

# Experimenting with Gen AI Capabilities to Support Dev Productivity

4<sup>th</sup> October 2023



# Our Presenters



**Liyana**

Lead Product Manager  
GDS  
SG Tech Stack



**Leon**

Product Manager  
GDS  
SGTS: SHIP-HATS



**Zaini**

Data Analyst  
GDS  
SG Tech Stack



**Janice**

Senior Software Engineer  
ACE  
LLM Stack

## Agenda

### 1. Overview

1. SGTS and Dev Productivity
2. SHIP-HATS and AI Vision
3. Gen-AI Productivity Tools

### 2. What we have been experimenting with?

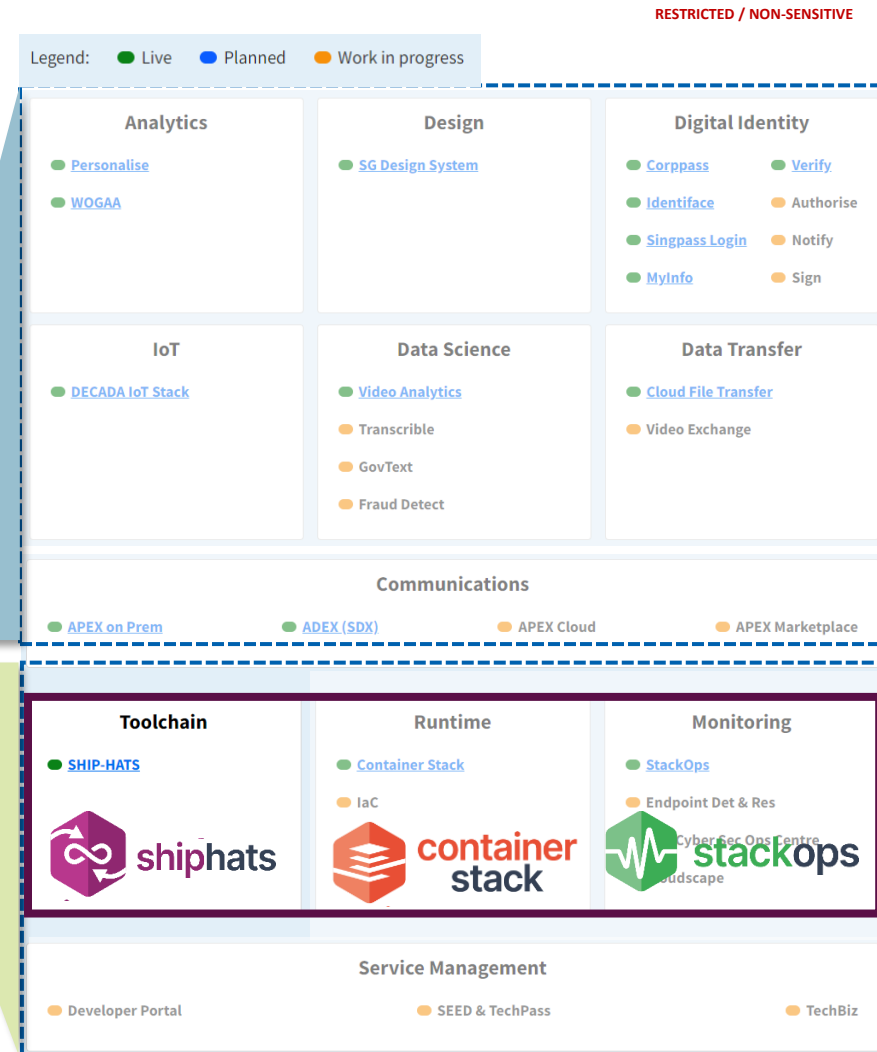
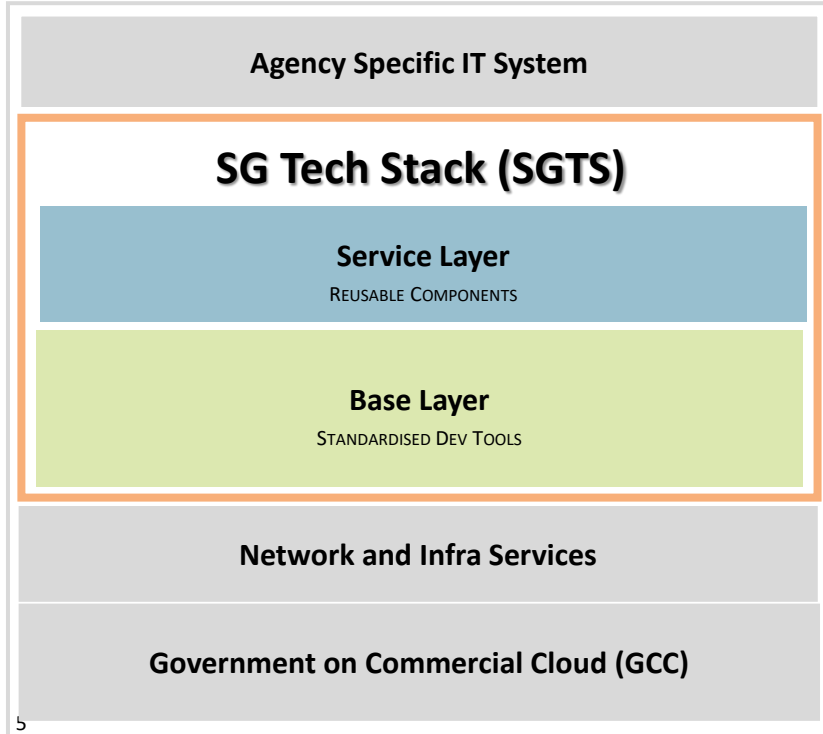
1. VS Code Extension for code suggestions
2. HiveMind Chatbot

### 3. LLM Stack

### 4. Q & A

# Stack – Government-in-a-Box

[www.developer.tech.gov.sg](http://www.developer.tech.gov.sg)



# SGTS aims to enable Dev productivity



shiphats



container  
stack



stackops

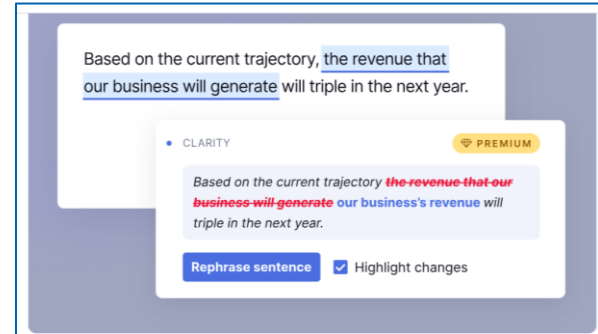
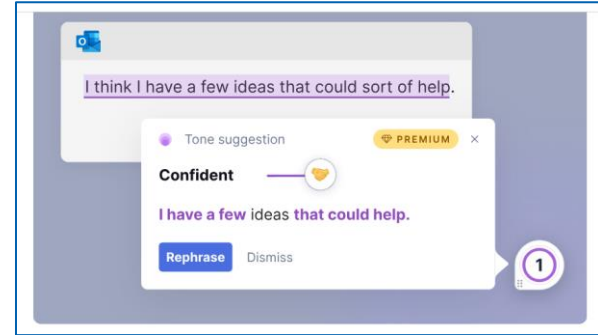
## Our Goal

- SGTS aims to improve dev experience and productivity by providing reusable components that are IM-compliant, in order to reduce cognitive workload on devs
- #platform engineering

# AI tools have been around for awhile



- Founded in 2009
- Started as a cloud-based typing assistant; reviews spelling, grammar, punctuation, clarity, engagement, and delivery mistakes
- Detects plagiarism, and suggests replacements for the identified errors
- Allows users to customize their style, tone, and context-specific language



# AI is not a panacea but will be the invisible norm for development teams



- Not new technology
- Artificial Intelligence, Machine Learning and Automation - an opportunity to explore more use-cases to reduce the cognitive workload for development and operation teams
- For Dev: Code Automation, Security Vulnerability Remediation etc.
- For Ops: Alert Summarization, Workflow Suggestions etc.
- More established use cases

