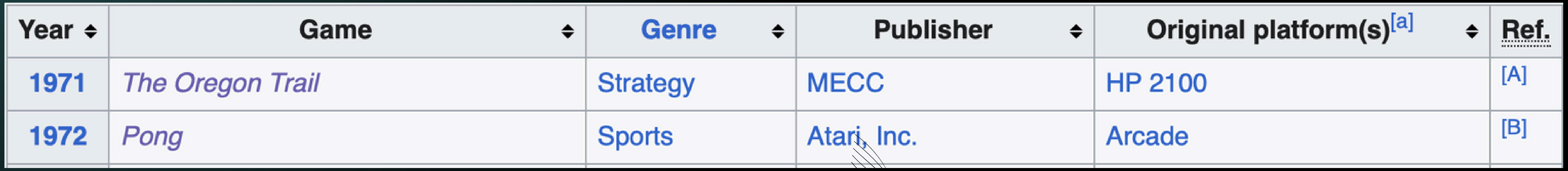
DSCI 510
Final Project

Sim Wang - University of Southern California

https://github.com/simwang-codes

Project Introduction - Best Video Games Analysis

Based on “List of video games considered the best” table on Wikipedia

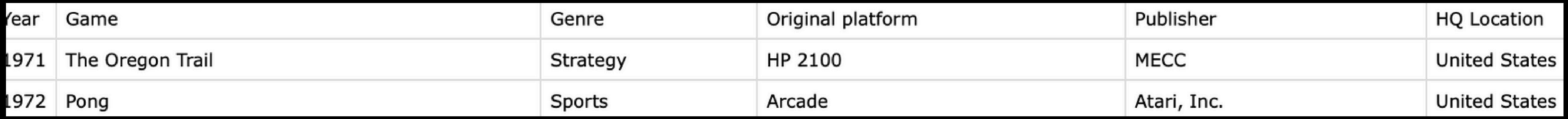


Year	Game	Genre	Publisher	Original platform(s)[a]	Ref.
1971	*The Oregon Trail*	Strategy	MECC	HP 2100	[A]
1972	*Pong*	Sports	Atari, Inc.	Arcade	[B]

- Built a web scraping ETL pipeline that extracts the table contents
- Scrapes through hyperlinks in the Game column to retrieve all text data from each game's Wikipedia page, and produce a .txt file
- Scrapes hyperlinks in Publisher column to retrieve the country

- Developed a RAG system with LangChain to embed .txt files into vectors and create a smart chatbot powered by OpenAI API for question answering

Results: Visualizations and a “Video Game Expert” RAG chatbot



ear	Game	Genre	Original platform	Publisher	HQ Location
1971	The Oregon Trail	Strategy	HP 2100	MECC	United States
1972	Pong	Sports	Arcade	Atari, Inc.	United States

→ CSV for visualisations

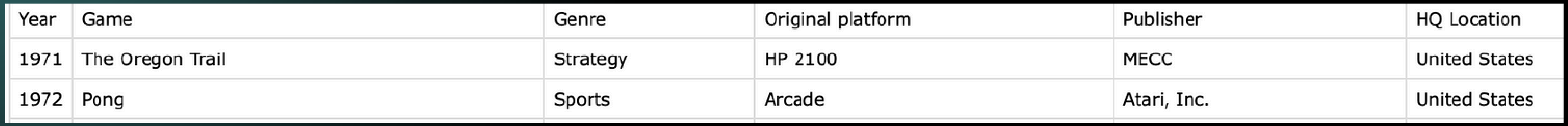
→ A cleaned .txt file with all games’ wiki description

→ A RAG AI Chatbot

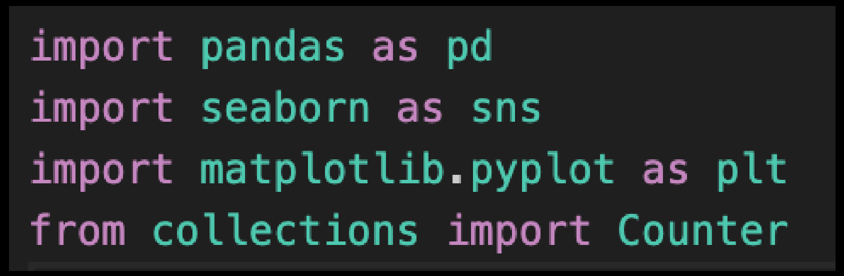
Data Sources

Data Sources	**Data Types**	**Details**
List of video games considered the best	Wikipedia HTML table	Contains 337 games praised by journalists and magazines
A text file containing text descriptions of 337 games	HTML text → .txt	Contains over 290k words and includes comprehensive details about every game
A table of games info	HTML table/text → CSV	Contains the year of production, genre, publisher, country, and platforms of the games

