Whitepaper: Agentic Architecture in Neuron7 Search

Amit Verma, Head of Engineering/AI
Neuron7.ai

Abstract: Neuron7's search tool is a cutting-edge, AI-powered solution that extends the concept of **Retrieval-Augmented Generation (RAG)** to provide **context-aware search** and **real-time adaptability**. By combining RAG with **entity disambiguation**, **LLMs**, **metadata enrichment**, **vision models**, and **user feedback loops**, Neuron7's search platform not only retrieves and generates content but also autonomously refines, filters, and adapts search results to deliver actionable insights. This whitepaper explores how Neuron7's search solution takes RAG to the next level by embodying **agentic architecture**, providing a system that perceives, makes decisions, acts, and adapts based on real-time input and evolving user needs.

1. Introduction

The landscape of enterprise search has evolved beyond traditional keyword-based search tools, which simply retrieve documents based on text matching. Today's advanced search solutions leverage **artificial intelligence (AI)** to provide not just information, but **contextual understanding** and **actionable insights**. Neuron7 Search takes this to the next level by combining **Retrieval-Augmented Generation (RAG)** with a wide array of advanced features such as **Named Entity Recognition (NER)**, **metadata enrichment**, **vision models**, and **real-time learning**. This system exemplifies **agentic architecture**, making decisions autonomously, taking actions based on those decisions, and continuously learning from feedback to enhance search accuracy and relevance.

Neuron7 Search does not merely retrieve data and generate content; it autonomously refines search results, filters data, and is designed to meet the needs of enterprises where information needs to be not just found, but also **understood**, **contextualized**, and **acted upon**.

2. What is Agentic Architecture?

Agentic Architecture refers to systems that are capable of **autonomous decision-making**, **adaptive behavior**, and **taking actions** based on environmental inputs and internal goals. These systems **perceive** inputs from the environment (such as data or user queries), **make decisions** based on that input, and **take action** to achieve specific objectives, all without requiring human intervention at every decision point.

Key characteristics of agentic systems include:

- **Autonomy**: The system can make decisions and take actions independently.
- **Goal-Oriented Behavior**: The system operates to achieve specific goals.
- **Perception**: The system perceives input from the environment (data or user queries).
- **Decision-Making**: The system makes decisions based on perceived data.
- **Action**: The system performs actions based on its decisions.
- **Adaptability**: The system learns and adapts to new inputs over time.

3. Neuron7 Search: Extending RAG with Agentic Architecture

3.1 RAG-Based Retrieval and Generation

At the core of Neuron7 Search is **RAG**—a process where:

- **Retrieval**: The system first searches the repository for documents or data that match the query.
- **Generation**: An LLM is then used to generate a coherent response or contextual answer based on the retrieved data.

While **RAG** provides a solid foundation for combining retrieval with generative capabilities, Neuron7 Search enhances this process by adding additional layers of **context-awareness**, **entity identification**, **vision models**, and **metadata enrichment** to ensure that the results are not only relevant but also highly actionable.

3.2 Autonomy in Data Extraction Using Vision Models

Neuron7's search tool autonomously extracts structured and unstructured data from **PDFs**, **documents**, and other formats with the added power of **vision models**. These models enable the system to **better understand** the underlying document content, including images, diagrams, and tables, which significantly improves data extraction quality.

- **Document Understanding**: Vision models process visual elements (images, tables, charts) within documents to extract context and meaning, allowing for better comprehension of complex or mixed-format documents.
- **Data Extraction Enhancement**: By understanding visual content in addition to textual information, the system improves its ability to **extract key data.**

This integration of vision models into data extraction makes the search tool more **efficient**, **accurate**, and **context-aware**, offering a cleaner input for further processing, such as filtering and generating responses based on that data.

3.3 Dynamic Named Entity Recognition (NER)

Unlike traditional NER systems, which typically rely on static checklists, dictionaries, and rules, Neuron7's advanced entity identification is **dynamic** and adapts as more data is processed.

