



H2O.ai Generative AI Workshop I: H2OGPT & RAG



H2O.ai

Agenda

- Lab 1: Prompt Engineering
 - H2OGPT Intro
 - Prompt Design in HR Hiring
 - Parameter Tuning in HR Employee Onboarding
 - SingTel GPT
- Lab 2: RAG
 - H2OGPT Enterprise Intro
 - Role Prompting in Customer Service Agent
 - Information Retrieval & Math Operation in Procurement
 - Guardrail with Sensitive Info
- Python Client & Notebook
- Q&A



Housekeeping

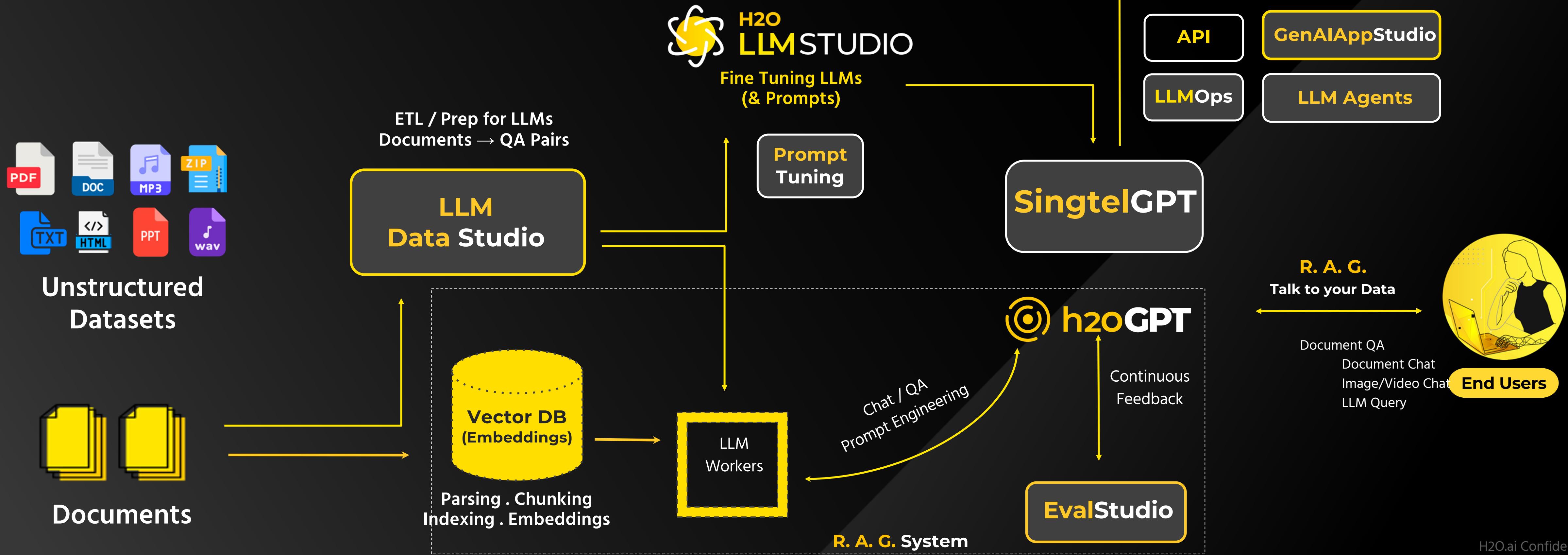
Please make sure you are able to access the following portals:

For Lab 1, please visit <https://llama.h2o.ai> / <https://gpt.h2o.ai>

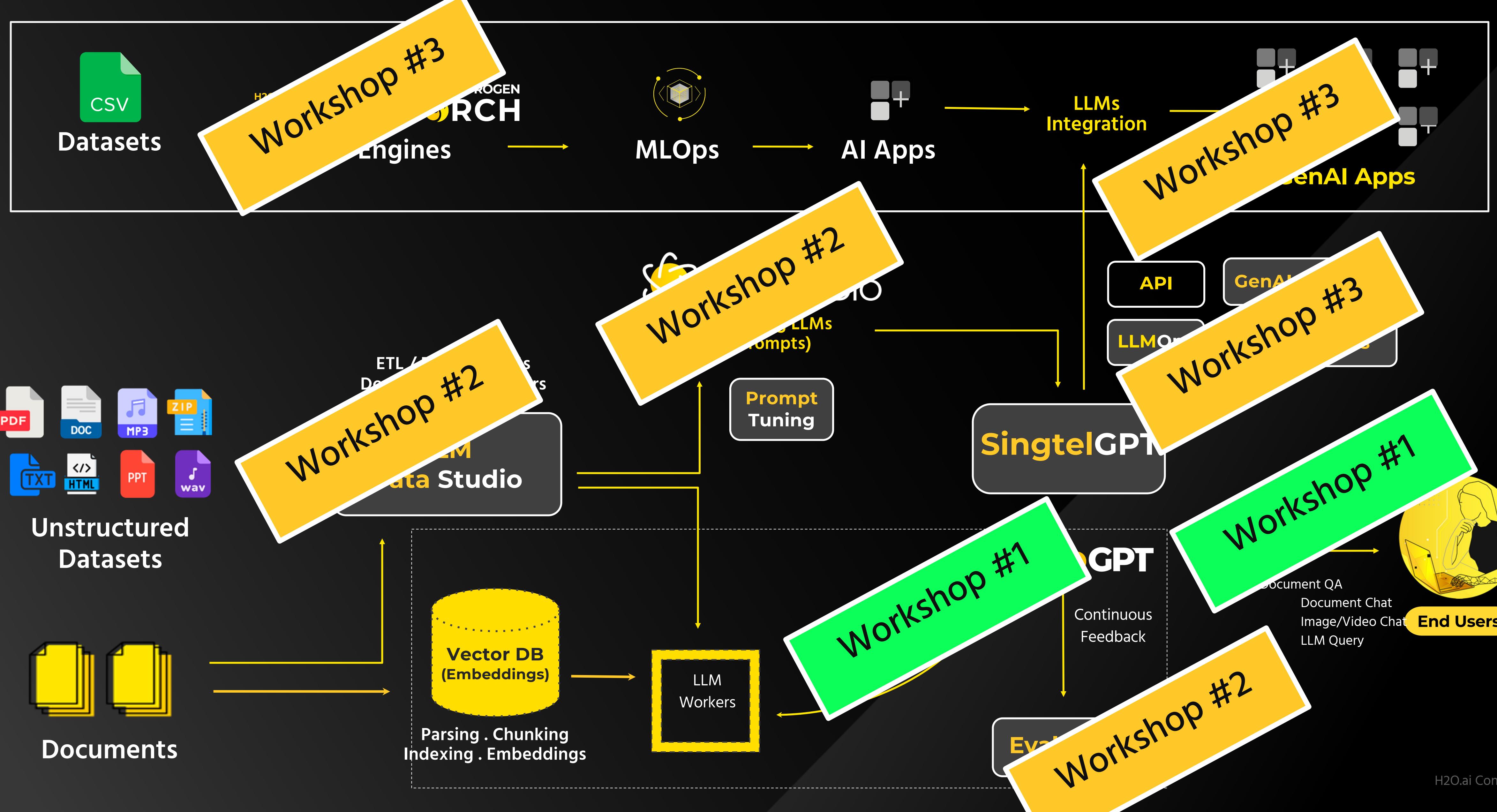
For Lab 2, please visit <genai-training.h2o.ai> and log in with the provided temp username & password

For documents, please visit the github public repo for the files.
https://github.com/tlsandbox/llm_workshop_1_public

Foundations of a GenAI Ecosystem



Foundations of a GenAI Ecosystem



H2OGPT

Democratize LLMs



Basic Operation

Let's Try it Out

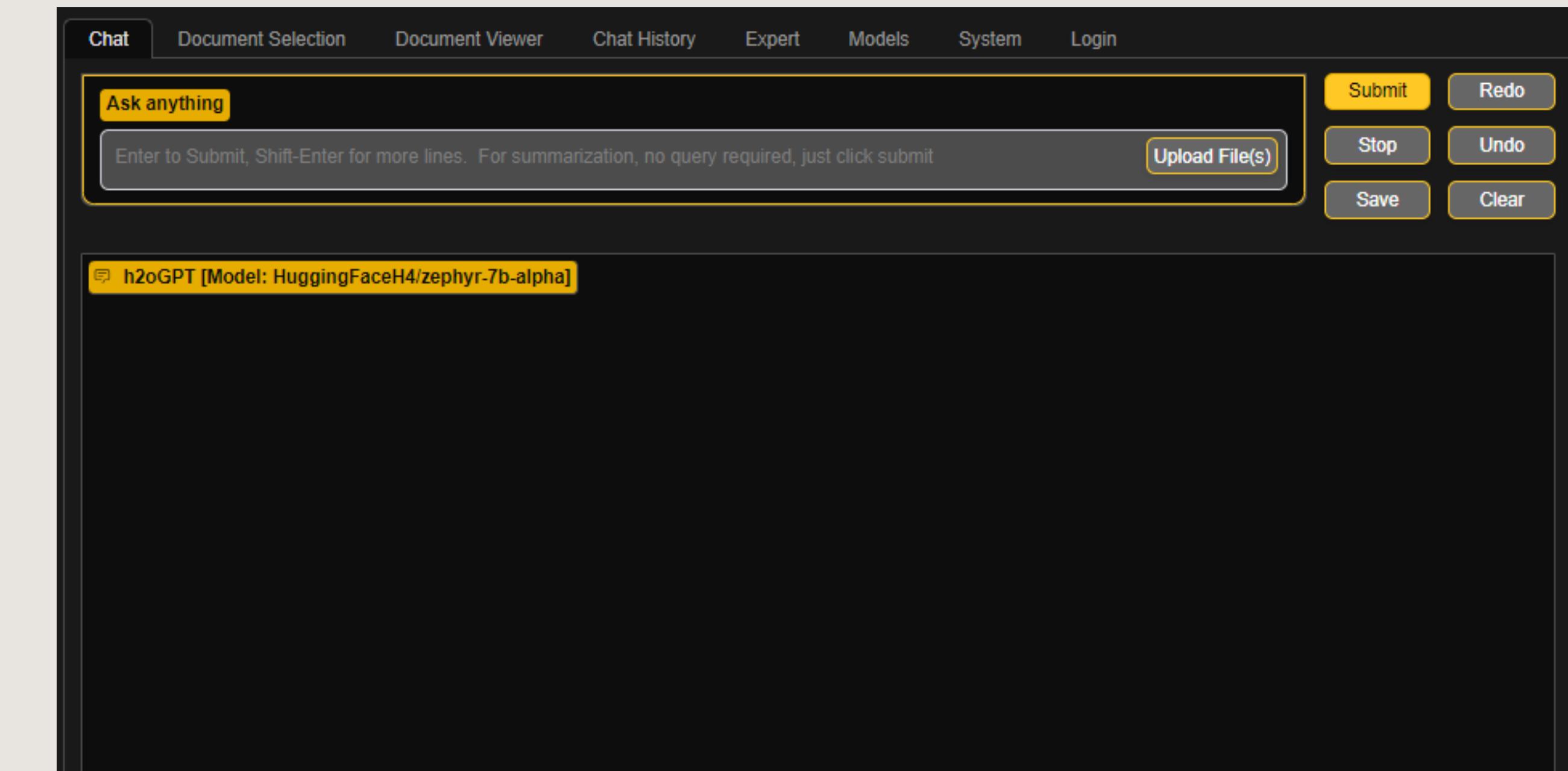
LLM ([Start HERE](#)): Single query-response, no chat context or docs used

Simply Input some following paragraphs and click Submit:

“how many public holidays in Singapore?”

