New York
Public
Library

NYU

NYPL & NYU ITP Fall 2024 Project

NYU Resources

Team:

- Simon Zeng
- Rahul Singhal
- Tanmay Khandelwal
- Fei Wang

Professor: Jean-Claude Franchitti

Teaching Assistant: Joanna Gilberti

NYU

NYPL Resources

Eric Shows, Zachary Peyton, Brent Reidy Stephen

A. Schwarzman Building

445 Fifth Avenue, New York, NY 10016

Email: ericshowsg@nypl.org

NYPL

Background

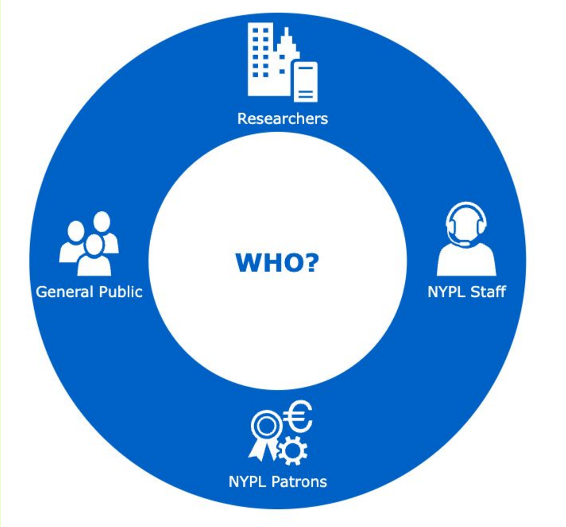
- The New York Public Library (NYPL) is one of the largest libraries in the world, offering access to millions of books, historical documents, and digital collections.
- Its mission is to inspire lifelong learning, advance knowledge, and strengthen communities by making information accessible to all.
- NYPL's digital collections play a critical role in preserving and democratizing access to cultural heritage, offering resources such as photographs, handwritten letters, and manuscripts.
- With the growing volume of digital content, there is a pressing need to enhance organizational and search capabilities to meet modern accessibility standards.
- Ensuring the accuracy of classification and metadata is vital for reducing barriers to discovery and maximizing the impact of digital archives.
- As user expectations evolve, tools like user-friendly Q&A systems and AI-driven classification are essential for maintaining NYPL's leadership in public knowledge accessibility.

Goals

- The current DigiSuite platform faces challenges such as inefficient search capabilities, manual metadata generation, and limited classification accuracy. To overcome these issues, we are implementing the following enhancements:
- **Intelligent Q&A System**:
  Developing a **Q&A bot** powered by **LLMs** to provide real-time, context-aware answers and dynamic follow-up support for users. Integrating **knowledge graphs** and **vector databases** to enable faster, context-driven access to relevant collections.
- **Improved Classification Accuracy**:
  Upgrading existing models to minimize false positives and ensure precise organization of digital assets.
- **Automated Metadata Generation**:
  Replacing manual, error-prone processes with enhanced classification models (**AT Gen**, **HWT Gen**) to improve accuracy and reduce manual effort.

Beneficiaries

- **Researchers**: Quickly locate historical documents, images, and letters with improved classification and tagging tools.
- **NYPL Patrons**: Browse collections like NYC landmarks and cultural artifacts effortlessly using enhanced metadata and a Q&A bot.
- **NYPL Staff**: Streamline operations and assist patrons with advanced classification tools and a Q&A bot.
- **Students**: Access organized digital collections for academic research, history projects, and assignments with easy-to-use tools.



Modeling

- **Improved Digisuite Classifier**: Window-level classification is now done to prevent downstream errors caused by incorrect image-level decisions.

- **Improved ATGen Alt-Text Generation:** The model leverages a pretrained LLM fine-tuned on the LAION dataset to create more detailed alt text for input images.