## Our Approach:

- Use-Case driven
- Experiment internally, and work with our partner providers

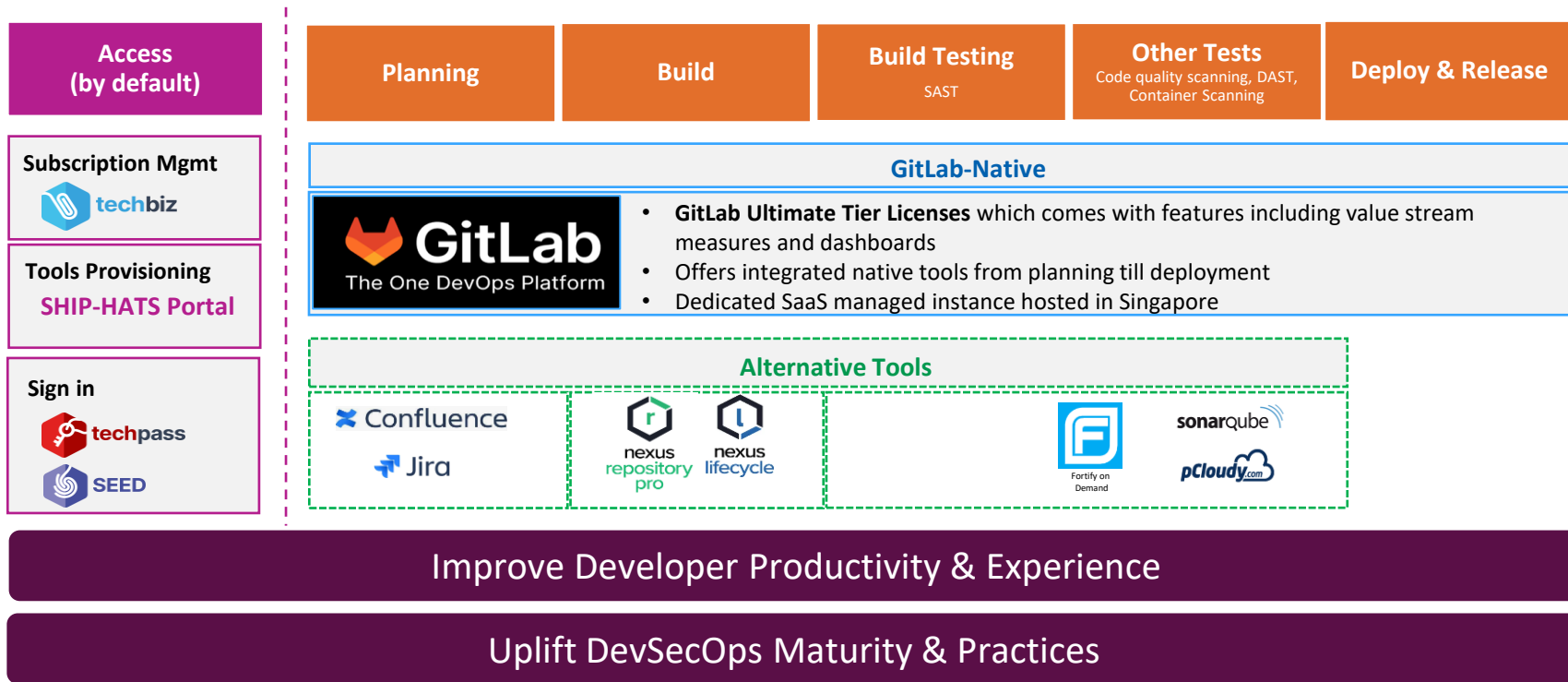
# SHIP-HATS AI Vision



# SHIP-HATS – Our Vision



# SHIP-HATS - Toolchain





GPT - 4

# Myth or Potential?

## The numbers speak for themselves

Research has found GitHub Copilot helps developers code faster, focus on solving bigger problems, stay in the flow longer, and feel more fulfilled with their work.

74%

Focus on more satisfying work

88%

Feel more productive

96%

Are faster with repetitive tasks

10X

A well-tuned AI innovation capability can deliver 10X or greater growth to an enterprise



A giant leap forward in developer productivity

57% faster

During the preview, Amazon ran a productivity challenge and participants who used Amazon CodeWhisperer were 57% faster than those who did not use CodeWhisperer.

27% more likely to succeed



## Efficient & secure software development

In a research paper looking into the impact of AI on developer productivity published in February 2023, [Peng et al.](#) found that developers who used AI-powered tools completed their tasks 55.8% faster than those who didn't. In this section, we'll look at how GitLab Duo can help teams reduce development time, improve developer productivity, and ensure secure software development.

# Promising gains with AI Tools

We would like to leverage potential of AI productivity tools to help developers:



## Achieve better “Flow”

Prioritizing cognitive load on what matters



## Improve Satisfaction

Enable devs to spend more time spent on satisfying work



## Expose Security Vulnerabilities early

during code development in IDE



## Gain Productivity

Shifting left on repetitive and less complex tasks



## Learn Rapidly

Learn new and unfamiliar code syntax quickly with code explanation and suggestions

*AI tools do not replace developers; rather it augments their capabilities*

# SHIP-HATS



# What have we been experimenting with?

## SHIP-HATS

### Coding Assistants



GitHub  
Copilot



GitLab  
Duo



SHIP-HATS  
VS-Code Extension

Commercial-off-the-shelf (COTS)

Customized Development

### Chatbots



Hivemind  
Chatbot

Customized Development

## Our Partners

### Tech Stack

LLM Stack



LaunchPad

Customized Development

# SHIP-HATS Gen-AI Productivity Tools Pilot

With GitHub Copilot & GitLab DUO



# Dev Productivity Tools we would like to evaluate

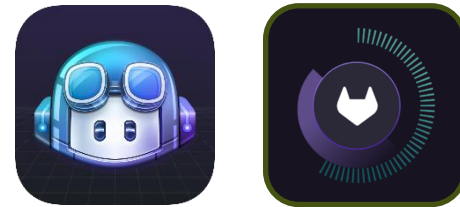
*Numerous tools; limited time and resources*



- Tabnine, VS Intellicode, Google Bard / Colab ... etc.



- Amazon CodeWhisperer
- ChatGPT-4+



- GitHub Copilot
- GitLab Duo

Increasing Synergy to scale with existing GovTech Tools / Market Leader for AI Code Generation

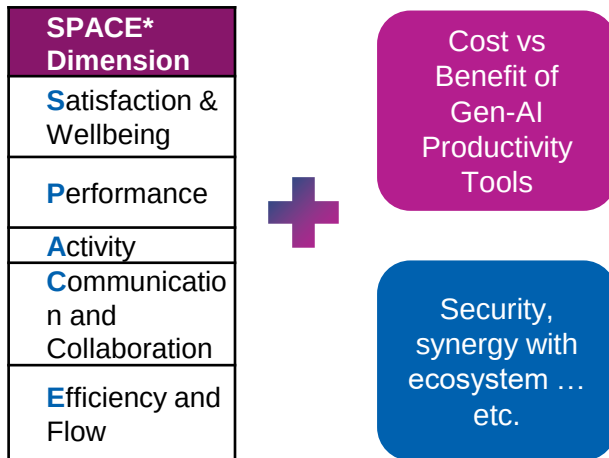
# Pilot: GitHub Copilot and GitLab Duo

Duration: Sep to Nov 2023 (We have seats available!)

## Key Features

- Prompt / Chat Q&A
- In-Editor Code Suggestions
- Code Explanation
- Vulnerability Explanation
- Test generation
- Prototyping
- Learning assistant
- Custom Private Models (No ETA)

## What is being evaluated?



\*Framework for measuring developer productivity

## Pilot Scope

x100



### Persona

- Software Developers
- DevOps engineers
- Mostly GovTech



### Data Classification

- Official (Closed) / Non-Sensitive
- We are pushing for Restricted



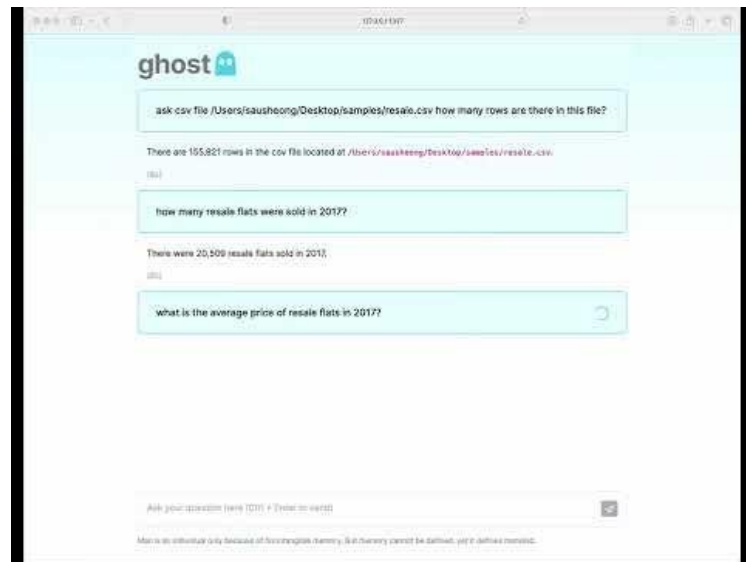
### Evaluation Medium

- Surveys
- Focus Groups
- Telemetry



*"What if a cyber brain could possibly generate its own ghost, create a soul all by itself?"*

