Dialectal Intent and Entity Extraction System (DIEES)

A comprehensive solution to dynamically scale intent and entities from user input across various dialects

Objectives

The main objectives of the project

1. Improve Accessibility

Enable seamless interaction for users speaking different dialects, thereby broadening the user base.

2. Increase Accuracy

Provide more accurate and contextually relevant responses, leading to better user satisfaction and engagement.

3. Enhance Customizability

Tailor the system to the client specific linguistic needs, making it versatile and adaptable to diverse applications.

Challenges

The main challenges in the project

1. The application is dynamic and context-specific due to user data.
2. Insufficient domain-specific data.
3. Frequent data drift over time.
4. Intent and entity data imbalance.

Data Description

Data Scope

General statistics about the data

21

Intents

Total number of predefined intents available by data annotation team

59

Entities

Total number of general entities associated with each intent extracted by data annotation team

14K

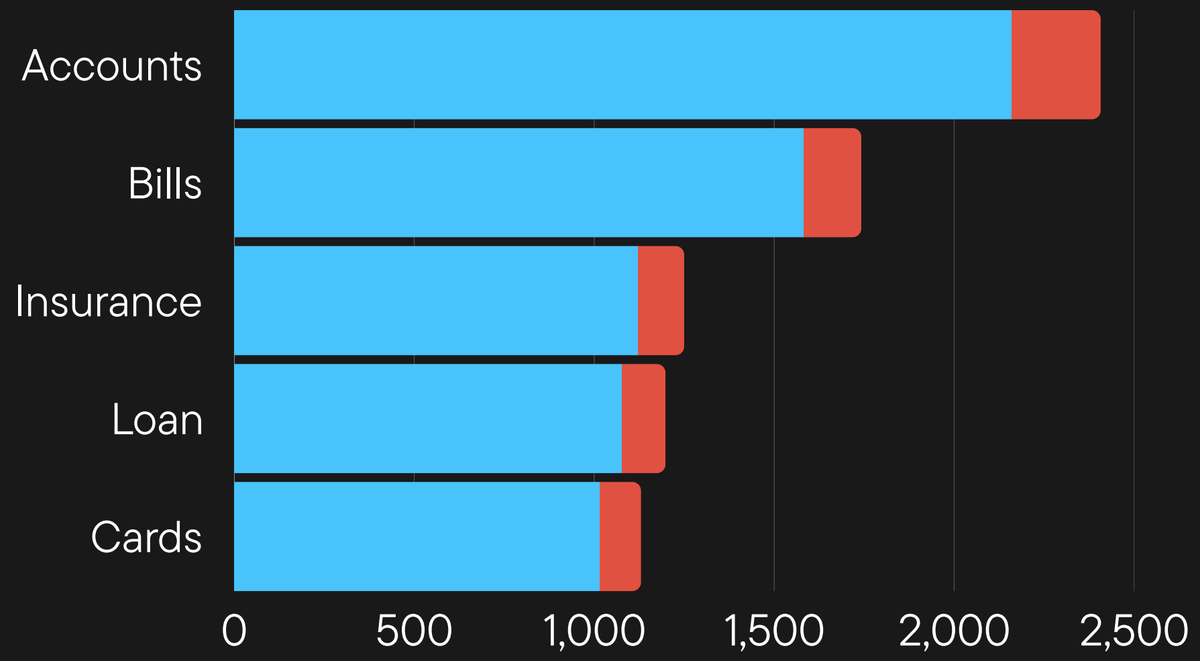
Variations

Total number of unique variations across all intents

Data Preparation

Data Scope

Intents in the first dynamically segregated intent groups



5

Number of Intents

Total number of predefined intents in the the first group of intents from the annotated data

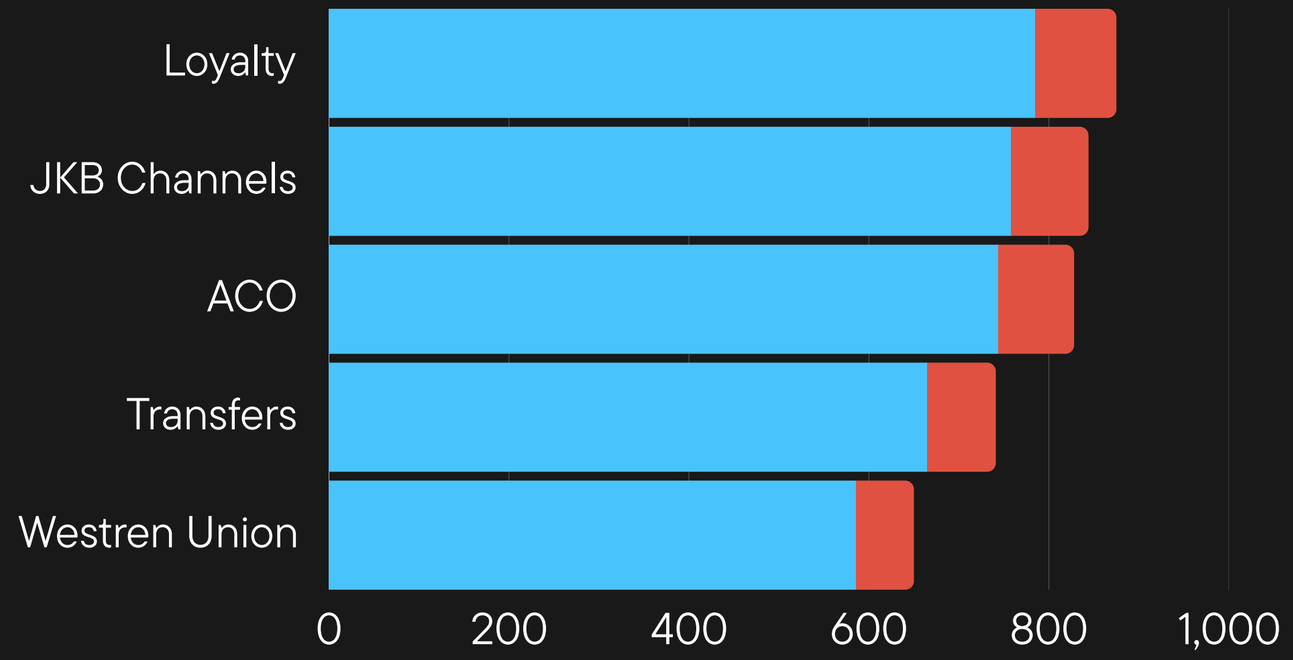
7K+

Number of Variations

Total number of variations in the annonated data

Data Scope

Intents in the second dynamically segregated intent groups



5

Number of Intents

Total number of predefined intents in the the first group of intents from the annotated data

