiple mechanism types from different papers — none of which a keyword ruleset or regex pattern can handle.
- Scientific questions routinely involve **contradictory findings across papers**, requiring the system to surface tension and uncertainty rather than returning a single "correct" answer.
- Vocabulary is highly variable: the same concept appears under different MeSH terms, brand names, gene aliases, and abbreviations across sources. Semantic embedding models bridge this gap; keyword matching cannot.
- **Judgment calls** — when to say "I don't have reliable information on this" rather than hallucinating — require probabilistic reasoning that deterministic systems cannot perform.

Large language models with Retrieval-Augmented Generation (RAG) are the appropriate architecture because they: (a) embed semantic meaning across vocabulary diversity, (b) synthesise coherent summaries from multiple retrieved chunks, (c) acknowledge uncertainty when context is absent, and (d) generate natural-language output that researchers can act on immediately.

2.3 Unstructured Scientific Data — Why ML Is Required

ScienceAI ingests and processes heterogeneous unstructured data formats that have no standardised schema:

Source	Format	Challenge
Scientific journals / PubMed	Full-text PDF, PubMed XML abstracts	Mixed layouts, mathematical notation, citation graphs
Conference proceedings	HTML pages, PDF slide decks	Inconsistent structure, figures, speaker notes
ClinicalTrials.gov	JSON / XML records	Regulatory vocabulary, status codes, eligibility criteria
Competitor intelligence	Press release HTML, news articles	Varying structure, marketing language, implicit claims

Traditional NLP approaches (TF-IDF, keyword extraction, named entity recognition with fixed schemas) fail because: domain jargon and abbreviations are not in general-purpose NLP dictionaries; sentence meaning changes with context across papers; and synthesis requires combining information across documents, not just extracting it from one. ML embedding models (used to build the Chroma vector store) learn latent semantic relationships from large corpora, enabling cross-source, cross-vocabulary retrieval that rule-based systems cannot achieve.

2.4 Defensible Competitive Moat

ScienceAI's long-term defensibility is built on a **three-layer compounding flywheel**, not just the technology stack:

1. **Domain-specific ingestion flywheel:** As the Langflow Sync Agent continuously ingests from scientific literature, conferences, clinical trials, and competitors, the Chroma vector store grows richer and more current than any static corpus. Each new source added increases retrieval quality — this is a data advantage that a generic RAG deployment over public data does not replicate. Institutional partnerships for pre-publication access would create further exclusivity.

2. **Implicit relevance feedback loop:** As researchers query, save favourites, re-listen to audio, and return for follow-up queries, their behaviour signals which content and formats are most valuable. This proprietary behavioural signal — what researchers actually find useful vs. what is merely retrieved — can be used to continuously improve embedding ranking and content curation. Generic tools (ChatGPT, Copilot) have no access to this domain-specific signal.

3. **Workflow lock-in:** Favourited articles, saved queries, audio libraries, and sync timestamps stored in the platform create switching costs. As researchers build institutional knowledge bases (custom sources, curated collections), migration becomes progressively harder. At the enterprise tier, SSO integration and team-shared knowledge bases deepen this lock-in further.

These three layers — data compounding, behavioural signal, and workflow integration — combine into a moat that pure technology replication cannot replicate.

2.5 Market Opportunity

- **TAM:** $4.2B — global academic and scientific information tools market
- **SAM:** $820M — AI-assisted research discovery and synthesis tools
- **SOM:** $35M — early adopter segment: AI-forward R&D teams in pharma, biotech, and top-tier universities
- **Growth Rate:** 34% CAGR in AI-assisted research tools (2024–2028)
- **Why Now:** GPT-5-class models enable reliable RAG synthesis; ElevenLabs TTS quality makes audio consumption viable; vector DB costs have dropped 90% in 2 years; researchers' AI comfort level has crossed the mainstream adoption threshold

Current Solutions and Gaps:

Tool	Strength	Gap
PubMed / Google Scholar	Comprehensive index	Search only, no synthesis or audio
Elicit / Consensus	AI paper summaries	No audio, limited custom sources, no RAG pipeline
Semantic Scholar	Citation graph	No conversational interface
ChatGPT	Natural language	No citation grounding, no custom corpus, hallucinates sources

