Multilingual Retrieval-Augmented Generation (RAG) System Report

OCR to Create Text File:

- Used ocr_pdf.py to extract text from hsc26_bangla_1st_paper.pdf using pytesseract and pdf2image with Poppler.
- Saved the output as output.txt with a DPI of 300 for better quality.
- Faced formatting issues (e.g., broken sentences, mixed numbers like "109475৩") and fixed them with regex preprocessing.

Initial Attempt with Gemini:

- Tried ChatGoogleGenerativeAI (Gemini) from langchain-google-genai with FAISS and HuggingFace embeddings.
- Encountered limitations, prompting a switch to another model.

Switch to GPT with OpenAI API:

- I switched to ChatOpenAI (gpt-4o) using langchain_openai.
- Installed openai and langchain-openai, then upgraded LangChain packages to resolve conflicts (e.g., langchain-core, langsmith).
- Fixed LLMChain validation errors by configuring ChatOpenAI with a temperature parameter.

Troubleshooting and Iteration:

- Added a custom prompt to improve interpretive answers for the map_reduce chain.
- Successfully answered factual queries like "কাকে অনুপমের ভাগ্য দেবতা বলে উল্লেখ করা হয়েছে?".

Testing and Refinement:

- Tested Bangla queries, noting success with factual questions but struggles with interpretive ones (e.g., "অনুপমের ভাষায় সুপুরুষ কাকে বলা হয়েছে?").
- Explored solutions like better prompts and document context improvements.
-

English Queries:

Bangla Queries

Query:

কাকে অনুপমের ভাগ্য দেবতা বলে উল্লেখ করা হয়েছে?

Answer:

অনুপম তার মামাকে ভাগ্য দেবতা এবং ভাগ্য দেবতার প্রধান এজেন্ট বলে উল্লেখ করেছে।

Groundedness Score: 0.65

Relevance Score: 0.78

Source Snippets:

- সেন এক জোড়া এয়ারিং এগিয়ে দেন সেকরার হাতে তা পরখ করে দেখার জন্য। কেননা সেটা ছিল অনুপমের মামার দেওয়া বিলাতি জিনিস, যাতে সোনার ভাগ আছে সামান্যই। গ. উদ্দীপকের সবিতা ও অপরিচিতা গল্পের কল্যাণী উভয়েই যৌ...
- চলবে নিচের কোনটি সঠিক ক। খ। গ।1, 1 ঘ। 11, ৯৪। হরিশের বর্ণনায় মেয়ের বাবার পরিচয় সম্পর্কে জানা যায়- .এককালে তাদের বংশে লক্ষ্মীর মঙ্গলঘট উপুড় করা ছিল . দেশে বংশমর্যাদা রক্ষা করে চলা কঠিন বলে পশ...
- স্পষ্ট কথা বলার মতো সাহস তার নেই। নিজের সিদ্ধান্ত সে নিজে নিতে পারে না। সেকরা দিয়ে গহনা যাচাই যে কনেপক্ষের অপমান তা অনুপম বুঝতে পারে না। এতে তার ব্যক্তিত্বহীনতার চরম প্রকাশ লক্ষ করা যায়। এসব দিক বিচ...
- আমি অন্নপূর্ণার কোলে গজাননের ছোটো ভাইটি। মতো তিনি আমাদের সমস্ত সংসারটাকে নিজের অন্তরের মধ্যে শুষিয়া লইয়াছেন। তাহাকে না খুঁড়িয়া এখানকার এক গণ্ডষও রস পাইবার জো নাই। এই কারণে কোনো-কিছুর জন্যই আমাকে ক...
- মধ্য দিয়ে গল্পের শেষাংশে কল্যাণীর শুচিশুভ্র আত্মপ্রকাশও ভবিষ্যতের নতুন নারীর আগমনীর ইঙ্গিতে পরিসমাপ্ত। অপরিচিতা মনস্তাপে ভেঙেপড়া এক ব্যক্তিত্বহীন যুবকের স্বীকারোক্তির গল্প, তার পাপস্খালনের অকপট কথাম...
- মিল আছে ক হরিশের খ মামার গ শিক্ষকের ঘ বিনুর ৪1 উক্ত চরিত্রে প্রাধান্য পেয়েছে - 1 দৌরাত্ম 1 হীনম্মন্যতা 11 লোভ নিচের কোনটি ঠিক ক।। ও খ। ও গ।1ও ঘ।1,1 ও ৫. অনুপমের বয়স কত বছর ক পঁচিশ খ ছাব্বিশ গ সা...
- বলাও তিনি আবশ্যক বোধ করিলেন না। কারণ, প্রমাণ হইয়া গেছে, আমি কেহই নই। তারপরে যা ইহল সে আমি বলিতে ইচ্ছা করি না। ঝাড়লণ্ঠন ভাঙিয়া-চুরিয়া, জিনিসপত্র লণ্ডভণ্ড করিয়া, ও কন্নট একসঙ্গে বাজিল না এবং অভ্র...
- হয়েছে ঠিকই কিন্তু গল্পের পরিণতিতে বৃত্তভাঙা ভিন্ন এক ব্যক্তি হিসেবে তাকে পাওয়া যায়।- মন্তব্যটি যুক্তিযুক্ত। প্রতিষ্ঠিত হয় যখন সে সীমাবদ্ধতার গণ্ডি পেরিয়ে অসীমের সন্ধান পায়। তখন মানুষের চিত্ত হয়...
- কোন চরিত্রের ইঙ্গিতবহ ক হরিশ খ বিনুদা গ মামা ঘ শম্ভুনাথ উত্তর গ ৩৭। শাওনের কোন কোন বৈশিষ্ট্য অনুপমের মধ্যে থাকলে অনুপমের বিয়েটা টিকে যেত ।. সাহসিকতা . ব্যক্তিত্ব . গভীর ভালোবাসা নিচের কোনটি সঠিক ক 7,...
- সংকলন গল্পসপ্তক-এ এবং পরে, গল্পগুচ্ছ তৃতীয় খণ্ডে ১৯২৭। অপরিচিতা গল্পে অপরিচিতা বিশেষণের আড়ালে যে বলিষ্ঠ ব্যক্তিত্বের অধিকারী নারীর কাহিনি বর্ণিত হয়েছে, তার নাম কল্যাণী। অমানবিক যৌতুক প্রথার নির্মম ...

