Product Requirements Document – Content-to-Insights Pipeline

1. Product Overview

Objective

Automate the capture of insights from long-form tech talks (e.g., YouTube videos) into a structured, searchable hub. Reduce research drag from hours of manual note-taking to a few minutes, while making knowledge reusable across teams.

Primary Users

- Analysts and consultants who track emerging tech topics.
- Students and professionals doing rapid knowledge capture.
- Teams who want to share "bite-sized" learnings without rewatching hours of content.

Problem Statement

Research from long talks is slow, repetitive, and inconsistent. Each new 60–90 min video costs ~4 hours to watch, summarize, and file. Insights are scattered across personal docs, making them hard to retrieve or share later.

2. Goals & Non-Goals

In Scope (MVP)

- Fetch video metadata and captions via YouTube API.
- Generate 3-line summary + 3–5 key learnings using LLM.
- Auto-tag videos by topic for future search.
- Push outputs into a Notion database for central access.
- Send daily digest of new insights via Slack/email.

Out of Scope (for now)

- Semantic video search across full transcripts.
- Multi-language caption translation.
- Collaboration features (comments, rating).

Why: The MVP focuses on speed of capture and searchability rather than deep media analysis.

3. User Stories

- *As an analyst, I want quick summaries of tech talks, so I can decide which ones are worth deeper study.*
- *As a researcher, I want key learnings tagged by topic, so I can retrieve relevant insights later without rewatching.*
- *As a team member, I want a daily digest of new insights, so I can stay updated with minimal effort.*

4. Solution Overview

Data Flow

- n8n workflow (7 nodes):

1. YouTube API fetches latest video IDs from subscribed channels/playlists.
2. Captions pulled and cleaned.
3. LLM (GPT-4o) generates summary, learnings, and tags.
4. Deduplication logic skips already-processed videos.
5. User preference toggle → choose summary vs learnings vs both.
6. Results pushed to Notion DB (searchable hub).
7. Slack/email digest sends daily updates.

Future Interactivity

- Chatbot interface → "What are 3 AI trends from last month's videos?"
- Web front-end (Streamlit/Glide) where users can input a video link and get structured insights.

5. Success Metrics

North-Star: Reduction in research time per video (from ~4 hrs → <10 min).

Drivers:

- **of videos processed per week.**
- % correctly tagged by topic.
- **of unique insights stored and retrievable in DB.**

Counters:

- % duplicate videos processed (target <5%).
- % hallucinated or low-quality learnings flagged by users.
- Digest open rate <50% (signals poor usability).

6. Discovery Insights

Conversations with early users/peers:

- *"I skip half the talks because I can't spare 4 hours to watch them."*
- *"Notes are scattered across Notion, Drive, and Slack — I can't find past insights when I need them."*
- *"Sometimes I just want the key takeaways, not a full summary."*

These shaped the need for deduplication, topic tags, and user-choice of output format.

7. Trade-Offs & Risks

Trade-Offs

- Chose Notion DB for simplicity and free tier, rather than building a custom search backend.
- Limited to English captions initially to keep processing accurate.
- Decided to prioritize deduplication logic before chatbot search to prevent noise.

Risks

- LLM outputs may misrepresent or hallucinate takeaways.
- Dependence on YouTube caption quality.
- Notion DB may limit scale if video volume grows.

8. Roadmap

- **v1 (MVP – Done):** Fetch captions, generate summaries/insights, push to Notion, send daily digest.
- **v2:** Add deduplication, toggle for summary vs learnings, improve formatting.
- **v3:** Build chatbot interface for query-based exploration of stored insights.
- **v4 (Stretch):** Multi-language support, semantic transcript search, team collaboration features.

9. Appendix

- Sample Notion DB schema.
- LLM prompt template (summary + insights + tags).
- Example daily digest output.