

PRD

askDr.ai

A RAG-powered multi-intent health information assistant

Author	Tabassum Khanum, Product Manager
Version	1.0 · June 2026
Status	Shipped — Live at ask-dr-ai.vercel.app
Repository	github.com/tabassum2507/askDr.ai

1. PRODUCT OVERVIEW

askDr.ai is a free, AI-powered health information assistant that uses Retrieval-Augmented Generation (RAG) to deliver cited, trustworthy answers across 8 health categories. Every answer is grounded in official medical sources — openFDA drug labels and MedlinePlus health guidelines — not the model's training data.

Vision

A world where anyone can access reliable, cited health information in seconds — without needing medical literacy, an internet connection fast enough to parse 20 search results, or the ability to distinguish credible sources from misinformation.

Product Principles

- **Grounded, not generated.** Every answer must trace back to an authoritative source. If the system doesn't have data, it says so.
- **Safety first, always.** Emergency detection runs before any other logic. The system never diagnoses and never prescribes.
- **Show your sources.** Citations are not optional. They are the trust mechanism.
- **Free and accessible.** The entire stack runs on free tiers. Cost is not a barrier to health information.

2. PROBLEM STATEMENT

User Pain Points

- **Unreliable AI answers:** Generic chatbots hallucinate medical information. A wrong answer about drug interactions or dosages can cause real harm.
- **Information overload:** Searching for 'headache causes' returns millions of results ranging from 'drink water' to 'brain tumor.' Users lack the medical literacy to filter.
- **No safety net:** Someone typing 'I can't breathe' into a search engine gets articles, not emergency guidance.

- **Language barriers:** India's health information gap is largest among non-English speakers and users unfamiliar with medical terminology.

Why Now

- RAG architecture has matured for production-grade cited-answer systems on free infrastructure.
- Open medical data (openFDA, MedlinePlus) provides authoritative, API-accessible health information at zero cost.
- LLM inference costs have dropped to zero for demo-scale usage (Groq free tier).

| 3. TARGET USERS

Primary: The Informed Patient

A 25–45 year old in India who searches for health information online before or after a doctor visit. They want to understand their medicines, check side effects, or make sense of lab results. They want information to have a better conversation with their doctor.

Secondary: The Caregiver

A family member managing health for an elderly parent or young child. They need quick, reliable answers about medicines, symptoms, and when to escalate. Speed and trustworthiness matter more than depth.

Anti-Persona (Out of Scope)

- Medical professionals seeking clinical-grade decision support
- Users looking for prescription fulfillment or telemedicine
- Users seeking mental health therapy (the app provides information and helpline numbers, not therapy)

| 4. USER STORIES

Medicines

- As a patient, I want to look up side effects of my prescribed medicine, so I know what to watch for.
- As a caregiver, I want to check if two medicines interact, so I can alert the doctor if needed.
- As a user, I want to photograph a medicine box and get information about it, so I don't have to type the name.

Symptoms & Health

- As a user feeling unwell, I want to describe my symptoms and learn possible conditions, so I can decide whether to see a doctor.
- As a parent, I want home remedy suggestions for my child's cold, so I can provide comfort while monitoring.

Mental Health

- As someone experiencing anxiety, I want to learn management techniques from trusted sources, so I feel more in control.

Reports & Diet

- As a patient, I want to understand what 'high cholesterol' means in my blood test, so I can ask my doctor the right questions.
- As someone managing diabetes, I want to know which foods help control blood sugar, so I can make better dietary choices.

5. REQUIREMENTS

P0 — Must Have

ID	Requirement	Rationale
P0-1	RAG-grounded answers with source citations on every response	Core trust mechanism
P0-2	Emergency safety screen that bypasses RAG for life-threatening queries	User safety is non-negotiable
P0-3	Medicine information lookup from openFDA drug labels	Most bounded, lowest-risk starting point
P0-4	Multi-category intent routing (8 health categories)	Different needs require different data
P0-5	Medical disclaimers on every response	Legal and ethical requirement
P0-6	No-diagnosis policy enforced in system prompt	Prevents most dangerous misinformation

P1 — Should Have

ID	Requirement	Rationale
P1-1	Medicine image recognition via vision model	Photo is easier than typing drug names
P1-2	Streaming responses (word-by-word)	Perceived performance
P1-3	Conversation memory for follow-up questions	Natural conversation flow
P1-4	Suggested starter questions per category	Reduces blank-screen anxiety
P1-5	Live openFDA API fallback for non-ingested medicines	Coverage without full ingestion

P2 — Could Have

- Guided symptom checker flow (multi-step structured input)
- Drug interaction checker with severity badges
- PDF lab report upload, parsing, and plain-language explanation
- Multilingual support (Hindi, Bengali, Tamil)
- Doctor/clinic locator via OpenStreetMap APIs

Out of Scope

- User accounts or authentication
- Prescription management or medication reminders

- Telemedicine or appointment booking
- Clinical-grade decision support

6. SUCCESS METRICS

Metric	Definition	Target	Measured By
Grounding Rate	% of RAG vs fallback answers	> 70%	Mixpanel: source property
Safety Pass Rate	Eval suite safety tests	100%	Automated eval script
Engagement	Messages per session	> 3	Mixpanel: Message Sent
Conversion	Landing → Message funnel	> 30%	Mixpanel: Funnel
Eval Score	Overall eval pass rate	12/12	eval.mjs

7. SAFETY & COMPLIANCE

Safety is not a feature — it is a constraint that shapes every design decision.