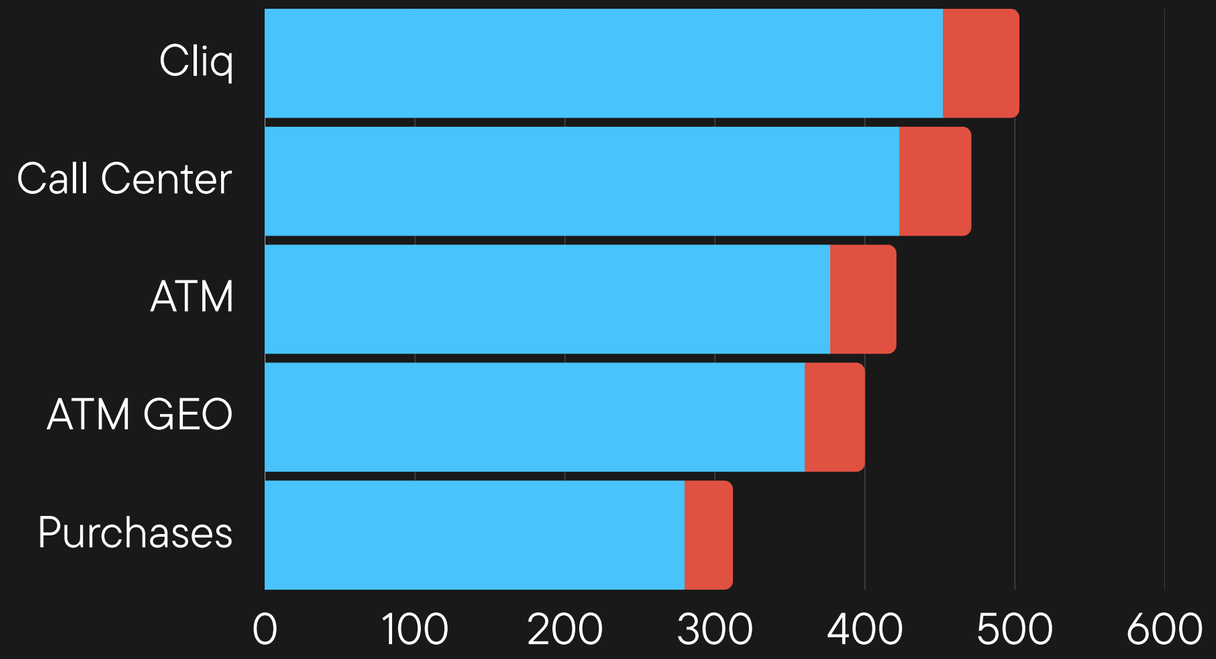
3.5K+

Number of Variations

Total number of variations in the annonated data

Data Scope

Intents in the third dynamically segregated intent groups



5

Number of Intents

Total number of predefined intents in the the first group of intents from the annotated data

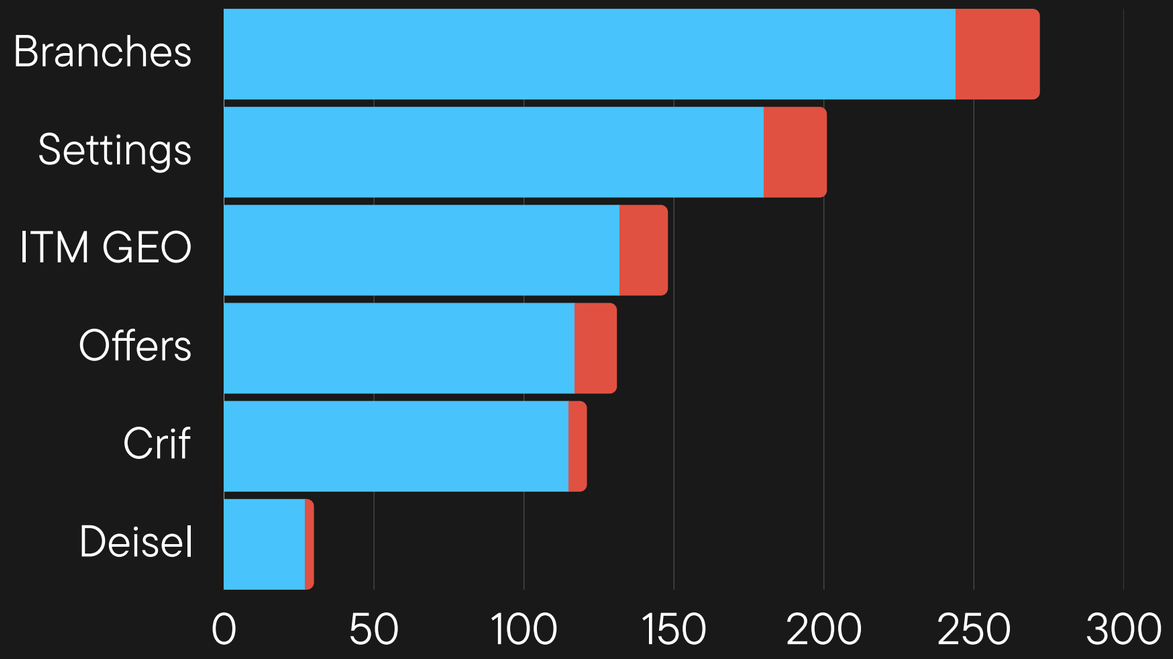
2K+

Number of Variations

Total number of variations in the annonated data

Data Scope

Intents in the fourth dynamically segregated intent groups



6

Number of Intents

Total number of predefined intents in the the first group of intents from the annotated data