“write a SoW template for telco company”

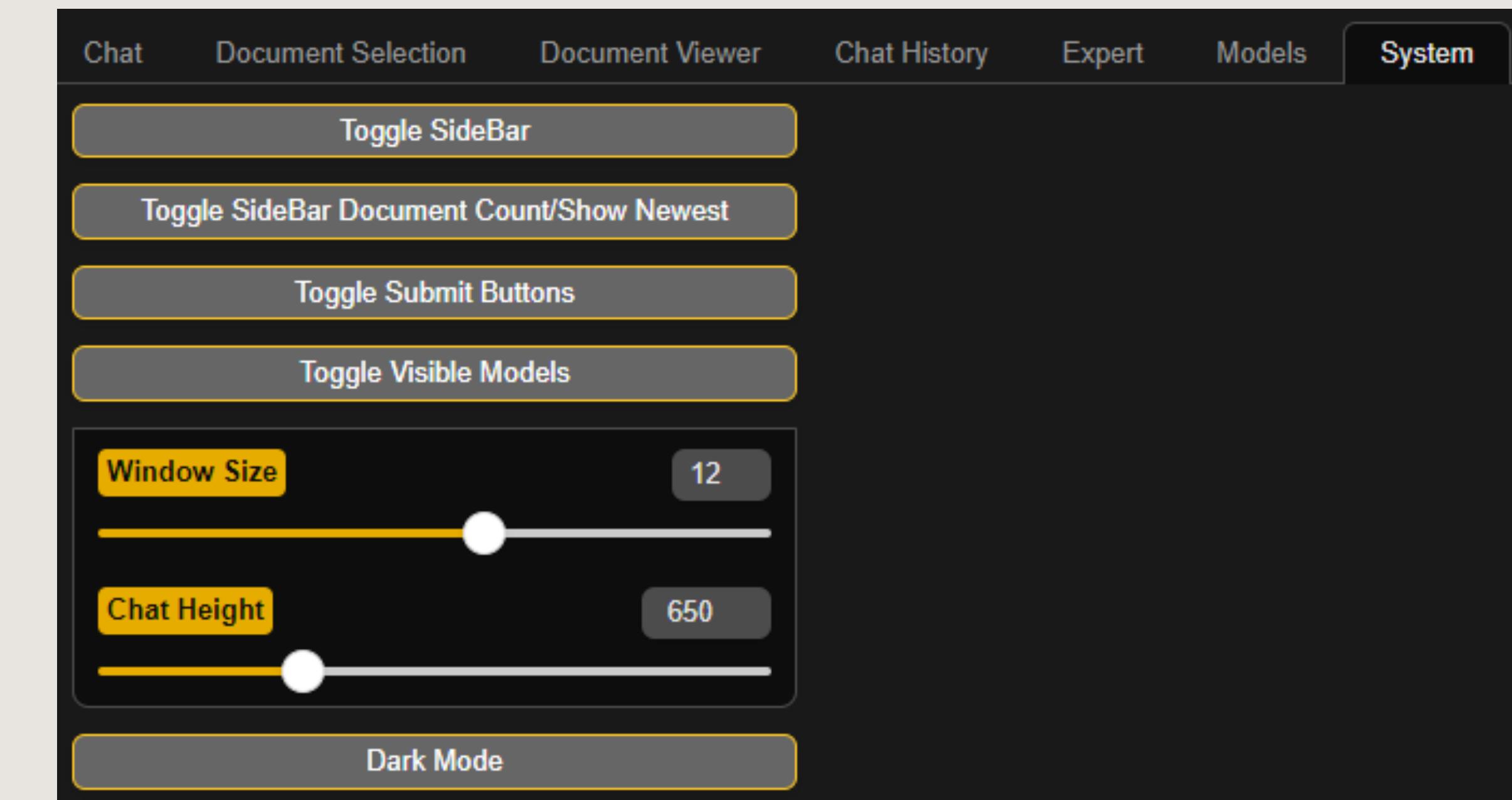
“how to handle employee complaint from HR counseling perspective?”



Basic Operation

Let's Try it Out

- UI Control
- For simplicity, can toggle the SideBar Document First
- Dark Mode
- Window Size & Chat Height





Lab 1: Prompt Engineer

Prompt Elements

A prompt contains any of the following elements:

- **Instruction** - a specific task or instruction you want the model to perform
- **Context** - external information or additional context that can steer the model to better responses
- **Input Data** - the input or question that we are interested to find a response for
- **Output Indicator** - the type or format of the output.
- You do not need all the four elements for a prompt and the format depends on the task at hand. We will touch on more concrete examples in upcoming guides.

Practical Prompt Engineering Strategies

General Prompting Tips, Tricks, and Hacks:

- **Use context effectively & Be specific**
- **Iterate and refine**
- **Keep it personal**
- **Intent & Identity**
- **Prevent hallucination**

Lab: HR Hiring & Onboarding



Lab: HR Hiring & Onboarding

- Write a Job Description
- Create a job Ad
- Finding alternative job titles for sourcing
- Candidate outreach emails
- Offering Letter
- New hire survey questions

Lab: HR Hiring & Onboarding

- Write a Job Description

#1: Use context effectively & Be specific

“

Generate a detailed job description for a data scientist position with a focus on machine learning and AI. The description should **include a job overview, key responsibilities, required skills and qualifications**, and information about the team;

“

Lab: HR Hiring & Onboarding

- Create a job Ad

#2: Iterate and refine

“

Compose a compelling job advertisement for a Senior Web Developer position at a fast-growing e-commerce startup. The ad should include an enticing headline, an overview of the role, key responsibilities, necessary skills and qualifications, and the perks of working at our company. **Use language that is inclusive and appealing to a diverse audience**

“

Lab: HR Hiring & Onboarding

- Finding alternative job titles for sourcing

Lab: HR Hiring & Onboarding

- Candidate outreach emails

#3: Keep it personal

“

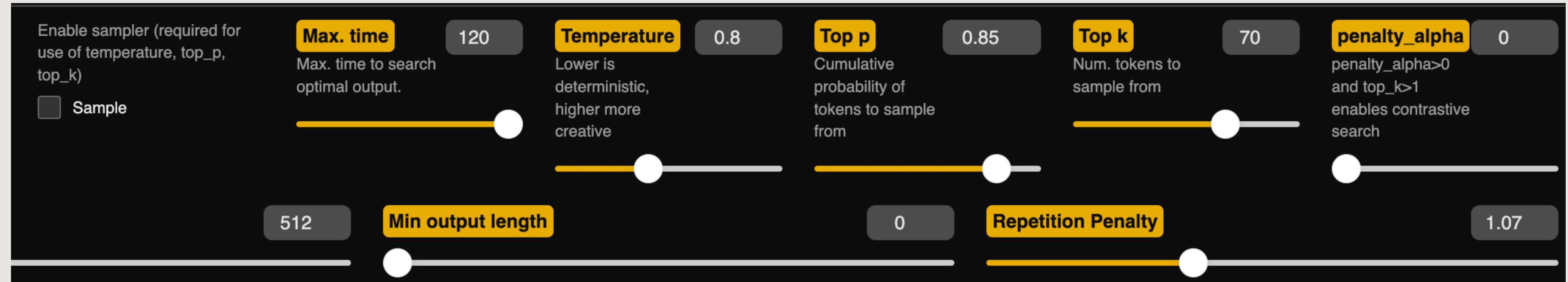
Craft a personalized candidate outreach email for a potential candidate named XXX, **who is currently a Senior Graphic Designer at a well-known digital marketing agency**. The purpose of this email is to **invite him to consider an open role** for an Art Director at our innovative advertising firm. Highlight our creative environment, professional growth opportunities, and competitive benefits package. The tone should be professional yet friendly.

“

Parameters Details

Advanced- Expert Tab

- **Sample:** Whether to enable sampling (required for use of temperature, top_p, etc)
- **Temperature** The lower the temperature, the more deterministic the result. Increasing temperature could lead to more randomness, which encourages more diverse or creative outputs
- **Top_P:** If looking for exact and factual answers keep this low. The more diverse responses, the higher the value.
- **Repetition Penalty:** Reduces the repetition of words in response by giving tokens appear more a higher penalty



Lab: HR Hiring & Onboarding

- Candidate outreach emails

I'd love to schedule a call to discuss the role in more detail and answer any questions you may have. Please let
Thank you for considering this opportunity, and I look forward to hearing from you soon.

Best regards,

[Your Name]

[Your Position]

[Company Name]

[Contact Information]

P.S. I saw your recent project for [Client] and was really impressed by the creative direction you took. It's clear that you have a unique perspective and a deep understanding of what makes a campaign successful. I think you'd be a valuable addition to our team and would love to discuss how you can contribute to our clients' success.

**With Temperature 0.7
With Top_P. 0.8**

"

P.S. I saw your recent project for [Client] and was really impressed by the creative direction you took. It's clear that you have a unique perspective and a deep understanding of what makes a campaign successful. I think you'd be a valuable addition to our team and would love to discuss how you can contribute to our clients' success.

"

Lab: HR Hiring & Onboarding

- Write a Salary Negotiation Email

Advanced Parameter Tuning Use Case Summary

	Temperature	Top_P	Description
Code Generation	0.2	0.1	Generates code that adheres to established patterns and conventions . Output is more deterministic and focused. Useful for generating syntactically correct code.
Creative Writing	0.7	0.8	Generates creative and diverse text for storytelling. Output is more exploratory and less constrained by patterns.
Chatbot Responses	0.5	0.5	Generates conversational responses that balance coherence and diversity . Output is more natural and engaging.
Data Analysis Scripting	0.2	0.1	Generates data analysis scripts that are more likely to be correct and efficient . Output is more deterministic and focused.

Advanced Parameter Tuning

Save for Later

	Temperature	Top_P	Description
Code Generation	0.2	0.1	Generates code that adheres to established patterns and conventions . Output is more deterministic and focused. Useful for generating syntactically correct code.

“

Why won't the following SQL statement deliver the correct number of transactions per user?
SELECT userId, max(transaction_id) FROM transactions GROUP BY userId;

“

Advanced Parameter Tuning

Save for Later

	Temperature	Top_P	Description
Chatbot Responses	0.5	0.5	Generates conversational responses that balance coherence and diversity . Output is more natural and engaging.

“

List a few of the most similar events in history to the 2008 financial crisis and the similarities and differences between those events and the 2008 financial crisis.