Results Summary - Visualisations

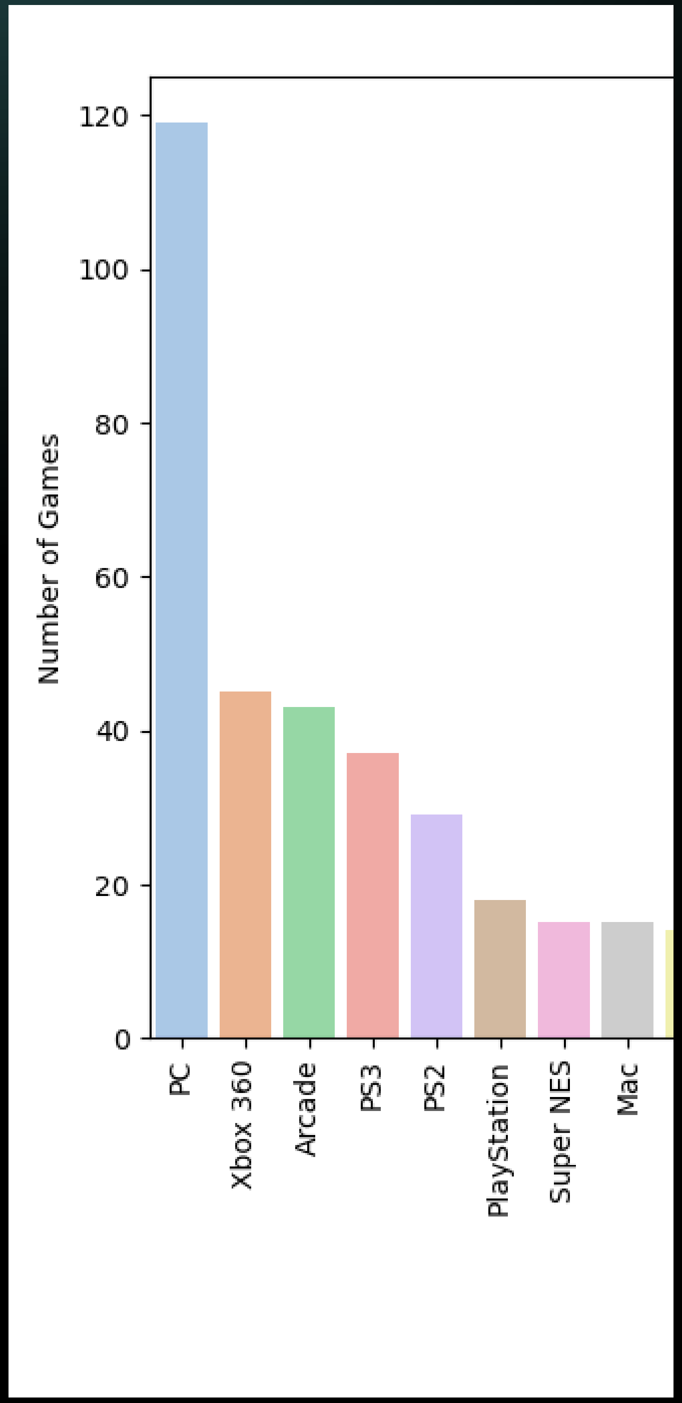


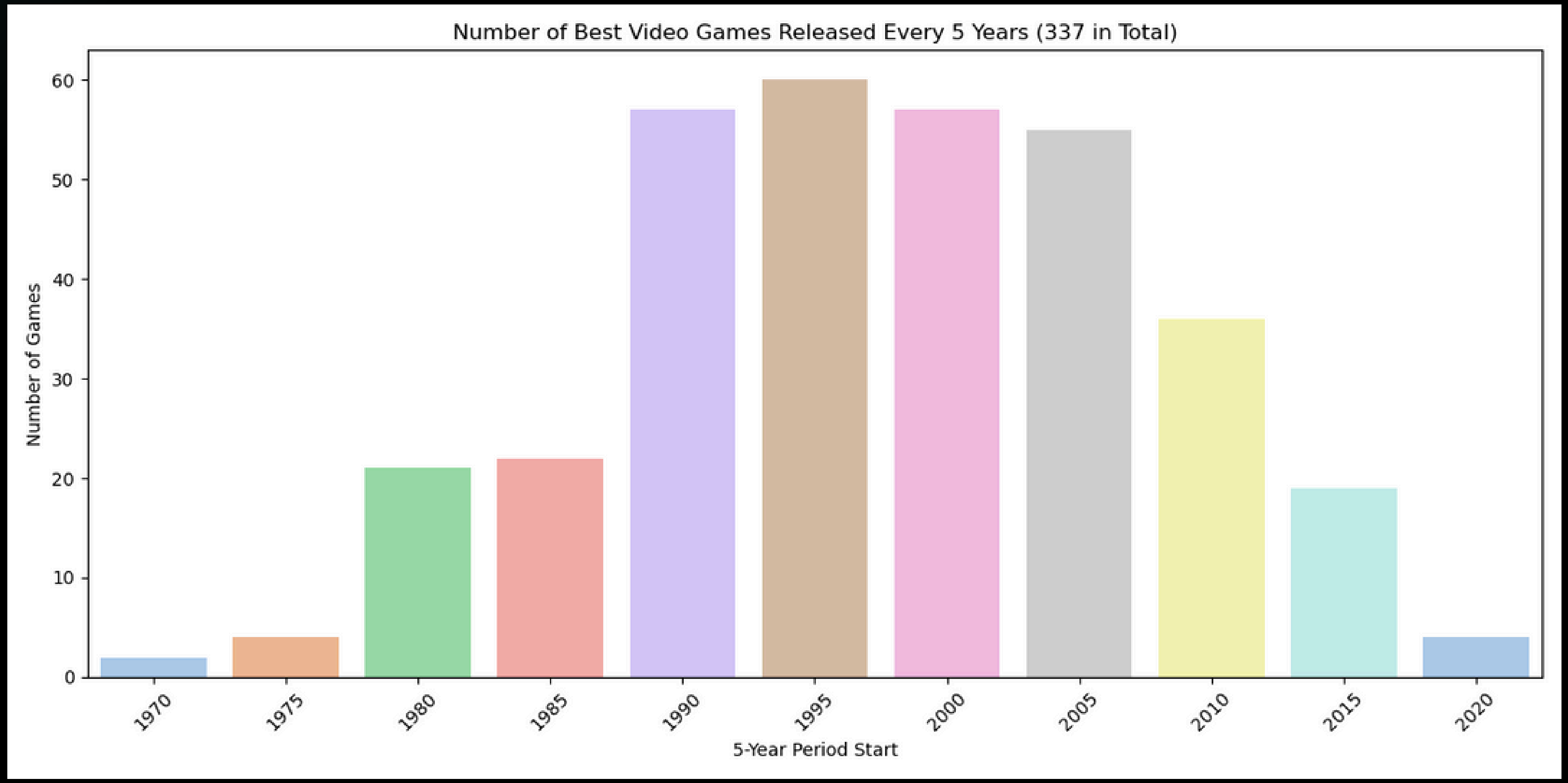
Year	Game	Genre	Original platform	Publisher	HQ Location
1971	The Oregon Trail	Strategy	HP 2100	MECC	United States
1972	Pong	Sports	Arcade	Atari, Inc.	United States



