Project Genesis & Design Rationale: RAG-Optimized Content Ingestion & Analysis Pipeline

Your Name

July 3, 2025

Executive Summary

This document outlines the design rationale for a two-tier AI content processing pipeline, integrated into the "Second Brain" knowledge management system. The pipeline ingests multi-modal content (YouTube videos/playlists, articles, PDFs, Microsoft Edge Spaces, OneNote pages/notebooks), producing structured, RAG-ready JSON summaries for a tri-database system (Supabase, Pinecone, Postgres with pgvector). The design prioritizes cost-efficiency, factual accuracy, and query flexibility, achieved through a multi-model workflow with automated quality assurance.

1 Design Evolution

1.1 Initial Goal

- **Objective**: Summarize diverse content into structured, metadata-rich formats for a personal RAG library.
- **Challenge**: Early prompts lacked database compatibility and scalability.

1.2 Two-Tier Model Strategy

- **Solution**: Tier 1 (Gemini 2.5 Flash/Lite) for cost-effective summarization; Tier 2 (Gemini 2.5 Pro) for validation and analysis.
- **Rationale**: Balances cost and capability, leveraging lighter models for initial processing and powerful models for deep insights.

1.3 Factual Accuracy

- **Problem**: Un-grounded models caused factual hallucinations (e.g., incorrect creator names).
- **Solution**: Mandated Grounding with Google Search for Tier 1, ensuring accurate metadata extraction.

1.4 Procedural Reliability

- **Problem**: Tier 1 models struggled with abstract sampling (e.g., 10–20% of text).
- **Solution**: Simplified commands (e.g., “Extract first 4, middle 4, last 4 sentences”) for reliable execution.

1.5 Automated QA Feedback Loop

- **Solution**: Tier 2 validates Tier 1 summaries, producing a validation$_{r}eportwithconfidencescores$
$Ensuresqualitywithoutmanualreview, enablingscalableprocessing.$

1.6 Final Pipeline

- **Architecture**: Tier 1 produces Markdown summary; Tier 2 validates, analyzes, and outputs JSON for Supabase/Pinecone/pgvector.
- **Rationale**: Single JSON output avoids additional API calls, optimizing cost and speed.

2 Finalized Architecture

- **Prompt #1 (Tier 1)**: Gemini Flash/Lite with Grounding, outputs Markdown summary.
- **Prompt #2 (Tier 2)**: Gemini Pro, validates summary, generates insights, outputs JSON.
- **Tri-Database**: Supabase (metadata), Pinecone (vectors), Postgres with pgvector (hybrid queries).