2.6 Business Case

- **Revenue Potential:** SaaS model at $49/mo individual / $299/mo team = $6M ARR at 1,000 team subscribers
- **Cost Savings:** Each researcher recovering 5 hrs/week × $80/hr = $20,800/yr per researcher
- **Strategic Alignment:** Positions ScienceAI at the intersection of two high-growth markets: AI tooling and scientific research automation
- **Risk of Inaction:** Microsoft (Copilot for Research), Elsevier, and well-funded startups are converging on this space; first-mover advantage with audio differentiation and domain ingestion depth matters now

---

3. Solution Overview

3.1 What We're Building

ScienceAI is a web application where a researcher types a natural language question and receives a synthesised, source-attributed AI answer drawn **exclusively** from a continuously updated vector knowledge base (Chroma DB). Responses are grounded — not hallucinated — because GPT-5-mini is constrained to retrieved context only; when no relevant context exists, the system says so explicitly rather than fabricating an answer. Researchers can also generate single-voice audio summaries or dual-voice podcast episodes from any research topic, enabling consumption during commutes or lab work without screen dependency.

3.2 User Flows (Input → Processing → Output)

Flow 1 — RAG Chat (primary)

> Researcher types query → Next.js `/api/chat-articles` → Langflow RAG Pipeline → Chroma DB semantic search (top-k chunks retrieved) → Chunks injected as context into GPT-5-mini prompt (with strict constraint: "answer only from provided context") → Token-by-token streamed response rendered in UI → Source document titles/URLs displayed alongside answer → Researcher can click sources to verify claims

Flow 2 — Article Discovery

Researcher types search terms → Next.js `/api/articles` → Chroma DB cosine-similarity search → Ranked article list returned (title, source, relevance snippet) → Researcher saves favourites to localStorage → Saved articles persist across sessions with sync timestamp

Flow 3 — Audio Summary

Researcher requests audio on a topic → Next.js `/api/generate-audio` → RAG context retrieved from Chroma DB → Synthesis prompt built from retrieved chunks → ElevenLabs TTS API (single voice) → MP3 streamed to browser / available for download → 3–5 minute coherent summary playable offline

Flow 4 — Research Podcast

Researcher requests podcast on a topic → Next.js `/api/generate-podcast` → Langflow Podcast Flow generates dual-voice dialogue script (host + expert framing) → ElevenLabs TTS renders two distinct voices → MP3 podcast (8–12 min) available for playback / download

Flow 5 — Automated Ingestion (background)

Scheduled trigger (daily) → Langflow Sync Agent → Fetches from 4 source types (PDFs, XML, JSON, HTML) → Chunks text into ~500-token segments → Generates embeddings via embedding model → Upserts into Chroma DB (deduplication by hash) → Sync timestamp updated in localStorage → Admin dashboard shows ingestion logs + freshness status

3.3 In Scope

Feature	Priority	Description
RAG-powered Chat	P0	NL Q&A grounded in retrieved chunks via Langflow + GPT-5-mini
Source Attribution	P0	Every answer surfaces originating documents; researchers verify claims
Article Semantic Search	P0	Semantic search over Chroma vector store via `/api/articles`
Automated Ingestion Pipeline	P0	Langflow Sync Agent pulling from literature, conferences, trials, competitors
Hallucination Prevention	P0	Hard constraint: retrieval-only context; explicit “no context” fallback
Streaming Responses	P1	GPT-5-mini token streaming for real-time display
Audio Summary (single-	P1	ElevenLabs TTS converts

Feature	Priority	Description
voice)		synthesis to MP3
Dual-Voice Podcast	P1	Two-voice dialogue format for deeper topic exploration
Favourites & Persistence	P1	localStorage-backed article favourites, theme, sync timestamps
Admin Ingestion Dashboard	P2	Logs, freshness status, error visibility for sync runs
Multi-source Configuration	P2	UI to add/remove ingestion sources without code changes

3.4 Out of Scope

- Native mobile app (iOS/Android) — Phase 2
- Collaborative workspaces and team annotations — Phase 2
- Full-text PDF upload and personal corpus management — Phase 2
- Real-time push alerts for new publications — Phase 3
- Fine-tuned domain-specific models — Phase 3
- User authentication and server-side session storage — Post-MVP (see Open Questions)

