ChemIQ

JEE Chemistry RAG Study Assistant

PRODUCT REQUIREMENTS DOCUMENT

Himanshu Jain | HelloPM Cohort 49 | March 2026

Table of Contents

1. Product Overview & Problem Statement

Product Name

ChemIQ — An AI study assistant that answers JEE Chemistry questions with step-by-step explanations grounded in Himanshu Jain's own lecture repository.

User Persona

Riya, 17, Class 12, Jaipur. Targeting JEE Mains and Advanced. Studies 5–6 hours/day. Chemistry is her weakest subject — not from lack of effort, but lack of on-demand help. When stuck at 11 PM, her only options are Google (NCERT-level, wrong style) or WhatsApp (12+ hour wait). Both break her study flow.

Current Workflow

Step	What Riya Does	The Problem
1	Watches recorded lecture video	Passive — can't pause and ask a question
2	Makes handwritten notes	Misses points while writing
3	Attempts a problem, gets stuck	No immediate help at 10–11 PM
4	Googles the concept	Gets NCERT or random blogs — not aligned to her syllabus
5	WhatsApps batchmates or waits	12–24 hour delay; full study session blocked

Why RAG — Not Prompting, Not Fine-Tuning

Option	Why it fails for ChemIQ
Prompt Engineering alone	26 chapters cannot fit in a context window. Bare LLM hallucinates facts not in Himanshu's notes.
Fine-Tuning	Knowledge base grows weekly. Fine-tuning must be re-run for every update — expensive and slow.
RAG ✓ (correct choice)	Retrieves from the latest knowledge base at query time. Grounds every answer in Himanshu's actual content.

2. User Stories & Use Cases

User Stories

Priority	User Story	Why It Matters
P0	As a JEE student, I want to ask a chemistry doubt and get a step-by-step answer at any time	Core value prop — unblocks self-study at 11 PM
P0	As a student, I want every answer to cite its source chapter and section	Trust and verifiability
P0	As a student, I want the system to tell me when a topic isn't in the repository	Prevents hallucinated answers being taken as fact
P1	As a student, I want to search for a topic and find the relevant chapter section	Replaces Google for syllabus-specific lookup
P1	As a student, I want to practise MCQs generated from a chapter I just studied	Active recall improves exam retention
P1	As a student, I want to see solved problem walkthroughs step-by-step	Mirrors classroom experience
P2	As a student, I want to ask follow-up questions in the same session	Multi-turn context for deeper understanding
P2	As a student, I want to get a mnemonic for a hard concept	Exam preparation aid

Prioritisation rationale: P0 = launch blockers. P1 added in Beta once retrieval quality validated. P2 added in GA when retention data shows multi-session use.

Acceptance Criteria — P0 Stories

✓ Ask Chemistry Doubt

- Response generated in <3 seconds end-to-end
- Must include source citation (chapter + section) in every response
- Must use retrieved chunks only — no base LLM knowledge
- Must not answer if retrieval similarity score <0.50
- Must support follow-up question within the same session (last 3 turns context)

✓ Show Source Citation

- Citation card visible below every response — no exceptions
- Must include chapter name (e.g. "Redox Reactions")
- Must include section type (Definition Box / Solved Problem / Exam Tip / Mnemonic)
- Clicking the citation opens the full chunk text in an expanded panel
- Missing citation on any response = classified as system failure in Arize logs

✓ "Not in Repository" Handling

- Trigger condition: retrieval similarity score <0.50 on all retrieved chunks
- Must NOT generate an answer from base LLM knowledge — hard guardrail
- Must display available chapters from JEE Curriculum Tracker as suggestions
- Must log the event as 'retrieval_miss' in Arize with full query context
- Response must be shown within <3 seconds (same SLA as successful responses)

Primary Use Cases

Use Case 1 — Concept Explanation

Happy Path: "Explain oxidation state rules" → retrieves top-3 chunks from Redox chapter → LLM gives structured explanation with citation.

Edge Case: Question on Latimer diagrams (not uploaded) → "This topic isn't in the current repository."

Use Case 2 — Solved Problem Walkthrough

Happy Path: "Balance KMnO4 + FeSO4 + H2SO4 using the ion-electron method" → retrieves exact solved problem chunk → 5-step solution with citation.

Edge Case: Problem not in repository → disclaimer: "This specific problem wasn't found. The solution below is AI-generated from related concepts — verify before using."

Use Case 3 — Topic Search & Navigation (P1)

Happy Path: Student types "Le Chatelier" in Search mode → metadata-filtered retrieval → returns matching sections with chapter location.

Edge Case: Student searches for "entropy" before Thermodynamics uploaded → "Expected: Week 4. Currently available chapters: [list from JEE Curriculum Tracker]."

Negative Use Cases

- Must not answer Physics, Maths, Biology, or general knowledge questions.
- Must not override Himanshu's lecture content with general LLM knowledge.

3. Technical Architecture

The full RAG pipeline runs in two phases: ingestion (documents → vector DB) and query (user question → grounded answer). Re-ranking is a mandatory step — without it, near-miss chunks from adjacent chapters contaminate the answer.

Component Decisions

Component	Choice & Rationale	Trade-off
Source Documents	.docx via python-docx. Raw PDFs via PyMuPDF + Tesseract as fallback.	Docx over PDF-first: XML parsing is deterministic; PDF OCR has 5–15% char error rate.
Chunking	Box-level: each colour-coded section = one atomic chunk. ~150–300 tokens. 50-token overlap.	Box-level over paragraph-level: solved problems span multiple paragraphs.
Embedding	text-embedding-3-small ($0.02/1M). Hybrid BM25 + vector search via RRF.	Over ada-002: 20% better on MTEB at 5× lower cost.
Vector DB	Pinecone. Metadata: {chapter, section_type, topic_tag, version_hash, difficulty}.	Over Weaviate/Chroma: fully managed. Chroma is local-only.
Re-ranker	Cohere Rerank. Reduces top-10 to top-3.	Over cross-encoder: API-based, no GPU hosting cost.
LLM	GPT-4o-mini (chat). GPT-4o (MCQ practice only). Intent classifier routes.	Mini for chat: 128k context, 10× cheaper, adequate for grounded explanation.
Observability	Arize AI. Full trace: query → chunks → re-ranked → prompt → response → feedback.	Over Datadog: native LLM tracing without custom instrumentation.