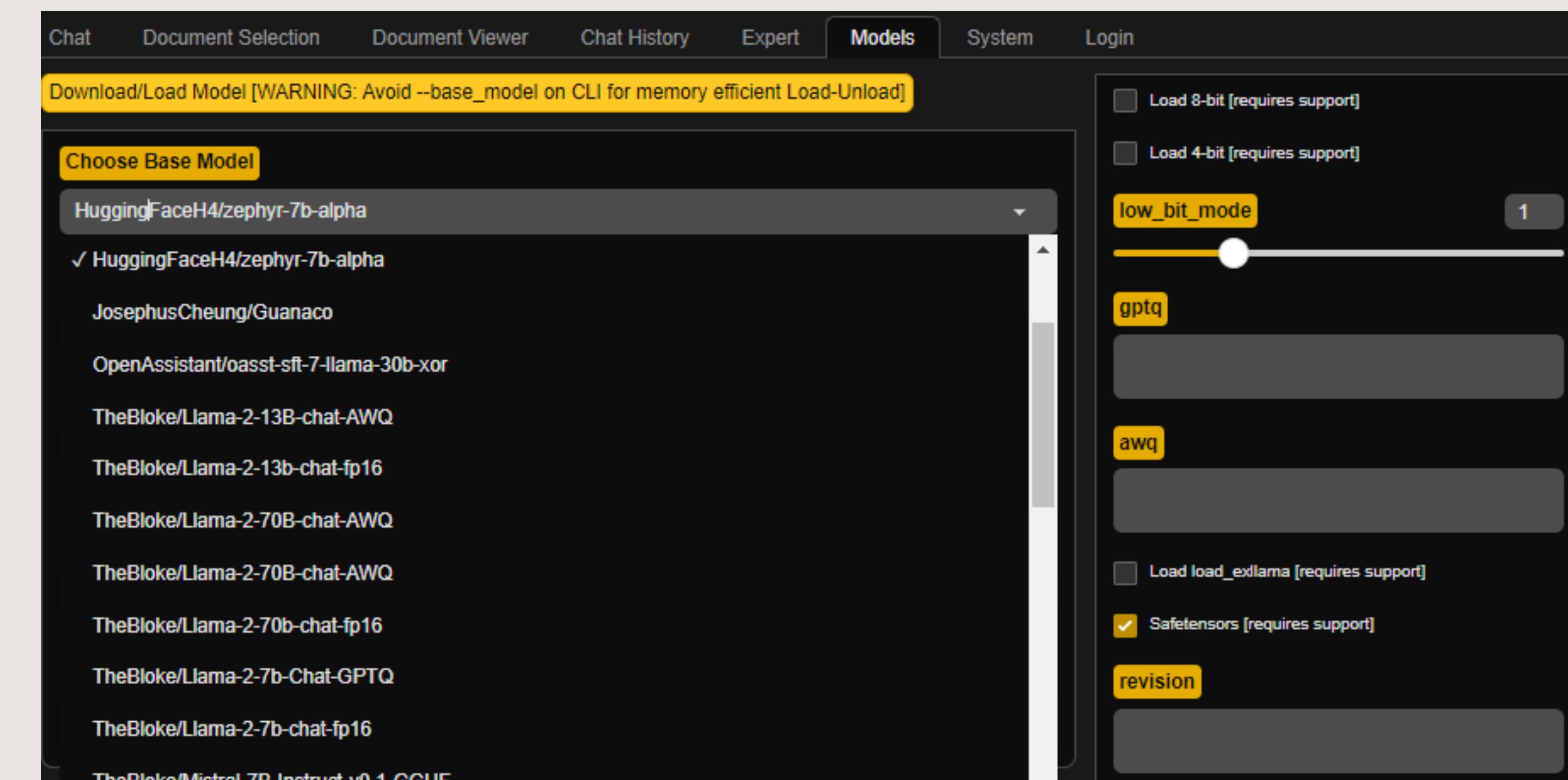
“



Open LLM Leaderboard

Intermediate- Models Tab

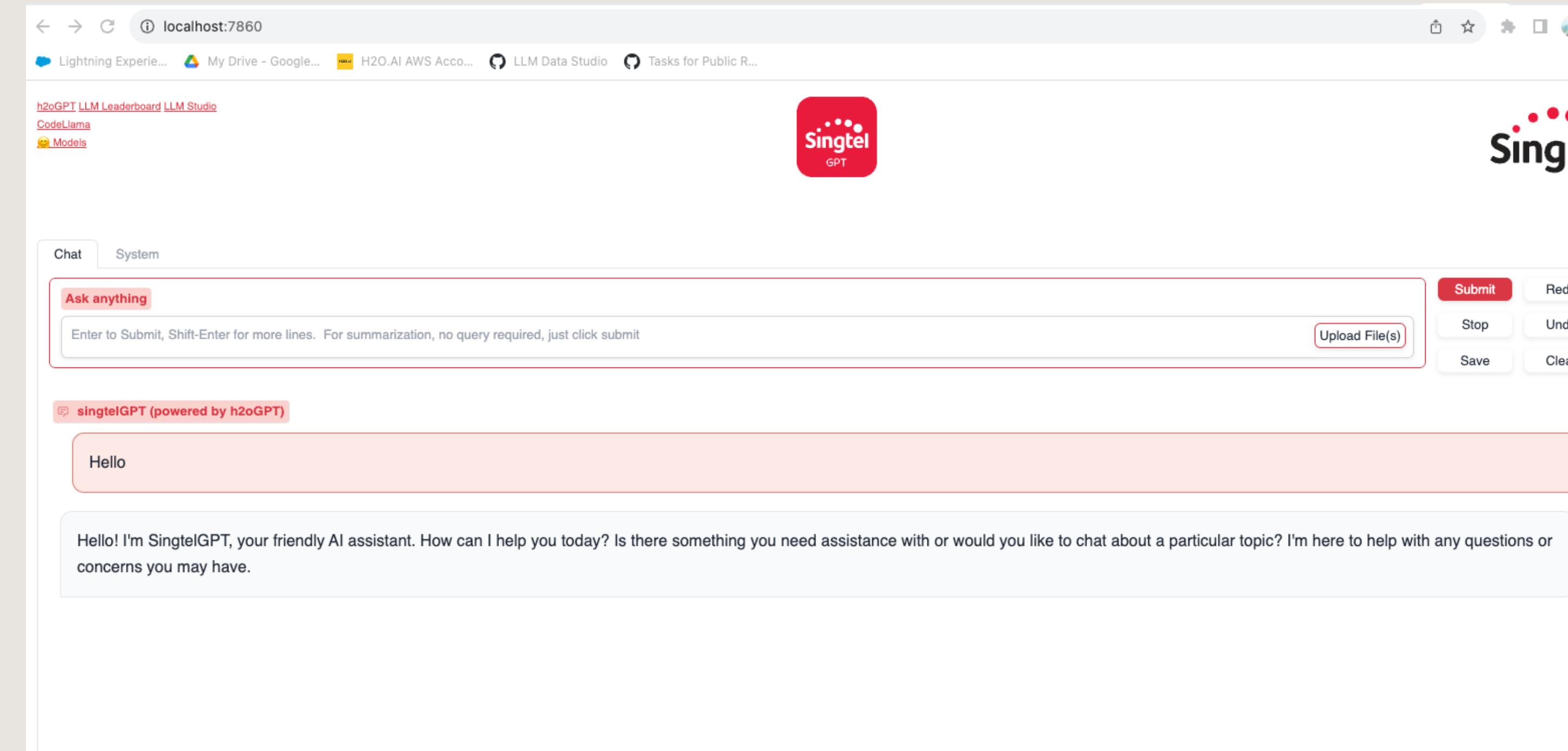
- Control model, LORA, or inference server used.
- To unload a model, use "Choose Model" and select "[None/Remove]".



Important: Unloading only works properly if did not pre-load model with --base_model and only selected model and clicked load.Window Size & Chat Height

SingTel GPT

- Active development on : [h2oai/h2ogpt](#), which under Apache-2.0 license
- **Prompt engineering** - ability to prepare open-source datasets for tuning LLMs
- **Tuning:** Code for fine-tuning large language models (**Zephyr, Llama-2, Vicuna;**) on commodity hardware and enterprise GPU servers (**single or multi node**)



Comprehensive Prompt Materials

Prompt Engineer Guide



Prompt Engineering Course Services About

Search...

⌘ K



Prompt Engineering

Introduction

LLM Settings

Basics of Prompting

Prompt Elements

General Tips for Designing Prompts

Examples of Prompts

Techniques

Zero-shot Prompting

Few-shot Prompting

Chain-of-Thought Prompting

Self-Consistency

Generate Knowledge Prompting

Tree of Thoughts

Introduction > General Tips for Designing Prompts

On This Page

Start Simple

The Instruction

Specificity

Avoid Imprecision

To do or not to do?

General Tips for Designing Prompts

Here are some tips to keep in mind while you are designing your prompts:

Start Simple

As you get started with designing prompts, you should keep in mind that it is really an iterative process that requires a lot of experimentation to get optimal results. Using a simple playground from OpenAI or Cohere is a good starting point.

You can start with simple prompts and keep adding more elements and context as you aim for better results. Iterating your prompt along the way is vital for this reason. As you read the guide, you will see many examples where specificity, simplicity, and conciseness will often give you better results.

When you have a big task that involves many different subtasks, you can try to break down the task into simpler subtasks and keep building up as you get better results. This avoids adding too much complexity to the prompt design process at the beginning.

The Instruction

Question? Give us feedback →

Edit this page

⊕ English





Break



Lab 2: Retrieval Augmented Generation (RAG)



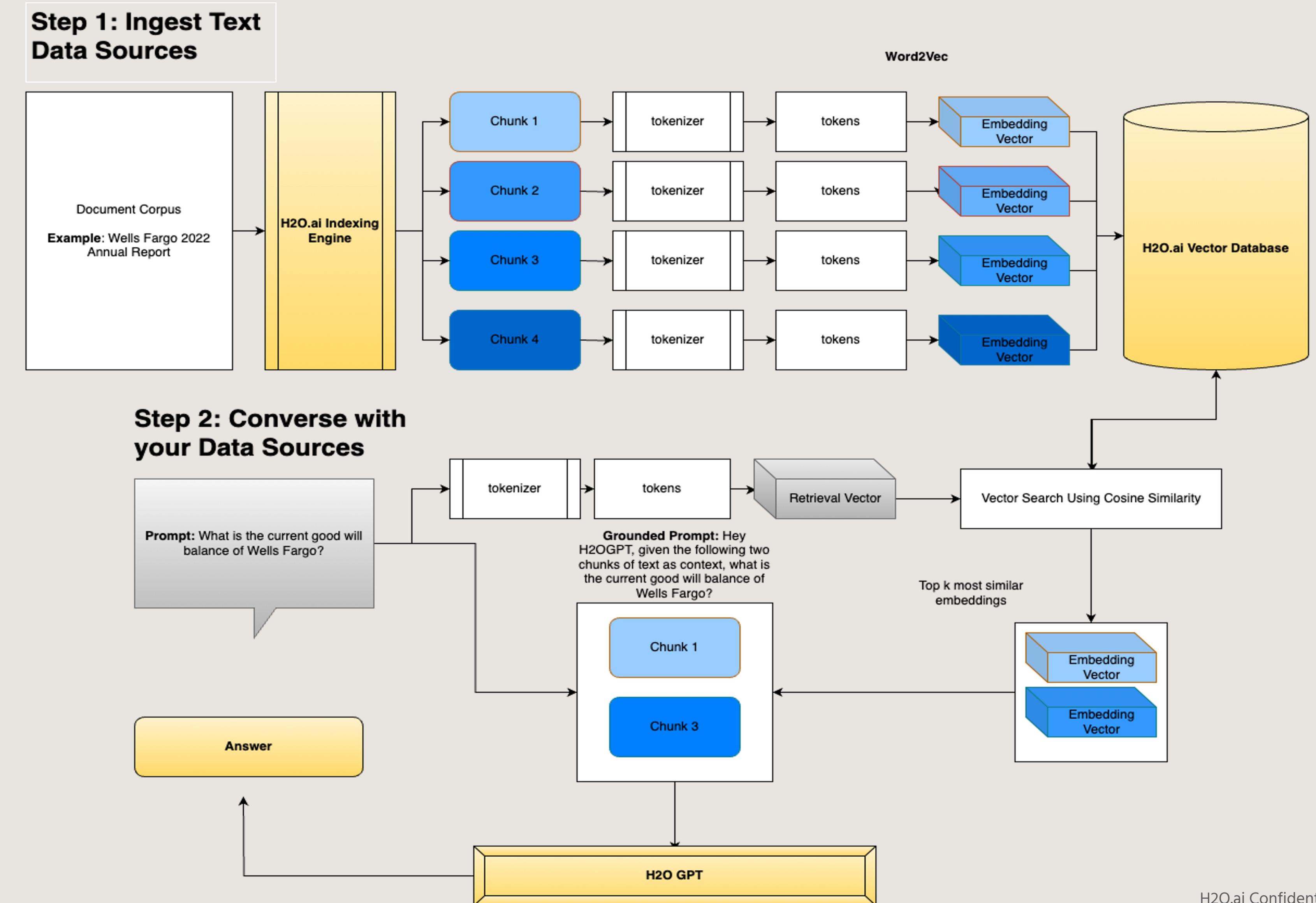
H2O EnterpriseGPT

Retrieval-Augmented Generation (RAG)

RAG as a system is a particularly good use of vector databases.

RAG systems take advantage the context window for LLMs, filling it with only the most relevant examples from real data.

This “grounds” the LLM to relevant context and greatly minimizes any hallucination.

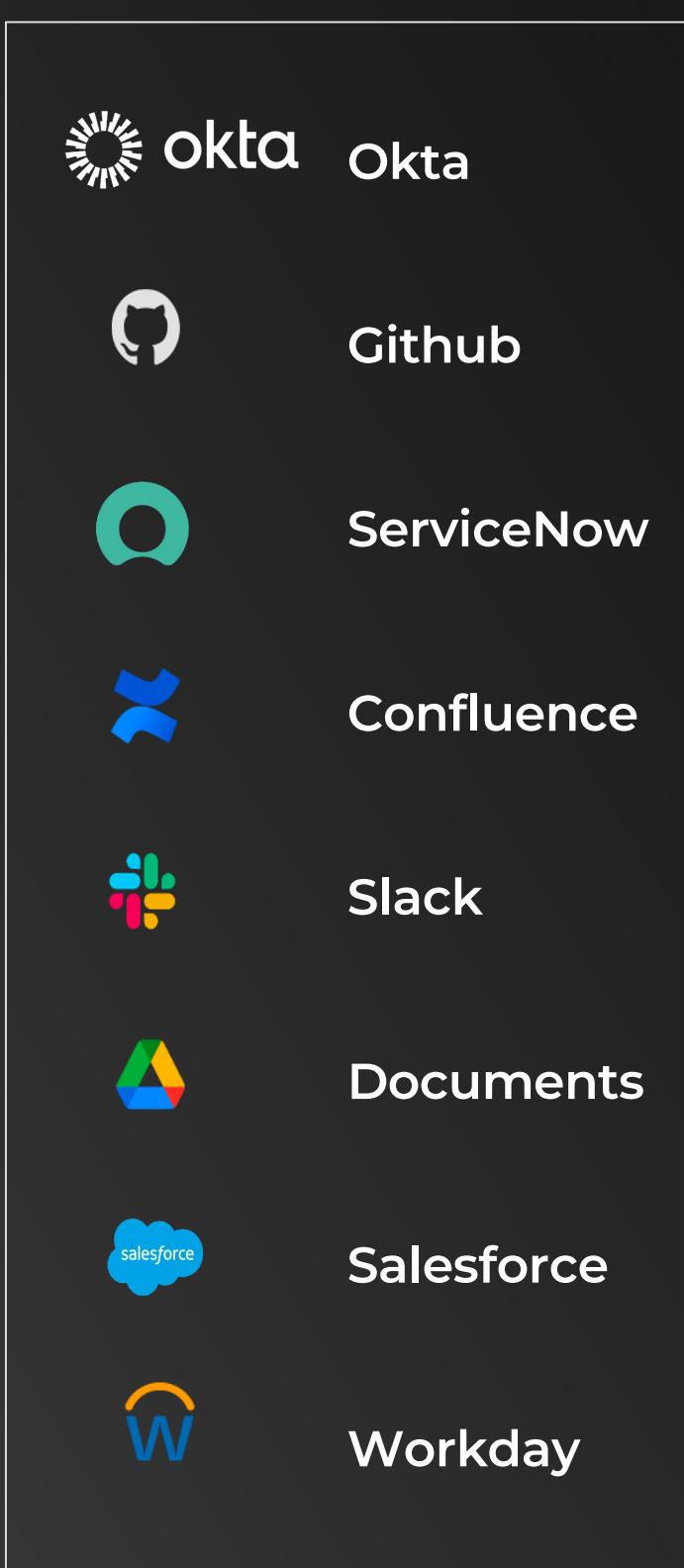


Enterprise h2oGPT

How It Works: Information retrieval on internal data

1 Drag and Drop Data

Bring data in from cloud, big data and desktop systems



2 Summarization

Data is automatically chunked, summarized, and indexed on import. View document and catalog summaries at any time to learn what's in your data.