3.5 MVP Definition

- **Core Features:** RAG chat with citations, article semantic search, audio summary, automated ingestion from 4 source types, hallucination prevention
- **Success Criteria:** End-to-end query → grounded answer < 5s; audio generation < 30s; zero hallucinated citations in 50-query manual test; sync freshness < 24 hours
- **MVP Target:** Q3 2026 Private Beta
- **Learning Goals:** Validate researcher trust in AI-synthesised answers when sources are shown; validate audio consumption habit formation in commute/lab context

---

4. User Stories & Requirements

4.1 User Stories

Story 1 — Literature Chat As a **biotech research scientist**, I want to **ask a natural language question about a research topic and receive a synthesised, source-cited answer**, So that **I can get to insight in seconds instead of manually reading 20 papers**.

Acceptance Criteria: - [ ] Answer is generated exclusively from retrieved Chroma DB chunks (no parametric hallucination) - [ ] At least 2 source documents are cited per response - [ ] Response begins streaming within 2 seconds of query submission - [ ] If no relevant context exists, system responds: “I don’t have reliable information on this in the current knowledge base” — no fabrication

---

Story 2 — Audio Commute As a **researcher with limited screen time**, I want to **convert a research topic into a listenable audio summary**, So that **I can stay current while commuting or doing bench work without staring at a screen**.

Acceptance Criteria: - [ ] Single-voice MP3 generated within 30 seconds - [ ] Audio is 3–5 minutes long and coherent for a standard topic - [ ] Download or in-browser playback available - [ ] Audio content is grounded in the same vector store as chat answers

---

Story 3 — Research Podcast As a **department head preparing for journal club**, I want to **generate a dual-voice podcast on a specific research theme**, So that **my team can listen to a conversational overview before our meeting**.

Acceptance Criteria: - [ ] Two distinct voices audible (host + expert framing) - [ ] Podcast covers problem context, key findings, and open questions - [ ] MP3 generated within 60 seconds - [ ] Topic grounded in retrieved literature with no fabricated statistics

---

Story 4 — Article Discovery As a **PhD student conducting a literature review**, I want to **search for articles semantically — not just by exact keywords**, So that **I find conceptually related papers even when terminology differs across sources**.

Acceptance Criteria: - [ ] Semantic search returns ranked results with title, source, and snippet - [ ] Results are sortable by relevance score - [ ] Articles can be saved to favourites and persist across sessions - [ ] Sync timestamp visible so I know how current the knowledge base is

---

4.2 Functional Requirements

ID	Requirement	Priority	Notes
FR1	System must generate answers grounded only in retrieved vector chunks — no parametric fill-in	P0	Core trust mechanism
FR2	Every response must include source attribution (document title and URL/reference)	P0	Required for researcher trust and verification
FR3	GPT-5-mini responses must stream token-by-token to the UI	P0	UX responsiveness

ID	Requirement	Priority	Notes
FR4	Chroma DB must be queryable via semantic similarity search with configurable top-k	P0	Powers chat and article search
FR5	Langflow Sync Agent must ingest from ≥ 4 source types on automated schedule	P0	Literature, conferences, clinical trials, competitors
FR6	System must return explicit “no context” response rather than fabricate when retrieval returns empty	P0	Hallucination prevention
FR7	`/api/generate-audio` must produce listenable MP3 via ElevenLabs TTS within 30 seconds	P1	Single-voice audio summary
FR8	`/api/generate-podcast` must produce dual-voice MP3 within 60 seconds	P1	Two distinct ElevenLabs voices
FR9	User favourites, theme preference, and sync timestamp must persist via localStorage	P1	Light persistence without auth at MVP
FR10	Sync timestamp must be displayed in UI so users can assess knowledge base freshness	P1	Transparency / trust
FR11	Admin dashboard must surface ingestion logs, per-source counts, and failure alerts	P2	Operational visibility
FR12	All API routes must implement rate limiting to prevent	P1	Cost control and reliability

ID	Requirement	Priority	Notes
	abuse and control LLM costs		

4.3 Agent Capabilities and Autonomy Boundaries

The Langflow Sync Agent operates as an autonomous background agent with defined human-review thresholds:

