Chapter 12 - Using APIs with Artificial Intelligence

The Overlap of AI and APIs

- Important Data Sources = API, Databases, Files
- How to inference a model? REST API
- Most AIs are cloud hosted and made available through APIs.
- Calling APIs directly from Generative AI application, like:
    - Retrieval Augmented Generation (RAG): calls APIs and other data sources and feeds the retrieved information to the LLM $\xrightarrow{\text{why?}}$ Help to solve the Knowledge Gap problem in LLMs
- Generative AI application uses LLMs to determine what API endpoints to use: Agentic Application
    - The LLMs make decision by interpreting definitions from OAS files, Python code, or API documentation

Designing APIs to Use with Generative AI and LLMs

- Is your API endpoint appropriate to use with an agentic generative AI application without additional safeguard in place?
    - The Providers of LLMs provide warnings such as:
        - Anthropic Claude 3: LLM should not be used on their own in high-risk situations
        - Google NotebookLM: LLMs may sometimes provide inaccurate information
        - ChatGPT: Check important info
            - Solutions:
                - Requiring human to approve tasks recommended by LLMs before executing
                - Reviewing logs of the functioning of the system
                - Combining multiple AI agents to review tasks before executing
                - Reviewing and filtering inputs and outputs to the models
                - Foundational practices of API management and security are required when LLMs use APIs
                    - Restrict the permissions provided to systems that include LLMs
- Limit the size of the data results:
    - Important for Cost and Accuracy:

        - Cost perspective: Model providers charge for processing tokens (chunk of text) → more data a model processes = greater the cost
        - Accuracy perspective: ChatGPT struggles to perform calculations from a very large datasets returned by APIs → limit size of the data = improves accuracy
            - Rather than returning all fields, return critical fields
            - add parameters, filters, pagination to narrow down the specific records in API call
- Make data structures consistent throughout the API:
    - More predictable API = More accurate AI
    - Re-using schemas inside API + defining them inside OAS file + use Pydantic (enforce schemas + publish them into OAS file)
- Provide a Software Development Kit (SDK):
    - Better endpoints + Customized API calls + Documentation
- Customize your OpenAPI Specification (OAS):
    - Endpoints in OAS file should have unique and clear operation IDs.
    - Customize OAS file with AI appropriate endpoints + detailed descriptions of each endpoint & its parameter that assist LLM in inferring their meanings
- Provide a separate endpoint for summary statistics:
    - Remove the guesswork of LLMs out
- Provide a search endpoint that doesn't rely on a record identifier:
    - LLMs are more comfortable with language than numbers

Arazzo to Define Multistep Processes

- Because AI agents are nondeterministic → difficult to ensure they use API correctly $\xrightarrow{\text{especially}}$ when you have multiple API calls $\xrightarrow{\text{Solutions}}$ Add information to the descriptions of each API endpoint in the OAS file or tool function → Description Job = explain to AI model how one endpoint related to others $\xrightarrow{\text{Problem}}$ It is still up to AI model to use them correctly
- Other Solution = Arazzo Specification: sequences of calls and their dependencies $\xrightarrow{\text{so}}$ you have a set of related calls $\xrightarrow{\text{which means}}$ You have a deterministic building block → add reliability to API usage

Defining Artificial Intelligence

> **Cole Stryker & Eda Kavlakoglu**

> Artificial Intelligence is technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity and autonomy..

- AI is a computer program that can have humanlike conversations and complete humanlike tasks
  - AI include Expert System: Complicated rule-based systems that can perform human like tasks
  - Modern AI = Machine Learning, LLMs, Generative AI and etc.
    - Generative AI: have the ability to Generate text, music, and videos base on text prompts, like:
      - ChatGPT
      - Copilot
      - Gemini
    - Problems of Generative AI: warnings about bias, hallucinations, mistakes, harmful content

Creating Agentic AI Applications

- AI agents are the forefront of AI research and development
- Agent is software that controls application flow using an LLM, the more autonomously the LLM controls the system, the more agentic the system is.
- Tools to Create Agent or Orchestrate multiple agent:
  - Autogen: Python, dotnet
  - CrewAI: Python
  - LangChain/LangGraph: Python
  - LlamaIndex: Python, Typescript
  - PydanticAI: Python
  - Vercel AI SDK: Typescript

Project Part III

- Chapter 13: Deploying a Machine Learning API
- Chapter 14: Using APIs with LangChain
- Chapter 15: Using ChatGPT to call your API

Additional Resources:

- AI and APIs — What 12 Experts Think The Future Holds
- Syntax Sunday: Custom API Wrapper for GPTs
- What is the Model Context Protocol (MCP)? - Model Context Protocol
- OpenTelemetry
- Is Your API AI-ready? Our Guidelines and Best Practices - Guide | Blobr Copy
- Arazzo-Specification
- Nordic APIs