- **Dynamic NER**: The system continuously improves over time, identifying and categorizing new types of entities based on user interactions and evolving data.
- **Real-Time Entity Recognition**: It allows the system to recognize **error codes**, **model numbers**, **part numbers**, **product names**, and other relevant entities as it processes data, adapting in real-time.

By incorporating **dynamic, adaptive NER**, Neuron7's search tool is capable of enhanced accuracy and **context-awareness**, refining the search results based on newly identified entities and patterns.

3.4 Query Parsing and Intent Understanding

Neuron7's search system also incorporates a **query parser** that understands the query grammar and **intent** behind user queries. This step enhances the system's ability to correctly interpret and respond to search requests.

- **LLM Cleanup**: The query parser works in conjunction with **Large Language Models (LLMs)** to clean and refine the user's query. LLMs help the system understand the underlying **intent** of the query by fixing grammatical issues or ambiguity, ensuring that the query is properly structured for retrieval.
- **Intent-Based Query Enhancement**: The system identifies any relevant filters or parameters that need to be applied. These can be based on recognized entities from the query (via NER) or automatically inferred from the LLM's analysis, resulting in a **richer, more precise query**.

This dynamic query parsing ensures that the system not only understands the search query but also optimizes it for better results before retrieval.

3.5 Decision-Making Based on AI and LLMs

Neuron7's search system incorporates **Large Language Models (LLMs)** to make sense of the extracted and identified data. These models enhance the search tool's ability to understand the context of both the query and the underlying data, allowing it to provide more accurate and relevant results.

- **Metadata Generation**: By leveraging LLMs, the search tool enriches the data with additional metadata, such as summarizing content, identifying key relationships, and adding insights that provide a clearer understanding of the search results.
- **Contextual Decision-Making**: The search tool decides which data is most relevant by interpreting the meaning behind the query and matching it with appropriate content in the repository.

This autonomous decision-making process ensures that the search tool provides more precise results, going beyond keyword matching to truly understanding the intent behind the search.

3.6 Action Through Search Refinement and Filtering

Once the tool processes the query and filters it based on NER and LLM-based enhancements, it takes action by refining the search criteria and filtering the repository based on the identified entities and enriched metadata. The search is therefore contextual and dynamic, adjusting based on real-time inputs and providing results that are tailored to the user's intent.

For example, if the user queries an **error code**, the system will focus on documents and resources directly related to that code, automatically prioritizing relevant results over irrelevant data. This action-oriented behavior is a core aspect of agentic systems.

3.7 Learning and Adaptability Through User Feedback

One of the defining features of agentic systems is their ability to adapt and learn from interactions with their environment. In Neuron7's search system, this learning is achieved through **user feedback**:

- **Relevance Feedback**: The system continuously improves based on user input, such as ranking preferences, and relevance ratings. This helps the system understand which results are most useful and adjust future search results accordingly.
- **Automated NER Refinement**: As the system processes more queries, it learns to improve its NER accuracy, better identifying critical entities and relationships within the data.

This feedback loop ensures that the system becomes increasingly efficient and accurate over time, adapting its behavior to provide better search results based on evolving user needs.

4. Conclusion: Neuron7 Search as an Agentic System

Neuron7 Search extends the foundational concept of Retrieval-Augmented Generation (RAG) by integrating autonomous decision-making, real-time adaptability, and actionable search results. By incorporating advanced NER, dynamic metadata enrichment, vision models for document understanding, query parsing, and continuous learning, Neuron7 Search embodies the principles of agentic architecture—perceiving, deciding, and acting autonomously to guide users through complex tasks and decision-making processes.

- **Autonomy**: The system autonomously processes queries, identifies entities, enriches data, and generates actionable results.
- **Context-Aware Decision-Making**: Using LLMs and NER, the system makes decisions that align with the user's intent, ensuring highly relevant search results.
- **Action-Oriented**: Neuron7 Search takes actionable steps, guiding users through their tasks and providing the next steps based on real-time data.
- **Learning and Adaptation**: Neuron7 Search continuously improves through user feedback, ensuring it adapts to evolving search needs and improves over time.

Neuron7 Search is more than just a search engine; it is an **agentic system** that **learns, adapts, and takes action**, transforming how enterprises search, discover, and act on information.