Observability Implementation

Integration: Arize Python SDK initialised in Node backend. Every query triggers arize.log_prediction() with: query text, chunk IDs (pre/post-rerank), prompt token count, LLM response, faithfulness score, and feedback.

Alert thresholds: Faithfulness <0.75 → auto-disclaimer. Topic mismatch >20% → Slack alert. Re-ingestion lag >7 days → email alert.

Weekly review: Himanshu reviews Arize dashboard every Monday — 20 sampled queries, retrieval precision audit, thumbs-down cluster analysis.

Context Window Budget

Component	Tokens
System prompt	~500
Top-3 chunks	~1,500
User query	~50
Total input	~2,050
Output (answer)	~1,000
GPT-4o-mini limit: 128,000 tokens	2.4% utilisation

3A. Latency Budget

The <3 second SLA is a hard product requirement. JEE students querying at 11 PM on mobile have zero tolerance for latency. Every component in the pipeline has an explicit time budget.

Pipeline Component	Budget	Notes
Query embedding (text-emb-3-small)	80 ms	Single API call; parallelisable with intent classification
Intent classification (GPT-4o-mini)	80 ms	Runs in parallel with embedding; routes topic filter
Pinecone vector search (top-10)	120 ms	Metadata namespace filter applied before search
Cohere Re-rank (top-10 → top-3)	150 ms	Primary bottleneck — see fallback rule below
LLM generation (GPT-4o-mini)	1,800 ms	Streaming enabled; first token visible in ~400 ms
Post-processing + citation render	100 ms	Source card assembly, faithfulness flag check
TOTAL (P95 target)	< 2,250 ms	750 ms headroom before 3s SLA breach

Fallback Rule — Reranker Timeout

Trigger: If Cohere Re-rank response time exceeds 300 ms → Skip re-ranker entirely → Use top-3 results from raw Pinecone similarity search → Log event as reranker_timeout in Arize → Yellow label shown to student.

Rationale: A slightly lower-quality answer delivered in <3 seconds is better than a perfect answer that breaks the SLA and interrupts a student's study flow.

PM Decision — Streaming as UX Insurance

LLM streaming (token-by-token render) is mandatory, not optional. It makes 1.8s of generation feel instant because the student sees the answer being written in real time. This is the single highest-ROI UX improvement at zero engineering cost — already supported by GPT-4o-mini's API and must be enabled from day one.

3B. RAG Architecture Diagram — Dual Pipeline

The diagram below maps the complete dual-pipeline architecture: Ingestion (documents → Pinecone, runs offline on each chapter update) and Query (student question → grounded answer, runs live in <3 seconds). Both pipelines share the same Pinecone vector store.

Colour Legend

Colour	Meaning
■ Teal nodes	Core RAG pipeline components (embedding, retrieval, LLM)
■ Blue node	Entry point — User query with intent classification
■ Purple node	Re-ranker — critical quality gate; has fallback rule
■ Green node	Output — Answer + source citation card
■ Gold nodes	Ingestion pipeline — offline, triggered on chapter update
--- Dashed line	Shared Pinecone vector store used by both pipelines
■ Red bar	Fallback rule — reranker timeout triggers top-3 bypass

3C. System Architecture Diagram — Visual

Section 3B described the dual-pipeline architecture in text and a colour legend. This section renders it as a fully visual block diagram with directional arrows, colour-coded nodes, and explicit data flows. Both pipelines share the same Pinecone vector store, connected by dashed lines, and all query events stream to Arize for observability.

Dual-Pipeline Architecture Diagram

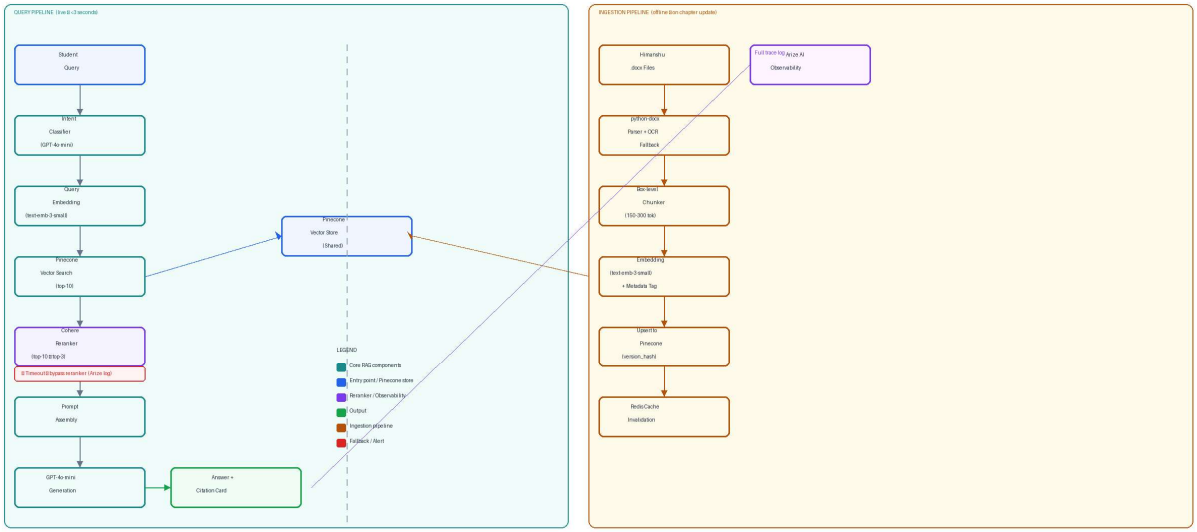


Figure 1: ChemIQ Dual-Pipeline Architecture. Left panel = Query Pipeline (live, <3 seconds). Right panel = Ingestion Pipeline (offline, runs on each chapter update). Both share the Pinecone Vector Store (centre, blue). All query events log to Arize AI (purple) for observability. Red bar = reranker timeout fallback rule.