- Ghost is a small footprint AI assistant, built to run locally in your computer. It is meant for developers and can be modified/customized easily to suit your needs.
- Ghost is built using Langchain.
- Out of the box:
  - Ghost is able to analyse CSV files, images and documents via RAG (Retrieval Augmented Generation)
  - Ghost is able to query relational databases
- **Call to action for the community:**
  - Discover more use cases related to ghost
  - Contribute improvements and feedback



[Innersource / AI and ML / Ghost · GitLab \(gitlab-dedicated.com\)](https://gitlab-dedicated.com)

# Experimental Features

HivemindBot (Telegram)  
HiveGPT (VSCode Extension)

Zaini Chia  
Data Analyst  
(GDS Central, Stack Pillar)



# AI & Dev Productivity: Internal Resources

- In addition to evaluating commercial tools like Gitlab Duo and Github Co-Pilot, we wanted to explore how AI can be used to help developers **make the most of our own internal resources**:
  - o SHIP-HATS Documentation,
  - o Innersource content (reference pipelines, modular templates)

## **Our hypothesis:**

Using Gen AI tools that are customized to connect to our internal resources will improve dev productivity and satisfaction

# Tools

| Item                                   | What is it?                                                                                                                                           |
|----------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>HivemindBot (SGTS)</b>              | A Telegram bot that is powered by GPT3.5 and augmented with our own internal knowledge base; <b>powered by <u>LaunchPad Studio</u></b>                |
| <b>SHIP-HATS<br/>VS Code Extension</b> | A custom-made VS Code Extension that can connect to multiple LLM back-ends. These backends can give code assistance / suggestions from within VS Code |
| <b>LaunchPad Studio</b>                | A low-code interface to build workflows that can be used to power AI applications; <b>powered by <u>LLM Stack</u></b>                                 |
| <b>LLM Stack</b>                       | A platform for quick prototyping and production launch of LLM-powered applications                                                                    |

# Background: What is an LLM?

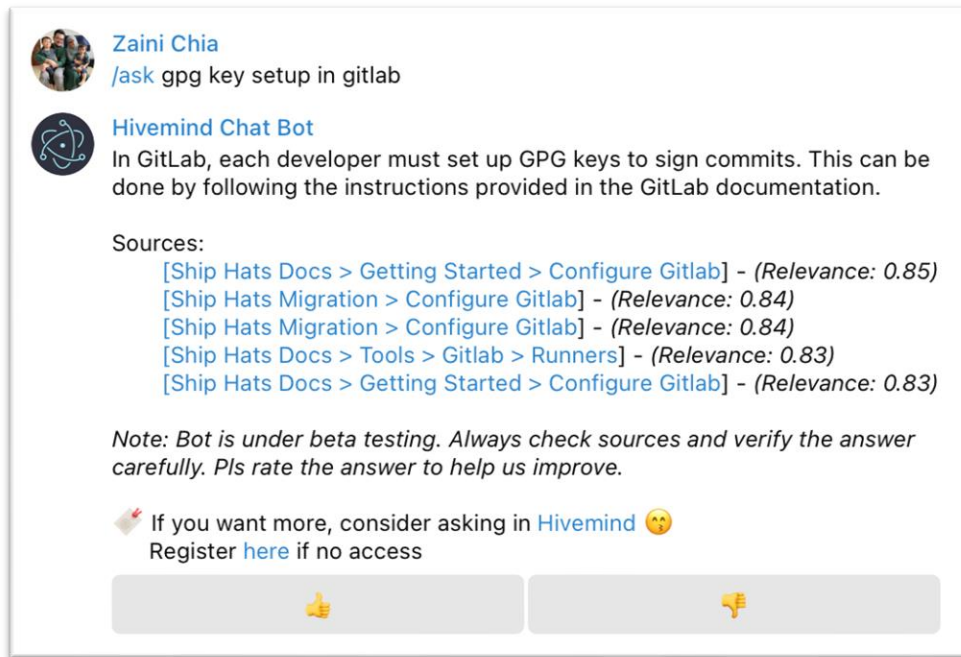
A large language model (LLM) is a deep learning algorithm that can perform a variety of natural language processing (NLP) tasks.

Large language models use transformer models and are trained using **massive datasets** — hence, large.

This enables them to **recognize, translate, predict, or generate text or other content.**

# HivemindBot

- Telegram Bot powered by GPT3.5 and paired with our internal knowledge base
- Instead of spending time trying to find relevant documentation, the Telegram bot can give users a quick answer, and provide relevant sources
- Supports users in their migration process to SHIP-HATS 2.0



 **Zaini Chia**  
/ask gpg key setup in gitlab

 **Hivemind Chat Bot**  
In GitLab, each developer must set up GPG keys to sign commits. This can be done by following the instructions provided in the GitLab documentation.

Sources:

- [[Ship Hats Docs > Getting Started > Configure Gitlab](#)] - (Relevance: 0.85)
- [[Ship Hats Migration > Configure Gitlab](#)] - (Relevance: 0.84)
- [[Ship Hats Migration > Configure Gitlab](#)] - (Relevance: 0.84)
- [[Ship Hats Docs > Tools > Gitlab > Runners](#)] - (Relevance: 0.83)
- [[Ship Hats Docs > Getting Started > Configure Gitlab](#)] - (Relevance: 0.83)

*Note: Bot is under beta testing. Always check sources and verify the answer carefully. Pls rate the answer to help us improve.*

 If you want more, consider asking in [Hivemind](#) 🙋  
Register [here](#) if no access



Zaini Chia

/ask gpg key setup in gitlab



Hivemind Chat Bot

In GitLab, each developer must set up GPG keys to sign commits. This can be done by following the instructions provided in the GitLab documentation.

Sources:

[[Ship Hats Docs > Getting Started > Configure Gitlab](#)] - (Relevance: 0.85)

[[Ship Hats Migration > Configure Gitlab](#)] - (Relevance: 0.84)

[[Ship Hats Migration > Configure Gitlab](#)] - (Relevance: 0.84)

[[Ship Hats Docs > Tools > Gitlab > Runners](#)] - (Relevance: 0.83)

[[Ship Hats Docs > Getting Started > Configure Gitlab](#)] - (Relevance: 0.83)

*Note: Bot is under beta testing. Always check sources and verify the answer carefully. Pls rate the answer to help us improve.*



If you want more, consider asking in [Hivemind](#) 🙄

Register [here](#) if no access



Sources with:

- Links
- Relevance scores

Thumbs up/down rating  
(for analysis & refinements)

# HivemindBot – Code Suggestions

- Most LLMs can already help with generic code suggestions: what is our value-add?
- SHIP-HATS team provides **curated reference pipelines and modular templates** for users to use and modify to their needs
- Hivemindbot leverages the versatility of LLMs together with these resources to help users set up and configure their pipelines



Zaini Chia

/ask I want to do a nexus iq scan on my node.js app, show me a sample. How do I integrate it to .gitlab-ci.yml? Include testing and building stages as well



Hivemind Chat Bot

You can use the `.nexus-iq-scan` template to perform a Nexus IQ scan on your Node.js app. Here's an example of how you can integrate it into your `.gitlab-ci.yml` file:

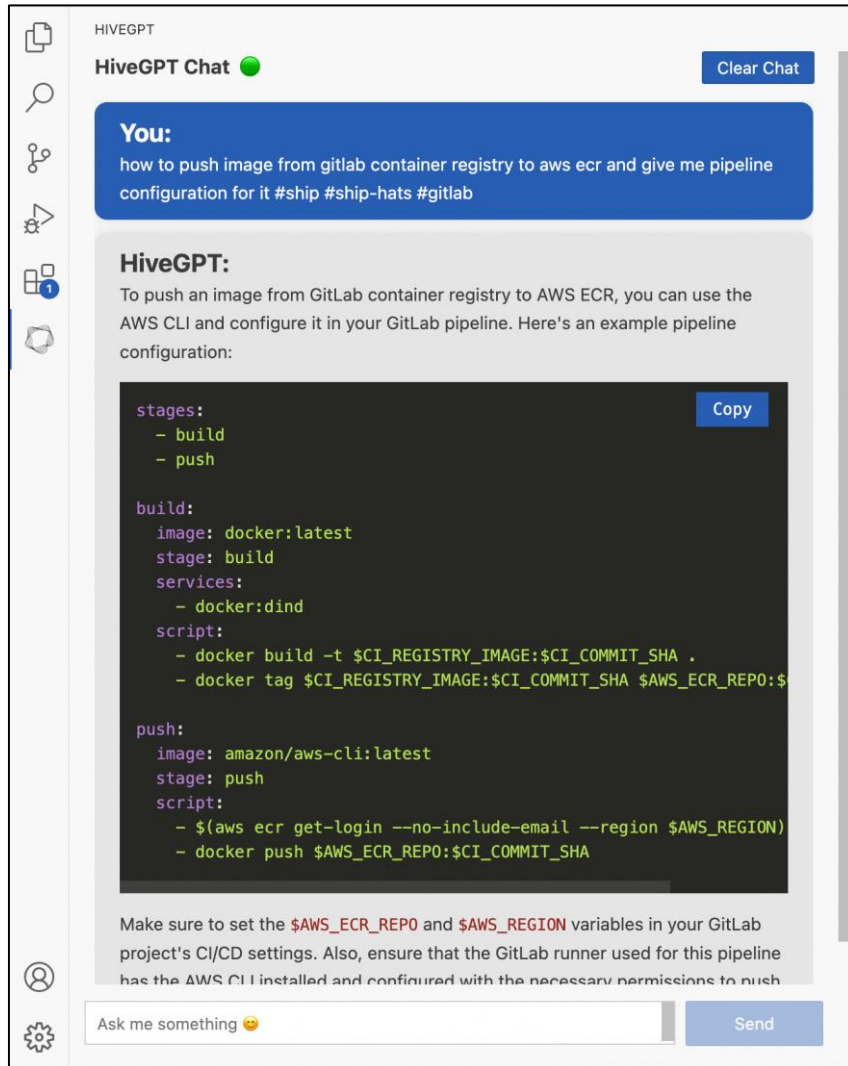
```
include:
  - local: "/templates/vars/.gitlab-ci-sonatype-vars.yml"
  - local: "/templates/vars/.gitlab-ci-pipeline-coe-vars.yml"
  - local: "/templates/.gitlab-ci-common.yml"

# Define stages
stages:
  - build
  - test
  - scan

# Build stage
build:
  stage: build
  script:
    - npm install
    - npm run build
  artifacts:
    paths:
      - dist/
```