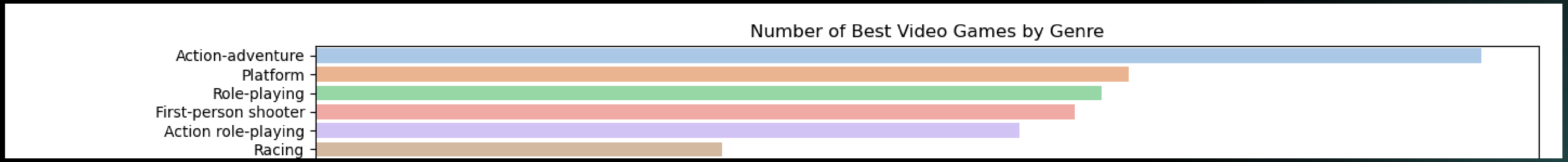
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from collections import Counter
```

- Visualized the distribution of top video games by genre, company location (countries), release period, and original platform.
- 337 games in total, range from 1970 ~2023



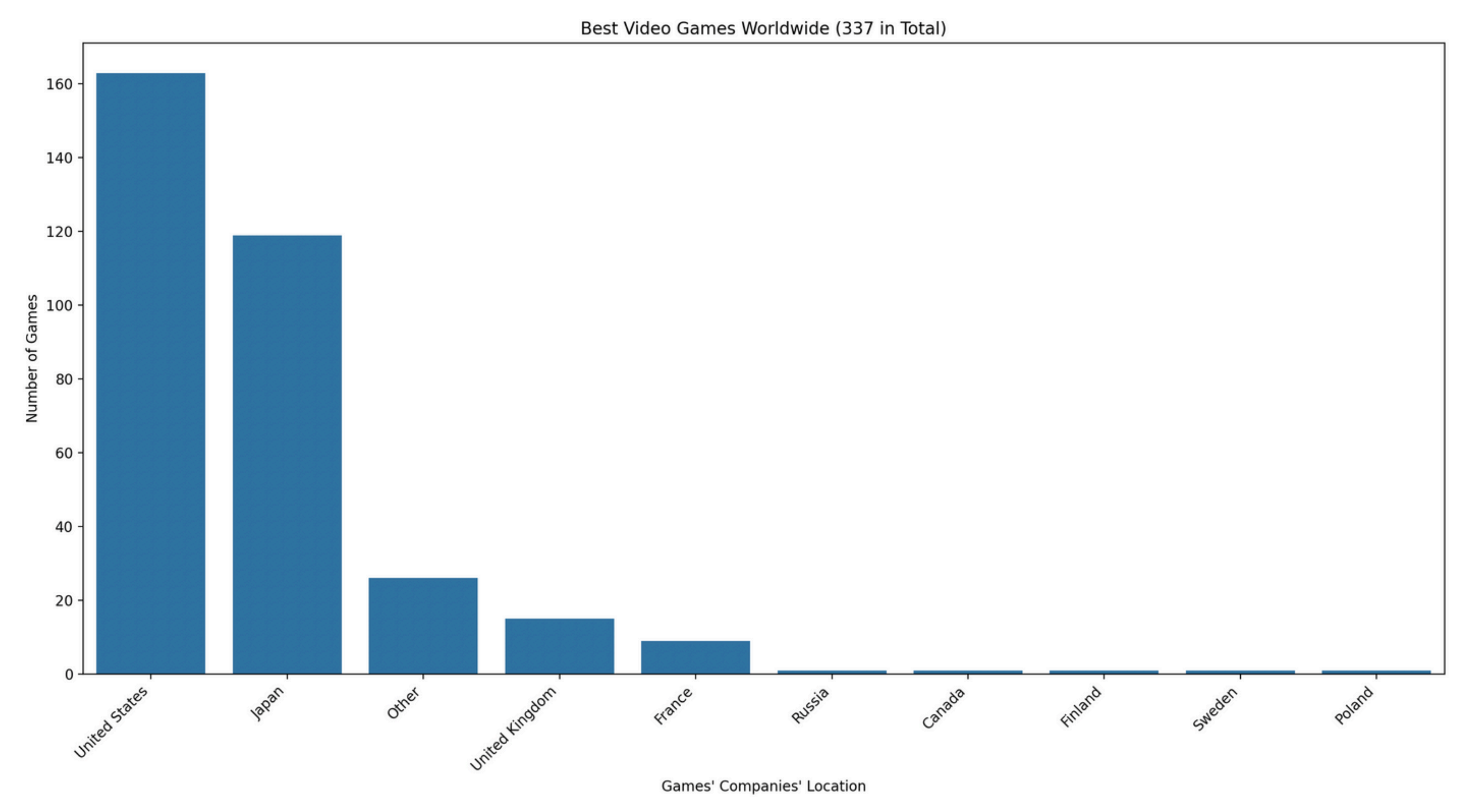


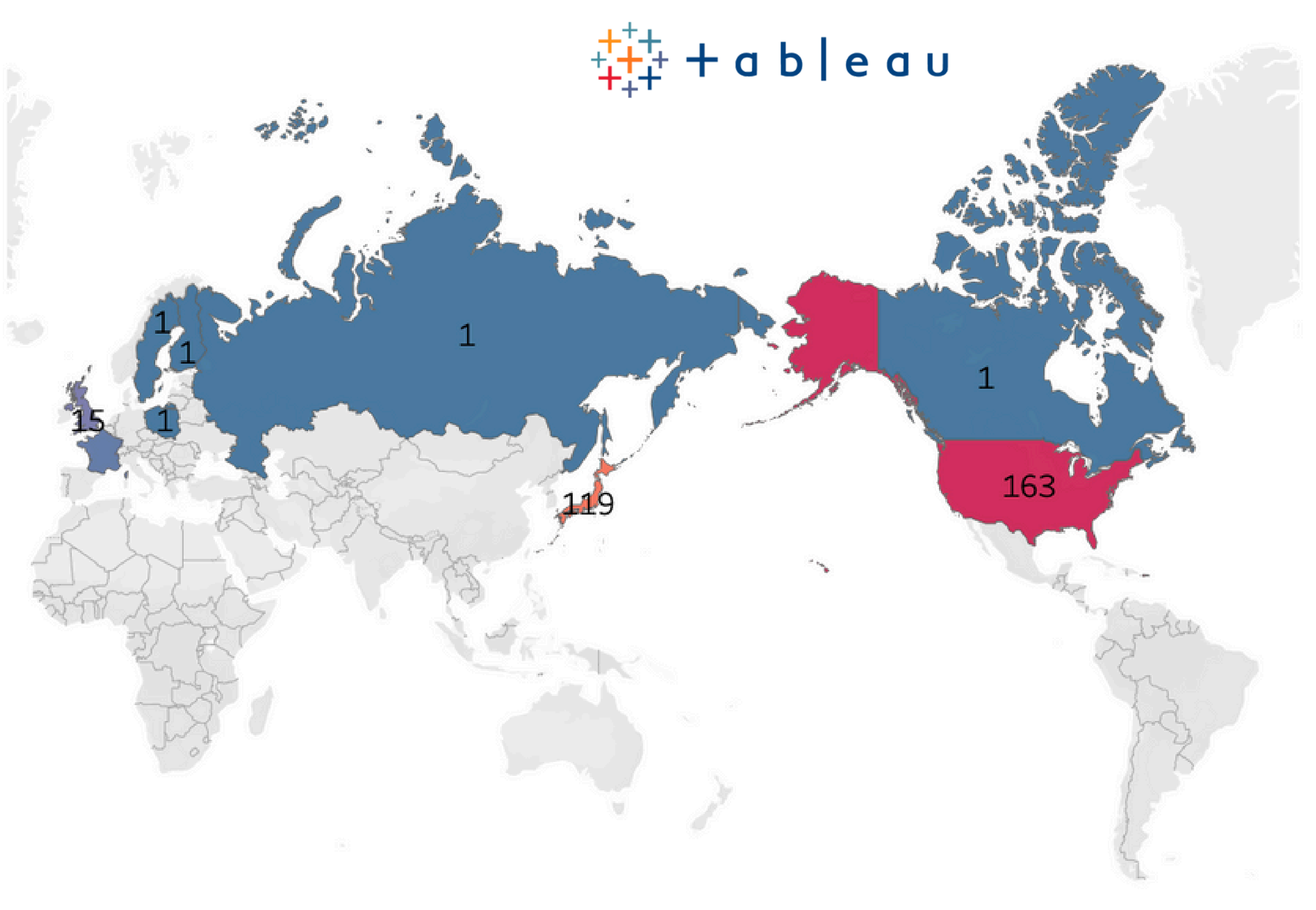
Number of Best Video Games Released Every 5 Years (337 in Total)



Number of Best Video Games by Genre

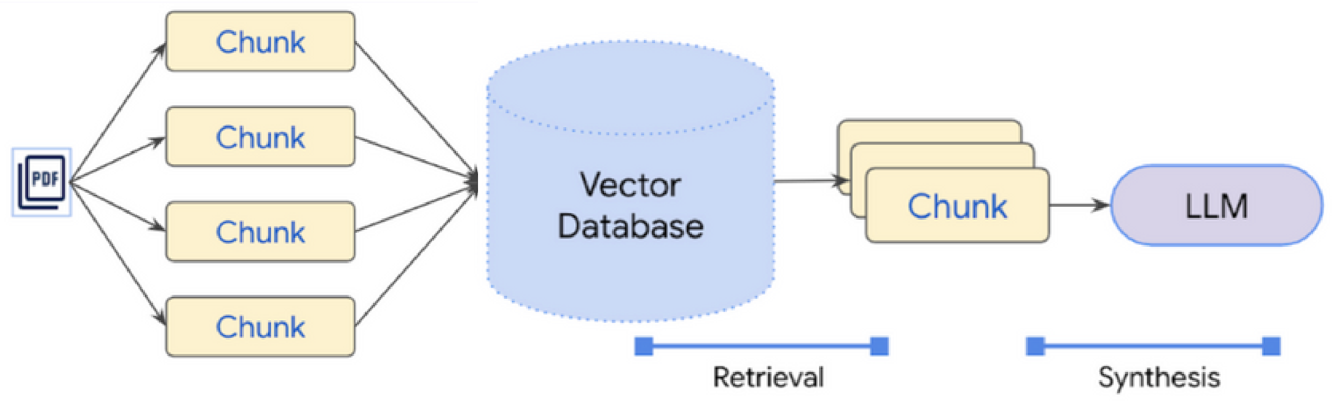
Results Summary - Visualisations