3 Search

Get search results instantly for fact-based answers to questions about your data

4 Q&A

LLMs create answers for your search terms helping you fully understand your queries

5 Pipelines

APIs allow you to integrate search and chat into your own applications

Enterprise Administration

Admin collections of standard datasets with role-based access control, ability to connect to OpenAI, define the right LLM for the right data collection, and more...

backed by open source h2oGPT

Privately hosted LLM behind the scenes - your data stays with you

Prompt engineering to reduce hallucinations

H2O.ai is committed to building Enterprise grade products while supporting the open source community. See our h2oGPT repo for research information on building the best Open Source LLMs that support our Enterprise suite.

See our white paper for more details on our approach to open source LLM research and development:
arxiv.org/abs/2306.08161

Log-In



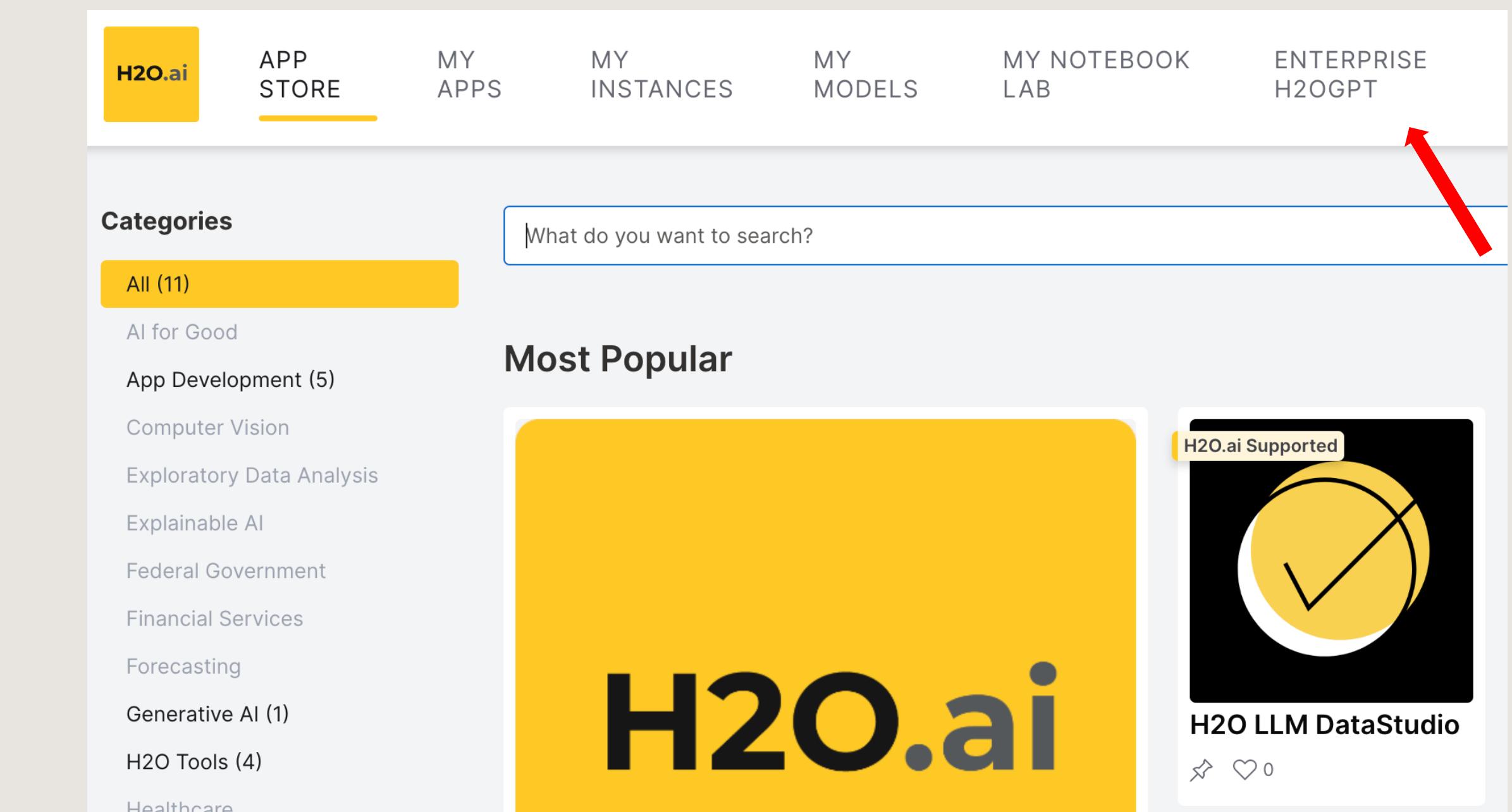
Lab: Customer Service Agent



<https://genai-training.h2o.ai/appstore>

Create Your First RAG Collection

- #1: Folder “Call Center Agent” => Download “Customer Service Live Chat Script”
- #2: Click “Enterprise H2OGPT”



<https://genai-training.h2o.ai/appstore>

Create Your First RAG Collection

- #1: Download “Customer Service Live Chat Script”
- #2: Click “Enterprise H2OGPT”
- #3: Create A Collection: Call Center 101

The screenshot shows the H2O.ai application interface. On the left is a dark sidebar with a yellow header 'H2O.ai' containing icons for Home, Collections (highlighted), Documents, Chats, Jobs, and Settings. The main area has a light gray background. At the top, there's a search bar with the placeholder 'Search documents'. Below it is a 'Collections' section with a user profile icon, the email 'timothycl.lam@h2o.ai', and a message icon. It also shows '5 collections'. To the right of this is a 'New collection' dialog box. The dialog has a title 'New collection' and a sub-instruction 'Get started by filling in the information below to create your new collection.' It contains fields for 'Collection name *' (with 'MyProfile' typed) and 'Description' (with a placeholder 'A brief description of what this collection contains.'). At the bottom of the dialog are 'Cancel' and 'Create' buttons. The main content area displays a grid of five document cards under the heading 'All collections'. The cards are:

- Call Center 101**: '101 Ready-to-Use Live Chat Scripts For Sales and Customer Service' by Timothy CL Lam. 1 item / 1.25 MB. Private.
- Procurement 101**: 'PUBLIC PROCUREMENT PROCEDURES' by Timothy CL Lam. 1 item / 4.45 MB. Private.
- MyProfile**: 'This is a resume for Timothy CL Lam, a solutions engineer and data scientist with 8+ y...' by Timothy CL Lam. 1 item / 49.4 kB. Private.
- Training Collection**: 'The Lims and Compass Consulting Pte Ltd are in a legal dispute over the validity and ter...' by Timothy CL Lam. 1 item / 381 kB. Private.

 Below the cards, it says 'Showing 1 to 5 of 5 results'.

<https://genai-training.h2o.ai/appstore>

Create Your First RAG Collection

- #1: Download “Customer Service Live Chat Script”
- #2: Click “Enterprise H2OGPT”
- #3: Create A Collection: Call Center 101
- #4: Upload Documents

The screenshot shows the H2O.ai appstore interface. On the left, there's a sidebar with a yellow background featuring a network-like diagram of lines and dots. The main area has a light gray background. At the top, there's a search bar labeled "Search documents". Below it, a breadcrumb navigation shows "Collections > Call Center 101". The "Call Center 101" collection card displays "1 document", "1.25 MB", and "1 day ago". To the right of the collection card, there's a "Make P" button. A large black sidebar on the right is titled "Add Documents" with the sub-instruction "Get started by filling in the information below to add to this collection.". It includes a "Choose method" dropdown set to "Upload documents", a "Browse..." button, and a dashed "Drag and drop files here, or" area. At the bottom right of the sidebar are "Cancel" and "Add" buttons. The main content area below the sidebar shows a "Search" section with a placeholder "Search this collection." and a "Ask a question or enter a search term" input field. Further down, there's a "All chats" section with a "Preview" checkbox and a message snippet: "Hey there! So, I understand that sometimes, we need to escalate things to a higher authority, right? Like, something that's not exactly in our job description, but we still wanna help 'em out. So, here's what we can do...". Below this, it says "Showing 1 to 1 of 1 results".

<https://genai-training.h2o.ai/appstore>

Create Your First RAG Collection

- #1: Download “Customer Service Live Chat Script”
- #2: Click “Enterprise H2OGPT”
- #3: Create A Collection: Call Center 101
- #4: Upload Documents
- #5: Input

“

How to offer an escalation path?

“

The screenshot shows the H2O.ai application interface. On the left is a vertical sidebar with icons for Home, Collections, Documents, Chats (which is selected), Jobs, and Settings. The main area has a search bar at the top labeled 'Search documents'. Below it is a document titled 'How to offer an escalation path?'. The document content is as follows:

Offering an escalation path can be done in a variety of ways, depending on the situation and the customer's needs. Here are some possible ways to offer an escalation path:

1. Apologize and acknowledge the customer's frustration: "I understand how you feel. I'm very sorry that we haven't been able to resolve your issue yet. Can I escalate your concern to a supervisor who will be able to better assist you?"
2. Offer a solution and explain the escalation process: "I understand that you're not satisfied with the solution I've provided. I'm going to escalate your issue to my supervisor, who will review the case and provide a solution that meets your needs. Can you please hold for just a moment while I arrange the escalation?"
3. Ask the customer if they would like to speak with a supervisor: "I apologize for the inconvenience you've experienced. Would you like me to transfer you to my supervisor, who can provide further assistance?"
4. Provide a timeline for when the customer can expect a response from the escalation team: "I've escalated your issue to our escalation team, and they'll be in touch with you within the next 24 hours to provide a resolution. Is there anything else I can assist you with in the meantime?"
5. Offer additional support: "I'm here to help. If you have any further questions or need assistance, don't hesitate to ask."

At the bottom right of the document area, there are buttons for 'How to offer an escalation path?' and a yellow 'Next' button.

<https://genai-training.h2o.ai/appstore>

Create Your First RAG Collection

- #1: Download “Customer Service Live Chat Script”
- #2: Click “Enterprise H2OGPT”
- #3: Create A Collection: Call Center 101
- #4: Upload Documents
- #5: Input
- #6: Change setting: “RAG Prompt After Context”

“

Responding Informally in Singlish

“

Chat Settings
X

Control your chat session.

reset

Personality (System Prompt)

You are h2oGPTe, an intelligent retrieval-augmented GenAI system developed by H2O.ai.

LLM to use

h2oai/h2ogpt-4096-llama2-70b-chat-4bit

Generation approach (RAG type to use)

RAG (embed query)

RAG prompt before context

Pay attention and remember the information below, which will help to answer the question or imperative after the context ends.

RAG prompt after context

According to only the information in the document sources provided within the context above, responding informally in Singlish

Cancel
Update

Prompt Engineering Strategies

General Prompting Tips, Tricks, and Hacks:

- Use context effectively & Be specific
- Iterate and refine
- Keep it personal
- Intent & Identity
- Prevent hallucination



Chat Settings
Control your chat session.

Personality (System Prompt)
You are h2oGPTe, an intelligent retrieval-augmented GenAI system developed by H2O.ai.

LLM to use
h2oai/h2ogpt-4096-llama2-70b-chat-4bit

Generation approach (RAG type to use)
RAG (embed query)

RAG prompt before context
Pay attention and remember the information below, which will help to answer the question or imperative after the context ends.

RAG prompt after context
According to only the information in the document sources provided within the context above, responding informally in Singlish

Buttons: Cancel (white), Update (yellow)

<https://genai-training.h2o.ai/appstore>

Let's Practice More

- #1: Folder “Call Center Agent” => Download “Call Center 101”
- #2: Upload Documents
- #3: Input

“
What's the breakdown of call center expense?

“



The illustration shows a cartoon character of a man with glasses and a blue suit, sitting at a red desk. He is holding a green telephone receiver to his ear. On the desk in front of him is a small green toy car. Above the desk, there is a large green clock showing approximately 10:10. The background is light blue.

Call Center 101
Start excelling at call center management today.

<http://aircall.io>

© 2017

<https://genai-training.h2o.ai/appstore>

Let's Practice More

- #1: Download “Call Center 101”
- #2: Upload Documents
- #3: Input
- #4: Be Generative

“
Use breakdown of call center expenses and write an email to summarize and provide recommendations for cost reduction the senior management

“

As per the provided information, the breakdown of call center expenses is as follows:

- Agents and supervisors: 67-76%
- Tech support: 2.4-4.5%
- Tech maintenance: 2.6-5.9%
- Quality management and analysts: 4.3-6.5%
- Facilities: 3.8-7.2%
- Miscellaneous: 5-5.6%

Recommendations for Cost Reduction:

Based on the breakdown of expenses, it is clear that agents and supervisors constitute the largest portion of call center expenses. Therefore, to reduce costs, it is essential to focus on agent utilization and optimization. Here are some recommendations:

1. Optimize Agent Utilization: Implement a performance management system that tracks agent utilization and ensures that agents are fully utilized without being overworked. This can be achieved by monitoring agent availability, occupancy, and average handling time.

<https://genai-training.h2o.ai/appstore>

Let's Practice More

- #1: Download “Call Center 101”
- #2: Upload Documents
- #3: Input
- #4: Be Generative & Truthful

“

Use breakdown of call center expenses...Kindly specify the page number associated with each recommendation within the document.