Operates autonomously (no human approval required): - Scheduled daily ingestion from configured source list - Chunking (500-token segments), embedding generation, and upsert into Chroma DB - Deduplication by content hash (skips already-ingested documents) - Retry on transient API failures (exponential backoff, max 3 retries) - Updating sync timestamp on successful run completion

Requires human review or approval: - Adding a new source type to the ingestion configuration - Any ingestion run that produces 0 new embeddings (potential source structure change — flagged in admin dashboard) - Citation accuracy audit trigger: if manual review detects > 5% mismatch rate, ingestion is paused pending investigation - Source list changes that may introduce GDPR-relevant personal data

Error handling behaviour: - API timeout or source unreachable: logs failure with source name, timestamp, and error code; continues with remaining sources; does not surface error to end users - Embedding model failure: queues documents for retry in next scheduled run; alerts admin dashboard - Chroma DB write failure: full run rolled back; previous sync state preserved; admin alerted

4.4 Non-Functional Requirements

- **Performance:** Chat first-token response < 2s P95; article search < 500ms P95; audio generation < 30s P95
- **Scalability:** 500 concurrent users at MVP; 5,000 at V1.0; Chroma DB to support > 10M vectors
- **Security:** API keys stored server-side as environment variables; rate limiting on all `/api/*` routes; no PII stored at MVP (localStorage only)
- **Reliability:** 99.5% uptime target; graceful degradation if ElevenLabs or OpenAI APIs are unavailable (cached responses where possible; user-facing error messages for live failures)
- **Usability:** WCAG 2.1 AA compliance; responsive at 1280px+ desktop; keyboard-navigable; audio player accessible without mouse
- **Compliance:** No patient-level data ingested (ClinicalTrials.gov public metadata only); GDPR-compatible data handling for EU researchers by V1.0

5. Go-to-Market Strategy

Launch Plan

- **Private Beta (Q3 2026):** 10–20 research teams from partner universities and biotech firms; white-glove onboarding; weekly structured feedback sessions targeting trust, audio adoption, and query satisfaction
- **Public Beta (Q4 2026):** Open waitlist; product-led growth via academic Twitter/LinkedIn; integration guides for existing research workflows
- **V1.0 Launch (Q4 2026):** Press outreach to science and tech publications (The Scientist, Nature News, TechCrunch); ProductHunt launch; conference presence at NeurIPS / Bio-IT World
- **Marketing:** 60-second demo video showing query → sourced answer → podcast flow; researcher testimonials vs. PubMed baseline; LinkedIn thought-leadership on "AI-assisted research"
- **Support:** In-app contextual docs, FAQ, community Slack for beta researchers, admin dashboard for sync health

Pricing

Tier	Price	Limits
Free	$0/mo	20 queries/day, text only, no audio
Researcher	$49/mo	Unlimited queries, 50 audio generations/mo
Team	$299/mo	Up to 10 seats, unlimited audio, custom source config
Enterprise	Custom	SSO, dedicated ingestion, SLA, audit logs

6. Metrics Framework

North Star Metric

Hours of active research work saved per researcher per week Target: 3+ hours/week by Q4 2026. This captures the platform's core value proposition — time — and correlates directly with engagement depth, willingness to pay, and renewal rates. Measured via: session-length analysis (proxy), onboarding survey (baseline self-report), and quarterly researcher survey during beta.

Metric Hierarchy

Tier	Metric	Unit	Baseline	Target
North Star	Research hours saved / researcher / week	Hours	—	3+ hrs

Tier	Metric	Unit	Baseline	Target
Primary	Time-to-insight P95	Seconds	< 5s (POC)	< 3s
Primary	Citation accuracy rate	% grounded	Manual check	≥ 95%
Secondary	30-day researcher retention	%	—	≥ 60%
Secondary	Audio feature adoption	% of WAU	—	≥ 40%
Secondary	Cost per query	USD	—	< $0.05
Secondary	Net Promoter Score (NPS)	Score	—	≥ 45
Health	API error rate (chat endpoint)	%	—	< 0.5%
Health	Vector store freshness lag	Hours	—	< 24 hrs
Health	Sync Agent success rate	% runs succeeding	—	≥ 98%

