### *Project Report: An n8n-Powered Multi-Agent "NotebookLM" System*

### 1. Introduction

The goal of this project was to create a sophisticated, multi-agent AI system on the n8n automation platform, designed to replicate the core functionalities of services like NotebookLM. By integrating a suite of specialized APIs and a vector database, we built a framework capable of ingesting, processing, storing, and creatively acting upon unstructured data from various sources, including text, documents, and web content. The system delivers relevant, context-aware responses and outputs, such as summaries, mindmaps, and podcasts.

### 2. Framework Overview

The system is architected as a modular, hub-and-spoke model within n8n, comprising a central orchestrator and multiple specialized agentic workflows:

* **Main Orchestrator Flow:** (Telegram Trigger -> Input Preparation -> AI Agent with multiple tools)
* **Specialized Agent Flows:**
  + Preprocessing Agent: Cleans and extracts text from any source.
  + Summarization Agent: Condenses text into summaries.
  + Mindmap/Podcast Agents: Generate creative outputs.
  + Storage Agent: Ingests and embeds knowledge into a vector database.
  + Retrieval Agent: Answers questions using the stored knowledge (RAG).

This design allows for both the structured processing of new information and direct, knowledge-based responses to user queries.

### 3. Purpose of the Report

This report analyzes the custom-built multi-agent framework, evaluates the performance of the integrated APIs within the n8n environment, and assesses the system's overall effectiveness in managing complex, multi-step AI tasks.

### 4. Features Comparison (Key Integrated APIs)

* **OpenRouter (for LLM Reasoning):**
  + Cloud-hosted API: Yes
  + Dockerized support: No (API only)
  + GUI/visualization: Yes (via n8n)
  + Summarized answers: Yes (model dependent)
  + Free trial: Pay-as-you-go with initial free credits
  + Emoji usage: Yes
* **Cohere (for Embeddings):**
  + Cloud-hosted API: Yes
  + Dockerized support: No (API only)
  + GUI/visualization: Yes (via n8n)
  + Summarized answers: Yes (has generation models, but used for embeddings here)
  + Free trial: Generous free tier
  + Emoji usage: Yes

### 5. Commercial License Details

* **OpenRouter:** Developer-friendly, pay-as-you-go model. Commercial use is allowed.
* **Cohere:** Generous free tier for development. Usage-based pricing for commercial applications.
* **ElevenLabs:** Free tier with character limits, subject to strict network checks. Paid plans are required for commercial use and to bypass restrictions.

### 6. Memory Management

* **Built-in Conversational Memory:** The main AI Agent can leverage n8n's Simple Memory node for short-term, session-based memory.
* **Session Storage:** User sessions are maintained via the chat\_id from the initial Telegram Trigger, ensuring responses are sent to the correct user.
* **Long-Term User Memory:** The system's long-term memory is powered by a **Supabase** vector database. The Storage Agent writes to this memory, and the Retrieval Agent reads from it, enabling persistent knowledge across sessions.

### 7. Web Search Tools

The system's "web search" capability is specialized for content extraction rather than general-purpose searching.

* **Tools Library:** The Preprocessing Agent acts as the primary web interaction tool.
* **Configuration:** It is configured to handle website URLs and YouTube links, extracting the main body of text or the video transcript.
* **Performance:** Performance is high for well-structured articles and transcripts but can be limited by the reliability of external transcript services or the complexity of target websites.

### 8. Documentation and Code

* **Versioning:** n8n workflows are JSON-based, allowing them to be version-controlled using tools like Git.
* **Documentation:** The primary method of documentation is through descriptive node naming (e.g., Prepare Agent Input, Generate SVG Mindmap) and detailed text in the Description fields of the agent's tools.

### 9. Toolkits and APIs

* **Prebuilt Toolkits:** The system leverages pre-built tool functionality within the AI Agent node, which can intelligently select and chain sub-workflows (e.g., calling Summarizer Agent after Preprocessing Agent).
* **API Support:** The entire system is an example of advanced API orchestration, connecting Telegram, OpenRouter, Cohere, Supabase, and various media generation services in a seamless flow.
* **Observability:** Compatibility with observability platforms is primarily through n8n's built-in execution logs, which provide detailed input/output data for each node in every run, simplifying debugging.

### 10. Workflow Analysis

n8n excels at supporting complex workflow design. The system's hub-and-spoke architecture allows for a clear separation of concerns, making each agent modular and independently testable. The visual editor makes it easy to trace the flow of data and conditional logic, which is essential for debugging multi-step agentic processes.

### 11. Reasoning and Thinking Tools

The core reasoning engine is the AI Agent, powered by a model hosted on **OpenRouter**. Its reasoning is guided by a detailed **System Prompt** that instructs it to differentiate between two main tasks: ingesting new information or retrieving existing knowledge. The agent demonstrates dynamic tool selection by choosing to call the Mindmap Agent or Podcast Agent based on keywords in the user's prompt.

### 12. Conclusion

n8n has proven to be a robust and highly capable platform for building and orchestrating a complex, multi-agent AI system. **For ingesting and processing new information, the Preprocessing -> Summarization -> Storage pipeline is highly effective.** **For answering questions, the RAG-based Retrieval Agent provides accurate, context-aware responses.** The system successfully integrates memory, reasoning, and a diverse set of tools into a unified, functional workflow.

### 13. Recommendations for Further Exploration

* **Integrate a General Web Search Tool:** Add a tool like SerpAPI to allow the agent to answer questions about topics not contained in its long-term memory.
* **Enhance Error Handling:** Build automated error-handling branches within the workflows to notify the user if an external API (like a text-to-speech service) fails.
* **Upgrade to Paid API Tiers:** For production use, upgrading services like ElevenLabs and hcti.io would remove free-tier limitations and increase reliability.
* **Implement Dynamic Routing:** Create more sophisticated logic in the Main Workflow to dynamically choose between different LLMs or tools based on the perceived complexity or category of the user's request.