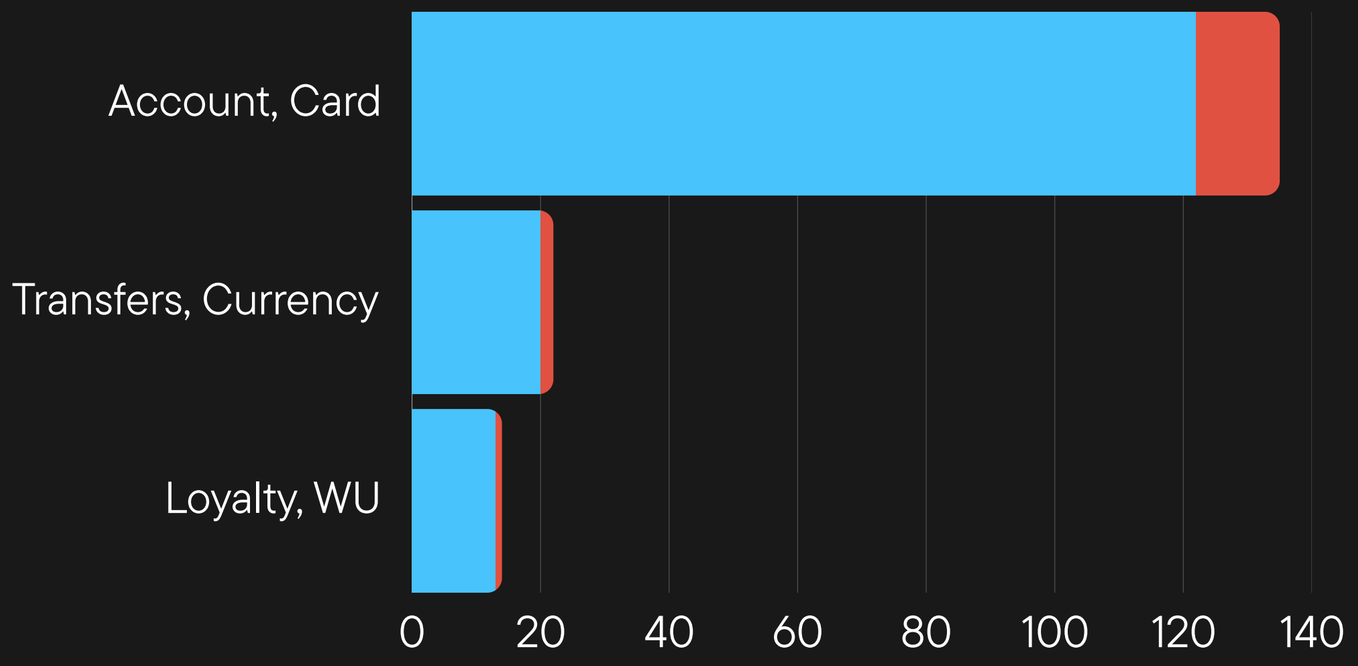
800+

Number of Variations

Total number of variations in the annonated data

Data Scope

Conflict intents combinations



3

Conflict Groups

Total number of predefined intents in the the first group of intents from the annotated data