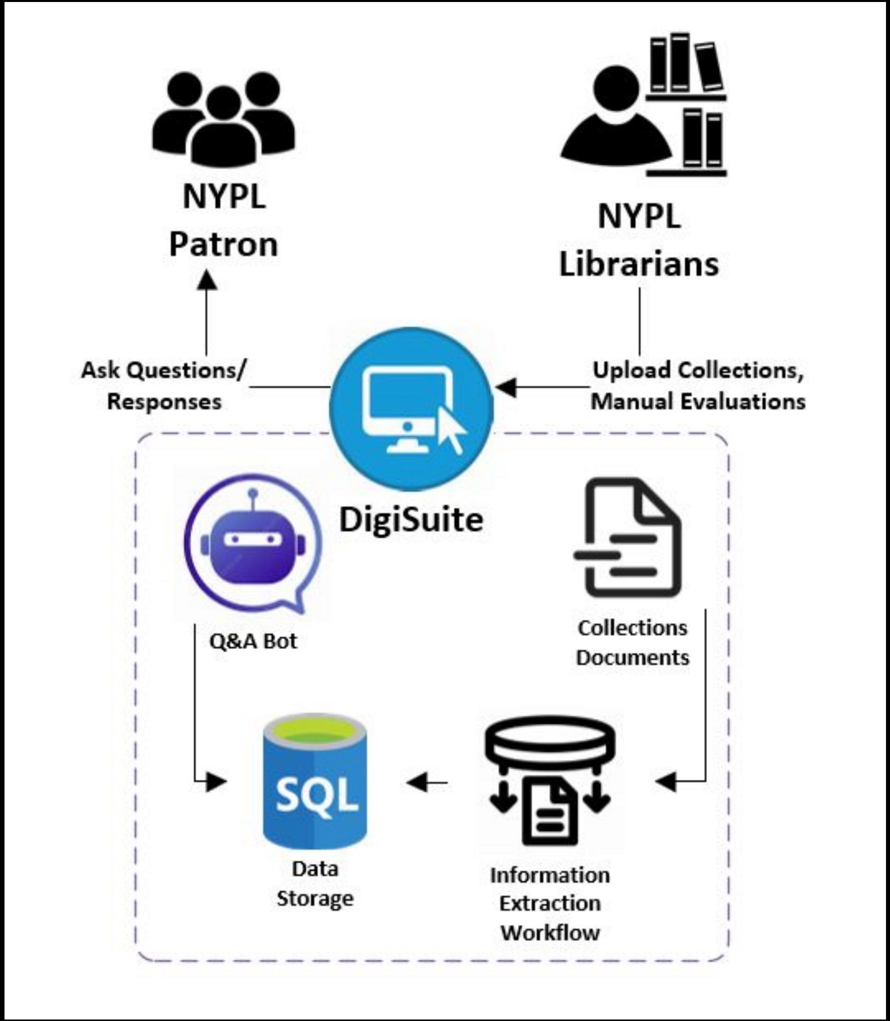
Q&A Bot

- **Smart Search**: The bot uses an LLM (Large Language Model) to redirect queries to the most relevant collection in NYPL's database.

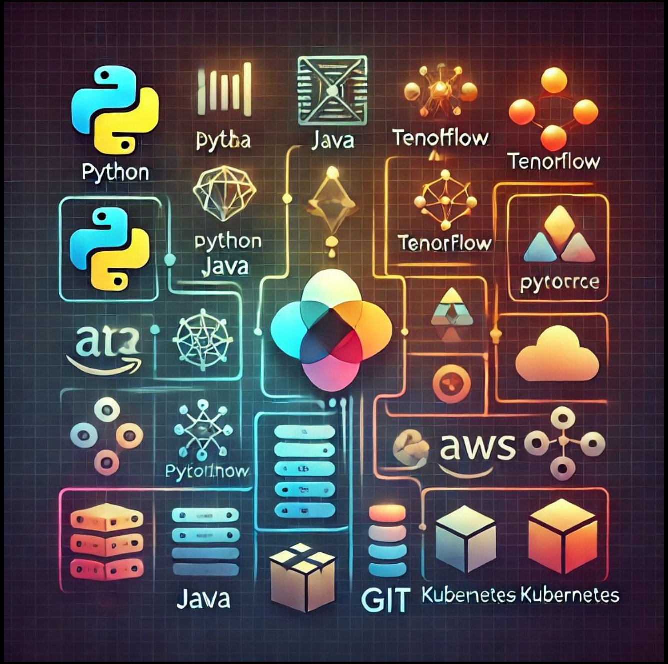
- **Contextual Data Retrieval**: It fetches and processes data, including ATGen Data and HWT Gen Data, with support from a knowledge graph and vector database for accurate results.

