Vrin

Memory & Context OS for AI agents & LLMs

Please reach out at:

530-204-8133 ● vedant@vrin.cloud

Presented by:

Vedant Patel

VRIN

Introduction

RAG was a great first step—now it's the **bottleneck**. Most teams wired an LLM to a vector store and hoped the model would do the heavy lifting: **multi-hop reasoning, cross-document synthesis, time awareness**. In practice, vector-only RAG **misses links, bloats context, and can't explain itself**. DIY RAG routinely costs companies **>$30k/month** and months of development to reach "good enough"

Vrin is the next step: the **Memory & Context OS for AI**. We replace ad-hoc pipelines with an insert-data-and-ask service:

- HybridRAG (graph + vector + rank fusion) that routes per query and proves every claim with citations and timestamps (**No Hallucinations** – a must for enterprise)
- Typed, auditable memory (provenance, TTL, conflict resolution) so answers are explainable by design.
- User-defined expertise—you set the domain expert once, we apply it everywhere.
- Production speed and governance out-of-the-box (SSO/SCIM, CBOM exports, VPC/on-prem).

Enterprise GenAI adoption is surging, but governance/reliability lag—VRIN fills that gap; See higher answer quality at lower cost
In short—**stop building brittle RAG plumbing; run on Vrin and ship reliable, evidence-backed answers**. (LLM tokens are cheap now; the value is routing, memory, and explainability.)

Problem and Solutions

Analyzing current market gap and exploring opportunities

Problem	Solutions
Vanilla RAG made Storage easy and Reasoning HARD!	VRIN's adaptive HybridRAG Router: per-request plan across graph, vector, lexical, cache, and long-context.
"Vector-DB + RAG" alone struggles with multi-doc synthesis and temporal reasoning; long contexts add cost/latency and still miss mid-context facts.	Typed Memory Engine: schematized, versioned memories with TTL, conflict resolution, provenance.
Enterprise adoption is exploding (GenAI used in 71% of companies by 2025), but governance and reliability lag.	Explainability & Governance: multi-hop evidence chains, citations, confidence; exports aligned to emerging standards
Teams glue prompts, RAG, and caches —brittle, expensive, non-auditable. Agents need durable memory, adaptive retrieval, and explainability	Cloud-based systems support business expansion effortlessly.

VRIN

Market Opportunity

Enterprises (TAM*)

$4.8 B

Small

Adoption – 40%
Workflows – 4
Attach – 50%

$25.2 B

Medium

Adoption – 60%
Workflows – 9
Attach – 70%

$72 B

Large

Adoption - 75%
Workflows – 12
Attach – 85%

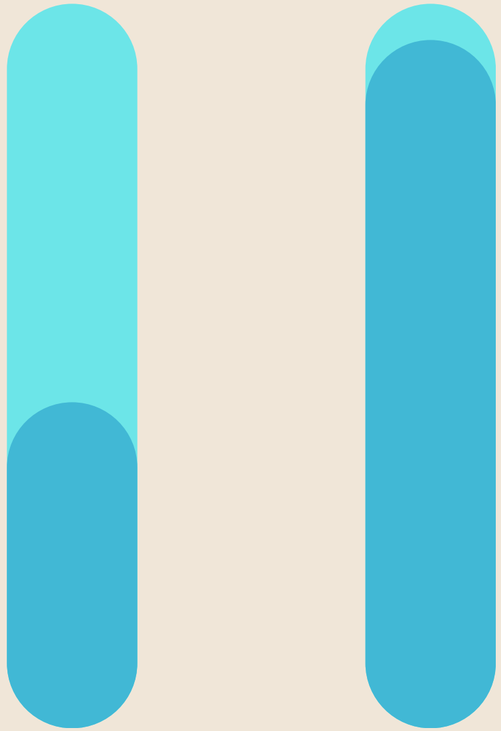
*TAM = (Enterprises × Adoption × Attach) × (Workflows × ACV/workflow)

