

Meeting Notes for SD-RAG-TEAM-16

These are the notes I took for Team 16. In case of any questions, please contact me at ngarablessing18@gmail.com or ngarable@iastate.edu

1 February 2025

Try out Ollama, AnythingLLM, and LM Studio.

8 February 2025

Compare all the Large Language Models you might think of eg, GPT 4.0, Gemma, Gemini, Perplexity AI, Grok, Llama, and make an evaluation sheet using Excel

15 February 2025

Compare reasoning models, eg, Deep Seek, GPT 4.5 and make an evaluation sheet using Excel.

22 February 2025

Plan for continue on the future week of our AI RAG Library search

- install n8n, by either using Docker desktop, or npm (npm will be easier to deal with host port)

<https://docs.n8n.io/hosting/installation/npm/>

- see the video for sample, I showed the n8n setup at the later part TRIMED - AI-search - Digital Repository - n8n .mp4

- test the demo from n8n, make it run

- try the workflow on RAG Movie recommendation

<https://n8n.io/workflows/2440-building-rag-chatbot-for-movie-recommendations-with-qdrant-and-open-ai/>

- modify the data preprocessing part. Use our metadata cab file instead.

- design preprocessing steps to run through each row from the csv file, process it and save the title, keywords, authors, and any others columns that you think important, to a vector database.

- use chat interface to search.

26 February 2025

Watch this tutorial

A good, complete A-Z tutorial on n8n.

<https://www.youtube.com/watch?v=uScURRX-Knc>

3 March 2025: Preparation of Demo1

This is a rough draft of the slides Blessing made

<https://docs.google.com/presentation/d/13MxCYDIWUliC42jetKBKyq1HeLa1qouqCkIbXSm9gSk/edit?usp=sharing>

5 March 2025

Workflow template: better sample n8n workflow - AI Powered RAG Chatbot for Your Docs + Google Drive + Gemini + Qdrant


Workflow template: This workflow is almost similar to what we are doing:

AI Powered RAG Chatbot for Your Docs + Google Drive + Gemini + Qdrant:

<https://n8n.io/workflows/2982-ai-powered-rag-chatbot-for-your-docs-google-drive-gemini-qdrant/>

another one, more simpler, but using pinecone instead of Qdrant: RAG Chatbot for Company Documents using Google Drive and Gemini

<https://n8n.io/workflows/2753-rag-chatbot-for-company-documents-using-google-drive-and-gemini/>

 AI-Powered RAG Chatbot with Google Drive Integration

This workflow creates a powerful RAG (Retrieval-Augmented Generation) chatbot that can process, store,

14 March 2025

Task for the next meeting (after break):

Watch the Team 16 meeting video

Each team member select your own 10 documents from

<https://dr.lib.iastate.edu/search?page=1&query=&f.department=Computer%20Science,equal&spc.page=1&scope=>

Each team member individual implementation of the tutorial, using Gemini or OpenAI, or other models (OpenRouter, Groqcloud, etc)

<https://n8n.io/workflows/2982-ai-powered-rag-chatbot-for-your-docs-google-drive-gemini-qdrant/> sign-up and use your own key for Google Drive, Qdrant. They are free anyway.

Each person approach of embedding document into vector database

Result of retrieval from chat models

Discussion of those results

No meeting this week due to Spring Break.

27 March 2025 - Week 10

Here are the next task list:

- try Gemini 2.5 pro experiment, and Gemini 1.5 Pro that have 2M large context window embedding model
- focus on Gemini model over OpenAI models, since we will be hosting on Google Cloud infrastructure, it would be faster.
- try make correct metadata to incorporate with document
- let's some reasoning model? compare the different
- try enforce (in the prompt, or in the system prompt) the chat response to structure with (clickable) URL links to the original documents. Since we are creating AI-based search, we want answer to look like Perplexity result