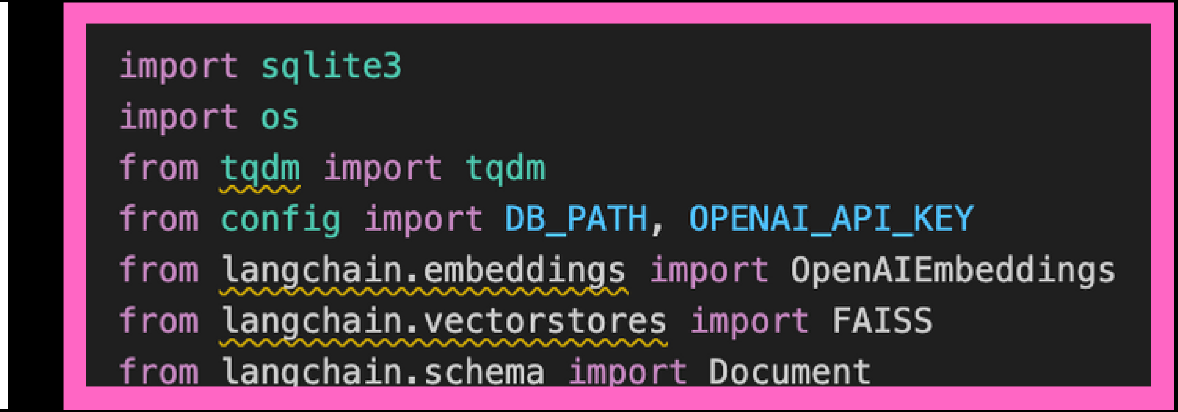




- The world's first video game *Tennis for Two* originated in the United States, following the emergence of companies like Atari and Nintendo, the U.S. and Japan became the most productive countries in the video game industry within the period of 1970 ~ 2023 (see diagrams in last page)
- However, most of the games on the *video games considered the best* list were selected by U.S. video game journalists and magazines, which might have included bias when selecting the "best video games."

Results Summary - RAG Chatbot



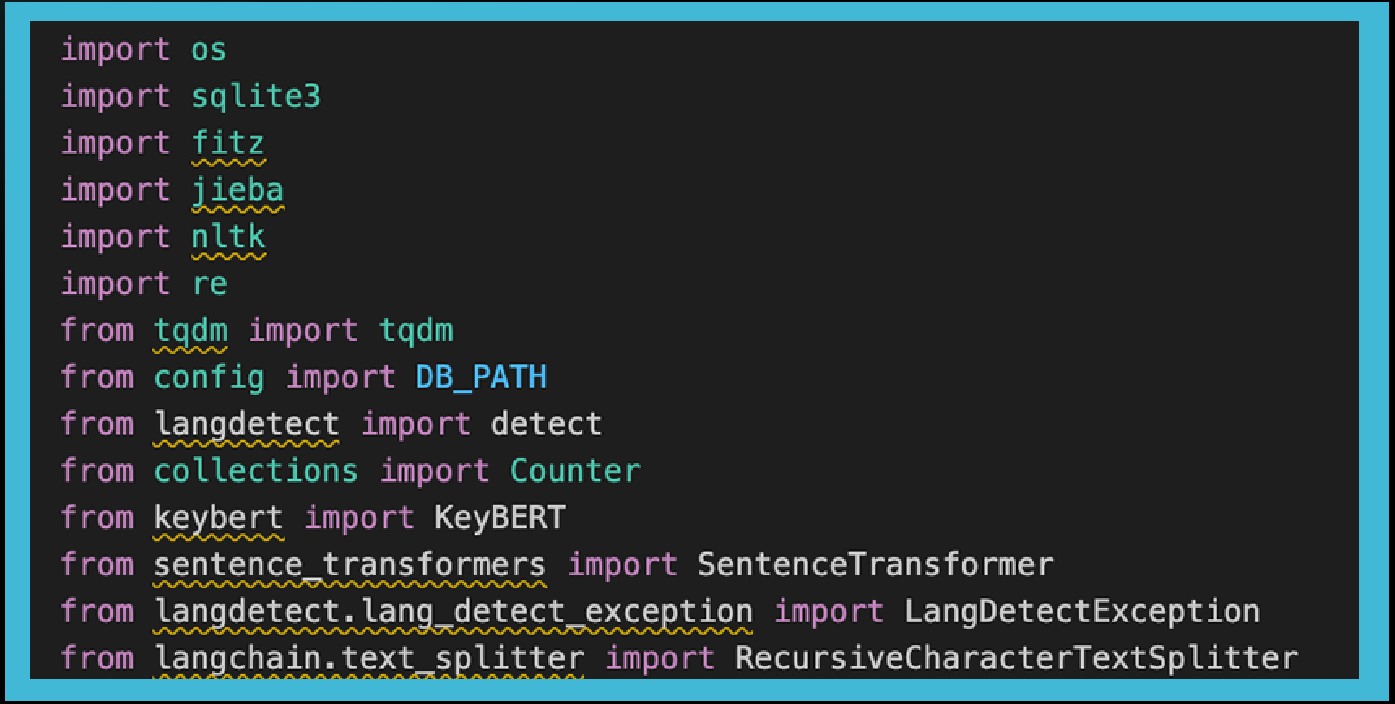