VRIN

Market Gaps and True Opportunistic value

Current – Vanilla RAG

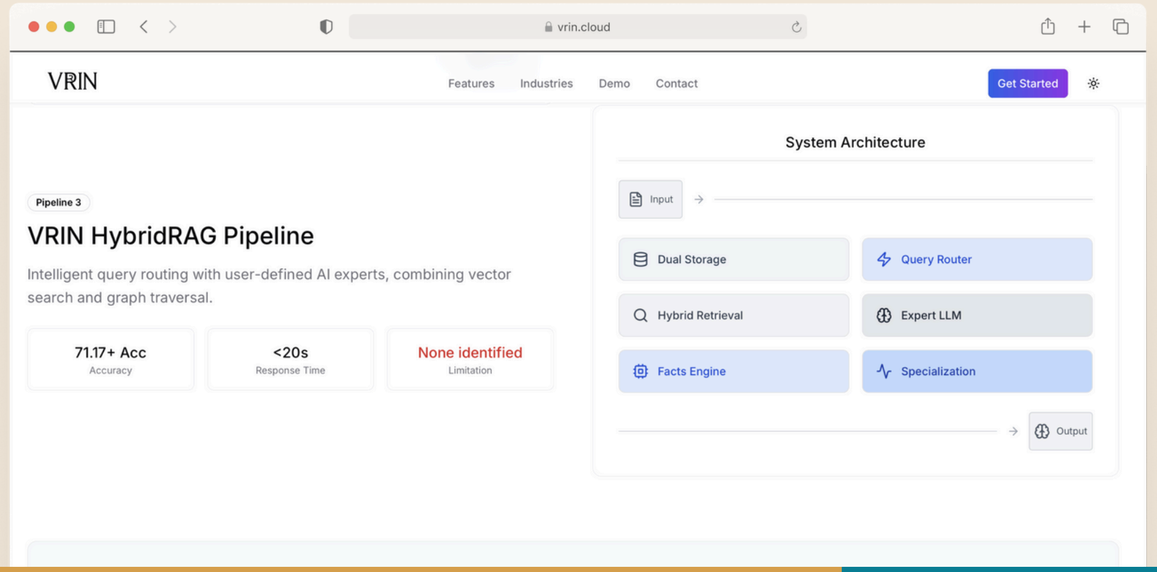
1. Build & babysit pipelines
2. Vector-only ≠ multi-hop synthesis
3. Long Context ≠ Reasoning
4. Total Cost $30k–$47k / month



Vrin's Complete State

1. Pre-built knowledge fabric
2. Insert records and destress
3. Get trustworthy answers by default
4. Total Cost $2.5k–$10k / month

What Vrin Actually Does



1

Specialize

Create your domain expert (policy, tone, schema

2

Insert Knowledge

Vrin extracts facts, dedupes, stores triples + vectors.

3

Query

Router selects best path; returns evidence-linked answer and a context bill of materials (CBOM)

4

Update Memory

Structured, explainable updates with lineage and TTL (Time to live).

VRIN

Technical Moat

Beyond “just RAG”: routing, memory, explainability

Differentiation

Primary functionalities include:

01. Adaptive Routing

Decides when to retrieve vs. expand context vs. cache)

02. Knowledge Graph

- Entities/relations
- Temporal context
- Multi-hop
- Vector/keyword fusion

03. User Specialization

Enterprises access expert advisory services, optimization strategies, and tailored solutions to enhance operational efficiency.