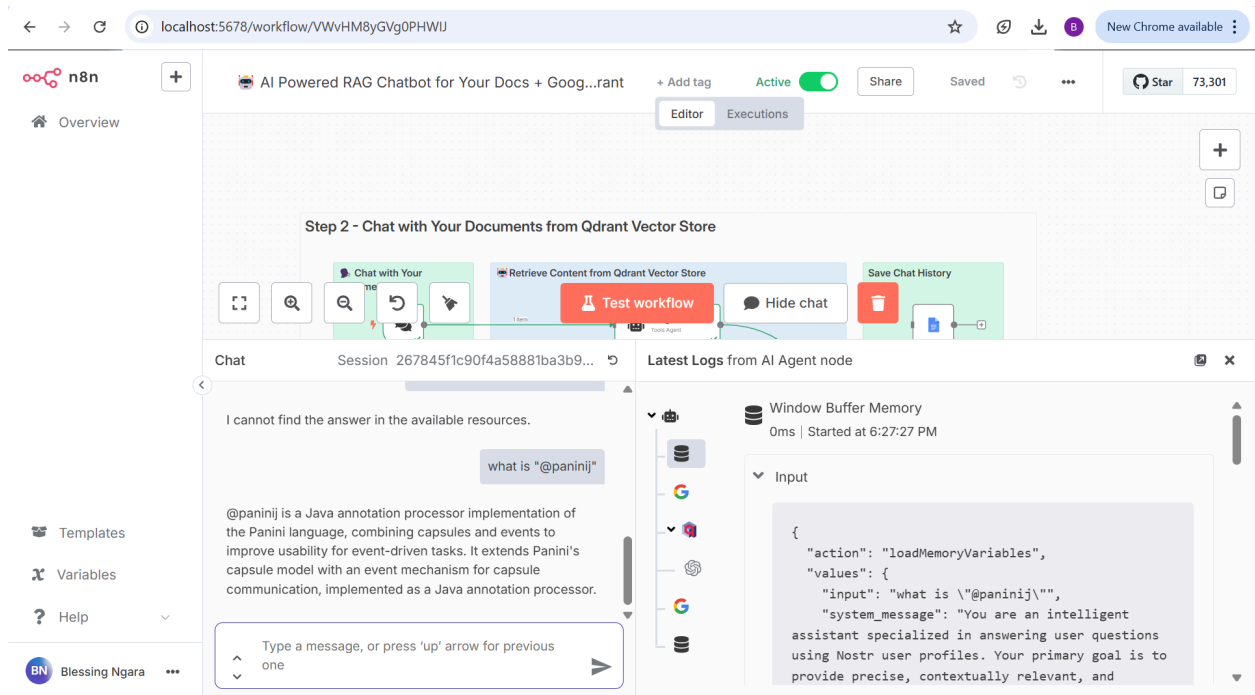
31 March 2025

Blessing Progress Update

Hi all! My workflow is now working. The maximum number of tokens is 8192. I can ask about specific authors from the 10 PDFs in Google Drive, e.g., "Who is Jack Maddox?" or about specific concepts, such as "What is @paninij?" I have attached some screenshots

The screenshot displays a web-based interface for an AI-powered RAG chatbot. The top navigation bar includes the user profile 'n8n', a '+ Add tag' button, an 'Active' toggle switch, 'Share', 'Saved', and a 'Star' button with a count of 73,301. Below the navigation, there are tabs for 'Editor' and 'Executions'. The main workspace shows a workflow diagram titled 'Step 2 - Chat with Your Documents from Qdrant Vector Store'. The workflow consists of several nodes: 'Chat with Your Documents', 'Retrieve Content from Qdrant Vector Store', 'Test workflow' (highlighted in red), 'Hide chat', and 'Save Chat History'. Below the workflow, there is a chat window with a session ID '267845f1c90f4a58881ba3b9...'. The chat history shows a user asking 'Hi, how can I help with Nostr or Damus user profiles?' and 'who is jackson maddox'. The chatbot responds with a detailed paragraph about Jackson Maddox. The user then asks 'who is hridayesh rajan?'. The chat input field contains the text 'Type a message, or press 'up' arrow for previous one'. To the right of the chat window, there is a 'Latest Logs from AI Agent node' section showing a 'Window Buffer Memory' log with an input object:

```
{  "action": "loadMemoryVariables",  "values": {    "input": "what is \@paninij\"",    "system_message": "You are an intelligent assistant specialized in answering user questions using Nostr user profiles. Your primary goal is to provide precise, contextually relevant, and
```



3 April 2025

Meeting notes

- continue to create a complete metadata file to compliment the pdf files. need some ID to match between them. Could be the original pdf file name or an ID.
- please to complete fill all metadata field (Title, Abstract, etc) into the csv file (like Tyler)
- try comparing Genimi 1.5 Pro (2m context windows) vs. Gemini 2.0 and Gemini 2.5, vs openAI model gpt-4o
- make search refer to a portion of the page (like Blessing work)
- make sure to have correctly clickable links back to the original documents on <https://dr.lib.iastate.edu/>

10 April 2025

- use the long id as primary key
- put all of your files together (those 10 documents of each person) in to a joined Google Drive folder
- decide and finalized the Qdrant format
- in the prompt, dictate the system prompt to display the "Title" of the document first, then "Authors", date, Subject Categories, Keywords, collections, Permanent Link (try make it click-able).
- start to join and use the single Qdrant Vector store and single API
- Continue on using python to extract all of the text from pdf file. Now with 40 files. Add them to the Qdrant

- Design set of prompts to test this RAG system. Read and follow this guideline:
<https://cloud.google.com/blog/products/ai-machine-learning/optimizing-rag-retrieval>

24 April 2025

please create a RAG testing question set, then test your RAG system. Manually give them score. **Ngara, Blessing [COM S]**
finish the Verifier model. that's a good idea. maybe create a loop of the verifier model decline the answer, ask user to confirm, then loopback to the beginning and generating a new message.
Manojkumar, Kausshik
run this system on the Google Cloud Munjuluri, Tara

1 May 2025

- Make the prototype work on <https://adisak2.app.n8n.cloud>
- sample chat interface
<https://adisak2.app.n8n.cloud/webhook/5f1c0c82-0ff9-40c7-9e2e-b1a96ffe24cd/chat>
- Ngara, Blessing [COM S] please link your chat interface on github to this endpoint
- Munjuluri, Tara please deploy your chat interface on Google Cloud, use Cloud Run service
- you all can duplicate and create your own version of workflow
- Only step 2 part that need to be active workflow. step 1 doesn't need to be active
- on step 1, add the process to check if the document is already in the Qdrant already, if yes, just skip to the next document. We don't need to keep embed the same document over and over.
- research on MCP, and MCP on n8n. It should help create a better way to communicate with LLM, and we shouldn't have to keep adding longer System Prompt

6 May 2025

Met in the Library and divided tasks for Demo3

Blessing: WHAT all was done? Group into MAJOR accomplishments and Minor parts. Which parts of the work does your team feel proud of?