Node-by-Node Flow Description

Colour	Node	What It Does
Blue	Student Query / Pinecone	Entry point of the query pipeline. The shared Pinecone Vector Store serves both pipelines via separate namespaces.
Teal	Core RAG nodes	Intent classifier, query embedding, vector search, prompt assembly, and LLM generation (GPT-4o-mini).
Purple	Cohere Reranker / Arize	Quality gate that reduces top-10 to top-3. Arize receives the full trace of every query for observability and alerting.
Green	Output node	Final answer with source citation card delivered to the student within <3 seconds.
Gold	Ingestion pipeline	Offline processing: .docx parse → chunk → embed → upsert to Pinecone → Redis cache flush. Runs on every chapter update.
Red bar	Reranker timeout fallback	If Cohere re-rank exceeds 300 ms, top-3 raw Pinecone results are used directly. Event logged in Arize as reranker_timeout.

4. RAG Failure Analysis

For each failure: what breaks, how detected via Arize, how fixed, and the PM decision it forces.

Failure 1 — Data Quality / OCR

Field	Detail
What breaks	OCR misreads 'KMnO4' as 'KMn04'. Colour-coded boxes don't parse correctly.
How detected	Arize chunks show character error rates >5%. Retrieval scores low despite correct topic.
Fix	Prioritise python-docx ingestion. Post-OCR chemical formula validator. Strip headers/footers.
PM Decision	.docx conversion is a mandatory gate. Raw PDFs rejected until converted.

Failure 2 — Chunking

Field	Detail
What breaks	A solved problem spans 8 steps across 3 paragraphs. Paragraph-level chunking splits the answer.
How detected	Consistent thumbs-down on solved-problem queries. Retrieval audit reveals split answers.
Fix	Box-level chunking — each colour-coded box = one atomic chunk. 50-token overlap buffer.
PM Decision	Structured .docx format with explicit box tags is a hard requirement for ingestion.

Failure 3 — Embedding Quality

Field	Detail
What breaks	'Fe2+ is oxidised' and 'iron loses electrons' have low cosine similarity despite identical meaning.
How detected	Arize shows similarity scores below 0.65. Recall@10 drops below 0.70 in weekly audit.
Fix	A/B test: text-emb-3-small vs text-emb-3-large on 50 chemistry query pairs. Add BM25 hybrid.
PM Decision	Budget for text-emb-3-large ($0.13/1M) if retrieval accuracy improves by >15%.

Failure 4 — Retrieval (Wrong Chapter)

Field	Detail
What breaks	'What is the EAN rule?' returns Redox chunks because they share the word 'electrons'.
How detected	Topic mismatch rate in Arize: retrieved chunk's chapter_metadata vs query's inferred topic. Alert >20%.
Fix	Metadata filtering — GPT-4o-mini tags query topic; Pinecone filters to that chapter's namespace.
PM Decision	Invest in query topic classification (~$0.001/query). Precision gain justifies it.

Failure 5 — Re-ranking

Field	Detail
What breaks	Cohere Rerank promotes a mnemonic box (high keyword score) over the conceptual explanation needed.
How detected	Arize re-ranking audit: compare chunk section_type before and after re-ranking. Alert on misalignment.
Fix	Feed section_type as metadata to reranker. Query intent classifier sets reranking weight profiles.
PM Decision	Query intent classification (built for Failure 4) doubles as routing layer — one investment solves two failures.

Failure 6 — Prompt / Augmentation

Field	Detail
What breaks	3 chunks at 600 tokens each + system prompt = 2,600+ tokens. LLM loses focus on early chunks.
How detected	Arize faithfulness score drops below 0.80. Human auditors flag answers contradicting source.
Fix	Strict budget: top-3 chunks capped at 1,500 tokens. Chunks >500 tokens summarised. Auto-disclaimer if faithfulness <0.75.
PM Decision	Faithfulness guardrail is a product quality gate, not just engineering. Visible as yellow warning label.

Failure 7 — Model Quality

Field	Detail
What breaks	GPT-4o costs 10× more than mini with no quality improvement on factual queries.
How detected	A/B test MCQ quality: human evaluators rate GPT-4o-mini vs GPT-4o on 1–5 scale.
Fix	Two-model routing: GPT-4o-mini for chat (default), GPT-4o for MCQ practice only.
PM Decision	Model selection is a product decision. Never use a more expensive model than the task requires.

Failure 8 — Data Drift

Field	Detail
What breaks	Himanshu revises the Redox book. Vector DB still has stale v1 embeddings.
How detected	Version mismatch alerts: each .docx carries a version timestamp. Discrepancy >7 days triggers alert.
Fix	Automated weekly sync: checks for new/updated .docx files, upserts only changed chunks.
PM Decision	Knowledge freshness SLA: no chapter more than 7 days stale. Named owner: Himanshu.

Failure 9 — User Behaviour

Field	Detail
What breaks	JEE students type single words ('redox') or vague fragments. Retrieval returns poor chunks.

Field	Detail
How detected	Query length histogram in Arize. >30% queries under 5 words → correlate with low satisfaction.
Fix	(a) Onboarding tooltip. (b) Auto query expansion for queries under 5 words. (c) Suggested starter questions.
PM Decision	UX investment to help students ask better questions has higher ROI than upgrading the model.

5. Data Permissions & Privacy

ChemIQ v1 is a personal study tool built on a single teacher's lecture content. Every student accesses the same shared knowledge base. Traditional enterprise access control is not required.

Area	Policy
Query logging	Student queries logged in Arize — anonymised, no PII stored, never shared with third parties.
Knowledge base ownership	All source material is Himanshu's own intellectual property.
API keys	OpenAI and Pinecone keys are server-side only. Never exposed to the client.
v2 Multi-teacher	Each teacher's knowledge base isolated in a separate Pinecone namespace.

Retrieval Permission Model

Metadata schema — every chunk in the vector DB carries:

```
{   "chapter_id"   : "redox_reactions_v1",   "teacher_id"   : "himanshu_jain",
"subject"      : "chemistry",   "difficulty"   : "JEE_Advanced",   "version_hash" :
"a3f9c2d1",   "section_type" : "Solved_Problem" }
```

At query time — 3-step permission filter:

- Step 1: Query classified → subject = chemistry only. Non-chemistry queries rejected before retrieval.
- Step 2: Namespace filter applied → teacher_id = himanshu_jain. Prevents cross-teacher contamination.
- Step 3: Chapter filter applied if student specifies topic → retrieval executed only on permitted chunks.