- **Interactive Responses**: Users can ask follow-up queries to refine answers, ensuring precise and dynamic results from the fetched collections.

Overall Architecture



Tech Stack



Frontend Application

Backend Application

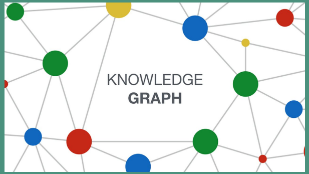
Large Language Modeling

Database

python

django

VectorDB



ubuntu Linux

Hugging Face

HTML CSS

OpenAI

LangChain

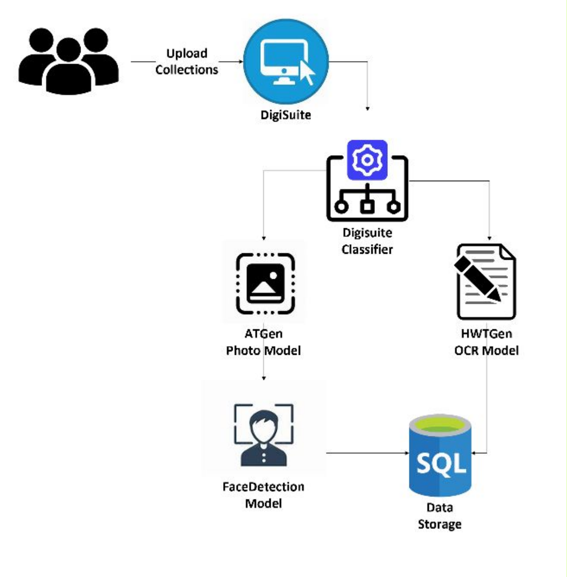
Modeling

Improvements on Existing Architecture

- **Digisuite Model Classification Errors**
  - Errors in the initial classification led to images being directed to the incorrect model, leading to poor data extraction from downstream models
- **Some collection documents contained both images and handwritten text.**
  - Image-level classifications oversimplify images, leading to information loss from sub-components.
- **Limited understanding of model performance**
  - There was no existing metric or test set – performance was manually evaluated by librarians

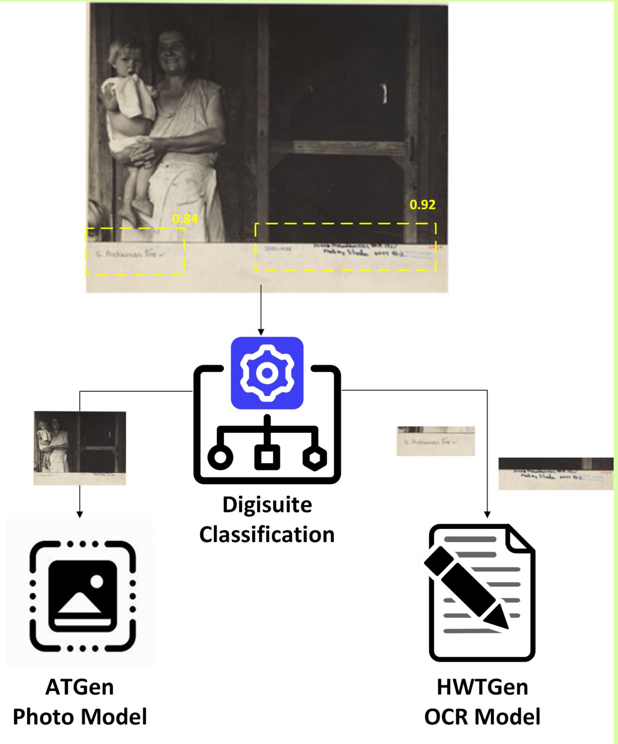
Hurrah! for Massa Lincoln
IMPETUOUS CHARGE OF THE FIRST COLORED REBEL REGIMENT

Primary Challenge



Solution

- **Window-Level Classification**
  - Images are segmented into “rectangles” for which the model predicts the presence of handwritten text
  - Individual rectangles containing handwritten text are sent to HWTGen for extraction
  - The rest of the image is sent to ATGen for alt-text generation. Results from both models are then concatenated together in the SQL database.