04. Explainability & Governance

- CBOM
- Per-claim evidence
- Policy masks
- Ready for audits

Current Offering

1 SDK & REST API

Create your domain expert (policy, tone, schema

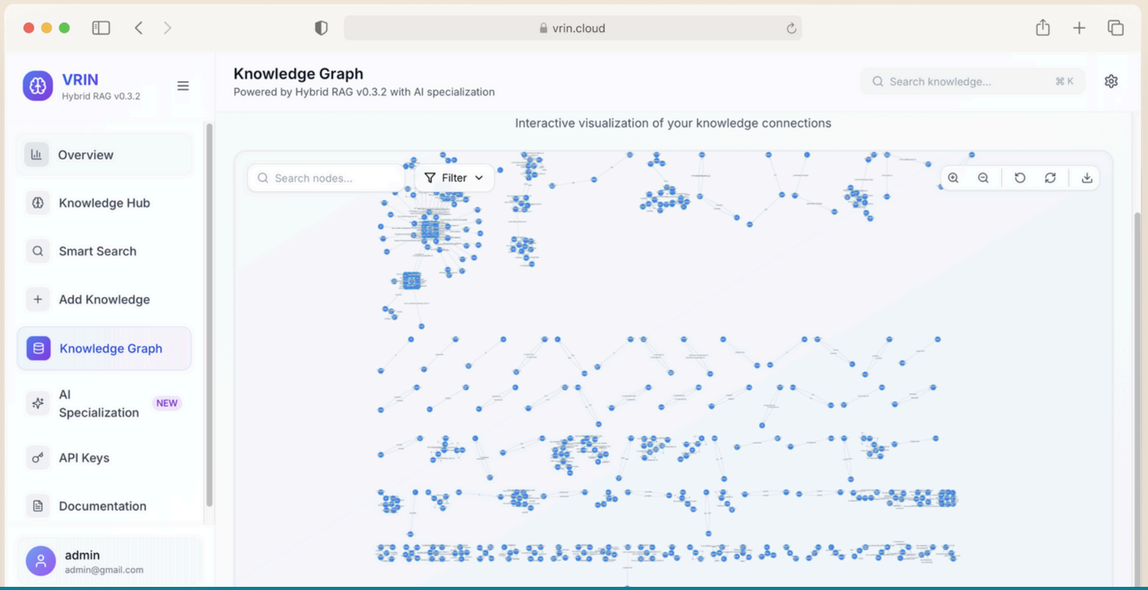
2 User Dashboard

Vrin extracts facts, dedupes, stores triples + vectors.

Basic SDK Usage

Copy

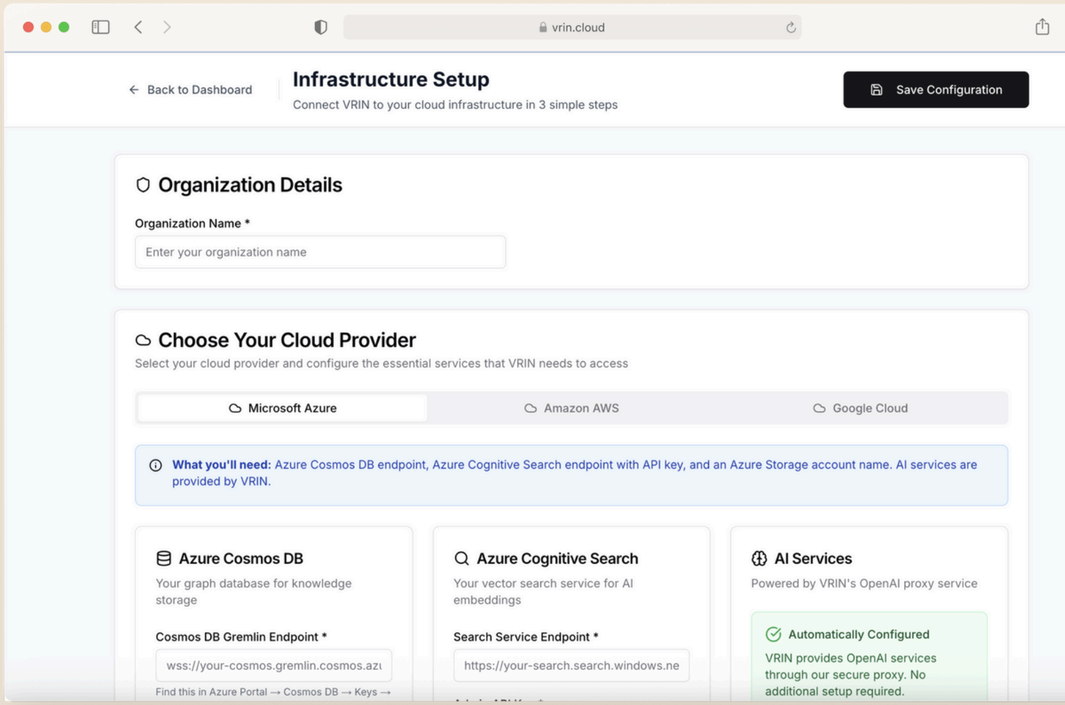
```
from vrin import VRINClient
import os

# Initialize client
client = VRINClient(api_key=os.getenv('VRIN_API_KEY'))

# Insert knowledge with smart deduplication
result = client.insert(
    content="The iPhone 15 Pro has a titanium body and USB-C port.",
    title="iPhone 15 Pro Specifications",
    tags=["apple", "iphone", "specifications"]
)

print(f"✅ {result['facts_extracted']} facts extracted")
print(f"📦 Chunk stored: {result['chunk_stored']}")
print(f"💾 {result['storage_details']}")
```