150+

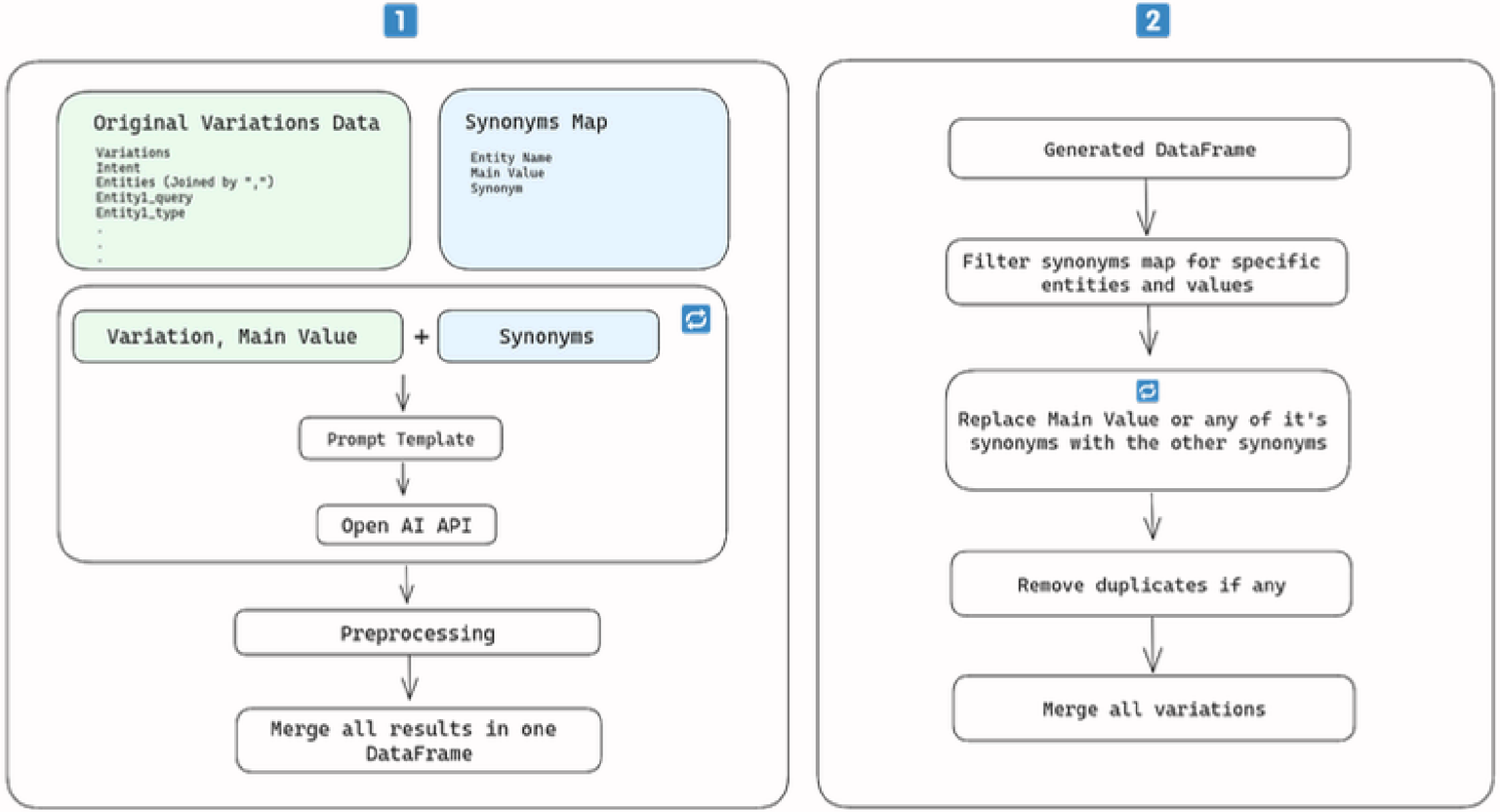
Number of Conflict Variations

Total number of variations in the annonated data

Data Generation

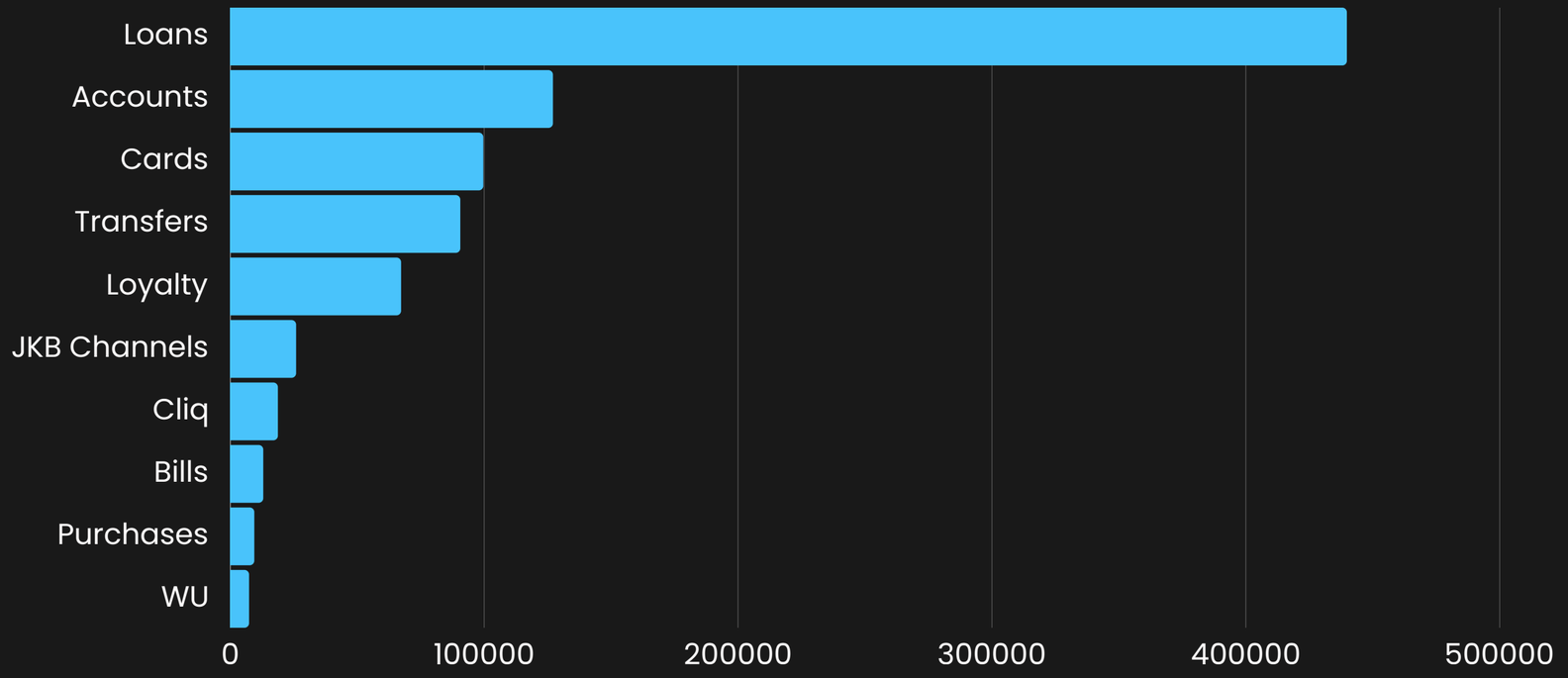
Synthetic Data Generation

Generating synthetic data using GPT-4



Synthetic Data Scope

Generating synthetic data using GPT-4



10

Number of Intents

Research areas mapped to each publication using advanced machine learning.

900K

Number of Variations

Total number of variations in the generated data

Data Governance

Data Governance