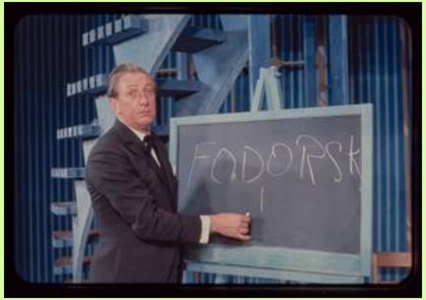


- **Test Set Creation**
  - A test set of ~950 images was manually created and labeled using NYPL collections, allowing for performance evaluation.
  - Previous DigiSuite classification model had a 59.5% accuracy (64.7% if removing images labeled “Both”)
  - After **hyperparameter optimization** on threshold values, the updated DigiSuite classification model has an **80.7% accuracy (+21.2%)**

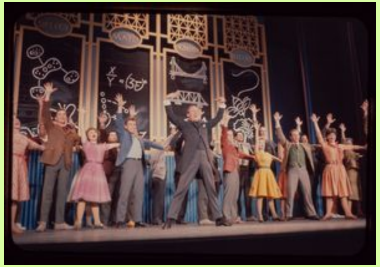
Solution

Secondary Challenge

- **ATGen Output Vagueness**
  - ATGen generates alt-text that was generic and unspecific, missing critical details in images for downstream knowledge base searches
- **Limited Understanding of Performance**
  - ATGen performance was based solely off of manual evaluation (and edits) performed by librarians.