Rule	Implementation
Emergency detection	Regex pre-RAG screen for chest pain, breathing, self-harm, overdose. Bypasses all processing.
No diagnosis	System prompt forbids ‘you have X.’ Phrases as ‘possible conditions to discuss with your doctor.’
No prescription	Never says ‘take X medicine.’ Provides information, always defers to doctor.
Mandatory disclaimer	Every response: ‘confirm with your doctor.’ Persistent UI disclaimer on every screen.
Citations required	RAG answers must show source citations. No citation = not trusted.
Privacy	No data stored. Images in-memory only. Anonymous device-level analytics.
Harmful content block	Queries about intentional misuse return crisis helpline information instead.

8. DATA ARCHITECTURE

Knowledge Sources

Source	Covers	Categories
openFDA Labels	Indications, dosage, warnings, side effects, interactions	Medicines
openFDA Live API	Any FDA-approved drug (real-time fallback)	Medicines (fallback)
MedlinePlus	Health topics, conditions, symptoms, self-care, nutrition	All other categories

Vector Database

- Supabase PostgreSQL + pgvector extension
- Schema: id, content, metadata (jsonb), embedding (vector(768)), created_at
- Index: HNSW with vector_cosine_ops

- Search: cosine similarity, category filter, threshold 0.3, top-k 5

Embedding Pipeline

- Model: Google Gemini gemini-embedding-001, outputDimensionality: 768
- Chunking: ~800 characters, 100 character overlap
- Rate limit mitigation: dual API key fallback with exponential backoff

9. TECHNICAL CONSTRAINTS

- **Zero cost:** Entire stack on free tiers — Supabase, Groq, Gemini, Vercel, Mixpanel.
- **JavaScript only:** Full-stack TypeScript/JavaScript. No Python in production.
- **Serverless:** Vercel functions — no persistent disk, 60s max execution.
- **Rate limits:** Gemini: 1,500/day. Mitigated with dual-key fallback.
- **Index limit:** HNSW max 2,000 dims. Gemini reduced to 768 via outputDimensionality.
- **No auth:** Anonymous device-level tracking only.

10. RELEASE PLAN

Phase	Scope	Ship Criteria
Phase 1	Medicine info: openFDA ingestion, single-intent RAG, safety screen, chat UI	End-to-end cited answer from correct drug label
Phase 2	All categories: MedlinePlus ingestion, intent router, category-filtered retrieval	All 8 categories return RAG-grounded answers
Phase 3	Advanced: image recognition, streaming, memory, live openFDA fallback	Photo of any medicine returns drug info
Phase 4	Polish: Mixpanel analytics, eval suite, landing page, deployment	12/12 eval, events flowing, live on Vercel

11. RISKS & MITIGATIONS

Risk	Severity	Mitigation	Status
Medical misinformation	CRITICAL	RAG grounding + citations + disclaimers	Mitigated
Missed emergency	CRITICAL	Pre-RAG safety screen + eval tests	Mitigated
LLM hallucination	HIGH	System prompt constraints + hallucination eval	Mitigated
API rate limits	MEDIUM	Dual-key fallback + retry logic	Mitigated
Limited knowledge base	MEDIUM	Live API fallback + graceful degradation	Partial
Privacy / data leak	HIGH	No storage, in-memory processing, anonymous analytics	Mitigated

| 12. FUTURE ROADMAP

- **V1.1 — Expanded knowledge:** Full openFDA (100K+ drugs), complete MedlinePlus, regional health data.
- **V1.2 — Multilingual:** Hindi, Bengali, Tamil. RAG in English, generation translates.
- **V1.3 — Report analyzer:** PDF upload, OCR, lab value explanation in plain language.
- **V1.4 — Clinic locator:** OpenStreetMap integration for nearby doctors and clinics.
- **V2.0 — Personalization:** User accounts, health profiles, medication reminders.

Built by Tabassum Khanum

Product Manager · AI-Native PM · [linkedin.com/in/tabassum-khanum](https://www.linkedin.com/in/tabassum-khanum)
ask-dr-ai.vercel.app · github.com/tabassum2507/askDr.ai