Various data governance considerations

1. Having a dedicated *data store*
2. Having a defined *data model*
3. Having a well defined *data lineage*
4. Having a unified *data labeling* tool

Data Stores

Suggested databases to store labeled data

SQL



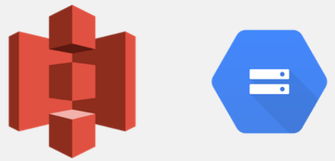
First option to store annotated data

NoSQL



Second option of storing annotated data

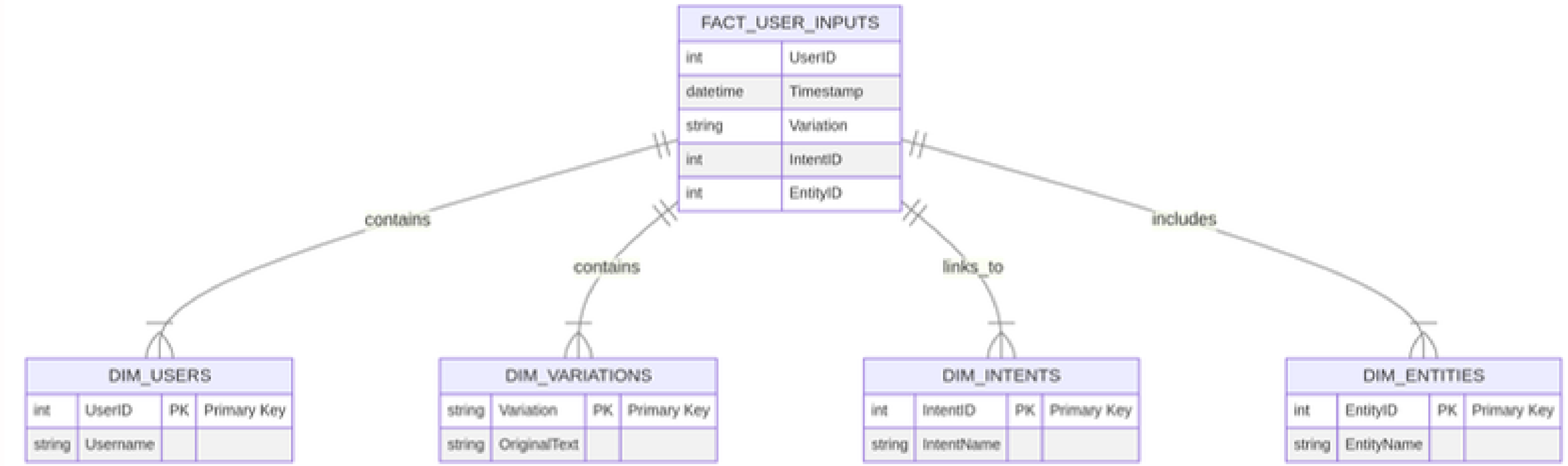
Data Repository



Important to have a separate data lake to store **training data** for the deployed models.