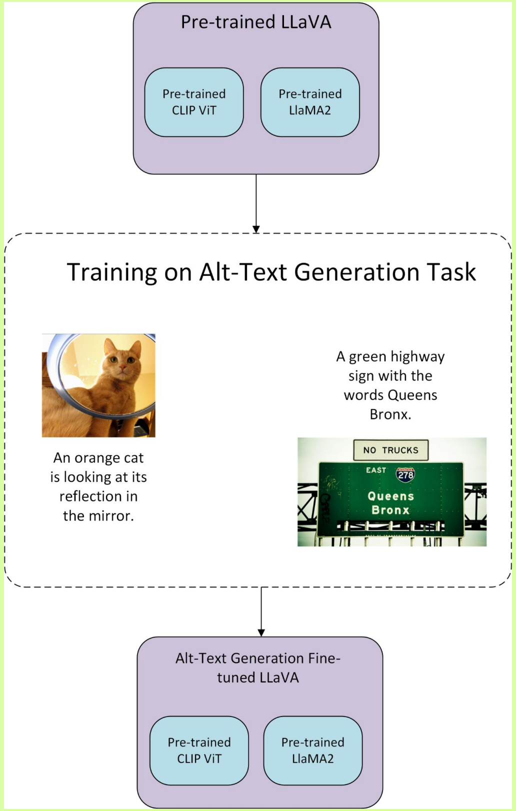
a man standing in front of a screen



a group of people standing next to each other

Solution

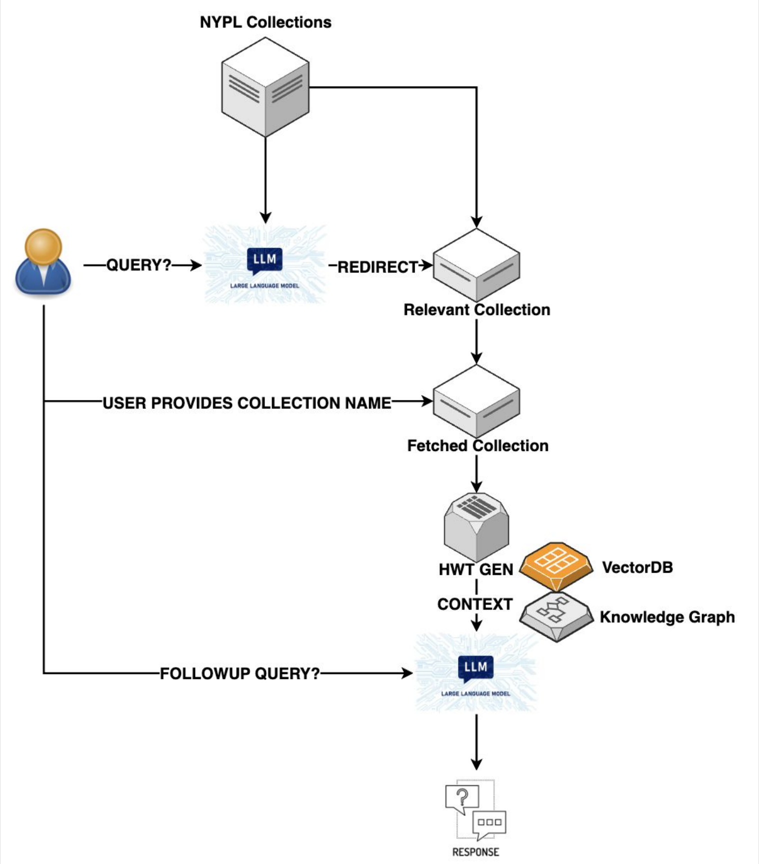
- **Model Architecture Modernization**
  - We integrate an **LLM** into ATGen. LLMs are pre-trained on massive datasets, giving them a sophisticated understanding of the world.
  - We use a **multimodal model** (LLaVA) to combine visual context/input with this advanced understanding to generate rich and meaningful descriptions
- **Fine-tuning**
  - The model is **fine-tuned** on the LAION-COCO dataset, which is known to give thorough descriptions of images.
- **Test Set Creation**
  - Model performance was evaluated on LAION-COCO dataset and tracked for future performance evaluations and comparisons.


Pre-trained LLaVA
Pre-trained CLIP ViT
Pre-trained LlaMA2
Training on Alt-Text Generation Task
An orange cat is looking at its reflection in the mirror.
A green highway sign with the words Queens Bronx.
NO TRUCKS
EAST
278
Queens
Bronx
Alt-Text Generation Fine-tuned LLaVA
Pre-trained CLIP ViT
Pre-trained LlaMA2

LLM ATGen Live Demo

Q&A Bot

NYPL Collection Question Answering System



Q&A Bot Architecture

Text Processing Pipeline

1. **Text Extraction**:
   Text is extracted from input collections using **PyPDF2** for PDF parsing.
2. **Text Chunking**:
   Extracted text is divided into manageable chunks of **1000 characters** with an overlap of **200 characters** using **LangChain's TextSplitter**.
3. **Embedding Generation**:
   Each chunk is embedded into dense vector representations using **Hugging Face's all-MiniLM-L6-v2** model.
4. **Knowledge Graph Integration**:
   - Extracted metadata is organized into a **Knowledge Graph** to provide relational insights.
   - The Knowledge Graph is queried to identify **context-aware relationships** between documents.
5. **Vector Store Creation**:
   Embeddings are stored using **FAISS** for efficient similarity search, enabling rapid and precise **semantic matching**.

- **Query Processing**
  - **Hybrid Retrieval**: Combines results from the **Vector Database (VectorDB)** and **Knowledge Graph** to enhance accuracy.
  - **Fallback Mechanism**: If no relevant matches are found, the system defaults to a **Large Language Model (LLM)** for response generation.
- **Conversational Memory**: Maintains query history using **LangChain's CustomConversationMemory**, ensuring context-aware and coherent follow-up responses.
- **Response Generation**
  - Integrates insights from the **Knowledge Graph**, **vector-based matches**, and generative text outputs to produce comprehensive answers.
  - Final responses include **generated text** along with relevant **source metadata** for transparency and traceability.

Query Processing Pipeline

Q&A Bot
Live Demo

Future Work

Contextual Enrichment

Integrate **pre-existing collection metadata** (collection name, time period, themes, etc.) into the context for alt-text generation to provide more descriptive and precise alt-text.

Dynamic Knowledge Extraction

Enable a feedback loop where new insights (like HWTGen's extracted handwritten text) are dynamically integrated into the context for re-evaluating or regenerating alt-text.

Multimodal Fusion

Combine multiple data inputs (images, extracted handwritten text, and metadata) into a **multimodal model pipeline** for richer context-aware alt-text generation.

Real-Time Collection Updates

Automate the pipeline to dynamically detect new collections and update the system with the latest data for all models.

Thank you!