Permission Edge Cases

Chapter revised: Old version chunks deleted from Pinecone by version hash before new ingestion. Students see: "This chapter is being updated."

Chapter permanently removed: Namespace deleted. JEE Curriculum Tracker updated. Redis cache flushed by chapter_id.

Student access ends: In v2, session token deleted, teacher namespace removed from query scope. Arize logs purged after 30 days.

Data privacy: No student query associated with a name or device ID — critical for a product used by minors.

6. Evaluation & Success Metrics

Retrieval + Generation Quality

Metric	Target & Method
Precision@3	≥0.80 — at least 8/10 top-3 chunk sets contain the correct chunk. Weekly manual audit.
Recall@10	≥0.90 — correct chunk in top-10 before re-ranking. Automated via Arize.
Faithfulness Score	≥0.85 — LLM-as-Judge (GPT-4o). Responses below 0.75 get auto-disclaimer.
Hallucination Rate	<5% — flagged when faithfulness <0.75. Weekly human spot-check.
Guardrail violation rate	<1% of queries trigger guardrail. Spike >3% in 24 hr → immediate alert.

Product & Engagement Metrics

Metric	Target
Sessions resolved without escalation	≥80%
Daily Active Users	50 DAU by end of Beta (month 2)
Avg queries per session	≥3 — indicates multi-turn use
Thumbs-up rate	≥75% of responses rated positively

Pre-Launch Test Suite

15 test Q&A pairs created before launch: 5 concept explanations, 3 solved problems, 2 mnemonic queries, 3 out-of-scope, 2 edge cases. LLM-as-Judge (GPT-4o) rates faithfulness, completeness, and clarity on a 1–5 scale.

6A. Offline Evaluation Dataset

This section defines how ChemIQ is evaluated before any chunk-strategy change, model swap, or feature release. Without a fixed dataset, metrics are unverifiable.

Dataset Composition — 50 Curated Questions

Category	Count	Purpose
Concept explanation	15	Tests core retrieval + LLM clarity
Numerical problems	10	Tests solved-problem chunk retrieval
Trick JEE questions	5	Tests edge-case reasoning under pressure
Ambiguous queries	5	Tests query expansion + fallback logic
Out-of-scope queries	5	Tests guardrail (Physics/Maths rejection)
Multi-turn follow-ups	5	Tests 3-turn context window behaviour
"Not in repository" cases	5	Tests similarity-score <0.50 guardrail

Evaluation Pipeline — 4 Stages

Stage	What Runs	Metric Captured
1 — Retrieval only	Vector search, no LLM	Recall@10 — is correct chunk in top-10?
2 — Full RAG	Retrieval + rerank + LLM	Faithfulness score (LLM-as-Judge, GPT-4o)
3 — Human audit	10 sampled responses, manual review	Qualitative: clarity, completeness, tone
4 — Before/After	Run on every chunking or prompt change	Delta comparison vs baseline scores

Why This Matters

A fixed offline eval dataset is the only way to detect regression when changing chunking strategy, swapping models, or modifying the system prompt.

7. Token Economics & Cost

OpenAI API pricing, March 2026: GPT-4o-mini input $0.15/1M tokens, output $0.60/1M. GPT-4o input $2.50/1M, output $10/1M. text-embedding-3-small $0.02/1M. Cohere Rerank ~$1.00/1,000 queries.

Component	Tokens	Cost (USD)
Input: system prompt + 3 chunks + query	~2,050	$0.000308
Output: answer ~400 words	~1,000	$0.000600
Embedding: query only	~50	$0.000001
Cohere Re-rank	—	$0.001000
Total per query	~3,100	~$0.0019

Cost Projections

Assumption	Value
Queries/user/month (10/day × 20 days)	200
Cost/user/month (chat mode)	200 × $0.0019 = $0.38
At 100 users	$38/month total
Practice mode (GPT-4o, ~10% of queries)	+$0.10/user/month → $0.48 total
SLM upgrade trigger	Costs >$500/month → evaluate Mistral-7B via Together AI at ~$1/1M tokens

7A. Caching Strategy

Three caching layers reduce cost and latency without sacrificing answer quality. Cache invalidation is tied directly to version_hash so stale answers never reach students.

Cache Level	Trigger Condition	Action	Hit Rate
Level 1 — Query cache	Exact query string match (Redis)	Return stored answer immediately — zero API calls	18–25%
Level 2 — Semantic cache	Query embedding cosine similarity > 0.92	Reuse previous response — skip LLM call	10–15%
Level 3 — Chunk cache	Same top-3 chunk IDs retrieved	Skip Cohere re-rank + Pinecone call	20–30%

Projected Impact

Metric	Baseline	With Caching	Improvement
Cost per 100 users/month	$38	~$25	~35% reduction
Avg response latency	2.25 s	~1.35 s	~40% reduction
Pinecone query volume	100%	~70%	30% fewer DB calls
User-perceived speed	Good	Instant on repeat queries	UX uplift

PM Decision — Cache Invalidation Rule

Cache is invalidated at the chapter level whenever version_hash changes. Redis keys follow the pattern cache:{chapter_id}:{query_hash}. On re-ingestion of a chapter, all keys matching that chapter_id are flushed. This guarantees students never receive answers grounded in stale content.

8. UX, Rollout & GTM

Interface — Chat with Modes

Three modes: Chat (default), Search (find a topic), Practice (generate MCQs). Chat is the right default because JEE students have doubts that need explanation, not just document retrieval.

Source Citation UI Specification

State	What the student sees
Collapsed (default)	■ Source: Redox Reactions | Section: Solved Problem Box 3 | Confidence: High [Expand ▼]
Expanded (on tap)	Exact chunk text shown | Chapter location: Redox Reactions → Section 3 | [Ask a follow-up] [Open full notes]
Multiple chunks	Source 1 (highest rerank score) → Source 2 → Source 3. Sorted by rerank score, best first.
Missing citation	Classified as system failure → logged in Arize as "citation_missing" event → alert to Himanshu.