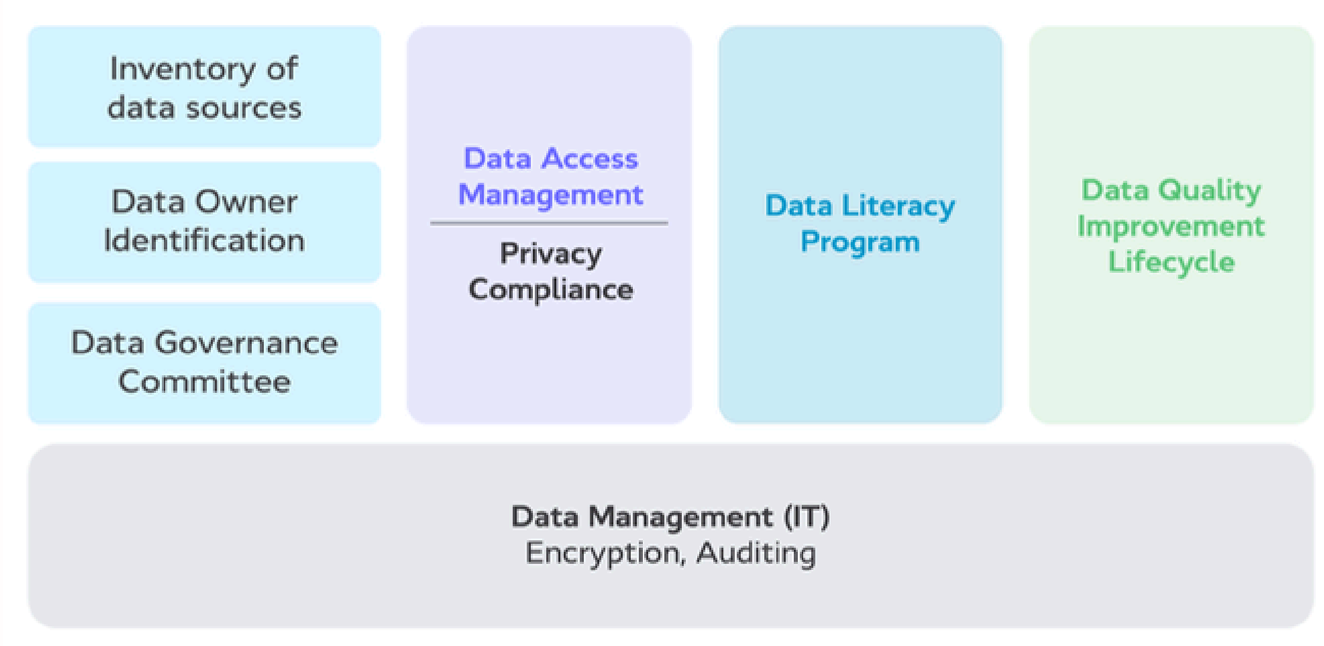
Data Model

Suggested data model for the development database



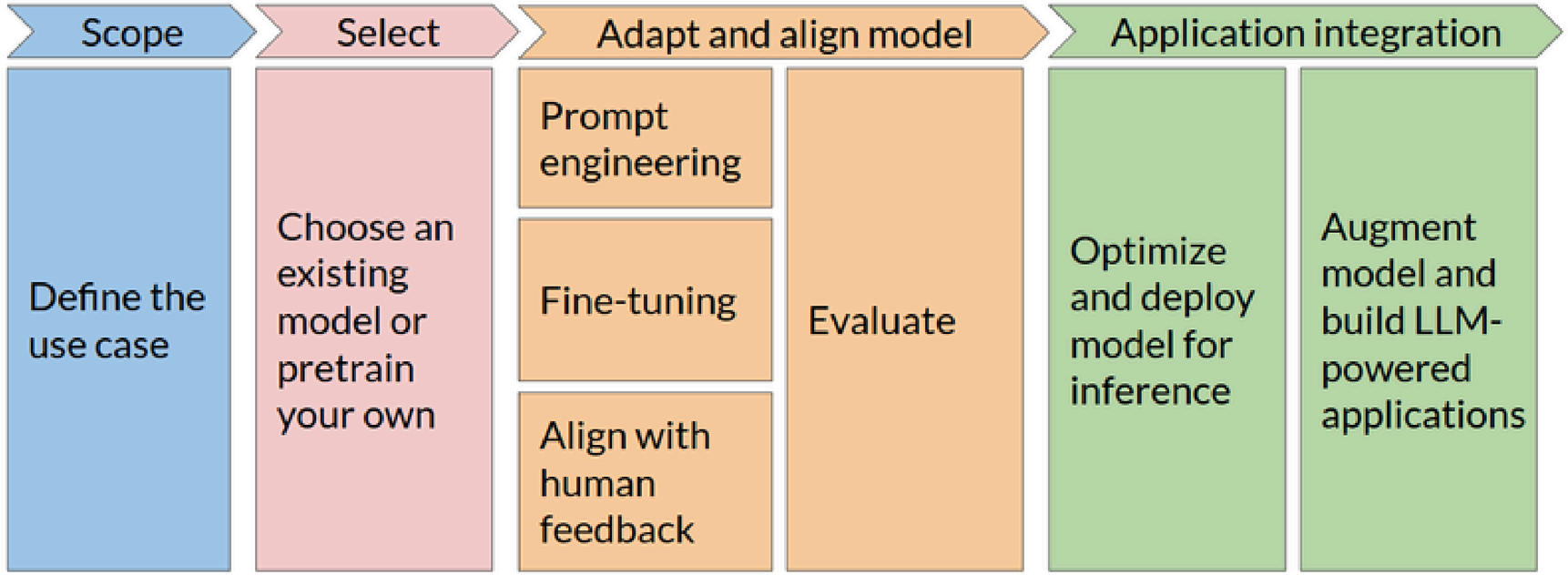
Data Lineage

The multiple layers and components of the project



Project Lifecycle

An LLM project has an iterative lifecycle

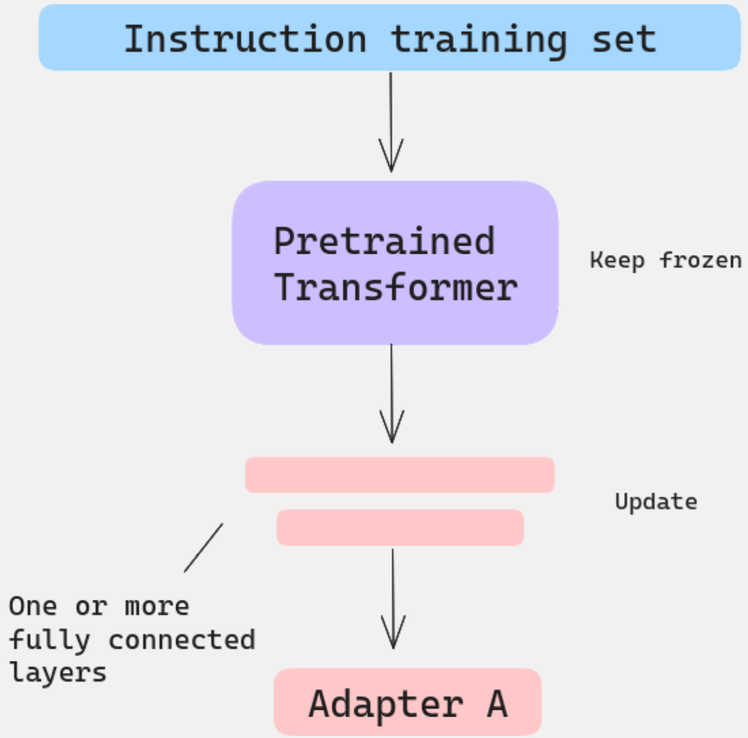


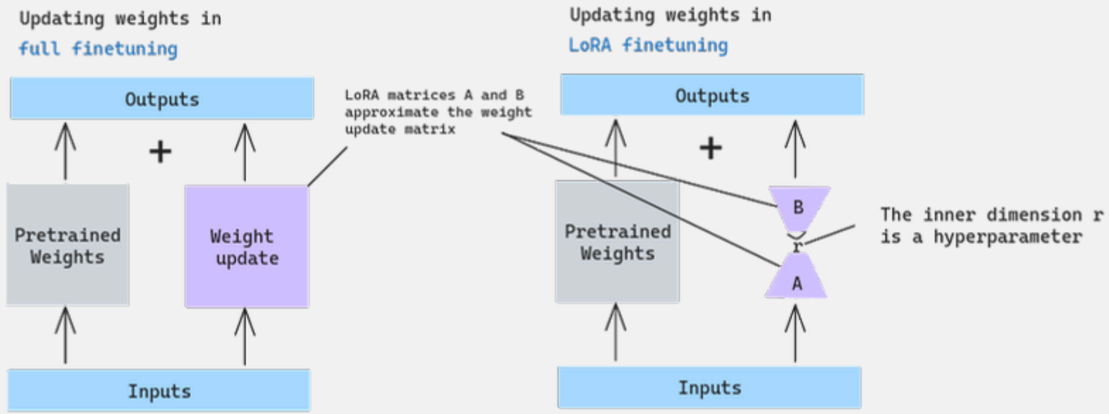
Methodology

Instruction Finetuning

The methodology used to finetune LLMs

Using Parameter Efficient Finetuning (PEFT) and Low Rank Adaptation (LoRA)