# SHIP-HATS VScode Extension

- A custom-made [modular extension](#) that can link to **multiple** LLM back-ends:
  - OpenAI
  - Oobabooga
  - LLM Stack
- When linked to our LaunchPad Studio, back-end, users can get the same responses as our HivemindBot
- The modularity increases the versatility of this extension, as users can get information / assistance from multiple sources



The screenshot shows a web-based chat interface for HIVEGPT. At the top, there's a header with 'HIVEGPT' and a 'Clear Chat' button. The chat area has a blue header 'HiveGPT Chat' with a green status indicator. The user's message is in a blue box: 'how to push image from gitlab container registry to aws ecr and give me pipeline configuration for it #ship #ship-hats #gitlab'. The assistant's response is in a light gray box, starting with 'HiveGPT:' followed by a helpful explanation. Below the text is a code block with a dark background and a 'Copy' button. The code is a GitLab CI pipeline configuration for pushing a Docker image to AWS ECR. At the bottom, there's a text box with a placeholder 'Ask me something' and a 'Send' button.

HIVEGPT

HiveGPT Chat

Clear Chat

**You:**

how to push image from gitlab container registry to aws ecr and give me pipeline configuration for it #ship #ship-hats #gitlab

**HiveGPT:**

To push an image from GitLab container registry to AWS ECR, you can use the AWS CLI and configure it in your GitLab pipeline. Here's an example pipeline configuration:

```
stages:
  - build
  - push

build:
  image: docker:latest
  stage: build
  services:
    - docker:dind
  script:
    - docker build -t $CI_REGISTRY_IMAGE:$CI_COMMIT_SHA .
    - docker tag $CI_REGISTRY_IMAGE:$CI_COMMIT_SHA $AWS_ECR_REPO:$CI_COMMIT_SHA

push:
  image: amazon/aws-cli:latest
  stage: push
  script:
    - $(aws ecr get-login --no-include-email --region $AWS_REGION)
    - docker push $AWS_ECR_REPO:$CI_COMMIT_SHA
```

Make sure to set the `$AWS_ECR_REPO` and `$AWS_REGION` variables in your GitLab project's CI/CD settings. Also, ensure that the GitLab runner used for this pipeline has the AWS CLI installed and configured with the necessary permissions to push

Ask me something

Send

# Demo

1. Hivemind Bot (SGTS)
2. SHIP-HATS VScode Extension





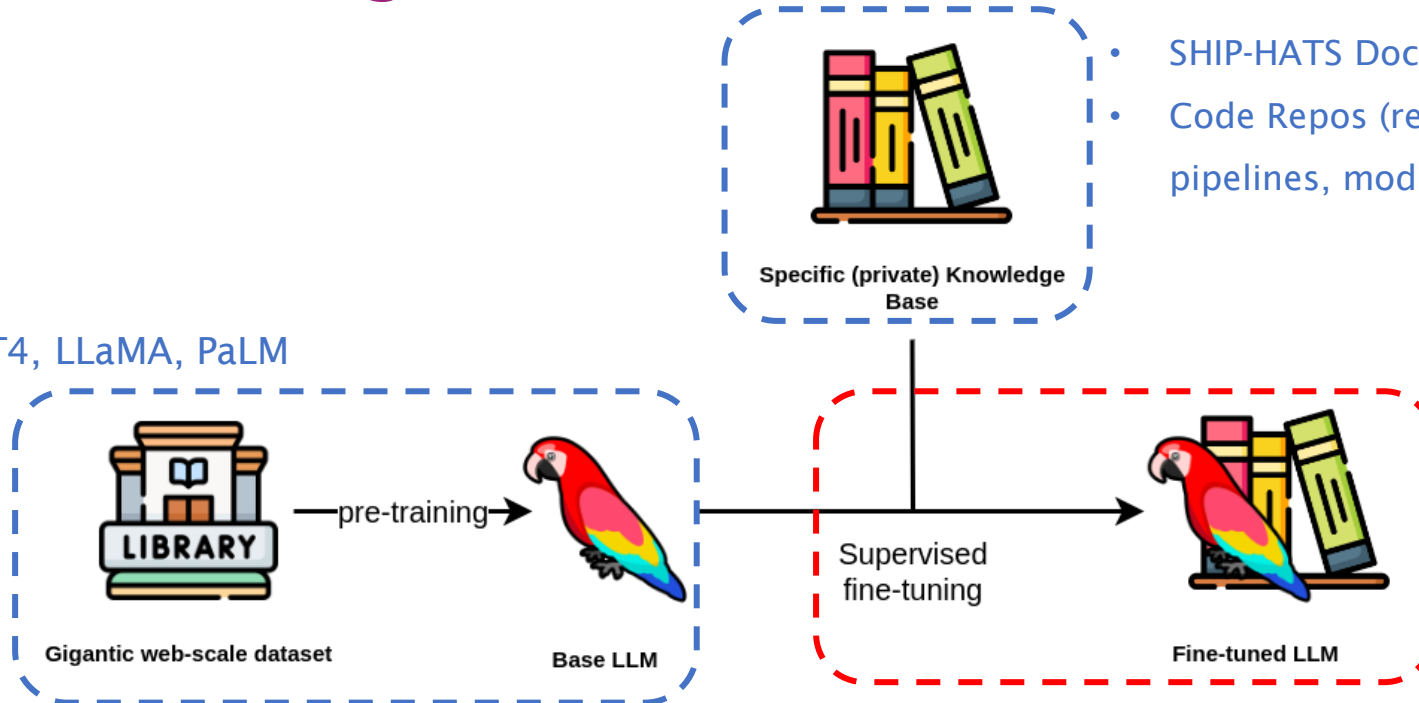
# What's happening behind the scenes?

# Background: LLM + Internal knowledge-base

- LLMs have excellent language skills, and vast general knowledge
- However, they:
  - will not have seen internal/private data
  - will not have knowledge of things that happened after its training
- Techniques for building applications with LLMs:
  - Fine-tune the LLM, or
  - Use Retrieval-Augmented Generation (RAG)