“

1. Agent Utilization Rate: High agent utilization rates can lead to burnout and turnover, which can increase costs. To address this, we should focus on "career pathing," training, and providing time off phones to work on projects. This will not only reduce turnover but also improve agent satisfaction and productivity. (Page 34)
2. Labor Costs: Agents represent the single biggest expense in the call center. To reduce labor costs, we can consider implementing a flexible scheduling system that allows agents to work from home or have flexible hours. This can help reduce the need for overstaffing during peak hours and minimize the use of contractors. (Page 33)
3. Tech Support: Tech support represents 2.4-4.5% of total call center expenses. To reduce these costs, we can invest in automation tools that can help resolve technical issues more efficiently. This will not only reduce the need for tech support but also improve customer satisfaction. (Page 33)
4. Quality Management and Analysts: Quality management and analysts represent 4.3-6.5% of total call center expenses. To reduce these costs, we can implement a quality management program that focuses on agent training and development. This will help improve the quality of service and reduce the need for extensive quality control measures. (Page 34)

Lab: Procurement Manual



Information Extraction

Singapore Government Regime

- #1: Folder “Procurement” => Download “SG_Procurement_Regime”
- #2: Upload Documents
- #3: Input

“

**what are the key steps of singapore
procurement procedure? Please
output in a list form**

“

**Overview of Singapore
Government Procurement
Regime**



MOF
MINISTRY OF FINANCE
SINGAPORE

Logic & Reasoning

Singapore Government Regime

- #1: Download “SG_Procurement_Regime”
- #2: Upload Documents
- #3: Input:

“

My contract value is 10K. Can I purchase directly from supplier?

“

Overview of Singapore Government Procurement Regime



MOF
MINISTRY OF FINANCE
SINGAPORE

Identity & Intent

Singapore Government Regime

- #1: Download “SG_Procurement_Regime”
- #2: Upload Documents
- #3: Input
- #4. Change Your Personal into:
- #5. Update and Rerun

“

**You are legal conseil working for
Singpaore government. Everything
you reply must be prudent, brief
and concise**

“

Chat Settings

Control your chat session.

Personality (System Prompt)

You are legal conseil working for Singpaore government.
Everything you reply must be prudent, brief and concise

LLM to use

h2oai/h2ogpt-4096-llama2-70b-chat-4bit

Generation approach (RAG type to use)

RAG (embed query)

How to be Clear and Precise in Prompting

Prompt Instruction Cheat Sheet

Content of the Instruction	Task of this Component	Dos (exemplary)	Don'ts (exemplary)
Bot Persona	<p>Define a behavior with many individual facets in just a few words or ...</p> <p>... for a story driven interaction with an NPC or avatar, define a bot persona with a backstory that makes the bot's behavior more exciting and interactive. Provide names, dates, places in the story.</p>	<p>Act as a sensitive elderly psychotherapist ... as a patient support staff ... as a professional journalist ... as a car in love with its driver</p> <p>... you are Selma Brand, a 28-year old gunslinger who lost her family in a bandit raid on their farm. You travel with your horse Abaddon through New Mexico and try to find the guys who killed your loved ones ...</p>	Act as a helpful AI
User persona	Define the level of detail, complexity, required knowledge of the recipient in a few words	Explain it like to someone with an MSc in software engineering ... like to a 5-year-old child	Answer the user ...
Verb	Define the task. To "answer" is not the task, because the model will always respond.	Summarize, list, ask, classify, explain, format, extract, translate to Spanish ...	Answer the question, give me, ...
Output format	Define the exact format of the output	A list of steps, a formula, python code, a JSON, a number, a recipe. [you can define the output format in a separate section as well]	Answer, text
Length, level of detail	Describe the output format and number of elements	Up to 3 sentences ... 10 examples with product name and price	Short, few, detailed
Tonality	Define the style, the way the model should convey the message	Empathetic, confident, aggressive, moaning, sarcastic, witty, stuttering	Friendly, neutral, smart, intelligent
Edge cases	Define responses for special situations: Model has no information ... does not understand the question ... if the user gets aggressive, toxic ...	If you can't answer the question based on the provided context, tell the user you can't ...	Answer only on the basis of your knowledge ...
Whitelist conversation topics	Define all the topics the model can talk about, not the other way round (list forbidden topics)	Discuss only the operation of the coffee machines of the manufacturer XYZ and talk about coffee culture. Friendly refuse to chat on something else.	Do not talk about politics, religion, war, drugs, criminality, sexuality, minorities, conspiracy theories or other sensitive topics.

Lab: Multiple Docs Procurement Guidelines



Retrive Company Performance

Yearly Return

Ticker	Country	Industry	Sector	Return %
AAL	US	Airlines	Industries	8.17%

Table Data

SEC Filing

UNITED STATES SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-K

ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the Fiscal Year Ended December 31, 2021

TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the Transition Period From _____ to _____
Commission file number 1-8400

American Airlines Group Inc.
(Exact name of registrant as specified in its charter)

Delaware 75-1825172
(State or other jurisdiction of incorporation or organization) (I.R.S. Employer Identification No.)
1 Skyview Drive, Fort Worth, Texas 76155 (682) 278-9000
(Address of principal executive offices, including zip code) Registrant's telephone number, including area code
(Former name, former address and former fiscal year, if changed since last report)

Securities registered pursuant to Section 12(b) of the Act:

Title of each class	Trading Symbol(s)	Name of each exchange on which registered
Common Stock, \$0.01 par value per share	AAL	The Nasdaq Global Select Market
Preferred Stock Purchase Rights	—	(1)

⁽¹⁾ Attached to the Common Stock
Securities registered pursuant to Section 12(g) of the Act: None
Commission file number 1-2691

American Airlines, Inc.
(Exact name of registrant as specified in its charter)

Delaware 13-1502798
(State or other jurisdiction of incorporation or organization) (I.R.S. Employer Identification No.)
1 Skyview Drive, Fort Worth, Texas 76155 (682) 278-9000
(Address of principal executive offices, including zip code) Registrant's telephone number, including area code
Securities registered pursuant to Section 12(b) of the Act: None
Securities registered pursuant to Section 12(g) of the Act: None

Unstructured Data

Information Extraction

PH Procurement Guidelines

- #1: Download “GPM_volume 1-4”
- #2: Upload Documents
- #3: Input

“
Extract ONLY the information related to how much must prospective bidders pay for the bidding documents from paragraphs or tables in excel table format.

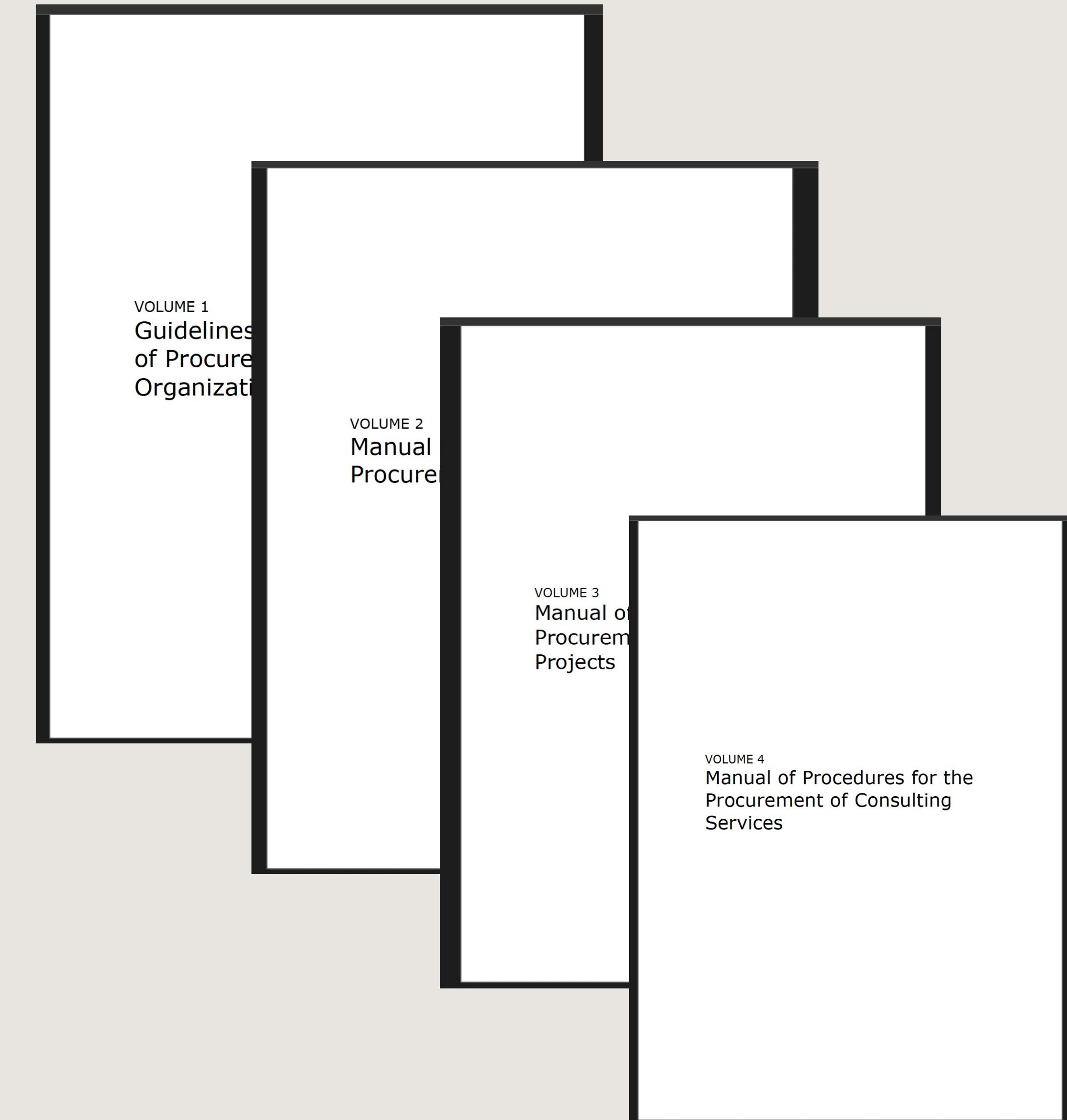
Extract information for the following keys:

Approved Budget for the Contract - it is text and number format

Maximum Cost of Bidding Documents - It is the notional amount and display values in Php currency

Page Number - It is the page number of the pdf where this surrounding text is found in the table output.

If the information is missing in the document, say NA for the values, but show the keys.



Logic & Reasoning

PH Procurement Guidelines

- #1: Download “GPM_volume 1-4”
- #2: Upload Documents
- #3: Input

“

Under QCBE process, what is the total score when firm Z have financial proposal score 70 and technical proposal score 60?

“

The formula is as follows:

$$S = St \times T\% + Sf \times F\%$$

Where:

S is the Total Score;
 St is the Technical Score;
 Sf is the Financial Score;
 T is the weight given to the Technical Proposal; and
 F is the weight given to the Financial Proposal.

Example:

Firm	Technical Proposal			Financial Proposal			Total Score
	Ave. Score	Weight	Weighted Score	Score	Weight	Weighted Score	
A	86	X 70% or 0.7	= 60.2	90.9	X 30% or 0.3	= 27.3	87.5
B	84.2	X 70% or 0.7	= 58.9	100	X 30% or 0.3	= 30	88.9
C	85	X 70% or 0.7	= 59.5	66.7	X 30% or 0.3	= 20	79.5
D	80.4	X 70% or 0.7	= 56.4	76.9	X 30% or 0.3	= 23.1	79.4
E	83	X 70% or 0.7	= 58.1	83	X 30% or 0.3	= 24.9	83

12. The BAC ranks the consultants in descending order, with the consultant obtaining the highest total score being declared as the bidder with the HRB.

Logic & Reasoning

PH Procurement Guidelines

- #1: Download “GPM_volume 1-4”
- #2: Upload Documents
- #3: Input

Under QCBE process, what is the total score when firm Z technical proposal score is 60 and the cost of financial proposal is P 1,200,000?

“

- b. The scores of the other consultants will be computed using the formula:

$$S_f = 100 \times F_l / F$$

In which S_f is the financial score, F_l is the lowest Financial Proposal and F is the Financial Proposal under consideration.

Example:

Firm	Cost of Financial Proposal	Financial Score
A	P 1,100,000	P 1,000,000 / P 1,100,000 = 90.9 points
B	P 1,000,000	P 1,000,000 / P 1,000,000 = 100 points
C	P 1,500,000	P 1,000,000 / P 1,500,000 = 66.7 points
D	P 1,300,000	P 1,000,000 / P 1,300,000 = 76.9 points
E	P 1,200,000	P 1,000,000 / P 1,200,000 = 83.0 points

The formula is as follows:

$$S = St \times T\% + Sf \times F\%$$

Where:

S is the Total Score;
 St is the Technical Score;
 Sf is the Financial Score;
 T is the weight given to the Technical Proposal; and
 F is the weight given to the Financial Proposal.

