Lab 1 – Product Description

John Hicks

Old Dominion University

CS411W

Dr. Sumaya Sanober

Feb 28, 2025

Version 5

Table of Contents

List of Figures

List of Tables

No table of figures entries found.

1 Introduction

Recent advances in Large Language Model (LLM) technology allow more generalized content generation based on foundation models. As such, LLM focused tooling has exploded onto the market (Uspenskyi, 2024). Despite these advances, the application of LLMs to the problem of turning natural language into Representational State Transfer (REST) Application programming interface (API) payloads has not seen a mature implementation yet.

Because of the lack of standardized frameworks for this use case, software developers are forced to learn LLM technology and build one-off solutions to generating API payloads from natural language. This requires extra cost in time and staff. Furthermore, progress on making applications use AI features is stopped by the risks involved in trusting LLMs' decision-making capabilities (Nexus, 2024; Tyen, 2024); any risk-aware solution to the REST API payload generation problem must include the opportunity for humans or business rules to validate the payload before it is sent.

The new software solution that fills these gaps is called CueCode.

2 Product Description

CueCode will offer a complete framework for application developers to integrate with a service for intelligently translating natural language to REST API payloads. CueCode has first-class support for humans and business rules in the loop of payload generation, as compared with other approaches to the problem. This will allow developers to begin using AI in a risk-aware manner while the technology is still improving, giving them a head start on preparing their applications and service offerings for the years to come.

2.1 Key Product Features and Capabilities

CueCode will translate natural language in text format into a series of correctly ordered REST API payloads, which a client application can then issue to the target API after further processing and/or human review of the API call suggestions.

CueCode will be a Web application that, with its supporting off-the-shelf services, offers a Developer Portal experience for configuring the application and another service for performing natural language to REST API translation suggestions. Developers can integrate their applications with CueCode by using the CueCode client library for their programming language of choice.

Getting an LLM to produce Web API payloads and validating them is a specialized task requiring skills that many Web and fullstack developers do not possess (Uspenskyi, 2024).

Since these developers are those most often building business applications, a solution for turning natural language into REST API payloads should take into consideration how easy it will be for developers with other skillsets to use the tool. CueCode gives a good foundation for Web and fullstack developers to turn natural language into REST API payloads in their applications.

CueCode will be a full framework and service to turn natural language into Web API payloads; nothing like it exists for arbitrary REST APIs. CueCode will work with any REST API defined with an OpenAPI specification (*OpenAPI Specification - Version 3.1.0 | Swagger*, n.d.).

Developers will upload their OpenAPI specification to CueCode's Developer portal, then answer questions about the API definition and structure as needed.

At runtime, the developer's application can make a request to the CueCode service (itself running over an HTTP Web API). The CueCode service will reply with the generated REST API

payload(s) corresponding to the natural language text input given. CueCode will provide client libraries to make integration with the CueCode service seamless.

2.2 Major Components (Hardware/Software)

The CueCode solution will consist of a backend Python application, Ollama service, PostgreSQL (Postgres) database with the pgvector extension, and a third-party identity provider. The Python application will require use of the SpaCy library, which allows natural language structuring and named entity recognition (*SpaCy · Industrial-Strength Natural Language Processing in Python*, n.d.).

Ollama is an application that allows communication with LLMs over a standardized HTTP API.

CueCode will require hardware capable of running the following systems separately:

- Ollama 3.1 (minimum) running the 70 billion parameter model (minimum)
    - The CS Systems group already runs Llama models (personal communication).
- A Python application using SpaCy, running on either CPU or GPU (*Install SpaCy · SpaCy Usage Documentation*, n.d.).

3 Identification of Case Study

The CueCode project will verify CueCode's effectiveness by performing a study of API payload generation against the commonly used Pet Store OpenAPI server specification (OpenAPI contributors, n.d.) in comparison to a standalone LLM server generating payloads for the same server. Evaluated in the study will be:

1. The ratio of generated payloads to expected generated payloads by exact match
2. The difference between the data objects generated by each system, with scoring weights assigned to the kinds of errors present (numerical, spelling, grammar)

3. The error rate for data dependencies between created objects
4. Invalid JSON generation rate

The study will then compare the results from the two kinds of systems. The standalone LLM will use a variety of prompting techniques:

1. Zero-shot prompt (*Zero-Shot Prompting | Prompt Engineering Guide*, n.d.)
2. Few-shot prompt (*Few-Shot Prompting – Nextra*, 2025)
3. Updating temperature parameter (*What Is LLM Temperature?*, 2024)

If this validation step shows the CueCode system does not perform well, Team RED will adjust the CueCode system's configuration and reevaluate, as part of the prototyping process.