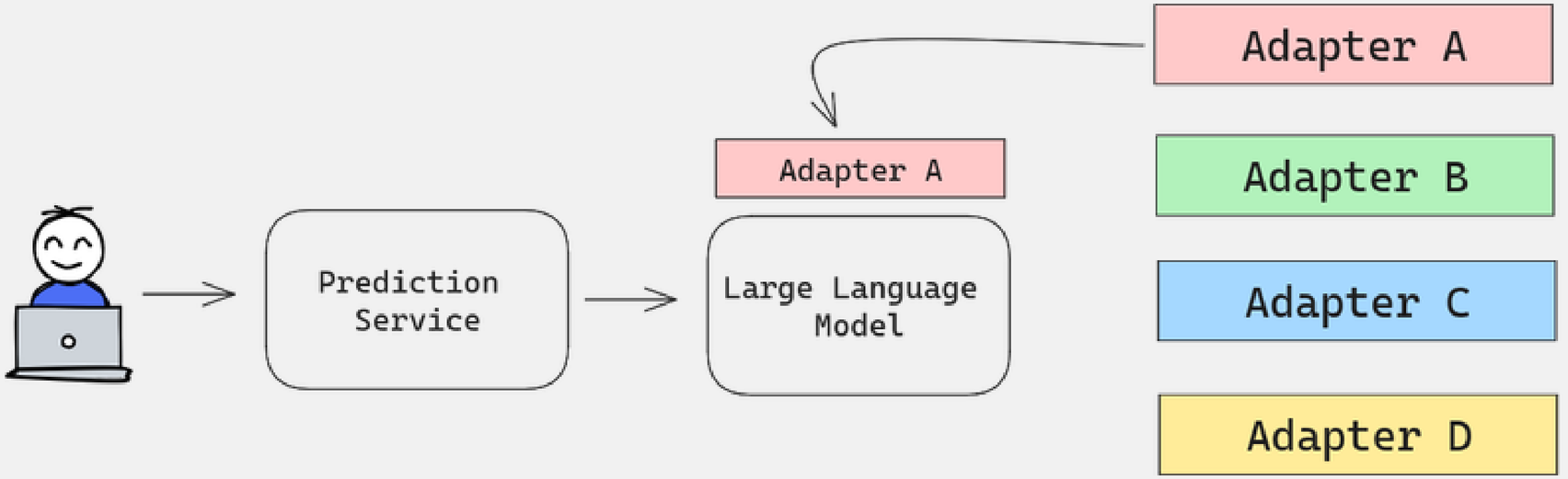




LoRA Adapters

Overview of the overall approach

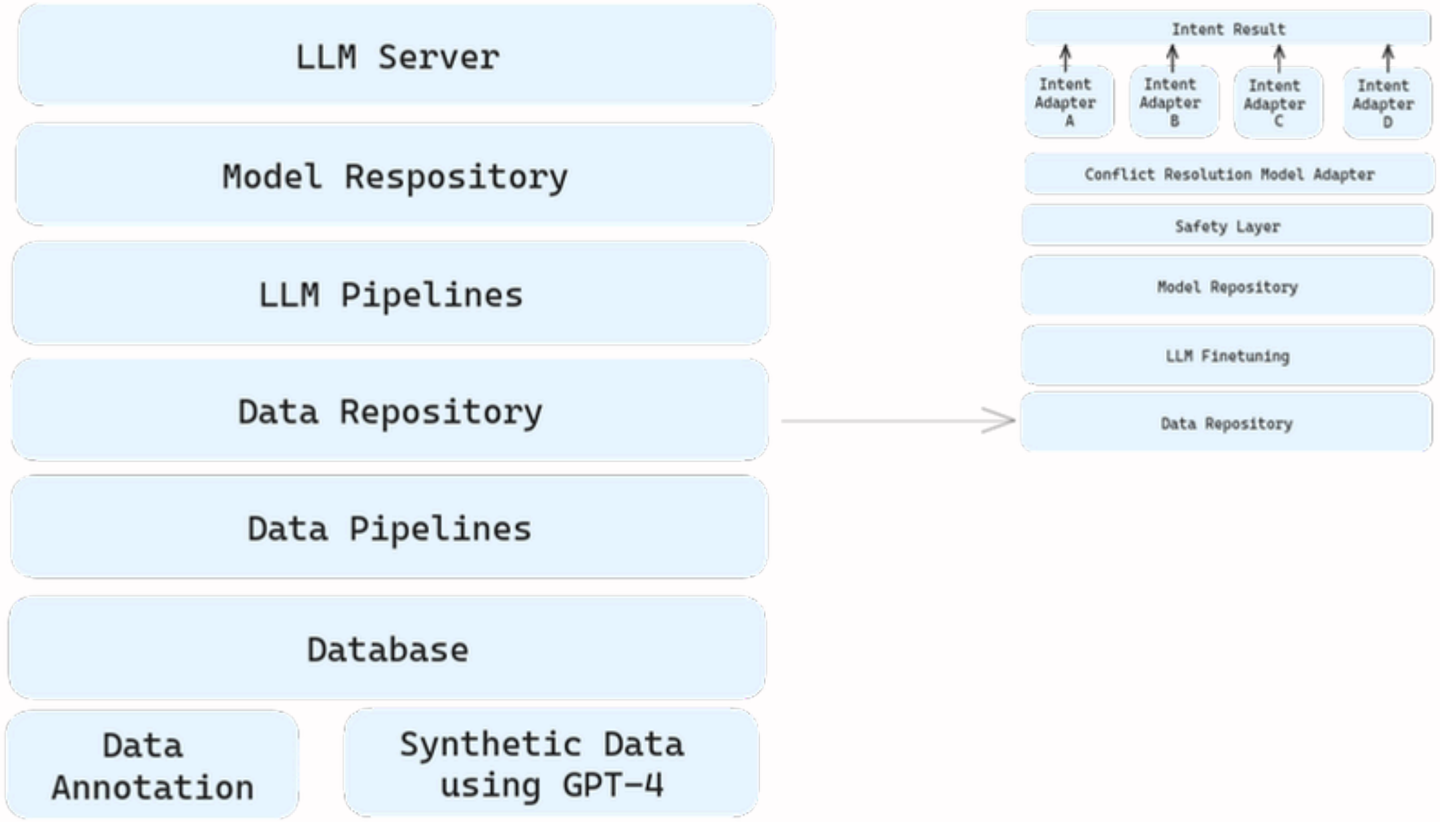
Using Arabic Finetuned Models



Solution Architecture

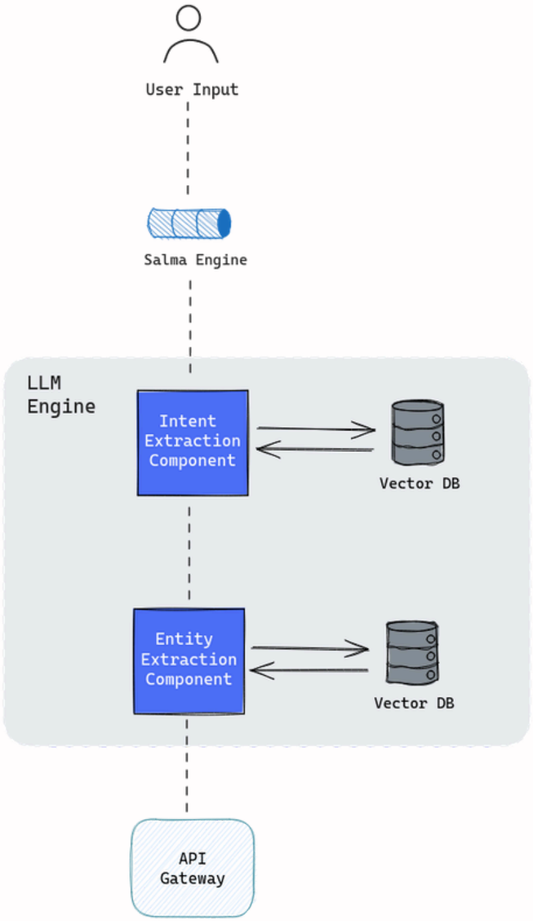
Solution Architecture

The multiple layers and components of the project



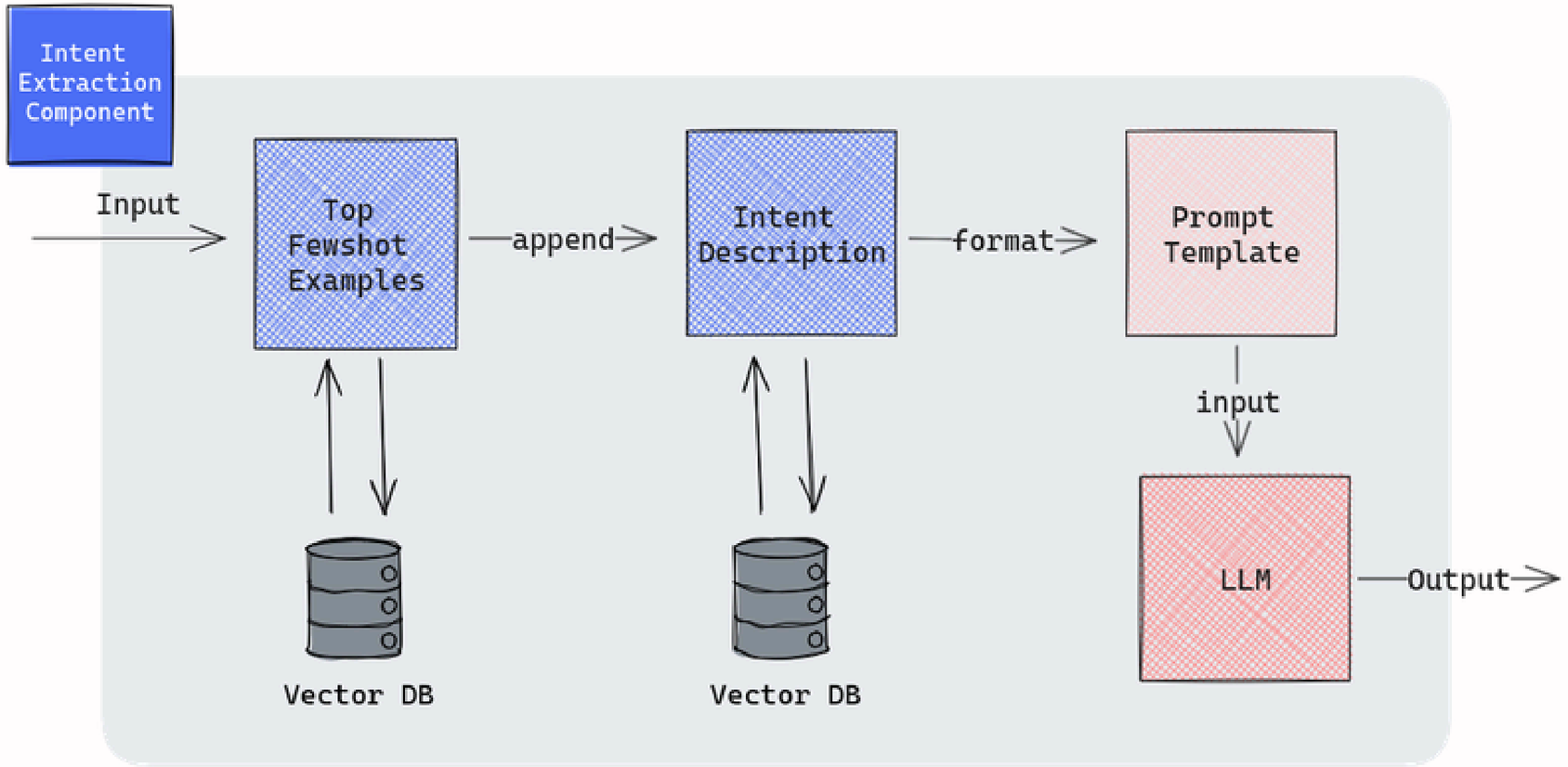
Solution Architecture