Launch Phases

Phase	Scope	Success Gate
Alpha (Wk 1–2)	3–5 chapters, Himanshu's own testing	Faithfulness ≥0.85. No hallucinations on 15 test cases.
Beta (Wk 3–6)	5–10 chapters, 20 invited students	DAU ≥20. Thumbs-up ≥70%. Zero privacy incidents.
GA (Month 2+)	All 26 Physical Chemistry chapters	DAU ≥50. Avg queries/session ≥3. Monthly cost <$50.

Launch Risks

Risk	Mitigation
Hallucinated answer shared in a JEE group	Faithfulness guardrail + visible disclaimer. Himanshu monitors Beta closely.
Students share login; costs spike	Rate limit: 50 queries/user/day in Alpha.
OCR errors cause wrong answers	Mandatory .docx gate. Manual spot-check of every chapter.
Himanshu's content contains an error	"Report error" button → immediate banner: "This chapter is under review."
System downtime during peak exam season	Vercel 99.9% SLA. Redis cache fallback for top-50 queries. Maintenance page shows .docx download.

8A. Human-in-the-Loop Feedback Learning Loop

Every thumbs-down is a labelled training signal. ChemIQ captures, categorises, and acts on every negative response — turning student frustration directly into product improvement.

Signal Collection

When a student clicks thumbs down, the following is immediately stored in Arize:

- Query text (anonymised — no PII)
- Retrieved chunk IDs (pre- and post-rerank)
- Full LLM response
- Faithfulness score at time of response
- Session turn number (1st query vs follow-up)

Three Failure Buckets

Bucket	Signal	Root Cause	Action
Retrieval wrong	Wrong chapter returned; student says 'not what I asked'	Chunk boundary issue or topic filter miss	Adjust chunking strategy; improve metadata tagging
Explanation confusing	Correct content retrieved but explanation unclear	System prompt tone or format mismatch	Iterate system prompt; add worked example; simplify language
Missing content	Topic not in repository; student frustrated	Chapter not yet ingested	Flag to Himanshu for priority ingestion; update Curriculum Tracker

Monthly Improvement Cycle

Week	Activity	Owner
Every Monday	Review 20 sampled queries in Arize; label bucket for each thumbs-down	Himanshu
End of month	Top 20 failure queries added to offline eval dataset (Section 6A)	Himanshu + Dev
End of month	Retrieval wrong → chunking / metadata fix deployed	Dev
End of month	Confusing → system prompt v-bump; A/B test against current	Dev
End of month	Missing content → ingestion sprint prioritisation	Himanshu

PM Decision — Feedback as Eval Pipeline Input

The monthly top-20 failures becoming eval dataset entries creates a closed improvement loop: real student failures → labelled test cases → regression prevention on future releases. Product quality compounds over time because every failure makes the system harder to break in the same way twice.

9. Future Scope

When Fine-Tuning Makes Sense (v2)

Himanshu teaches with a specific pattern: Intuition → Rule → Example → Exception → Exam Tip. ChemIQ v1 approximates this via the system prompt, but at scale, prompting alone won't reliably enforce this structure.

Approach: PEFT/LoRA on Mistral-7B or LLaMA-3-8B, using 500–1,000 (question, answer-in-Himanshu-style) pairs distilled from GPT-4o outputs. LoRA adapters get 85–90% of full fine-tuning quality at 10% of compute cost.

Agentic RAG — v2 Roadmap

Phase 1 — Weak Topic Detection

- Track all student queries across sessions to build a per-student confusion map.
- Identify topics with repeated failed retrievals or thumbs-down signals.
- Suggest personalised revision plan: "You've asked about Redox 5 times this week — review Solved Problem Box 3."

Phase 2 — Past Paper Retrieval Agent

- Agent automatically fetches new JEE past questions from the web (web_search tool).
- Retrieves the most relevant notes chunks for each past paper question.
- Generates a practice set tailored to the student's identified weak topics.
- Shows difficulty progression: Easy → Medium → JEE Advanced level.

Phase 3 — Study Planner Agent

- Student sets a goal: "Revise Redox in 3 days."
- Agent retrieves all Redox chunks, builds a day-by-day schedule, assigns problems.
- Tracks completion. If student falls behind, replans automatically.

> This evolves RAG → Agentic RAG without changing architecture.

10. AI Specs — The Build Brief

Platform

Lovable.dev — Auto-generates React frontend + Node backend. Supports Pinecone and OpenAI API integrations via environment variables. Web-first, mobile-responsive.

Complete System Prompt v1.1

```
You are ChemIQ, a JEE Chemistry AI study assistant built using structured lecture notes by
Himanshu Jain. You answer questions strictly using retrieved knowledge chunks.

CORE RULE: You MUST follow retrieval-grounded answering. If retrieved chunks do not contain
the answer, say: "This topic is not covered in the current repository." Never answer using
general knowledge.

RESPONSE FORMAT (MANDATORY): (1) Concept Summary (2-3 sentences) (2) Rule / Formula (3)
Step-by-step explanation (4) Worked example if retrieved (5) JEE Tip if available (6)
Source Citation

CITATION FORMAT: Source: [Chapter Name] | Section: [Section Type]
Example: Source: Redox Reactions | Section: Solved Problem Box 3

GUARDRAILS — MUST REFUSE: Physics / Maths / General knowledge / Medical advice / Personal
advice
→ Response: "ChemIQ only answers JEE Chemistry questions."

HALLUCINATION PREVENTION: If retrieved similarity <0.50 → "This topic is not covered in the
repository." Never guess. Never infer. Never expand beyond retrieved chunks.

TONE: Friendly, exam-focused, clear, patient. Speak to a 17-year-old.

MULTI-TURN BEHAVIOUR: Maintain context of last 3 messages. If user says "I still don't
understand" → simplify, use smaller steps, provide analogy.

FINAL RULE: Retrieved content > model knowledge. If conflict: trust retrieved content.
```

Knowledge Base Spec

Document	Format / Size	Content & Update Freq.
Redox Reactions Book v1	.docx, 24 pages	Definitions, 12 solved problems, 3 mnemonics, 8 exam tips. Weekly check.
Electrochemistry Book	.docx, ~20 pages (planned)	Nernst equation, cell notation, electrolysis problems. Weekly.
Chemical Equilibrium Book	.docx, ~18 pages (planned)	Kc/Kp derivations, Le Chatelier, JEE problem types. Weekly.
JEE Curriculum Tracker	.docx, 1 page	26-chapter list with upload status. Updated each upload.