In addition to the quantitative study above, the two fictional user personas describe the people who would benefit from CueCode's development. The system will be designed with the following users in mind:

- Steve, a fullstack developer, needs to integrate text-to-API-payload features but finds he needs to roll his own solution and understand NLP and LLM technology.
- Case study of Patricia, who needs to make an appointment at a hospital whose booking system already uses REST APIs.

Two general use-cases are supported by CueCode (Figure 1 below):

1. Human review of suggested API calls
2. Batch processing of textual data, issuing API calls without human review.

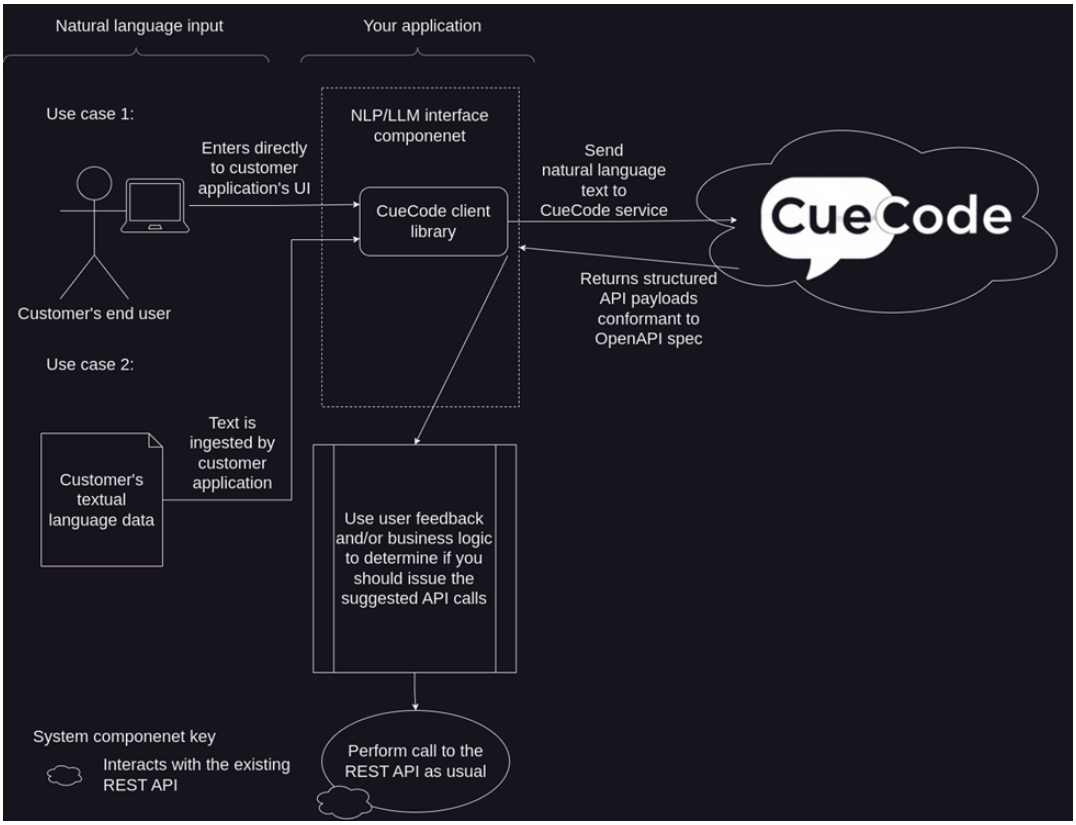

Figure 1 Two use-cases showing CueCode's API payload generation concept.

Use case 1, human review of suggested API payloads, gives end-users the opportunity to verify that CueCode's interpretation of the natural language is correct. While in most cases the user would not view raw API payloads, this use case represents the opportunity for CueCode's client applications to format suggested API responses in a way that allows users to be "in the loop" of payload generation.

Use case 2 represents when human review of the generated payloads is not required, but there are still deterministic rules that need to be applied by traditional software after the payload generation process.

Both of the use cases reduce the risk of using LLMs to generate API payloads.

4 Product Prototype Description

The CueCode prototype will limit the scope of features to those required to turn natural language to REST API calls. Other potentially beneficial features will be excluded to focus on delivering a working prototype of the core CueCode functionality.