7. Risks & Mitigations

Risk	Probability	Impact	Mitigation
LLM hallucination erodes researcher trust	Medium	High	Hard retrieval-only constraint (FR1); explicit “no context” fallback (FR6); source attribution on every answer (FR2)
ElevenLabs API latency or downtime	Medium	Medium	Queue audio requests; progress indicator shown during generation; cached audio for common queries
Chroma DB vector store becomes stale or outdated	Medium	High	Automated daily sync; freshness timestamp visible to users; admin alert on 0-embedding runs
OpenAI API cost scaling with user	High	Medium	Per-query token budgeting; response

Risk	Probability	Impact	Mitigation
growth			length limits; query result caching for common searches
GDPR / data privacy concerns from EU institutions	Low	High	No user PII stored server-side at MVP; legal review before EU outreach; GDPR-compatible data handling in V1.0
Competitor (Microsoft, Elsevier) launches similar product	Medium	Medium	Accelerate audio differentiation and domain flywheel; build researcher community moat early; focus on workflow integration depth that large players deprioritise
Source licensing concerns for ingested content	Low	High	Legal review of each source before production ingestion; scrape only publicly available metadata where full-text rights are unclear

8. Timeline & Milestones

Milestone	Date	Deliverables	Success Criteria
POC Complete	April 2026 ✅	Working demo, architecture + data flow diagrams, capstone presentation	All 6 API routes functional; end-to-end demo completed
Beta Prep	June 2026	Auth layer, usage limits, rate limiting, monitoring, error handling	Zero unhandled crashes in 48-hour soak test
Private Beta	August 2026	Onboard 10 research teams; weekly feedback cadence	≥ 7/10 teams report measurable time savings vs. baseline
V1.0 Launch	November 2026	Public release, pricing tiers, support docs, admin	500 WAU; < 0.5% error rate on chat;

Milestone	Date	Deliverables	Success Criteria
		dashboard	NPS ≥ 45
Enterprise Tier	Q1 2027	SSO, custom source ingestion, SLA, audit logs	First paying enterprise customer; ≥ 98% Sync Agent success rate

9. Team & Resources

Role	Allocation
Product Manager	1 × 100%
Full-Stack Engineer (Next.js)	2 × 100%
AI / ML Engineer (Langflow, RAG)	1 × 100%
Design / UX	1 × 50%
QA Engineer	1 × 50%
DevOps / Infra	0.5 FTE

Infrastructure Budget (Monthly at Beta Scale):

Category	Monthly Cost	Notes
OpenAI API	~$800	~500 users × avg 20 queries/day
ElevenLabs TTS	~$400	~5,000 audio generations/mo
Chroma DB / hosting	~$300	Vector store + retrieval infrastructure
Next.js hosting (Vercel/AWS)	~$200	API routes + CDN
Total (Beta)	**~$1,700/mo**	Scales ~3× at V1.0 launch

10. Open Questions

1. **Authentication:** Will MVP require user accounts, or is localStorage-only acceptable for private beta? Impacts privacy posture and personalisation roadmap.
2. **Source licensing:** Are all ingested sources (journals, trial databases) freely redistributable at scale? Legal review needed before production ingestion.
3. **Audio IP:** Who owns AI-generated audio content produced from third-party scientific literature — researcher, platform, or original authors?
4. **Ingestion frequency:** Should sync run on a fixed schedule (daily default) or event-driven (new publication detected)? Event-driven is more current but significantly more complex.
5. **Citation evaluation at scale:** How do we systematically measure citation accuracy beyond manual spot-checking as the knowledge base grows to millions of vectors?

11. Assumptions Made

- Primary user persona is a researcher at a university, pharma, or biotech company with at least basic AI tool familiarity
- “Clinical trial data” refers to publicly available metadata (ClinicalTrials.gov), not proprietary patient-level data
- POC architecture (Next.js 16 + Langflow + Chroma + ElevenLabs) is the target stack for V1.0 with incremental hardening
- OpenAI GPT-5-mini provides sufficient reasoning quality for research synthesis; model upgrades are a P2 roadmap item
- ElevenLabs TTS quality is acceptable without custom voice cloning at MVP
- Mobile-responsive web is sufficient for initial market validation; native app not required for V1.0
- All market size figures ($4.2B TAM, 34% CAGR) are estimates based on publicly available reports and should be validated with primary market research before fundraising use