Overview of the LLM Engine



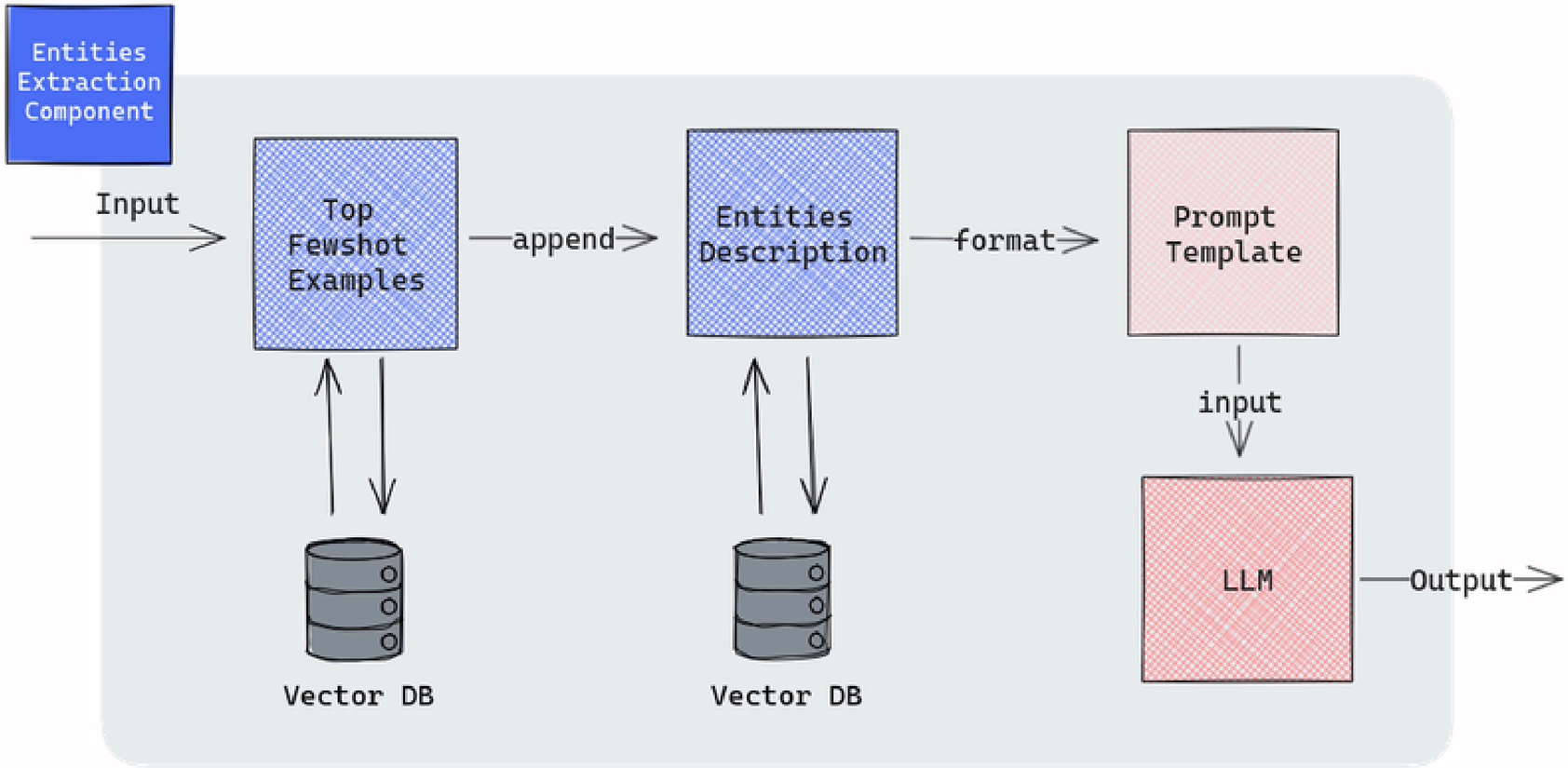
Solution Architecture

Overview of the LLM Engine



Solution Architecture

Overview of the LLM Engine



Results

LLM + RAG Results

On a sample of 4 intents and 8k records without any fine-tuning

Intent	Accuracy (%)	Predicted Instances
Loans	90	1797
Bills	83	1664
Cards	72	1439
Purchases	64	1282