4.1 Prototype Architecture (Hardware/Software)

The prototype will focus on REST APIs defined with OpenAPI specifications, using JSON content types only.

The application will use 12-factor application development practices, to allow for containerization and other modern application development and deployment practices (Adam Wiggins, 2017). By following 12-factor application development practices, the CueCode prototype will be horizontally scalable by design (Adam Wiggins, 2017). To make deployment and development easier during prototyping, the Python backend will group all functionality in one Python application, separated by modules.

The LLM, database, and authentication services will run as separate applications within the CueCode system.

Customers will connect their applications to CueCode via the client libraries supplied by CueCode, using authentication credentials they obtain from the Developer Portal Web application that developers can use to configure CueCode for use with their OpenAPI-defined REST APIs.

4.2 Prototype Features and Capabilities

CueCode's feature set will limit the kinds of API endpoints against which CueCode can be configured. For example, GraphQL is another API format that CueCode would eventually support if it were a real product.

Further research may also require hypermedia be included in the target API's responses or that the OpenAPI specifications used have certain properties to define relationships between target API's entities, to ease implementation for CS411.

Depending on the team's work capacity, the prototype might or might not include non-core features that could aid in the hypothetical commercialization of CueCode, such as an idea to create a marketplace where commonly used Web APIs (e.g., Google Drive) can have their CueCode configuration shared among CueCode users, making a Web 2.0 content sharing dynamic (Sean Baker, personal communication Oct 2024).

As discussed below, CueCode will provide authentication only by API key, rather than accounting for other use cases such as the integration of frontend applications.

4.3 Prototype Development Challenges

Prototyping CueCode has presented "known unknowns" and "unknown unknowns".

It can be known before prototyping that CueCode algorithm is highly dependent on the quality of results from the LLM.

Because LLMs are non-deterministic (Nexus, 2024; Uspenskyi, 2024), it may prove difficult to instruct an LLM for a task so specific as selecting an API endpoint. Thurs, we face an open question of whether CueCode should allow the LLM choose which REST API endpoint to generate data, or for CueCode to implement a cosine similarity search algorithm, similar to work done by Zafin's engineers and others (Mark Needham, 2023; Zafin, 2023).

However the LLM is prompted, it will return a text response that needs to be validated to confirm the API payload conforms to the OpenAPI specification. A combination of prompt template and LLM Function calling can be used to ensure that the generated API payload is

compliant with the OpenAPI specification provided enforcement (Mark Needham, 2023; *Microsoft/Prompt-Engine*, 2022/2024; *Stanfordnlp/Dspy*, 2023/2024).

Determining the relationships between entities will challenging. Doing so involves not just parsing natural language into a structured grammatical parse tree using Spacy (*Linguistic Features · SpaCy Usage Documentation*, n.d.), but it also requires mapping that structure to the API's entity relationship structure, as best CueCode can determine the API's structure from the API spec.

Another challenge will be developing a sorting algorithm for ensuring that the order in which API calls are made doesn't invalidate them because of unmet data dependencies. This algorithm will require using placeholders for data not knowable until after an API call is made. For example, this would happen when creating an entity, A, and several other entities related to it. The algorithm would need to ensure that the API request to create A would be issued prior to any requests creating entities that are related to A.

Apart known open questions, the team has also learned from encountering unexpected questions and challenges. The team keeps architectural decision records (*Architectural Decision Records (ADRs)*, n.d.) to document when important changes are made to the design that affect the system's architecture.

First of these changes was to move from using a command line interface (CLI) for uploading OpenAPI specifications to using a Flask web page. This, the team decided, would allow them to focus their efforts on Web development, without having to learn another tool for CLI development.

Second, the team decided not to implement short-lived JSON Web Tokens (JWT) to enable secure client-side integrations with CueCode; instead, the CueCode prototype will support only API key authentication.

Third, the team's initial OpenAPI example choice, NextCloud, proved not to have an OpenAPI specification that was thoroughly documented enough to serve as a ready-made example for input to CueCode (Nextcloud contributors, n.d.). After considering alternatives, the team decided to use the Pet Store example API (OpenAPI contributors, n.d.), which is thoroughly documented.

Fourth, a few updates to the team's original database schema were needed to align the CueCode data model's central themes those in the OpenAPI specification format.

As seen by a few of our most challenging development questions, CueCode will remove much complexity from the fullstack developer's code, enabling natural language interaction with REST APIs in a reusable, operationalized framework. CueCode will help developers leverage the generation capabilities of LLMs, while also controlling the risks thereof. This allows developers to humanize APIs, without the headache.