# Fine Tuning

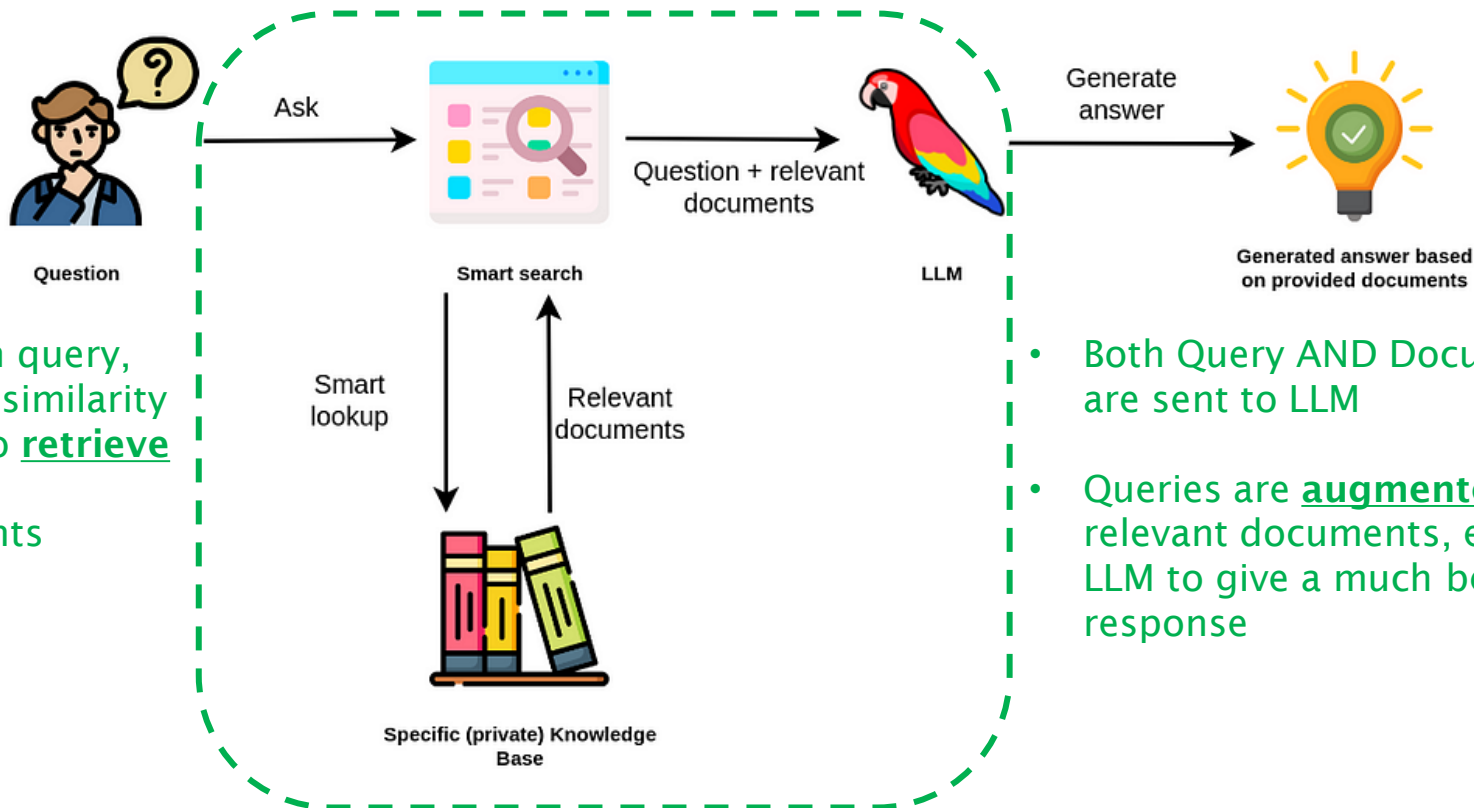
GPT4, LLaMA, PaLM



- SHIP-HATS Documentation
- Code Repos (reference pipelines, modular templates)

- Computationally expensive
- Resource intensive
- Needs to be repeated every time data is updated

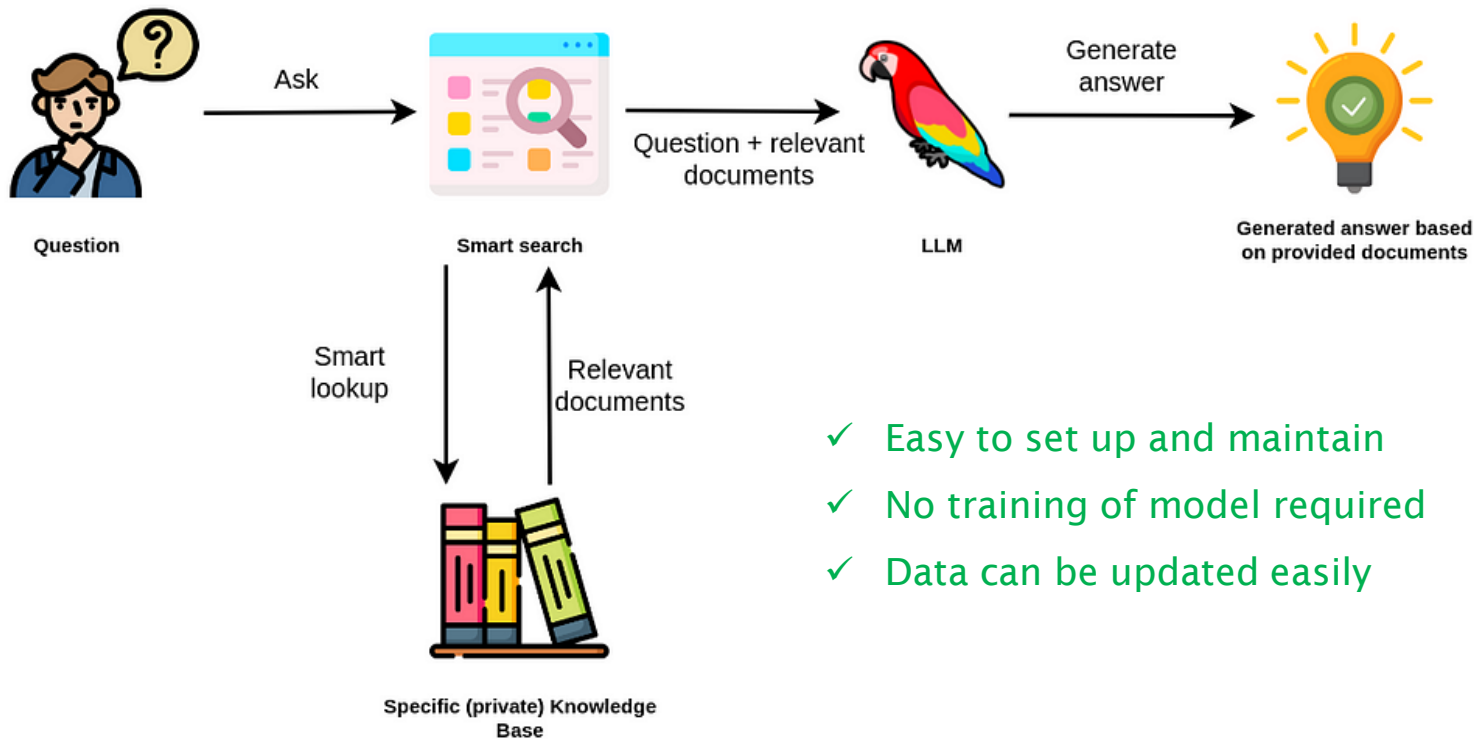
# Retrieval-Augmented Generation



- Based on query, first use similarity search to retrieve relevant documents

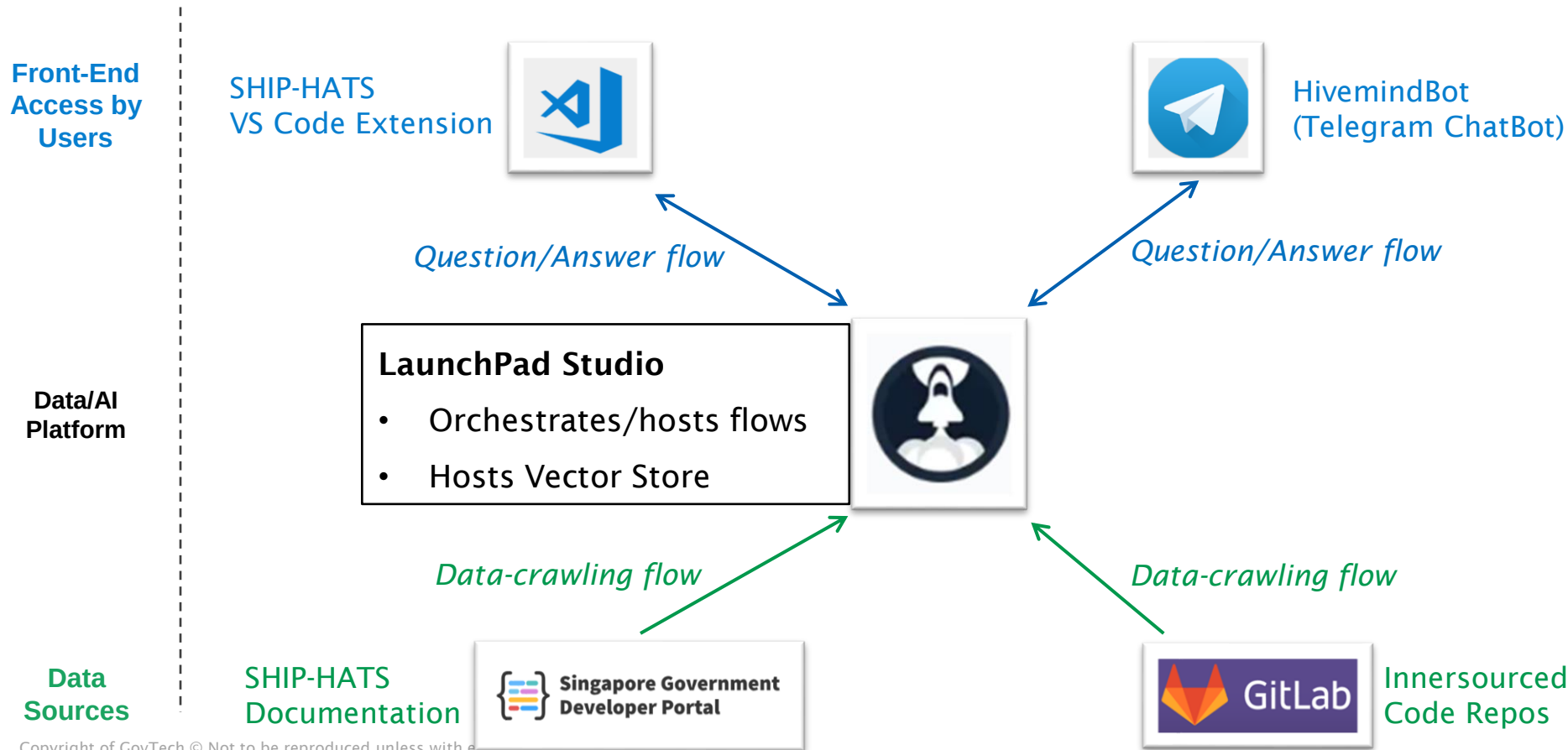
- Both Query AND Documents are sent to LLM
- Queries are augmented with relevant documents, enabling LLM to give a much better response