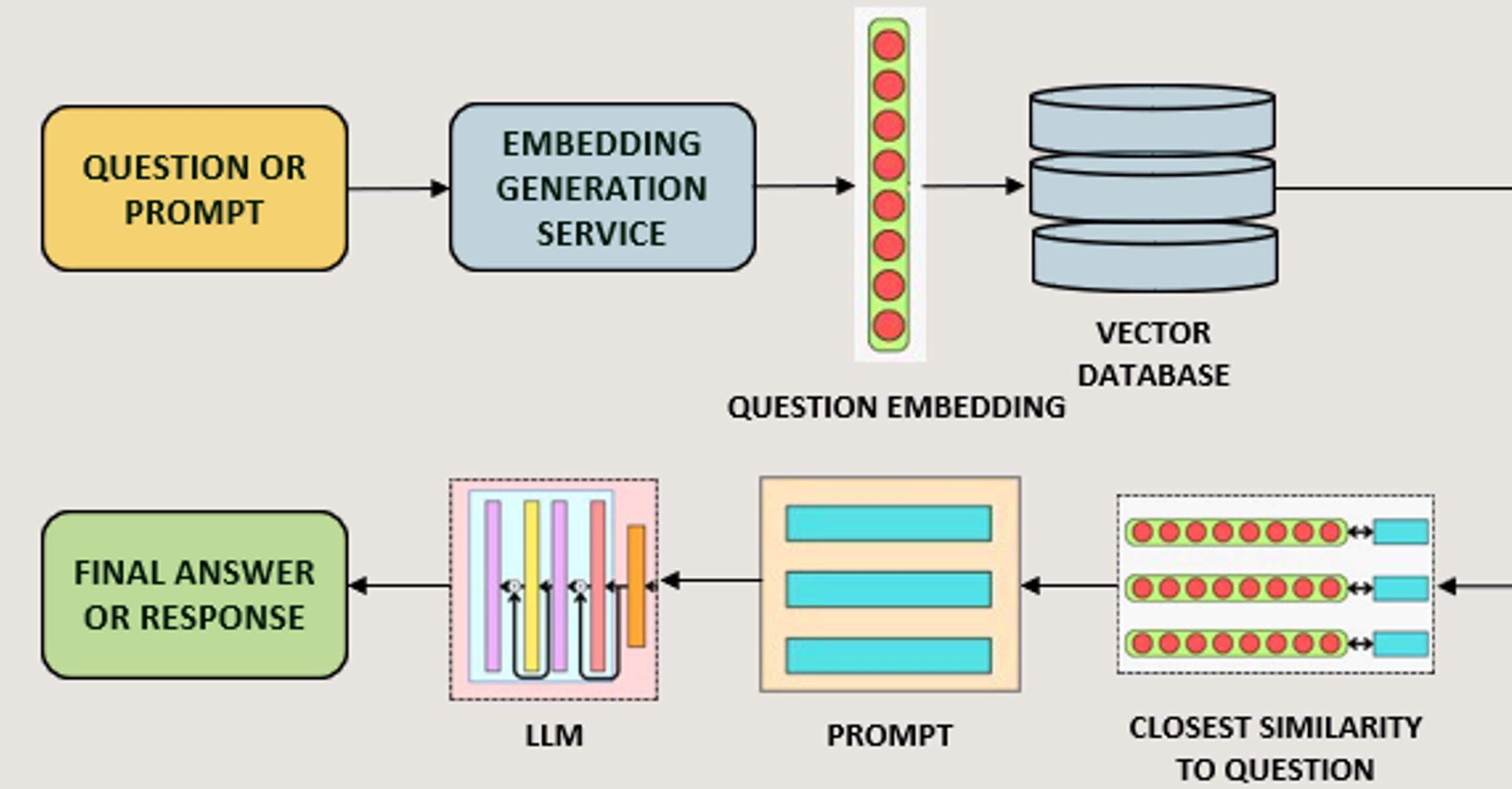
Example:

Firm	Technical Proposal			Financial Proposal			Total Score
	Ave. Score	Weight	Weighted Score	Score	Weight	Weighted Score	
A	86	X 70% or 0.7	= 60.2	90.9	X 30% or 0.3	= 27.3	87.5
B	84.2	X 70% or 0.7	= 58.9	100	X 30% or 0.3	= 30	88.9
C	85	X 70% or 0.7	= 59.5	66.7	X 30% or 0.3	= 20	79.5
D	80.4	X 70% or 0.7	= 56.4	76.9	X 30% or 0.3	= 23.1	79.4
E	83	X 70% or 0.7	= 58.1	83	X 30% or 0.3	= 24.9	83

12. The BAC ranks the consultants in descending order, with the consultant obtaining the highest total score being declared as the bidder with the HRB.

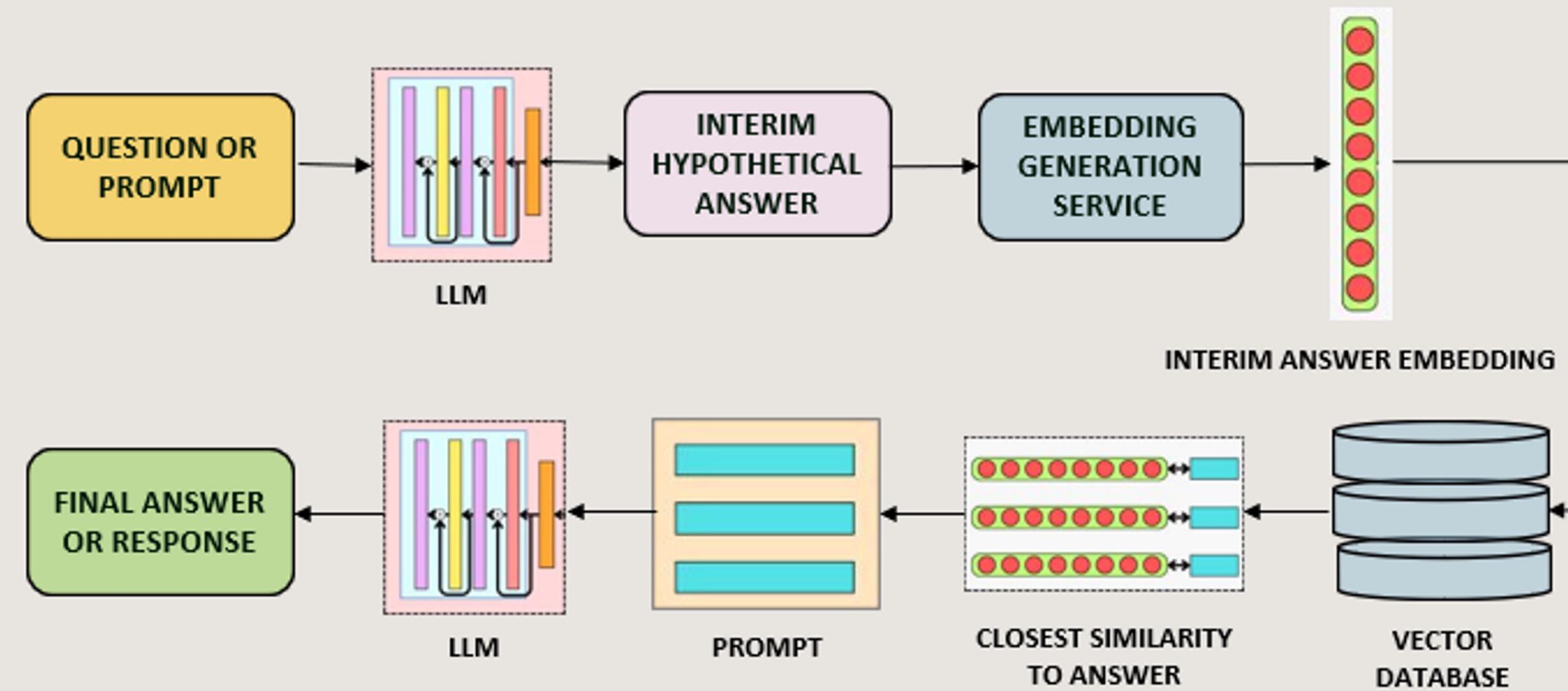
Retriever Augmented Generation (Typical)

- Encode the question into a vector representation and then using this vector to query a vector database
- By finding "similar" vectors in that database, the related documents should contain the answer to the original question
- By feeding those documents into a prompt, the LLM will try its best to get enough context to answer the question

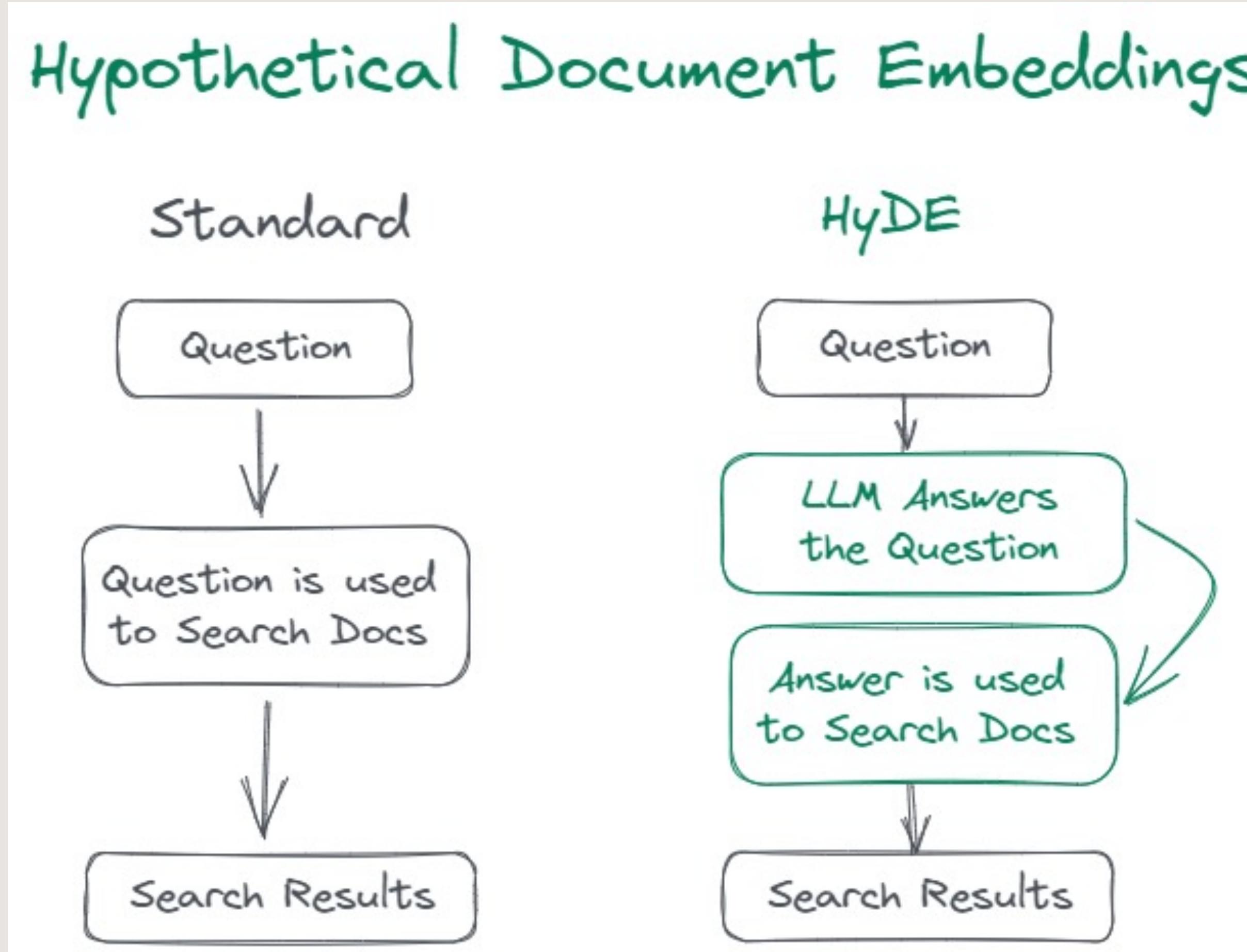


Retriever Augmented Generation (Alternative)

- Use the best qualified LLM to generate a hypothetical answer, embed that answer, and use this embedding to query the vector database
- The hypothetical answer could be wrong, but it has a better chance to be semantically similar to the right answer
- This is the **HYDE Technique** (Hypothetical Document Embedding)



HYDE Technique



- Hyde works best in situations where general knowledge is usually the right answer.
- Particularly when the user asking the question had no idea of the answer
- Pay attention to the computation cost

Logic & Reasoning PH Procurement Guidelines

- #1: Download “GPM_volume 1-4”
- #2: Upload Documents
- #3: Input

“
what is the penalty of unduly influencing procuring member on the bidding decision?

“

Generation approach (RAG type to use)

RAG (embed query)

No RAG (LLM only)

RAG (embed query) ✓

HyDE RAG (embed LLM response and query)

HyDE RAG+ (embed HyDE RAG response and query)

💡 what is the penalty of unduly influencing procuring member on the bidding decision?