Flexible Hybrid Cloud for Enterprise

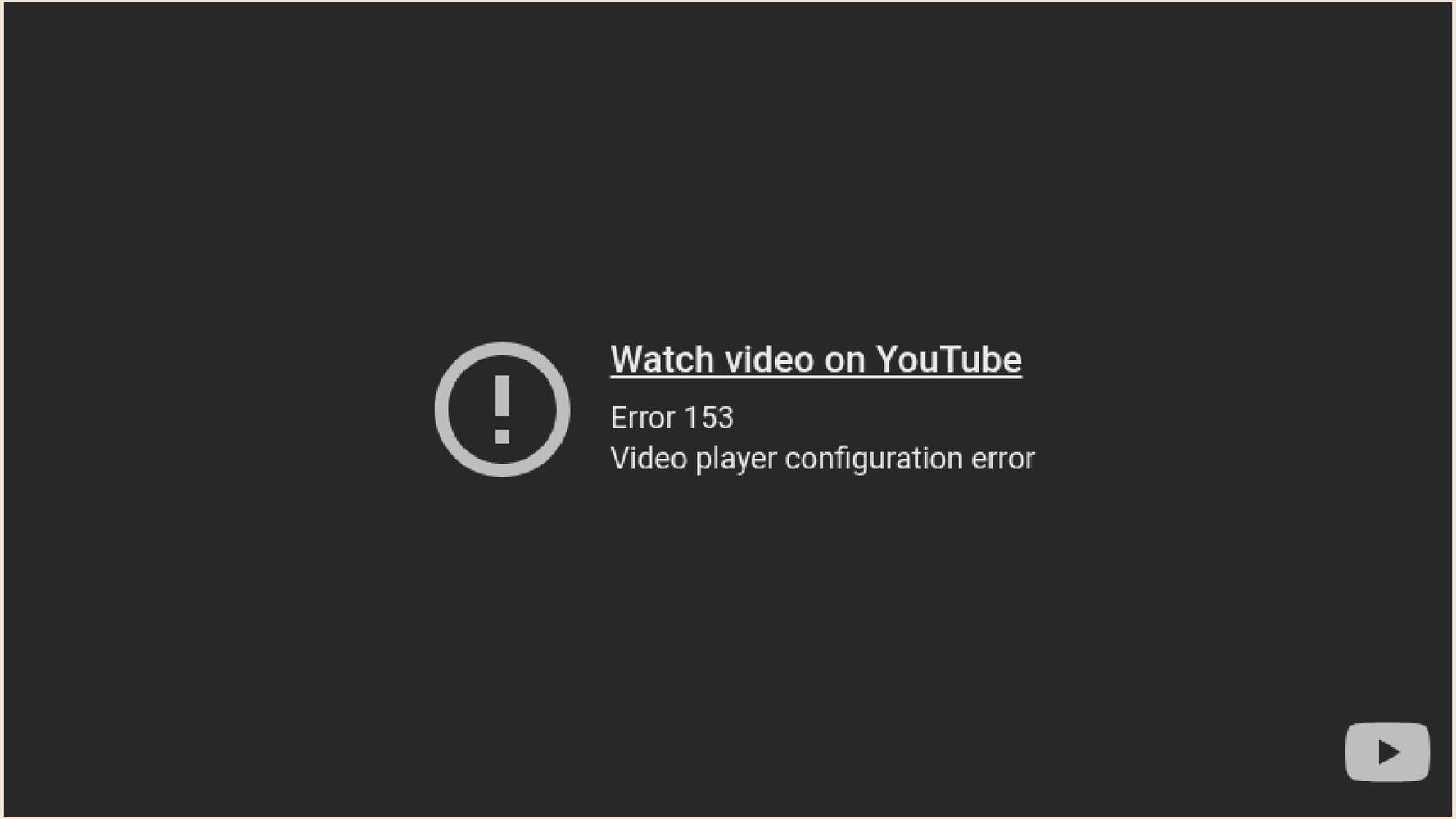
Deployment Models

Model	Data Location	Networking
Air-gapped	On-prem only	No external egress
VPC-isolated	Customer cloud	PrivateLink/PrivateEndpoint
Hybrid Explicit	Per-request (private/public)	Client-controlled routing