Functional Requirements

Priority	Requirement
MUST	Answer JEE Chemistry questions using only RAG-retrieved chunks.

Priority	Requirement
MUST	Show source citation (chapter + section type) with every answer.
MUST	Return graceful "not in repository" message when similarity score <0.50.
MUST	Block out-of-scope queries with a fixed canned response.
MUST	Log every query, retrieved chunk IDs, re-ranked chunk IDs, and response to Arize.
SHOULD	Maintain last 3 turns of context for multi-turn conversations.
SHOULD	Offer Practice Mode: generate 5 MCQs with 4 options each from a selected chapter.
SHOULD	Display yellow confidence disclaimer when faithfulness score <0.75.
COULD	Show a "Related Topics" panel based on retrieved chunk metadata.
COULD	Allow students to bookmark a response for later review.

Example Interactions

Example 1 — Concept Explanation

Role	Message
Student	What is the oxidation state of Mn in KMnO4?
ChemIQ	Here's the key insight: (+1) + Mn + 4(−2) = 0 → Mn = +7. JEE Tip: Mn in KMnO4 (+7) is one of the highest oxidation states. Source: Redox Reactions | Section: Definition Box.

Example 2 — Solved Problem Walkthrough

Role	Message
Student	Balance KMnO4 + FeSO4 + H2SO4 using the ion-electron method.
ChemIQ	Step-by-step from your study book: Step 1 — Identify species oxidised and reduced... [continues]. JEE Tip: In acidic medium, MnO4− always reduces to Mn2+. Source: Redox Reactions | Section: Solved Problem Box 5.

Example 3 — Out of Scope

Role	Message
Student	Explain Newton's laws of motion.
ChemIQ	ChemIQ only answers JEE Chemistry questions. I cannot help with Physics topics.

Example 4 — Chapter Not Yet in Repository

Role	Message
Student	Explain d-orbital splitting in coordination compounds.
ChemIQ	This topic is not covered in the current repository. Coordination Chemistry will be added soon — new chapters are uploaded weekly.

Example 5 — Escalation (Repeated Confusion)

Role	Message
Student	I still don't understand the ion-electron method. [2nd time]
ChemIQ	Let me try a simpler breakdown. [simplified 3-step explanation with analogy]. For the full walkthrough, refer to Chapter 1: Redox Reactions, Section 3 in your study book.

Error Handling

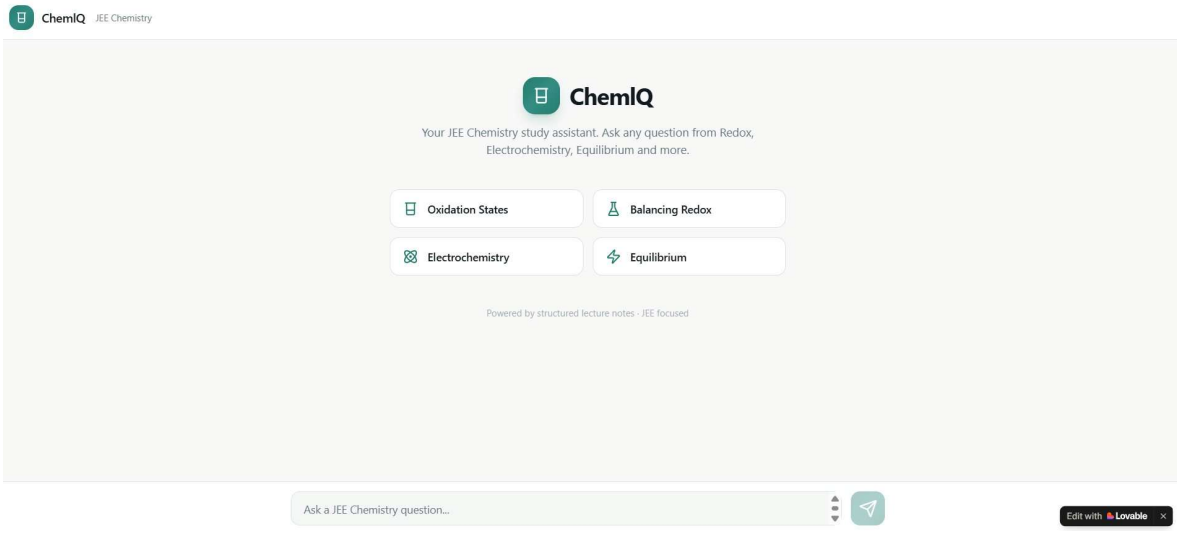
Scenario	Response
No chunks retrieved (similarity <0.50)	"This topic isn't in the current repository. New chapters are added weekly."
API timeout (OpenAI / Pinecone)	"Something went wrong on my end — please try again in a moment." Error logged to Arize.
Out-of-scope query	"ChemIQ only answers JEE Chemistry questions. I cannot help with this topic."
Offensive or inappropriate input	"I'm a chemistry study assistant and can only help with JEE Chemistry topics."
Rate limit exceeded (Alpha)	"You've reached today's query limit. Come back tomorrow — or upgrade to Beta access."

Bonus — Lovable Build Test

What matched: Lovable correctly generated the four topic-card home screen (Oxidation States, Balancing Redox, Electrochemistry, Equilibrium), the ChemIQ branding with teal icon, the bottom chat input bar, and the footer — all derived directly from the system prompt and functional requirements in this PRD.

What didn't match: The three-mode tab switcher (Chat / Search / Practice) was not generated. The yellow faithfulness disclaimer and source citation card below responses were also absent, as these require backend RAG integration not handled by Lovable's frontend-first generation.