```
import sqlite3
import os
from tqdm import tqdm
from config import DB_PATH, OPENAI_API_KEY
from langchain.embeddings import OpenAIEmbeddings
from langchain.vectorstores import FAISS
from langchain.schema import Document
```

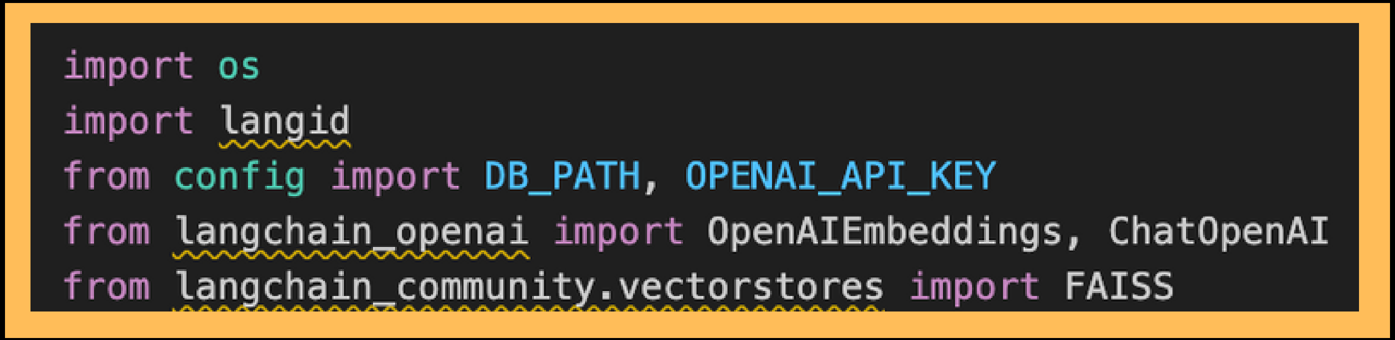
KeyBERT

LangChain

OpenAI

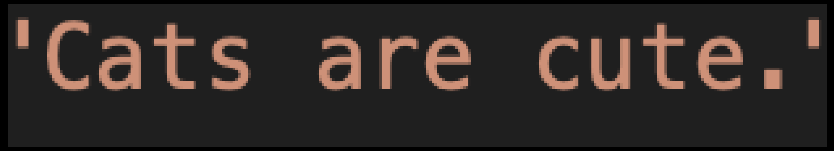