5 Glossary

API Payload (informal): Information that is sent together with an API request or response. This data, which can be organized in JSON or XML forms, usually includes the details needed by the client to comprehend the answer or by the server to carry out an action.

CueCode Developer Portal: A web-based platform that allows easy API creation with NLP-generated requests and gives developers access to CueCode's tools, API configuration, and integration workflow management.

HTTP Header: Additional metadata, such as the content type, authentication information, or caching instructions, are transmitted with HTTP requests and answers. Headers give context, which improves communication.

HTTP (Hypertext Transfer Protocol): The protocol that specifies the format and transmission of messages between web clients and servers. The type of request is determined by the HTTP methods (GET, POST, etc.).

Hypermedia – inter-linked content on the Internet. In the context of REST APIs, hypermedia allows REST APIs to be more or less RESTful, as defined by Roy Fielding and following authors (*What Is Hypermedia?*, n.d.).

Representational State Transfer (REST): A set of design guidelines for networked apps that use stateless, cacheable, and consistent HTTP processes to facilitate interaction. Through the use of common HTTP techniques, REST allows clients to communicate with servers by modifying resources that match an expected structure.

URL (Uniform Resource Locator): A web address that indicates where a resource is located on the internet. Protocol (such as HTTP/HTTPS), domain, and resource path are all included in URLs. They are necessary in order to access and consult internet resources.

6 References

Adam Wiggins. (2017). *The Twelve-Factor App*. https://12factor.net/

Architectural Decision Records (ADRs). (n.d.). Architectural Decision Records. Retrieved February 17, 2025, from https://adr.github.io/

Few-Shot Prompting – Nextra. (2025, January 7). https://www.promptingguide.ai/techniques/fewshot

Install spaCy · spaCy Usage Documentation. (n.d.). Install SpaCy. Retrieved November 8, 2024, from https://spacy.io/usage

Linguistic Features · spaCy Usage Documentation. (n.d.). Linguistic Features. Retrieved November 8, 2024, from https://spacy.io/usage/linguistic-features

Mark Needham. (2023, July 26). *Returning consistent/valid JSON with OpenAI/GPT*. https://www.youtube.com/watch?v=lJJkBaO15Po

Microsoft/prompt-engine. (2024). [TypeScript]. Microsoft. https://github.com/microsoft/prompt-engine (2022)

Nextcloud contributors. (n.d.). *OCS API*. Retrieved February 1, 2025, from https://docs.nextcloud.com/server/28/developer_manual/_static/openapi.html#/

Nexus, P. (2024, July 16). *Large language models make human-like reasoning mistakes, researchers find*. Tech Xplore. https://techxplore.com/news/2024-07-large-language-human.html

OpenAPI contributors. (n.d.). *Petstore OpenAPI specification*. Retrieved February 17, 2025, from https://petstore3.swagger.io/

OpenAPI Specification—Version 3.1.0 | Swagger. (n.d.). Retrieved September 10, 2024, from https://swagger.io/specification/

SpaCy · Industrial-strength Natural Language Processing in Python. (n.d.). Retrieved September 26, 2024, from https://spacy.io/

Stanfordnlp/dspy. (2024). [Python]. Stanford NLP. https://github.com/stanfordnlp/dspy (2023)

Tyen, G. (2024, January 11). *Can large language models identify and correct their mistakes?* Google Research. http://research.google/blog/can-large-language-models-identify-and-correct-their-mistakes/

Uspenskyi, S. (2024, September 19). *Large Language Model Statistics And Numbers (2024)—Springs*. https://springsapps.com/knowledge/large-language-model-statistics-and-numbers-2024, https://springsapps.ai/blog/large-language-model-statistics-and-numbers-2024

What Is Hypermedia? (n.d.). Smartbear.Com. Retrieved November 8, 2024, from https://smartbear.com/learn/api-design/what-is-hypermedia/

What is LLM Temperature? | IBM. (2024, December 17). https://www.ibm.com/think/topics/llm-temperature

Zafin, E. (2023, August 15). Bridging the Gap: Exploring use of Natural Language to interact with Complex Systems. *Engineering at Zafin*. https://medium.com/engineering-zafin/bridging-the-gap-exploring-using-natural-language-to-interact-with-complex-systems-11c1b056cc19

Zero-Shot Prompting | Prompt Engineering Guide. (n.d.). Retrieved February 17, 2025, from https://www.promptingguide.ai/techniques/zeroshot