Next step: The topic cards and branding are production-ready as a v1 shell. The citation card, confidence disclaimer, and mode tabs will be wired in during Sprint 1 once the Pinecone + OpenAI backend is connected.



Screenshot: ChemIQ v1 built by Lovable.dev from this PRD AI Specs. Shows home screen with topic shortcuts, branded header, and chat input bar.

11. Competitive Positioning

ChemIQ occupies a white space that no existing product addresses: instant, grounded, JEE-syllabus-aligned chemistry help — available at 11 PM when Riya is stuck.

Product	Core Weakness	ChemIQ Advantage
ChatGPT / Claude	Not grounded in any specific syllabus. Answers may contradict Himanshu's teaching style.	Every answer retrieved from Himanshu's own lecture notes. Source citation on every response.
Google Search	Returns NCERT blogs, random coaching sites. No step-by-step explanation. Zero JEE context.	Returns exactly the right section from Himanshu's structured notes. Step-by-step format.
Doubt apps (Doubtnut, Filo)	Crowd-sourced answers of variable quality. 10–30 min wait for live tutor.	Zero wait time. Grounded in a single trusted source. Free during Beta.
YouTube (recorded lectures)	Passive consumption. Cannot pause and ask a follow-up.	Active Q&A on the same lecture content. Multi-turn context. Topic search.
WhatsApp batchmates	12–24 hour response delay. Inconsistent answer quality. Breaks study flow.	Answer in <3 seconds. Consistent quality enforced by guardrails.
Photomath / Mathway	Maths and Physics focused. No JEE Chemistry coverage.	Chemistry-only depth. Full conceptual explanation, not just the answer.

Strategic Positioning Statement

ChemIQ is the only product that combines three properties simultaneously:

1. **Instant — answer in <3 seconds, no human in the loop**
2. **Grounded — every word traceable to Himanshu's verified lecture content**
3. **JEE-optimised — format, tone, examples, and tips tuned for exam success**

This combination is only possible through RAG on a structured, teacher-curated knowledge base. No existing product replicates it. Moat deepens with every chapter added.

12. A/B Experimentation Plan

ChemIQ will run structured A/B experiments before any significant change to chunking strategy, embedding model, re-ranking, or system prompt is promoted to 100% of traffic.

Experiment Framework

Element	Definition for ChemIQ
Control	Current production configuration (chunking / model / prompt in use today)
Treatment	Proposed change — one variable changed at a time
Traffic split	80 / 20 during Beta (low user base). Move to 50/50 at GA (≥50 DAU)
Minimum runtime	7 days minimum — enough for weekday + weekend query patterns
Sample size	≥200 queries per arm before reading results (avoid peeking bias)
Primary metric	Faithfulness score (LLM-as-Judge, GPT-4o). Secondary: thumbs-up rate
Decision rule	Treatment wins only if primary metric improves ≥5% AND no regression on secondary
Rollback trigger	If treatment arm faithfulness drops >0.75 in first 48 hrs → auto-revert

Planned Experiments — Prioritised Backlog

ID	Hypothesis	Control	Treatment	Metric	Threshold
EXP-01 Chunking	Box-level chunking improves solved-problem retrieval	Paragraph-level chunks (300 tok avg)	Box-level chunks (colour-coded box = 1 chunk)	Recall@10	≥ +0.05
EXP-02 Embedding	text-emb-3-large improves retrieval precision on chemistry synonyms	text-emb-3-small ($0.02/1M)	text-emb-3-large ($0.13/1M)	Precision@3	>15% improvement
EXP-03 Hybrid Search	BM25 + vector hybrid improves recall on formula queries	Vector-only search	BM25 + vector (RRF fusion)	Recall@10	≥ +0.05
EXP-04 System Prompt	Shorter prompt reduces hallucination	System prompt v1.1 (~500 tok)	Condensed v1.2 (~300 tok)	Faithfulness	≥ +0.03
EXP-05 Query Expansion	Auto-expanding short queries improves retrieval for vague inputs	Raw query passed to Pinecone	GPT-4o-mini expands query before embed	Thumbs-up rate	≥ +8 pp

Rollout Decision Process

1. Step 1 — Define before starting: hypothesis, control, treatment, traffic split, sample size, primary metric, decision threshold, rollback trigger.
2. Step 2 — Run for ≥7 days: collect ≥200 queries per arm. No early peeking.
3. Step 3 — Read results: if primary metric improvement ≥ threshold AND secondary metric stable → promote treatment to 100%.
4. Step 4 — If inconclusive: extend for another 7 days OR increase traffic to 50/50.
5. Step 5 — If treatment loses: discard, document learnings, add failure case to eval dataset (Section 6A).

PM Decision — One Variable at a Time

ChemIQ will never run more than one experiment simultaneously. With a small Beta user base (~20 students), compound experiments cannot be interpreted — if faithfulness improves, we need to know which change caused it. Speed of learning matters more than speed of shipping at this stage.

13. Moat Strategy

The competitive positioning section states “no existing product replicates this.” That is true today, but defensibility must be designed, not assumed. ChemIQ’s moat is built across four compounding layers — each one harder to replicate than the last.

Moat Layer 1 — Data Moat

ChemIQ’s knowledge base is not scraped from the web — it is built from Himanshu Jain’s proprietary lecture notes: colour-coded boxes, solved problems, exam tips, and mnemonics in his exact teaching style. This content is not publicly available and cannot be purchased. A competitor building a similar product would need to start from zero and negotiate access to a teacher’s intellectual property.

Moat deepens as chapters grow: At launch: 1 chapter (Redox Reactions). At GA: all 26 Physical Chemistry chapters. Each chapter added increases retrieval coverage, answer quality, and the uniqueness of the dataset. The vector store at 26 chapters represents thousands of hours of curated teaching that no competitor can replicate without Himanshu’s direct participation.

Flywheel: Every thumbs-down from a student (Section 8A) becomes a labelled data point. Over time, the offline eval dataset (Section 6A) grows into a proprietary benchmark that is impossible for an outside team to replicate without access to ChemIQ’s user base.

Moat Layer 2 — Switching Cost

A student who uses ChemIQ daily for 4 weeks accumulates session history, bookmarked responses, a personalised weak-topic map (Phase 1, Section 9), and a study rhythm built around the tool. Switching to a generic chatbot means losing all of this context and returning to answers that are not aligned to their teacher’s specific style, notation, and exam tips.