```
import os
import sqlite3
import fitz
import jieba
import nltk
import re
from tqdm import tqdm
from config import DB_PATH
from langdetect import detect
from collections import Counter
from keybert import KeyBERT
from sentence_transformers import SentenceTransformer
from langdetect.lang_detect_exception import LangDetectException
from langchain.text_splitter import RecursiveCharacterTextSplitter
```



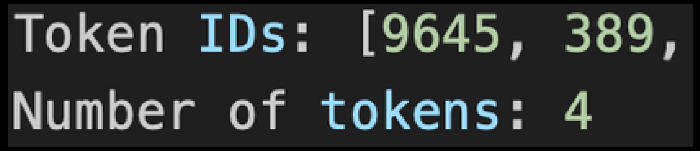
```
import os
import langid
from config import DB_PATH, OPENAI_API_KEY
from langchain_openai import OpenAIEmbeddings, ChatOpenAI
from langchain_community.vectorstores import FAISS
```

1. Read text content in .txt file that stores all 337 games' wiki info
2. Split text into chunks of 700 words
3. Extract the top 5 keywords/tags (the most semantically important words, normally nouns) from each chunk
4. Embed each chunk and keywords/tags into vectors, then store them in Facebook AI Similarity Search (FAISS)'s index file
5. Take user's input question, embed the question into vectors, and conduct mathematical calculation to find the top chunks and keywords that might answer the user's question
6. Push the user's question and the returned chunks/keywords into the GPT-4 model to generate a natural language answer



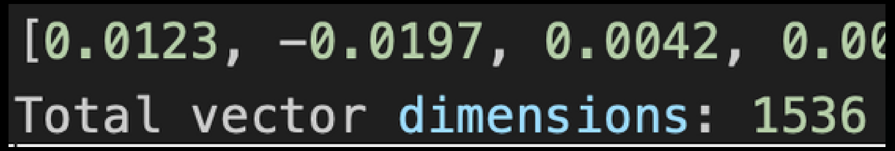
```
'Cats are cute.'
```

→