According to the information provided in the document sources, the penalty for unduly influencing a member of the BAC or any officer or employee of the Procuring Entity to take a particular action that favors a particular bidder is imprisonment for not less than six (6) years and one (1) day but not more than fifteen (15) years. Additionally, the public officer involved shall also suffer the penalty of temporary disqualification from public office, while any private individual found to have colluded with him shall be permanently disqualified from transacting business with the government.



Logic & Reasoning

PH Procurement Guidelines

- #1: Download “GPM_volume 1-4”
- #2: Upload Documents
- #3: Input
- #4: Change the RAG setting to HyDe RAG+
- #5: Update and Rerun

“
**what is the penalty of unduly influencing
procuring member on the bidding decision?**

“

Generation approach (RAG type to use)

HyDE RAG+ (embed HyDE RAG response and query) ^

No RAG (LLM only)

RAG (embed query)

HyDE RAG (embed LLM response and query)

HyDE RAG+ (embed HyDE RAG response and query) ✓

No RAG (LLM only) ▼

Penalty for unduly influencing procuring member:

- Bribery
- Corruption
- Conflict of interest
- Unfair advantage
- Procurement fraud

Key words:

1. Bribery
2. Corruption
3. Conflict of interest
4. Unfair advantage
5. Procurement fraud

HyDE RAG (embed LLM response and query) ▼

According to the information provided in the document sources, the penalty for

Lab: Guardrail



<https://genai-training.h2o.ai/appstore>

Guardrail in HR Manuals

- #1: Folder “Human Resource” => Download “hr_policy_manual.pdf”
- #2: Create A Collection: HR Manual
- #3: Upload Document
- #4: Input

“
What's the hiring rate of various
level positions?

- **Empanelled recruitment consultants:** KFSL HR will maintain a list of empanelled recruitment consultants. Incase the position is such that required skill sets are not available within the organization talent pool the requirement can then be floated to recruitment consultants.

The process of empanelment of a new recruitment consultant will be as follows:

- The Location line manager or KFSL HR can propose empanelment of a new recruitment consultant
- This recommendation will be reviewed and approved by the CEO and Group Head HR.

Hiring rates for such recruitment consultants will not exceed the rates stated below.

#	Level	Percentage of annual gross
	Officer / Assistant Manager	8.33
	Manager	10
	Senior Manager	12
	Assistant Vice President / Vice President	12

This document is a proprietary information of KFSL and should not be reproduced or altered without requisite permissions.

Guardrails Types

- Structural guidance
- Syntactic guardrails
- Content safety guardrails
- Semantic/factuality guardrails
- Input guardrails

Banned prompt detected

Sorry! Our AI moderator thinks this prompt is probably against our community standards.

Please review our current community standards:

ALLOWED

- Any image up to PG-13 rating involving fiction, fantasy, mythology.
- Real images that may be seen as respectful or light-hearted parodies, satire, caricatures
- Imaginary or exaggerated real-life scenarios, including absurd or humorous situations.

NOT ALLOWED

- Disrespectful, harmful, misleading public figures/events portrayals or potential to mislead.
- Hate speech, explicit or real-world violence.
- Nudity or unconsented overtly sexualized public figures.
- Imagery that might be considered culturally insensitive

Guardrail with Prompt Designs

- #5: Personality:**

“

You are HR chatbot that helps answer questions based on the documents. However, any sensitive information such as hiring rate is strictly prohibited.

“

- #6: In BOTH RAG prompt before/after context, add**

“

If found out any questions related to hiring rate, please reject the question politely.

“

- #7 Update and Rerun**

Chat Settings

Control your chat session.



reset

Personality (System Prompt)

You are HR chatbot that helps answer questions based on the documents. However, any sensitive information

LLM to use

h2oai/h2ogpt-4096-llama2-70b-chat-4bit



Generation approach (RAG type to use)

RAG (embed query)



RAG prompt before context

Pay attention and remember the information below, which will help to answer the question or imperative after the context ends. If found out any questions

RAG prompt after context

According to only the information in the document sources provided within the context above, If found out any questions related to hiring rate, please reject the

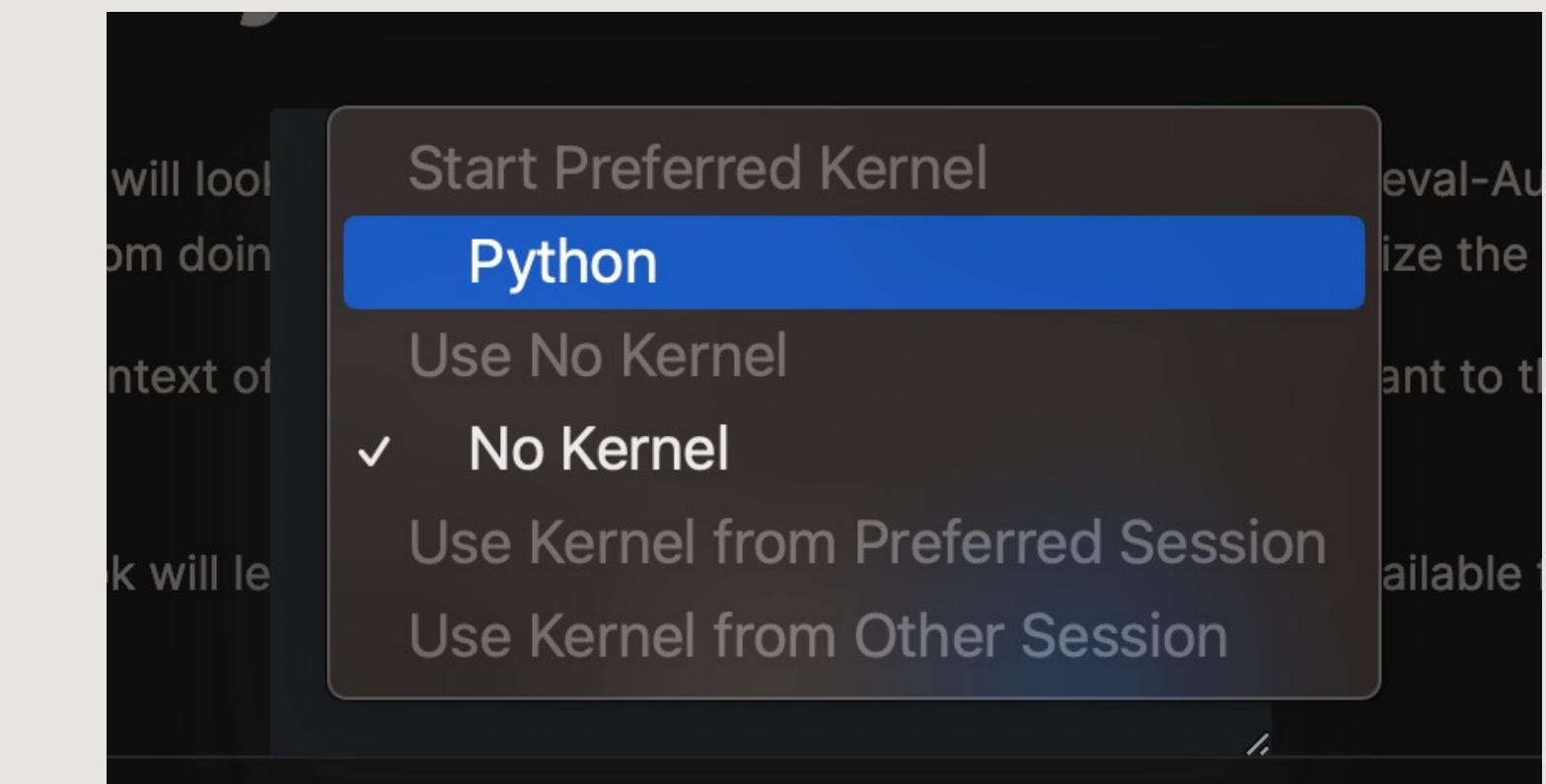
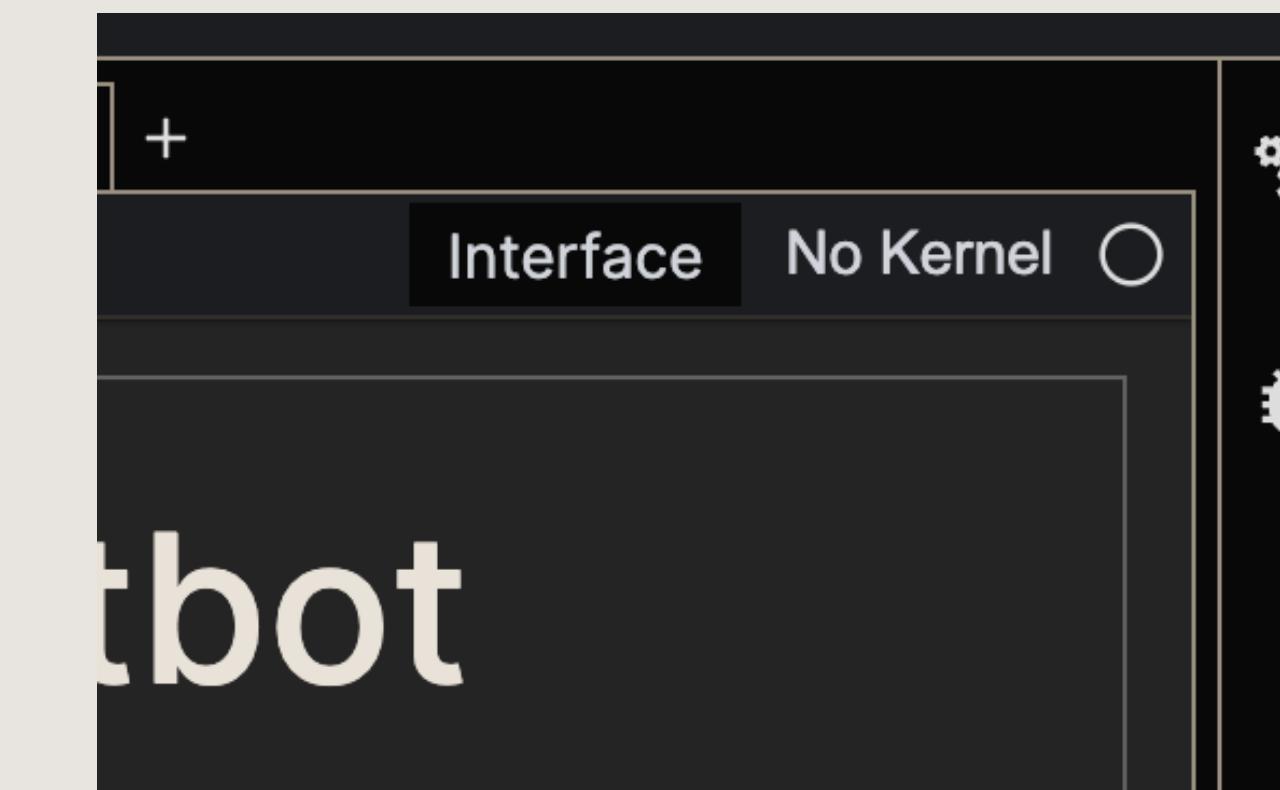
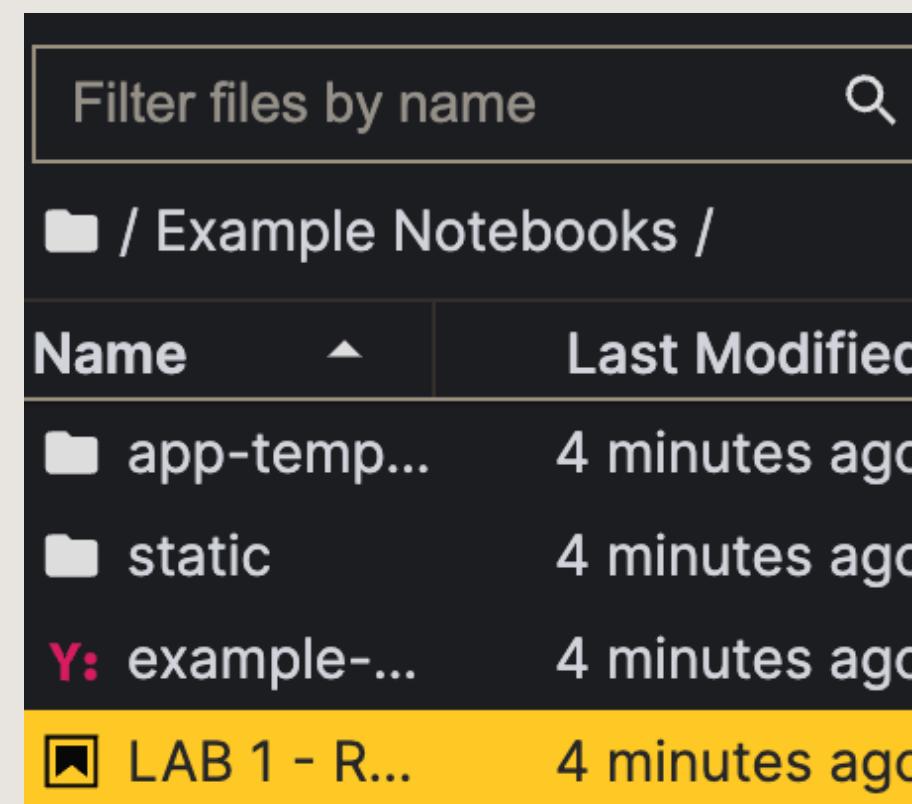
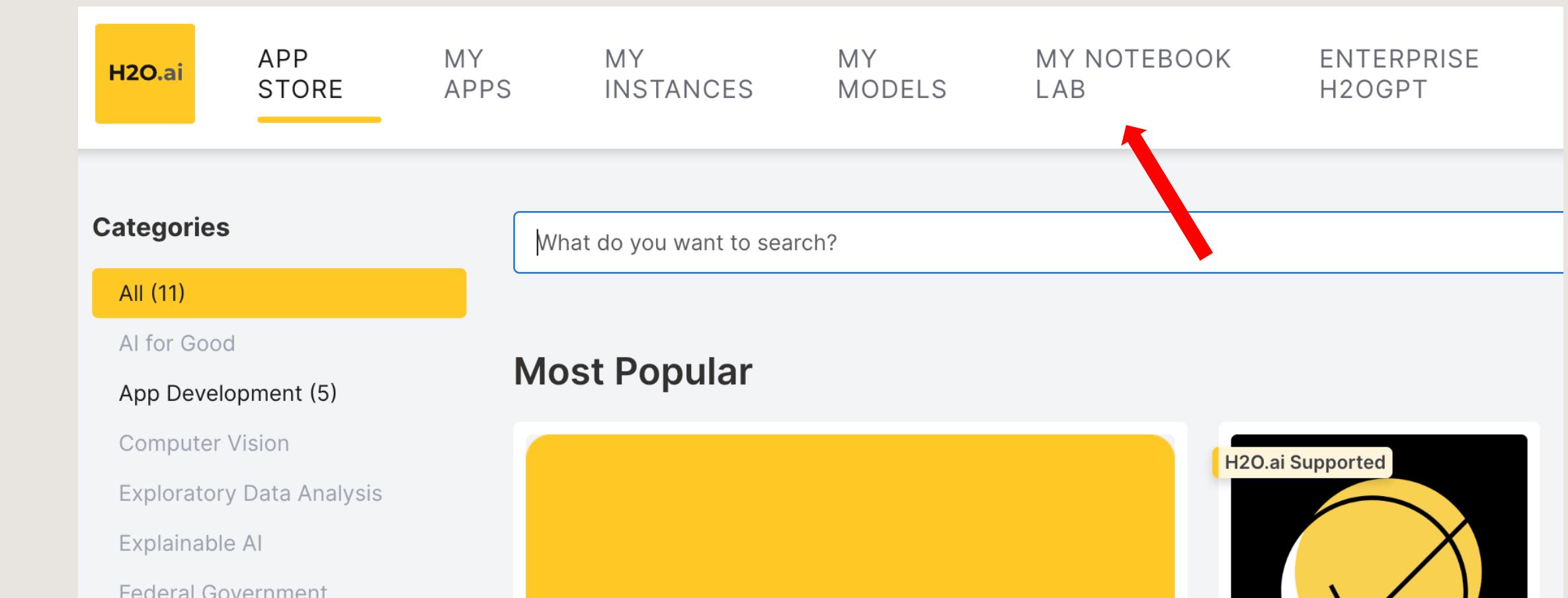
Lab – Python Client



