PRD — Content Concierge AI

PoC 1: API-First Agentic Architecture (Hybrid RAG Reserved)

---

1. Purpose of This PRD

This document defines the **AI system and AI-backend technical blueprint** for PoC 1 of the Content Concierge. It intentionally focuses **only** on AI responsibilities, orchestration, and retrieval logic. All non-AI backend concerns (authentication, session handling, UI rendering, core infra) are explicitly out of scope.

PoC 1 prioritizes **speed, safety, and clarity** by relying on structured APIs and LLM reasoning, while reserving architectural hooks for future Hybrid RAG (Traditional RAG + GraphRAG).

---

2. AI System Goals

In Scope

- Generate 2–3 personalized, text-only investment insights per user session
- Use portfolio, goal, and activity context to drive relevance
- Retrieve market data and financial context from approved external APIs
- Enforce citation-first output for all non-internal claims
- Maintain explainability and auditability at every step

Out of Scope

- Advice, recommendations, or action nudges
- Media generation (video/audio)
- UI logic
- User authentication and authorization

---

3. High-Level Architecture

Pattern: Agentic orchestration using **LangGraph**

The system is modeled as a deterministic graph of AI nodes, each with a single responsibility. State is explicitly passed between nodes to ensure traceability.

---

4. LangGraph Agent Design

4.1 Agent State (Canonical)

```
{
  "user_context": {"portfolio": {}, "goals": {}, "activity": {}},
  "market_context": {},
  "candidate_insights": [],
  "final_insights": [],
  "citations": []
}
```

4.2 Graph Nodes

1. **Context Ingestion Node**
2. Inputs: UserContext API (portfolio, goals, activity)
3. Output: Normalized user_context
4. Logic: Validation + summarization only (no inference)
5. **Insight Hypothesis Node (LLM)**
6. Task: Propose 3–5 *candidate insight themes*
7. Example: asset concentration, goal drift, market trend relevance
8. Output: candidate_insights (themes only, no claims)
9. **Market Data Retrieval Node (Tools)**
10. Tools: Alpha Vantage / equivalent financial APIs
11. Task: Fetch factual data aligned to candidate themes
12. Output: market_context with source metadata
13. **Insight Synthesis Node (LLM)**
14. Task: Convert facts + user context into textual insights
15. Constraints:
    - Educational tone
    - No prescriptive language
    - No numeric fabrication
16. **Citation Assembly Node**

17. Task: Attach citations to each insight
18. Rule: Every external claim must map to ≥1 source
19. **Ranking & Pruning Node**
20. Task: Select top 2–3 insights based on relevance heuristics

---

5. Retrieval Strategy (PoC 1)

5.1 Data Sources

Internal (Assumed APIs): - Portfolio composition - Goal progress - Activity summary

External: - Market data APIs (prices, flows, indicators) - Financial news APIs (headline + metadata only)

No document ingestion or embedding stores are used in PoC 1.

---

6. LLM Responsibilities & Guardrails

Allowed

- Theme generation
- Contextual explanation
- Natural language synthesis

Disallowed

- Investment advice
- Future predictions
- Source fabrication

LLM outputs must be reproducible given the same inputs.

---

7. Observability (AI-Specific)

- Prompt versions
- Tool calls & parameters
- Raw API responses
- Final insight + citations

All logged per session for auditability.

---

8. Reserved Extension Points (Not Implemented)

Traditional RAG Hook

- Placeholder retrieval node accepting documents
- Interface defined, implementation skipped

GraphRAG Hook (TigerGraph)

- Entity resolution node (assets, sectors, themes)
- Graph traversal node (disabled in PoC 1)

---

9. Success Criteria (AI-Only)

- Factual accuracy of insights
- Zero uncited external claims
- Deterministic graph execution
- Latency within acceptable SLA