```
Token IDs: [9645, 389,
Number of tokens: 4
```

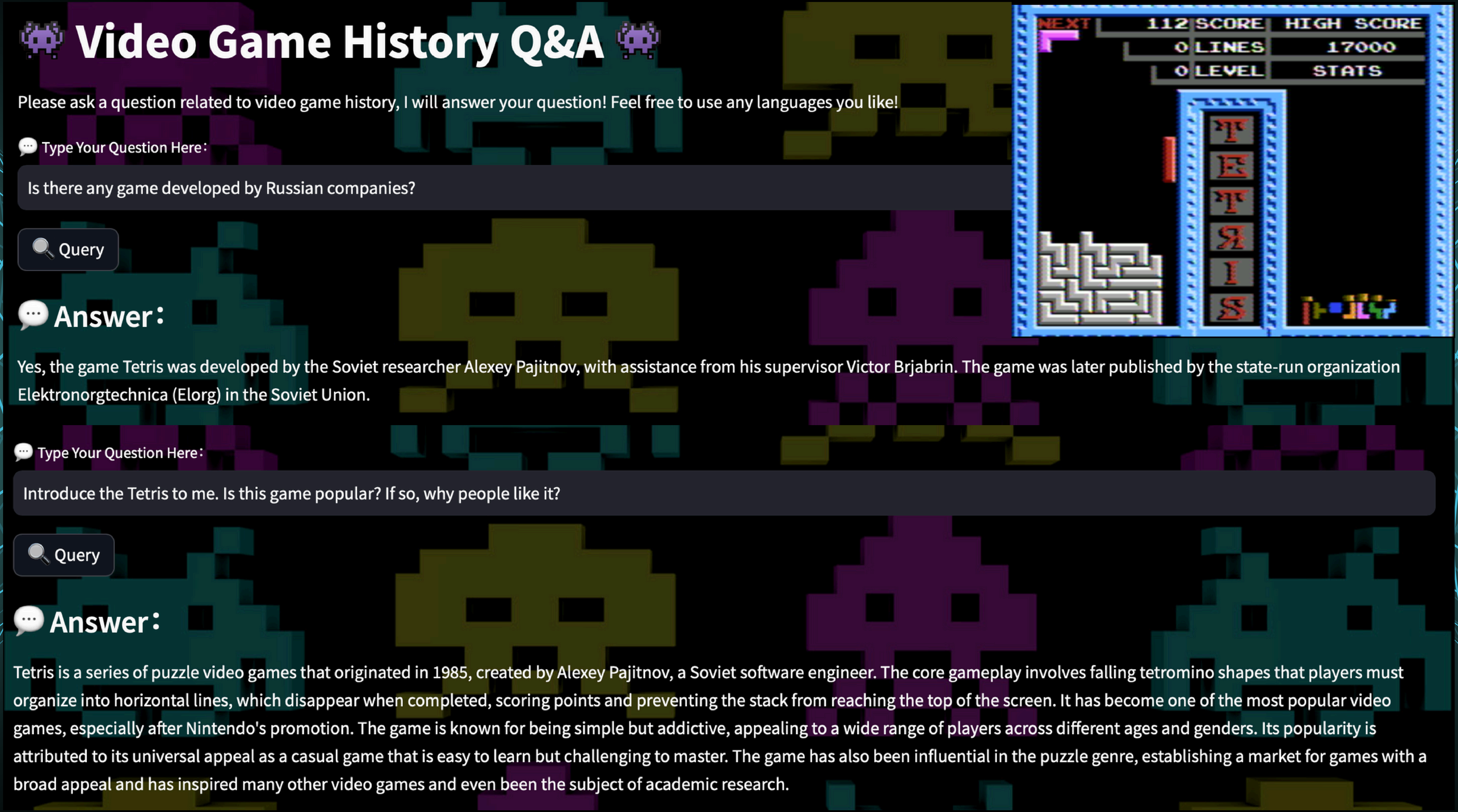
→