Ask another query

Figure 1

Query:

অনুপমের বয়স কত বছর?

Answer:

অনুপমের বয়স ছাব্বিশ বছর।

Groundedness Score: 0.62

Relevance Score: 0.74

Source Snippets:

- কারণে গ গয়না নিয়ে মনোমালিন্যের কারণে ঘ কনের বাবার আত্মগরিমার কারণে ৭৫। আমার পুরোপুরি বয়সই হলো না কথাটি দ্বারা কী বোঝানো হয়েছে ক তরুণ বয়সী খ অপরিণত বয়সী গ অতি নির্ভতশীল ঘ চিন্তায় অপরিণত ৭৬। তামা...
- চলবে নিচের কোনটি সঠিক ক।. খ।. গ।1, 1 ঘ।. 11, ৯৪। হরিশের বর্ণনায় মেয়ের বাবার পরিচয় সম্পর্কে জানা যায়- .এককালে তাদের বংশে লক্ষ্মীর মঙ্গলঘট উপুড় করা ছিল . দেশে বংশমর্যাদা রক্ষা করে চলা কঠিন বলে পশ...
- সেনাপতি কার্তিকেয় অপূর্ব শোভা পায়। বড়ো হয়েও অনুপম কার্তিকের মতো মায়ের কাছাকাছি থেকে মাতৃআজ্ঞা পালনে ব্যস্ত থাকে। তাই পরিণত বয়সেও তার স্বাধীন ব্যক্তিত্বের বিকাশ ঘটে না। পাঠ্য গল্পের অনুপম পরিবারত...
- সেন এক জোড়া এয়ারিং এগিয়ে দেন সেকরার হাতে তা পরখ করে দেখার জন্য। কেননা সেটা ছিল অনুপমের মামার দেওয়া বিলাতি জিনিস, যাতে সোনার ভাগ আছে সামান্যই। গ. উদ্দীপকের সবিতা ও অপরিচিতা গল্পের কল্যাণী উভয়েই যৌ...
- তেইশ, এখন হইয়াছে সাতাশ। এখনো আশা ছাড়ি নাই, কিন্তু মাতুলকে ছাড়িয়াছি। নিতান্ত এক ছেলে বলিয়া মা আমাকে ছাড়িতে পারেন নাই। তোমরা মনে করিতেছ, আমি বিবাহের আশা করি না, কোনো কালেই না। আমার মনে আছে, কেবল স...
- সংখ্যা ৬। অপরিচিতা গল্পে নায়কের বয়স কত বলা হয়েছে ক ২৮ বছর খ ২৬ বছর গ ২৭ বছর ঘ ২৫ বছর ৭ তবু ইহার বিশেষ মূল্য আছে এখানে কীসের মূল্যের কথা বলা হয়েছে ক জীবনের খ মরণের গ কর্মের ঘ ধর্মের ৮। ছেলেবেলায় প...
- আমি অন্নপূর্ণার কোলে গজাননের ছোটো ভাইটি। মতো তিনি আমাদের সমস্ত সংসারটাকে নিজের অন্তরের মধ্যে শুষিয়া লইয়াছেন। তাহাকে না খুঁড়িয়া এখানকার এক গণ্ডষও রস পাইবার জো নাই। এই কারণে কোনো-কিছুর জন্যই আমাকে ক...
- যাইতাম, কিন্তু মামাকে দমানো শক্ত। তিনি শম্ভুনাথবাবুর চুপচাপ ভাব দেখিয়া ভাবিলেন, লোকটা নিতান্ত নিজীব, একেবারে কোনো তেজ নাই। বেহাই-সম্পদায়ের আর যাই থাক, তেজ থাকাটা দোষের, অতএব মামা মনে মনে খুশি হইলেন।...
- মিল আছে ক হরিশের খ মামার গ শিক্ষকের ঘ বিনুর ৪1 উক্ত চরিত্রে প্রাধান্য পেয়েছে - 1 দৌরাত্ম 1 হীনম্মন্যতা 11 লোভ নিচের কোনটি ঠিক ক।। ও খ। ও গ।1ও ঘ।1,1 ও ৫. অনুপমের বয়স কত বছর ক পঁচিশ খ ছাব্বিশ গ সা...
- হাতে লইয়া দেখিলেন, এই এয়ারিং দিয়াই কন্যাকে তাহারা আশীর্বাদ করিয়াছিলেন। মামার মুখ লাল হইয়া উঠিল। দরিদ্র তাহাকে ঠকাইতে চাহিবে কিন্তু তিনি ঠকিবেন না এই আনন্দ-সম্ভোগ হইতে বঞ্চিত হইলেন এবং তাহার উপরেও ...