VRIN

Product Demo



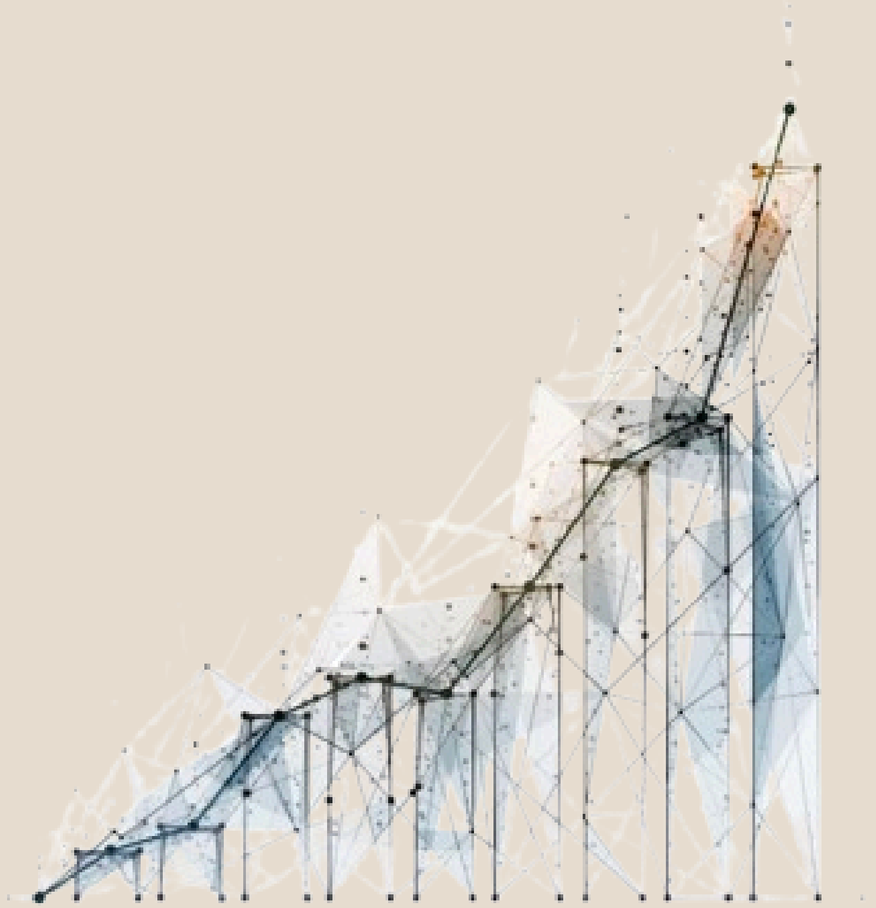
Watch at: https://www.youtube.com/embed/yzKWQXbtrrA

Traction

Landed a pilot customer
Health LLM – Research Group at UC Davis

Next Goal
Get paid customers

Market Expansion
Enter five new industry domain markets successfully.



VRIN

Competition

Category	They're great at	Gaps for enterprises	Vrin's advantage
Vector DBs / Enterprise search (e.g., Pinecone, Elastic)	Indexing, scaling search	Not a **reasoning/governance** layer; weak multi-hop provenance	**Adaptive routing** across graph/search/long-context; **per-claim citations**; **CBOM**
RAG frameworks (LlamaIndex, LangChain)	Dev libraries, fast protos	Not a **production control plane**; no audit artifacts	**Runtime + SDK + console**; typed memory; policy masks; **SaaS + VPC** deploys
Model-vendor memory/agents	Convenience, tight model coupling	**Vendor lock-in**; limited cross-source provenance; privacy limits	**Vendor-neutral**; customer-hosted indexes; **BYOK**; **evidence-backed**
KG platforms (Neo4j, Stardog)	Graph modeling & queries	Not end-to-end context + retrieval; little rank-fusion	**HybridRAG**: **graph + vector + long-context + cache** with rank-fusion
DIY pipelines	Tailored control	**Time-to-value**, ongoing **headcount**, audit gaps	Drop-in **control plane**; faster, cheaper, auditable; **days not quarters**