```
[0.0123, -0.0197, 0.0042, 0.00
Total vector dimensions: 1536
```

← Embedding Workflow

Results Summary - RAG Chatbot



👾 Video Game History Q&A 👾

Please ask a question related to video game history, I will answer your question! Feel free to use any languages you like!

💬 Type Your Question Here:

Is there any game developed by Russian companies?

🔍 Query

💬 Answer:

Yes, the game Tetris was developed by the Soviet researcher Alexey Pajitnov, with assistance from his supervisor Victor Brjabrin. The game was later published by the state-run organization Elektronorgtechnica (Elorg) in the Soviet Union.

💬 Type Your Question Here:

Introduce the Tetris to me. Is this game popular? If so, why people like it?

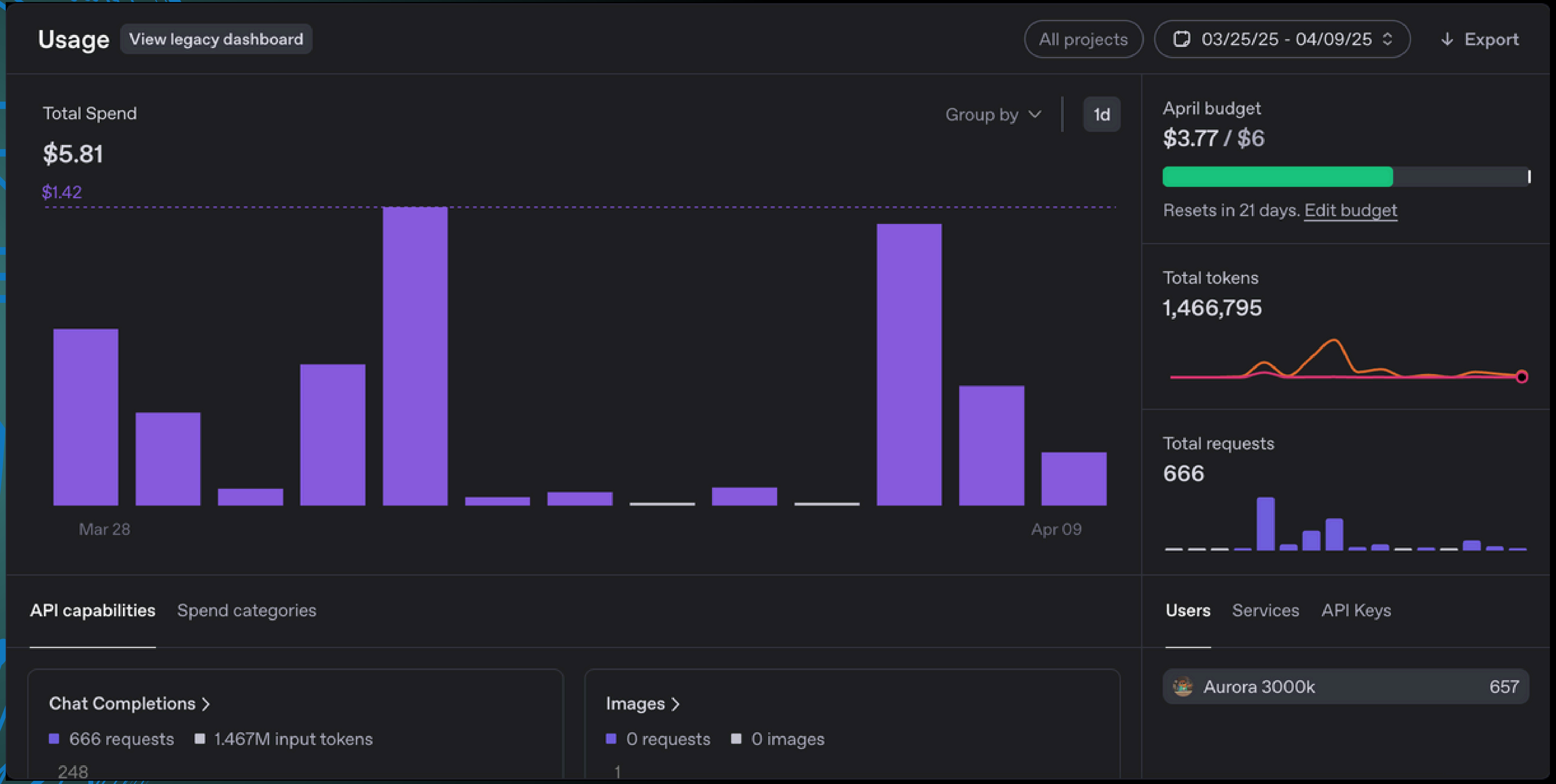
🔍 Query

💬 Answer:

Tetris is a series of puzzle video games that originated in 1985, created by Alexey Pajitnov, a Soviet software engineer. The core gameplay involves falling tetromino shapes that players must organize into horizontal lines, which disappear when completed, scoring points and preventing the stack from reaching the top of the screen. It has become one of the most popular video games, especially after Nintendo's promotion. The game is known for being simple but addictive, appealing to a wide range of players across different ages and genders. Its popularity is attributed to its universal appeal as a casual game that is easy to learn but challenging to master. The game has also been influential in the puzzle genre, establishing a market for games with a broad appeal and has inspired many other video games and even been the subject of academic research.

Challenges

KeyBERT



Usage

03/25/25 - 04/09/25

Total Spend
$5.81

April budget
$3.77 / $6
Resets in 21 days. Edit budget

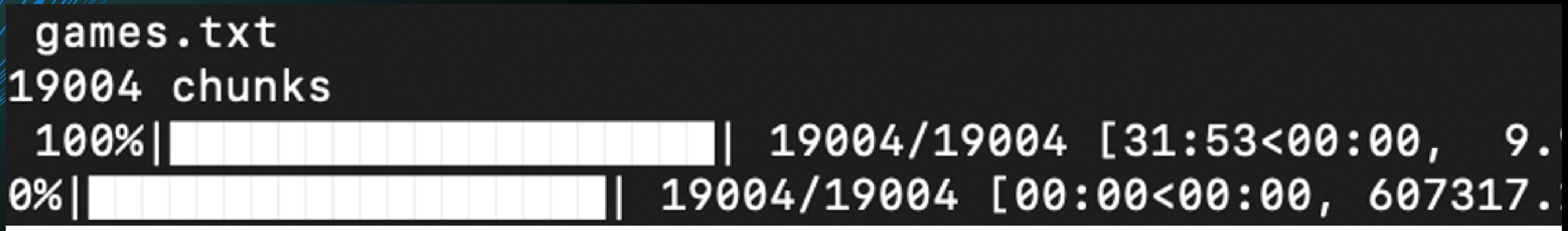
Total tokens
1,466,795

Total requests
666

Chat Completions
666 requests · 1.467M input tokens

Images
0 requests · 0 images

Aurora 3000k 657



```
 games.txt
19004 chunks
 100%|████████████████████| 19004/19004 [31:53<00:00,  9.
0%|███████████████████| 19004/19004 [00:00<00:00, 607317.
```

KeyBERT is slow:

- it first embeds chunks, then use vector index to extract keywords

GPT-4 is slow and expensive:

- takes longer time than less complex gpt models

Scalability:

- if you have 40 million words to process, it will be slow

Thank You

Sim Wang - University of Southern California - DSCI 510