# Why RAG?



- ✓ Easy to set up and maintain
- ✓ No training of model required
- ✓ Data can be updated easily

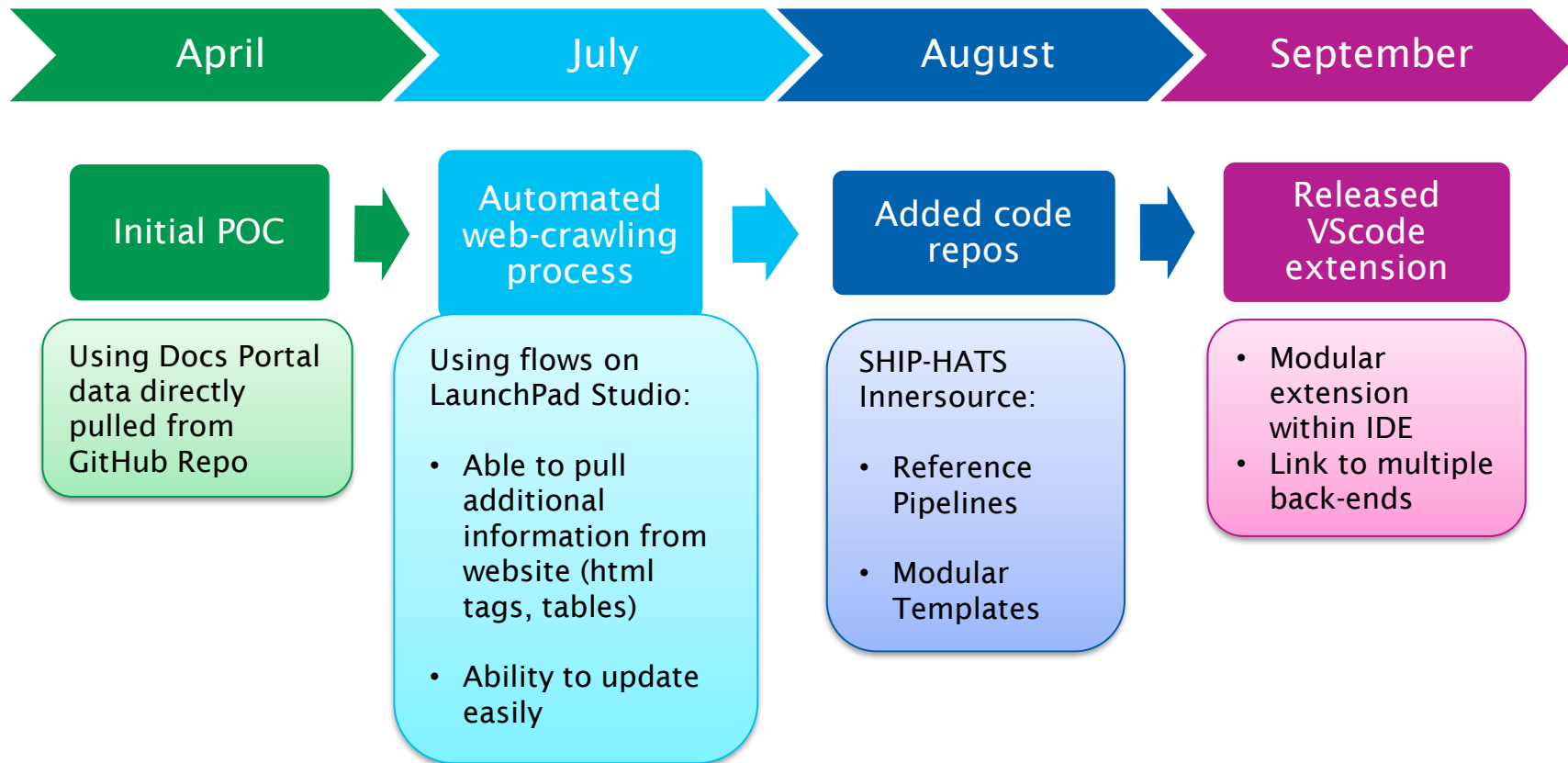
# Our RAG Setup



What have we learnt?

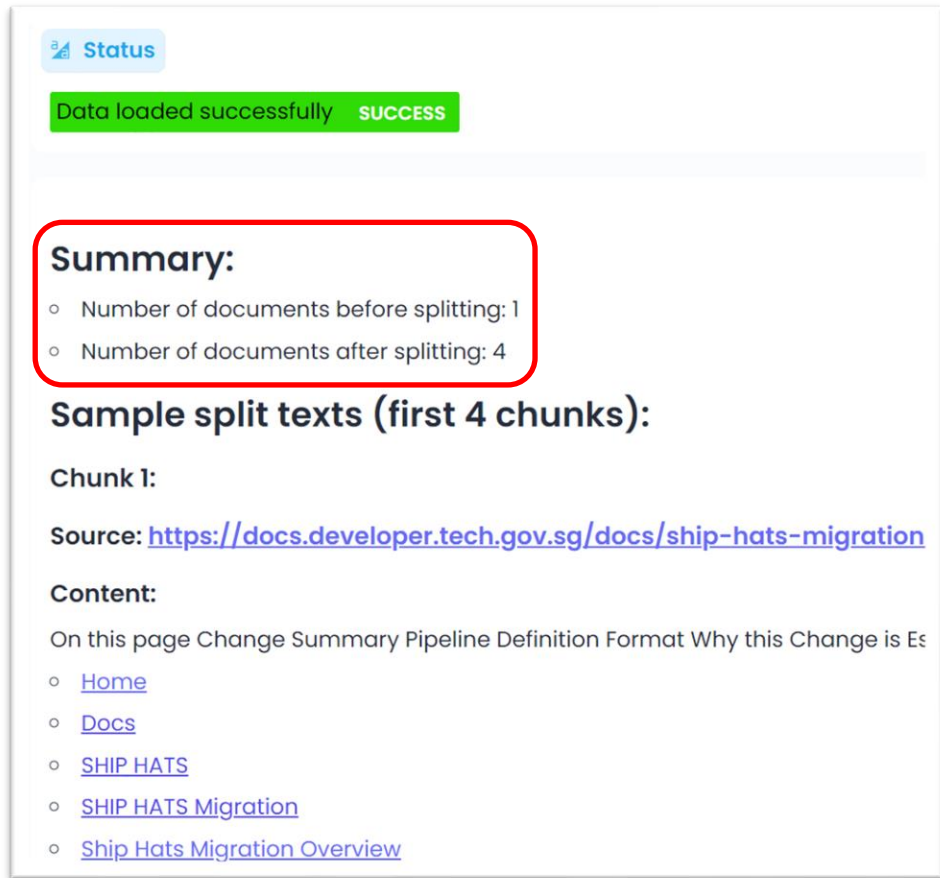


# Our Journey – An iterative process



# Lesson 1: Importance of Data Preparation

- Data Management and Quality: Importance of proper preparation and indexing of data
- When including additional data sources, test to see:
  - If it achieves intended purpose,
  - If it changes how responses are given



The screenshot shows a web interface with a 'Status' section at the top, indicating 'Data loaded successfully' with a green 'SUCCESS' badge. Below this, a red-bordered box highlights a 'Summary' section containing two bullet points: 'Number of documents before splitting: 1' and 'Number of documents after splitting: 4'. Underneath the summary is a section titled 'Sample split texts (first 4 chunks):' followed by 'Chunk 1:' and its source URL. The 'Content' section shows the beginning of a document about a change in the pipeline definition format, followed by a list of links: Home, Docs, SHIP HATS, SHIP HATS Migration, and Ship Hats Migration Overview.