VRIN

Revenue Model

Features	Builder (Free)	Team ($2.5k–$3.5k/mo)	Business ($8.5k–$12k/mo)	Enterprise ($60k–$250k+/yr + usage)
Data scope	100k chunks / 100k edges	2M chunks / 3M edges	10M chunks / 15M edges	Custom (100M+ chunks; 150M+ edges)
Queries/mo	5k	100k	500k	Custom (SLA'd)
HybridRAG	Shared	Dedicated indices	Dedicated + VPC peering	Private/VPC or on-prem
Memory & CBOM	Basic	Full CBOM & TTL	Full + exports	Full + auditor packs
Security	API keys	Basic RBAC	SSO/SAML + SCIM	SSO/SAML, SCIM, data residency
Connectors	CSV/S3	+ Postgres/Drive	+ Slack/Jira/Confluence	All + custom
Support	Community	Email (48h SLA)	Priority (8–12h SLA)	Dedicated TAM & DSE
Add-ons	—	Extra storage/queries	Compliance exports, private LLM	On-prem, managed upgrades

VRIN

Go-To-Market

1

Ideal Customer

- CIO/CDO
- Head of Platform/ML
- AI/ML/Search teams
- VP Engineering

2

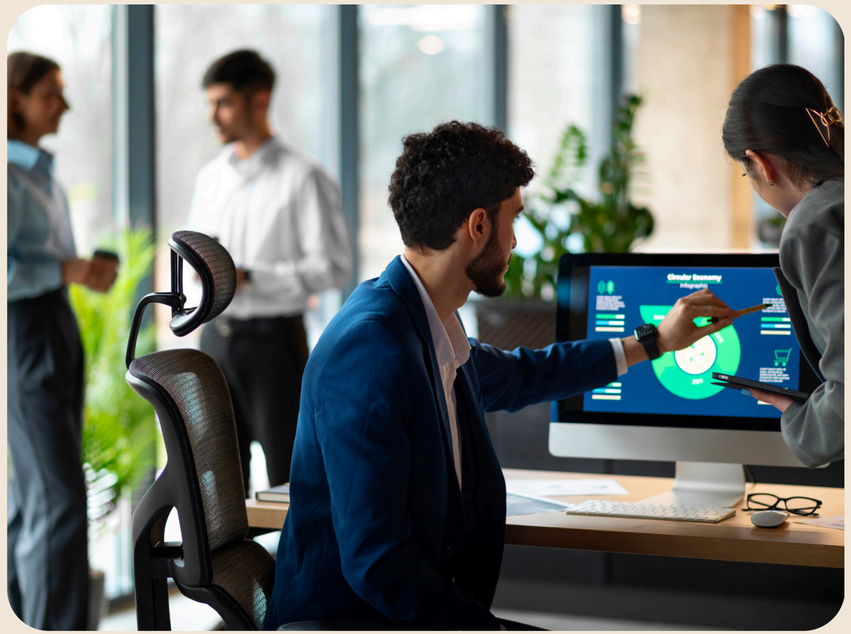
Offer (No Brainer Pilot → Expansion)

Target 1 workflow → Measure KPIs →
Expand to multiple workflows

3

How we Reach them

Founder-led outbound to teams with
AI workflow



Team

VEDANT PATEL

Founder & CEO

- 2+ years of research at UC Davis on RAG, graph reasoning, temporal memory, and agentic workflows
- Solo founder who shipped a production system end-to-end and secured the first real pilot

Looking Ahead

0-3 months

Eval & Observability console; CBOM v1; SSO/MFA/RBAC; 3–5 pilots; ≥100k routed req/day

4-6 months

Self-optimizing router v1; Graph+Vector 2.0; connectors pack; first on-prem/VPC; 10+ paying tenants.

7-12 months

SOC2 I→II; CBOM v2 (claim-level grounding UX); Expert Specialization Library; multi-region scale; $1M+ ARR run-rate target

We welcome your questions and feedback

Please reach out at:

530-204-8133 ● vedant@vrin.cloud