Switching Cost Factor	Why It Creates Lock-In
Teacher alignment	Answers match Himanshu’s exact notation, structure, and exam tips. Switching means re-learning a different explanation style mid-exam-prep.
Personalised weak-topic map (v2)	Query history builds a per-student confusion map. Abandoning ChemIQ means starting this map from scratch on any alternative.
Study habit integration	Daily 11 PM study sessions anchored to ChemIQ become a habit that is behaviourally difficult to replace — especially for high-stakes JEE preparation.

Moat Layer 3 — Network Effects

ChemIQ has two network effect mechanisms that grow with user scale.

Data network effect: Every query that generates a thumbs-down is a labelled training signal (Section 8A). More users → more failure signals → faster retrieval improvement → better answers → more users. A competitor launching later with fewer users cannot replicate this improvement velocity.

Social network effect (Beta → GA): JEE students study in tight peer groups. When Riya uses ChemIQ and shares a cited answer in her batch WhatsApp, she becomes a distribution channel. Early Beta students who succeed in JEE become the most credible testimonials for the next cohort. The product spreads through existing social trust networks, not paid acquisition.

Moat Layer 4 — Teacher Ecosystem

ChemIQ v1 is built on a single teacher. The long-term moat is a multi-teacher platform where each teacher brings their own student base. The teacher ecosystem creates a supply-side lock-in that is structurally difficult to replicate.

Teacher lock-in: Once a teacher’s knowledge base is ingested (26 chapters, structured and tagged), the cost of switching to a different platform is extremely high. The teacher would need to re-ingest, re-structure, and re-tag all their content — plus lose access to ChemIQ’s existing student analytics and feedback loop.

Ecosystem flywheel: Teacher A’s students trust Teacher A. When Teacher A is on ChemIQ, their students join ChemIQ. Teacher B sees Teacher A’s students getting results and wants the same tool for their students. Each additional teacher strengthens the platform’s subject coverage, student density, and brand authority in the JEE coaching market.

Moat Summary — Four Compounding Layers

Layer 1 — Data: proprietary teacher content + growing labelled eval dataset. Layer 2 — Switching cost: teacher alignment, habit, and personalised study history. Layer 3 — Network effects: data flywheel from student feedback + social spread through JEE peer groups. Layer 4 — Teacher ecosystem: supply-side lock-in that becomes harder to displace with each additional teacher onboarded. Each layer reinforces the others. A competitor would need to replicate all four simultaneously to match ChemIQ’s defensibility at scale.

14. Monetisation Strategy

Section 7 defines the cost structure. This section defines how ChemIQ recovers those costs and builds a sustainable revenue model. The strategy follows a three-phase progression: Free → Student Freemium → Teacher SaaS, with a B2B coaching integration track running in parallel from Phase 2.

Phase 1 — Free (Alpha + Beta)

No monetisation during Alpha (Wk 1–2) or Beta (Wk 3–6). The goal is product-market fit validation, not revenue. Cost per user is ~$0.48/month (Section 7) — at 20 Beta users, total cost is under $10/month, fully absorbable as product development expenditure. Free access also removes friction for the first cohort of students who are the most critical early validators.

Phase 2 — Student Freemium Model (GA, Month 2+)

Tier	Free	ChemIQ Pro — ₹99/month
Daily query limit	10 queries/day	Unlimited
Practice Mode (MCQs)	3 sets/week	Unlimited + weak-topic targeting
Study Planner (v2)	Not available	Personalised revision planner included
Cost to ChemIQ	~$0.10/user/month (capped queries)	~$0.48/user/month; ₹99 (~$1.20) = ~$0.72 gross margin/user

Freemium conversion hypothesis: A student who uses ChemIQ during the free tier and hits the 10 query/day limit during an active study session is the highest-intent upgrade candidate. The paywall is triggered at the moment of highest need — the same 11 PM scenario that defines the product's core value proposition.

Phase 3 — Teacher SaaS Model (v2)

In v2, ChemIQ becomes a platform that any teacher can use to deploy their own AI study assistant. Himanshu Jain is both the first customer and the proof of concept for this model.
Teacher SaaS pricing: ₹2,999/month per teacher. Includes: private knowledge base namespace, up to 30 chapters, Arize observability dashboard, weekly quality report, and white-label branding ("Himanshu's AI Assistant" instead of ChemIQ). At 10 teachers: ₹29,990/month (~$360/month) in recurring revenue against ~$50/month in infrastructure costs.
Teacher value proposition: The teacher pays once to deploy a 24/7 AI assistant that handles all repetitive doubt-clearing, freeing their time for live sessions and advanced problem-solving. ChemIQ's quality guarantees (faithfulness ≥0.85, source citation on every answer, hallucination rate <5%) are the teacher's protection against reputational risk from wrong AI answers.

Phase 3B — B2B Coaching Institute Integration

The highest-value B2B customer is a coaching institute (Allen, Resonance, FIITJEE, Aakash) with 500–5,000 enrolled students and multiple teachers per subject. ChemIQ becomes an institute-wide AI doubt-clearing layer, integrated into their existing LMS.

B2B Package	Detail
Institute Starter	₹9,999/month. Up to 3 teachers, 500 students, 2 subjects. White-label branding. Monthly quality report.
Institute Scale	₹29,999/month. Unlimited teachers + students. API integration with institute LMS. Dedicated Arize dashboard. SLA: 99.9% uptime.
Revenue at 1 Scale customer	₹29,999/month (~$360) covers all infrastructure costs for 100+ individual users. Single B2B deal is worth 250+ individual Pro subscriptions.

PM Decision — Revenue Sequencing
Phase 1 (free) builds trust and validates product-market fit. Phase 2 (freemium) proves willingness to pay at the individual level. Phase 3 (Teacher SaaS + B2B) is the actual business. The sequencing is intentional: do not attempt B2B sales until at least one teacher has used ChemIQ themselves and can

testify to its quality. Himanshu Jain’s own use of the product is the most credible sales asset for the first B2B conversation.