Status

Data loaded successfully SUCCESS

**Summary:**

- Number of documents before splitting: 1
- Number of documents after splitting: 4

**Sample split texts (first 4 chunks):**

Chunk 1:

Source: <https://docs.developer.tech.gov.sg/docs/ship-hats-migration>

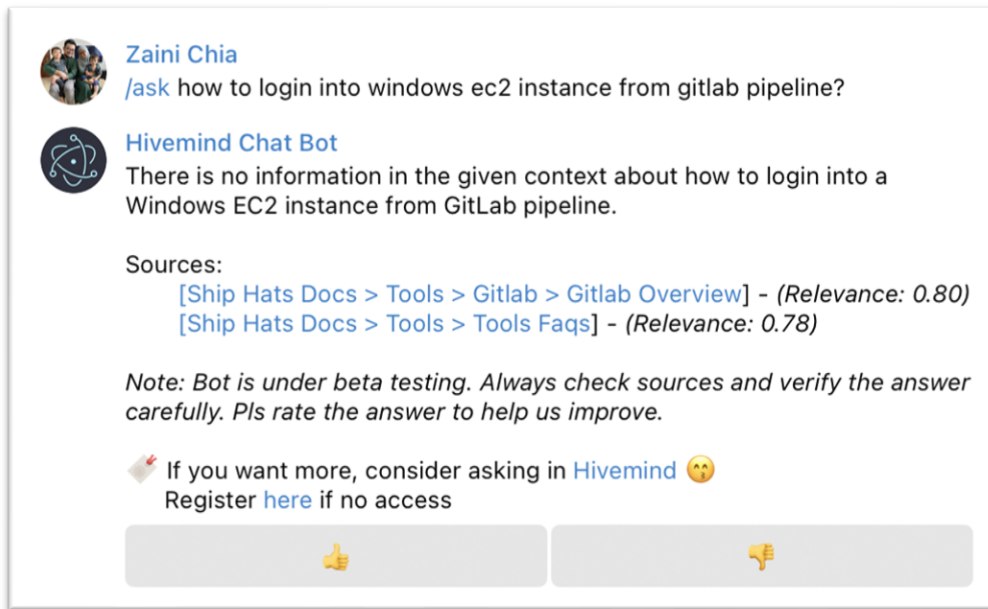
Content:

On this page Change Summary Pipeline Definition Format Why this Change is Es

- [Home](#)
- [Docs](#)
- [SHIP HATS](#)
- [SHIP HATS Migration](#)
- [Ship Hats Migration Overview](#)

## Lesson 2: RAGs limited by internal knowledge base

- RAG is useful for accuracy, but may reduce the versatility of the bot
- It essentially **limits the knowledge base** from which the LLM can provide answers from
- Our current HivemindBot may not answer generic coding questions well



 **Zaini Chia**  
/ask how to login into windows ec2 instance from gitlab pipeline?

 **Hivemind Chat Bot**  
There is no information in the given context about how to login into a Windows EC2 instance from GitLab pipeline.

Sources:  
[\[Ship Hats Docs > Tools > Gitlab > Gitlab Overview\]](#) - (Relevance: 0.80)  
[\[Ship Hats Docs > Tools > Tools Faqs\]](#) - (Relevance: 0.78)

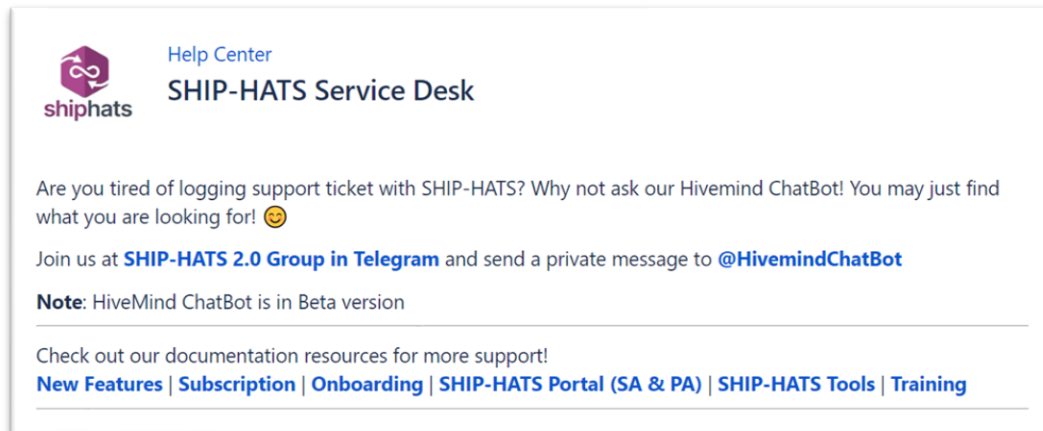
*Note: Bot is under beta testing. Always check sources and verify the answer carefully. Pls rate the answer to help us improve.*

 If you want more, consider asking in [Hivemind](#) 🙄  
Register [here](#) if no access

## Lesson 3: Usefulness of AI tools must be validated

- Attempted to get feedback but getting users to respond was a challenge
  - Improved UI for rating
- Being more intentional in understanding metrics to understand impact
  - e.g. reviewing whether the bot helped to reduce service tickets



# What's next?

- Improving on RAG: allowing bot to use its own knowledge base OR searching the internet for answers
  - Need to consider accuracy, sensitivity
- Additional data sources: other SGTS Products such as StackOps, Container Stack
  - Testing is required to see how additional data sources will impact chatbot responses; may be better to have different chatbots for different use cases

# Revisiting our Hypothesis

## Our hypothesis:

Using Gen AI tools that are customized to connect to our internal resources will improve dev productivity and satisfaction

- There is some promise as we see more people trying out the bot and giving user feedback
- More needs to be done to enrich the data source and to collect user feedback over a longer period to fully validate the hypothesis

# LLM Stack



## Today's Challenges

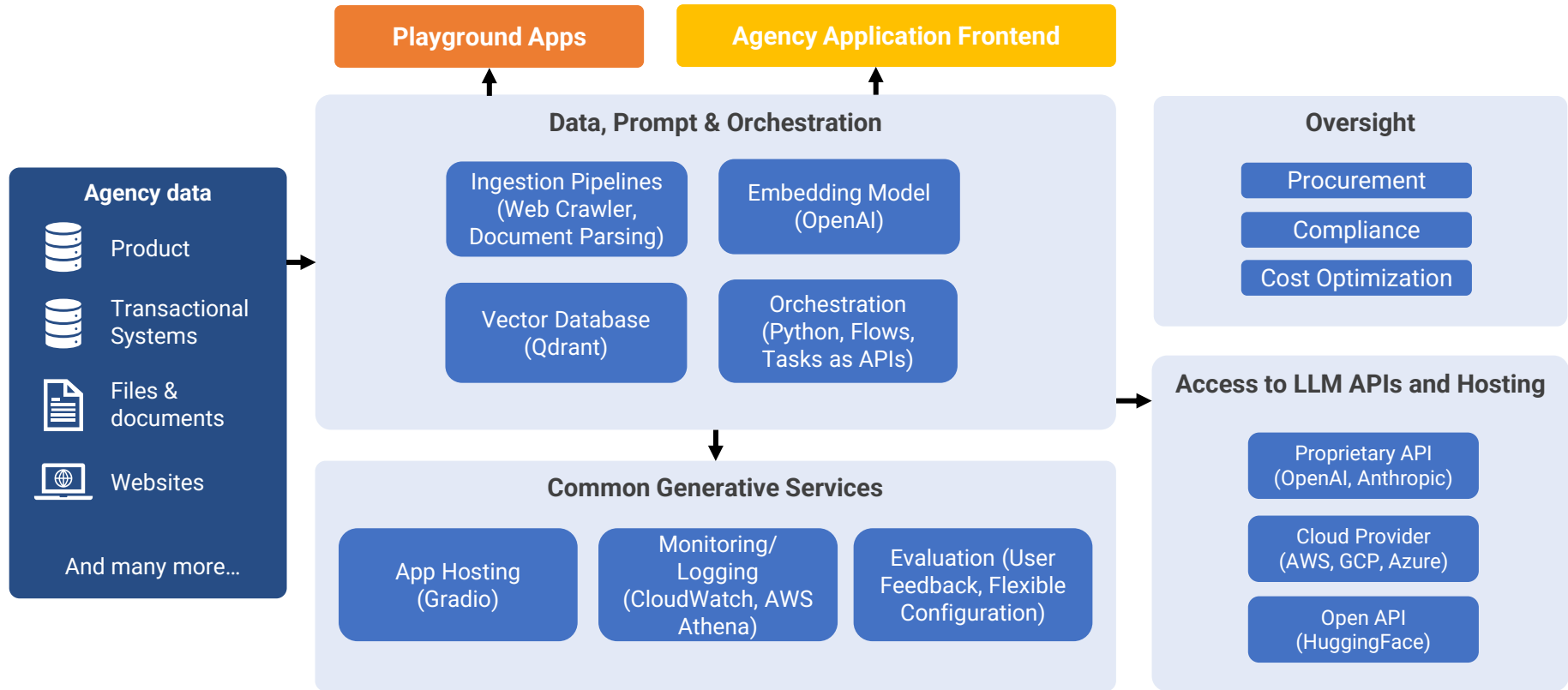
### Developers

- “It’s quite tedious to launch LLM components to production after building a PoC”
- “I would like to have an environment of sorts to test some LLM ideas I have”
- “I want to evaluate different models and settings to improve accuracy, but it takes a lot of effort.”