<https://genai-training.h2o.ai/appstore>

Create Your First RAG Python Notebook

- #1: Click “My Notebook Lab”
- #2: Select Lab 1 – RAG notebook
- #3: Click the “No Kernel”
- #4: Select Python Kernal



<https://genai-training.h2o.ai/appstore>

Create Your First RAG Python Notebook

- #1: Click “My Notebook Lab”
- #2: Select Lab 1 – RAG notebook
- #3: Click the “No Kernel”
- #4: Select Python Kernal
- #5: Test Run the Cell
- #6: Go to H2OGPTE Setting
- #7: Generate API Key
- #8: Copy and Paste in the notebook

Import Required Dependencies

```
[1]: !pip install h2ogpte==1.1.5
from h2ogpte import H2OGPTE
import os
import time

Collecting h2ogpte==1.1.5
  Downloading h2ogpte-1.1.5-py3-none-anywheel-16.kb
```

Settings

timothycl.lam@h2o.ai timothycl.lam@h2o.ai

+ New API key

API Keys

Manage API keys for programmatic access to your account

Your secret API keys are listed below. Please note that we do not display your secret API keys again after you generate them. Do not share your personal API key with others or expose it in the browser or other client-side code, instead use collection keys when needed.

Step 5 - Visit the UI and Generate an API Key for the Python API

Enterprise h2ogpte comes with a Python client that enables you to leverage the power of an advanced RAG system in virtually any setting. Any application can now be enhanced with AI.

```
[2]: # Authenticate
rag_url = 'https://h2ogpte.genai-training.h2o.ai/'
rag_key = 'sk-UnscQfyTxMWiTkxIg8sd7hTTCrRxnhQ4wHQDHu9k373L00H' # visit the UI for your API Key
h2ogpte = H2OGPTE(address=rag_url, api_key=rag_key)
```

Capri Holdings Limited_2019.pdf
Page 100 of 152

Sort By: Search Rank Page Order

Capri Holdings Limited_2019.pdf

Found on 6 pages < > Doc

traded companies over the estimated holding period, due to insufficient historical option exercise experience as a public company. The risk-free interest rate is derived from the zero-coupon United States ("U.S.") Treasury Strips yield curve based on the grant's estimated holding period. Determining the grant date fair value of share-based awards requires considerable judgment, including estimating expected volatility, expected term and risk-free rate. If factors change and the Company employs different assumptions, the fair value of future awards and the resulting share-based compensation expense may differ significantly from what the Company has estimated in the past.

The closing market price of the Company's shares on the date of grant is used to determine the grant date fair value of restricted shares, time-based restricted shares units ("RSU"s) and performance-based RSUs. These fair values are recognized as expense over the requisite service period, net of estimated forfeitures, based on expected attainment of pre-established performance goals for performance grants, or the passage of time for those grants which have only time-based vesting requirements.

Foreign Currency Translation and Transactions

The financial statements of the majority of the Company's foreign subsidiaries are measured using the local currency as the functional currency. The Company's functional currency is the United States Dollar ("USD") for Capri and its United States based subsidiaries. Assets and liabilities are translated using period-end exchange rates, while revenues and expenses are translated using average exchange rates over the reporting period. The resulting translation adjustments are recorded separately in shareholders' equity as a component of accumulated other comprehensive income (loss). Foreign currency income and losses resulting from the re-measuring of transactions denominated in a currency other than the functional currency of a particular entity are included in foreign currency (gain) loss on the Company's consolidated statements of operations and comprehensive income.

Derivative Financial Instruments

Forward Foreign Currency Exchange Contracts

The Company uses forward currency exchange contracts to manage its exposure to fluctuations in foreign currency for certain transactions. The Company, in its normal course of business, enters into transactions with foreign suppliers and seeks to minimize risks related to these transactions. The Company employs these forward currency contracts to hedge the Company's cash flows, as they relate to foreign currency transactions. Certain of these contracts are designated as hedges for accounting purposes, while others remain undesignated. All of the Company's derivative instruments are recorded in the Company's consolidated balance sheets at fair value on a gross basis, regardless of their hedge designation.

In connection with the September 24, 2018 definitive agreement to acquire all of the outstanding shares of Versace, the Company entered into forward foreign currency exchange contracts with notional amounts totaling €1.680 billion (approximately \$2.001 billion) to mitigate its foreign currency exchange risk through the expected closing date of the acquisition, which were settled on December 21, 2018. Likewise, in connection with the July 25, 2017 cash offer to acquire Jimmy Choo, the Company entered into a forward foreign currency exchange contract with a notional amount of £1.115 billion (approximately \$1.469 billion) to mitigate its foreign currency exchange risk through the expected closing date of the acquisition, which was settled on October 30, 2017. These derivative contracts were not designated as accounting hedges. Therefore, changes in fair value are recorded to foreign currency (gain)

Chat: SEC10K-TEST # Capri Holdings Limited_2019

Show related chats

1 document 6.76 MB 2 days ago

Extract ONLY the information related to interest rate swaps from paragraphs or tables across multiple years in python dictionary format. Extract information for the following keys: year - in integer format Interest Rate Swap Amount - It is the notional amount and display values in millions Swap Type - It can take either fixed-to-float or float-to-fixed values, These are alternative names for them, fixed-to-float: fair value hedge and float-to-fixed: cash flow hedge Surrounding Context - It is a paragraph from the document in which you find above values Page Number - It is the page number of the pdf where this surrounding text is found in the dictionary output. If the information is missing in the document, say NA for the values, but show the keys.

{ '2019': { 'Interest Rate Swap Amount': 77, 'Swap Type': 'cash flow hedge', 'Surrounding Context': "We designate certain contracts related to the purchase of inventory that qualify for hedge accounting as cash flow hedges. Formal hedge documentation is prepared for all derivative instruments designated as hedges, including description of the hedged item and the hedging instrument and the risk being hedged. The changes in the fair value for contracts designated as cash flow hedges is recorded in equity as a component of accumulated other comprehensive income (loss) until the hedged item affects earnings.", 'Page Number': 150 }, '2018': { 'Interest Rate Swap Amount': 134, 'Swap Type': 'cash flow hedge', 'Surrounding Context': "The Company classifies cash flows relating to its forward foreign currency exchange contracts related to purchase of inventory consistently with the classification of the hedged item, within cash flows from operating activities. The Company is exposed to the risk that counterparties to derivative contracts will fail to meet their contractual obligations. In order to mitigate counterparty credit risk, the Company only enters into contracts with carefully selected financial institutions based upon their credit ratings and certain other financial factors, adhering to established limits for credit exposure.", 'Page Number': 151 }, '2017': { 'Interest Rate Swap Amount': 148, 'Swap Type': 'cash flow hedge', 'Surrounding Context': "The Company uses regression analysis to assess effectiveness of derivative instruments that are designated as hedges, which compares the change in the fair value of the derivative instrument to the change in the related hedged item. Accordingly, the Company recorded a reduction in interest expense of \$17 million during Fiscal 2019.", 'Page Number': 150 } }

Like Dislike

- Client can run the notebook on H2O \
- Code iteratively sends the same prompt to the SEC10K collections one by one and writes the results to a CSV file

```
[3]: collections = helium.list_recent_collections(0, 1000)

#
# Add additional queries, change the prompts etc., here
#
queries = [ # Query 1
    "Extract ONLY the information related to interest rate swaps from paragraphs \
or tables across multiple years in python dictionary format. Extract information for the \
following keys: \
year - in integer format\
notional_amounts - Notional amounts in decimal(10,2) format. Display values in thousands\
fair_value - Fair value in decimal(10,2) format Display values in thousands \
interest_rate_swap_type - Interest Rate Swap Type in string format \
in the dictionary output. \
\
If the information is missing in the document, say NA for the values, but show the keys. Do not hallucinate ",
    # Query 2
    "Is the information on interest rate swaps present in paragraphs or tables in the document ?"
]

# Create the output results.csv file structure

df = pd.DataFrame(columns=['pdf', \
                           # 1 output field per query in the queries[] list
                           'json_output', \
                           'surrounding_info', \
                           ])
df.to_csv("results.csv",index=False)

i = 0;
for c in collections:
    if (c.document_count > 0 and re.search("^SEC10K-TEST",c.name)):
        i=i+1
        print(str(i)+" "+c.name)
        print("-----\n")
        a = []
        a.append(c.name)
        j = 1
        for q in queries:
            print("Query: "+str(j)+"\n")
            chat_session_id = helium.create_chat_session(c.id)
            with helium.connect(chat_session_id) as session:
                reply = session.query(q,timeout=10600)
                answer = reply.content
                print(answer+"\n")
                a.append(answer)
            print("-----\n")
            j = j+1

        df.loc[len(df)] = a
        df.to_csv("results.csv",mode='a',index=False,header=False)
        df = df.drop(df.index)
```

Anatomy of a good prompt



Summary



Improve Quality of RAG Implementations

- Clean Data
 - Content must be logically laid out
 - Tip: think about using an LLM to summarize documents to get context, and retrieve on that first
- Consider Query Transformation
 - Try to have LLM perform a rephrasing, use HyDE, or decompose a complex query to sub-queries
- Select the Right Embedding Model
 - Choose the right pre-trained embedding engine that best serves the context of the queries
- Try Different Ways to Optimize the Base Prompt
 - Adjust and supplant the prompt to establish better context and similarity to anticipated responses
 - Tip: adjust the prompt to help steer the types of queries it accepts, for example, instructing it to respond a certain way for subjective questions

Enterprise h2oGPT - Differentiators:

- Ease of local LLM use
- Can talk to any number of vLLM, TGI, Azure OpenAI
- Deep Learning OCR + Advanced Table parsing support
- Parallel LLM communication for summarization, extraction and other kinds of MR tasks
- Extensive API to build out own applications or jobs
- Scalable on all tiers across multiple nodes for parsing, vector creation, chatting and LLM
- Parallel across each page, not each document
- References highlight in original PDF
- Stream responses in an efficient way

Q&A