Ask another query

Figure 2

English Queries

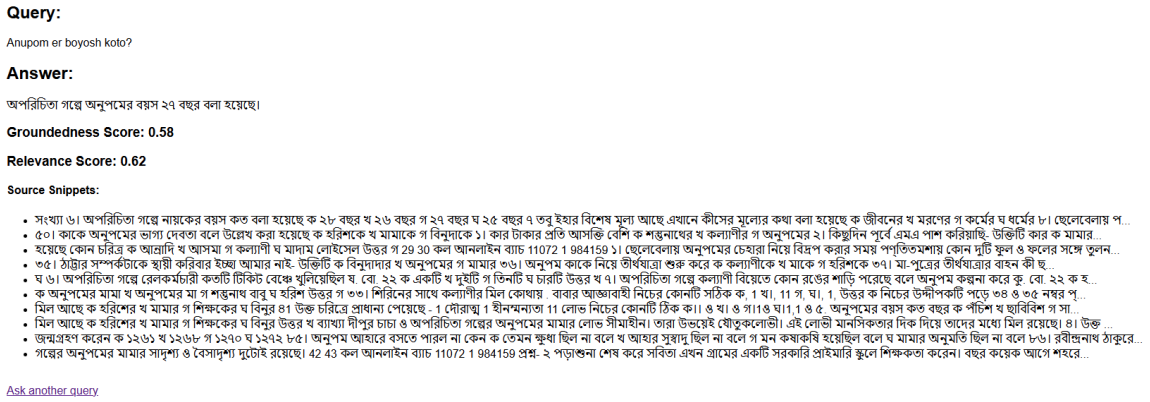

Figure 3

Used Tools and Libraries

- OCR: pytesseract and pdf2image (in ocr_pdf.py) with Poppler.
- FastAPI: Web app framework.
- LangChain: RAG implementation (langchain, langchain-openai, etc.).
- OpenAI: ChatOpenAI (e.g., gpt-4o).
- HuggingFace: sentence-transformers/paraphrase-multilingual-mpnet-base-v2 for embeddings.
- FAISS: Vector store.
- pdfplumber: PDF text extraction.
- NLTK: Text preprocessing.
- NumPy & scikit-learn: Similarity calculations.
- Pillow: Image handling for OCR.
- uvicorn: ASGI server.

API Documentation

- **/ (GET)**: Shows a query form.
- **/query (POST)**:
    - Params: query (string), language (e.g., "bn", "en").
    - Response: HTML with answer, scores, and snippets.
- **/reinit (GET)**: Re-initializes RAG chain, returns JSON {"message": "RAG chain re-initialized"}.

Evaluation Matrix

- **Groundedness Score**: Support from documents (0-1, cosine similarity mean).
- **Relevance Score**: Best query-chunk match (0-1, cosine similarity max).

Must Answer Questions

- **Text Extraction Method**:

    - Used pytesseract and pdf2image in ocr_pdf.py with Poppler for scanned PDFs.
    - Faced formatting issues (e.g., broken text), fixed with regex preprocessing.
- **Chunking Strategy**:
    - RecursiveCharacterTextSplitter (size 1500, overlap 300).
    - Works well for semantic retrieval with fragmented text, though larger chunks could help.
- **Embedding Model**:
    - sentence-transformers/paraphrase-multilingual-mpnet-base-v2.
    - Chosen for Bangla support and semantic accuracy, captures meaning via contextual vectors.
- **Query Comparison**:
    - Cosine similarity with FAISS storage.
    - Chosen for robustness and speed in high-dimensional spaces.
- **Meaningful Comparison**:
    - Embeddings map query and chunks to a shared space; custom prompt aids inference.
    - Vague queries return "দুঃখিত, এই প্রশ্নের উত্তর নির্ধারণ করা সম্ভব নয়।"
- **Relevance of Results**:
    - Relevant for factual queries, not interpretive ones.
    - Improvements: Larger chunks, xlm-roberta-base embedding, or more document context.