## What is the Central LLM Stack?

- A platform for quick prototyping and production launch of LLM-powered applications
- Ready-made reusable templates to accelerate LLM powered apps in minutes
- Suite of components to boost LLM Ops including quick model evaluation and user feedback on response

# Technical Overview



# Benefits

## Accelerate

- Procurement of LLM services, including legal clearance
- Scalable IM8-compliant infrastructure, up to Confidential (Cloud-eligible)
- Managed vector store service

## Optimize

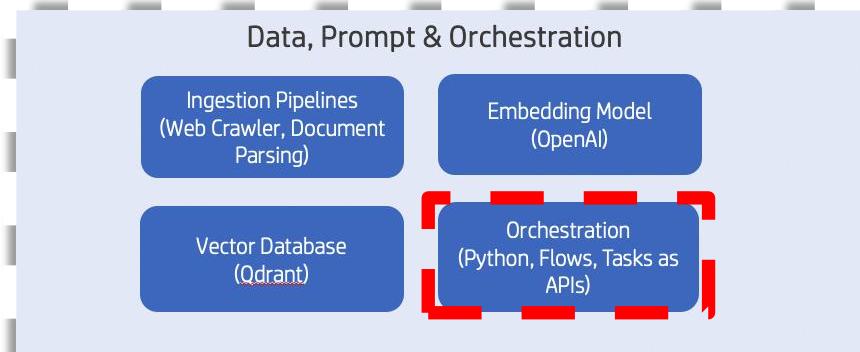
- Easily access & evaluate different LLMs incl. Open AI, Google, Open Source
- LLM Comparison Tool, to compare responses between different models
- User Feedback On Response

## Empower

- Enable users (DSAI champions, SWEs, etc.) through ease of prototyping and discoverability
- Useful apps & bots within minutes
- Enable chaining of reusable components for LLM Workflow

# Components - Orchestration

- **Flows and Tasks** – each output is an input to another block
- **Ability to connect several Tasks together for Flow archetypes**
  - o LLM Proxy
  - o Document Parsing
  - o Question Answering
    - With translation
    - With documents
  - o Vector Store Management
  - o Search
  - o Summarization



# Reusable Components – Q&A

LLM Component :

Embeddings



VectorStoreRetriever

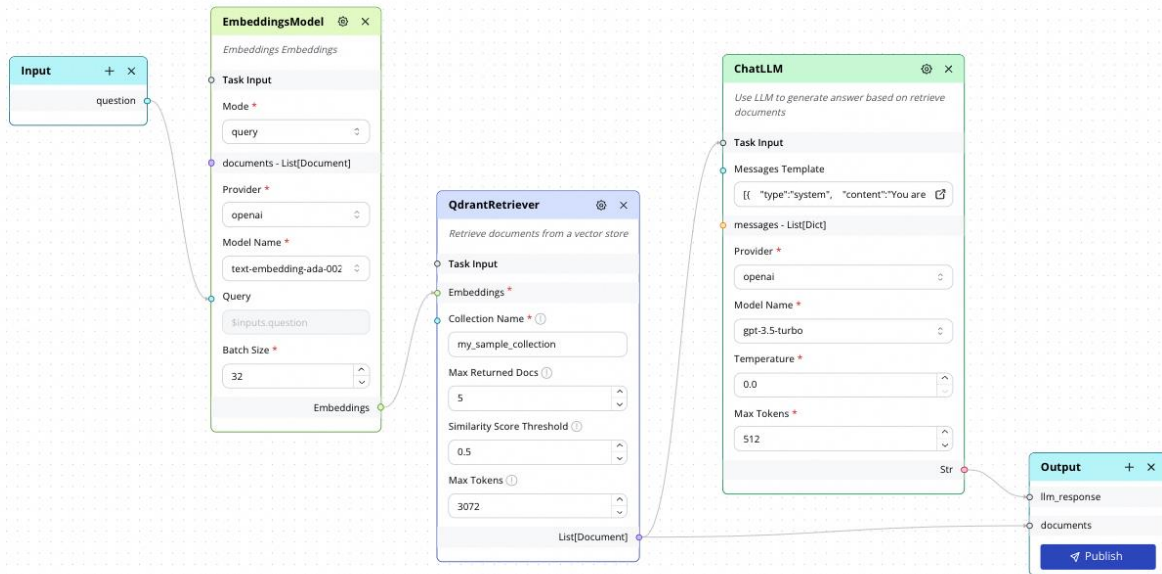


LLM

Question and Answer  
(using Retrieval  
Augmented Generation)

## Use Cases

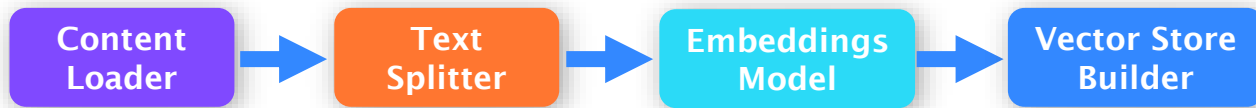
- SHIP-HATS Support Bot
- LaunchPad AI Bot



# Reusable Components – Vector Store Building



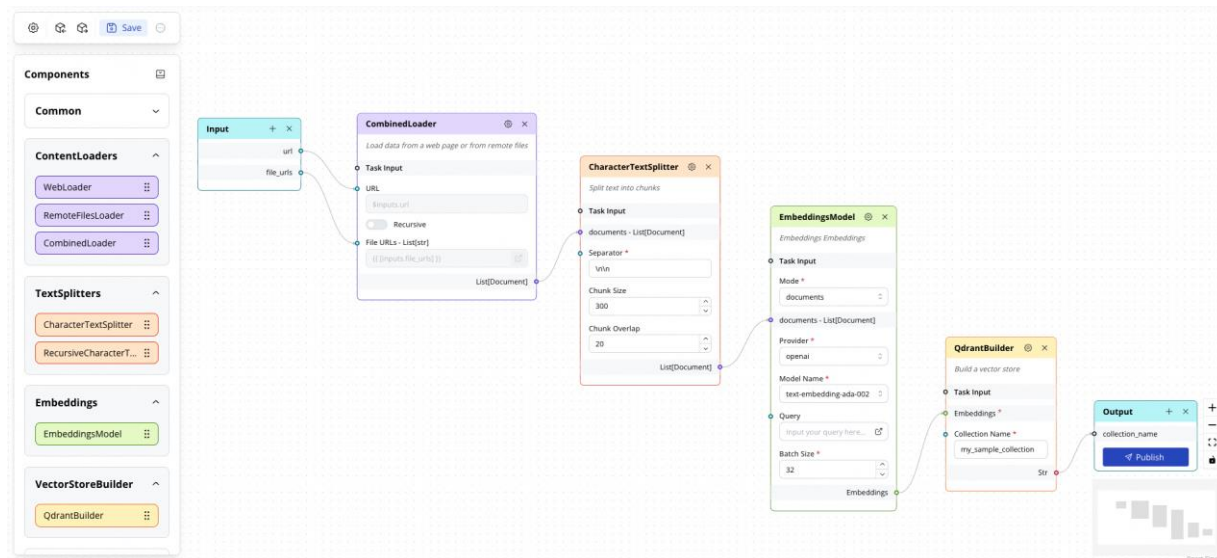
LLM Component :



Allows uploading of files or scraping through websites and then storing in a vector store database for subsequent workloads such as Q&A, topic modeling.

## Use Cases

Agency needs to use vector store for own purposes



## Contact us



<https://go.gov.sg/llm-stack-interest>

## Reach out to:

- Janice Tan ([janice\\_yp\\_tan@tech.gov.sg](mailto:janice_yp_tan@tech.gov.sg))
- Dominic Phoon  
([dominic\\_phoon@tech.gov.sg](mailto:dominic_phoon@tech.gov.sg))

# Round-up

- **Copilot / Duo Evaluation:** stay tuned for part II next year – reach out if interested
- **Ghost:** ([Try it out today!](#))
- **Hivemind Chatbot** ([Learn more](#))
- **SHIP-HATS VS-Code Extension** ([Learn more](#))
- **LLM Stack** ([contact](#))

Thank You

