INTERPOL INSTRUCTOR DEVELOPMENT COURSE

BUILDING A FUNCTIONAL RETRIEVAL-AUGMENTED GENERATION (RAG) SYSTEM WITH OPEN WEB UI

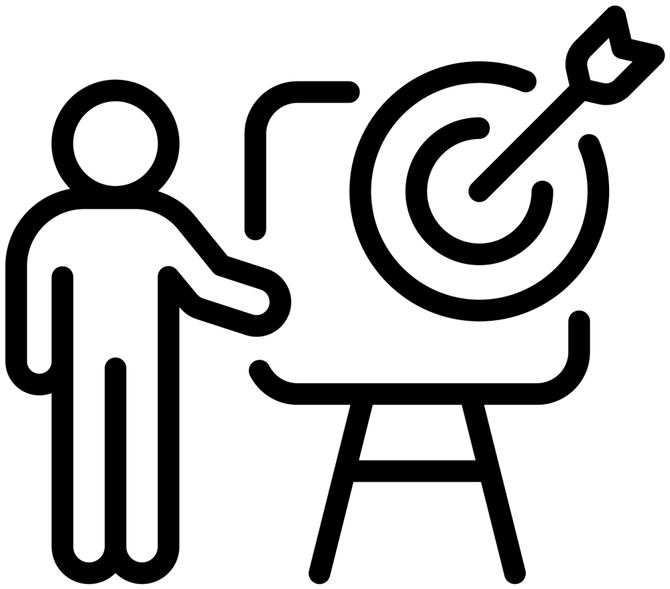
AI AWARENESS PROGRAM FOR SPECIAL BRANCH MALAYSIA

DSP Ts. NORZAIDI BAHARUDIN
Certified MQA Trainer
Auditor Technology & Technical Accreditation Council (TTAC) MBOT

"I BELIEVE AI IS GOING TO CHANGE THE WORLD MORE THAN ANYTHING IN THE HISTORY OF HUMANITY. MORE THAN ELECTRICITY"
– DR. KAI-FU LEE, FORMER PRESIDENT & CEO OF GOOGLE CHINA

GOALS & OBJECTIVE

USING THE OPEN WEBUI INTERFACE, PARTICIPANTS WILL IMPLEMENT A FUNCTIONAL RAG SYSTEM CAPABLE OF RETRIEVING RELEVANT CONTEXT FROM UPLOADED DOCUMENTS TO ACCURATELY ANSWER USER QUERIES.



01

Given access to the live Open WebUI environment, participants will list at least three main features of the interface and explain the basic components of a RAG system, as demonstrated during a guided walkthrough.

02

Using the live Open WebUI interface, participants will configure the essential components of a basic RAG pipeline (e.g., document collection & embedding settings), successfully saving and running the configuration as verified by the instructor.

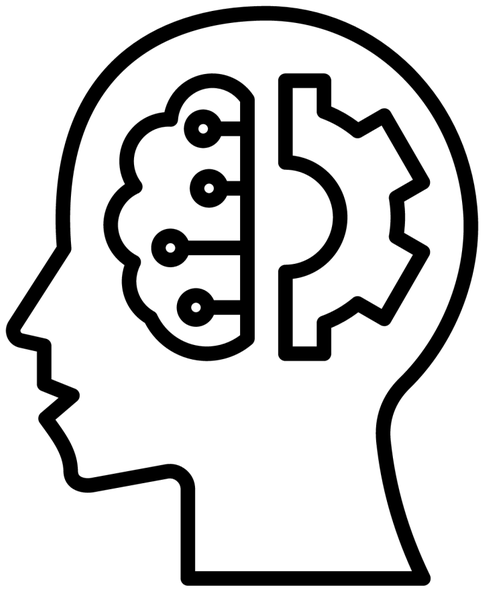
03

After uploading example documents in the Open WebUI, participants will verify and demonstrate that each file has been correctly assigned to the correct collection and that embeddings are successfully generated, as shown in the interface status/logs.

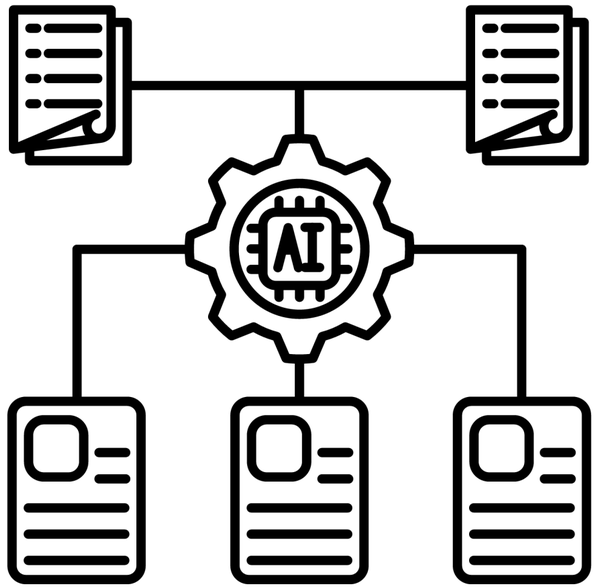
04

Given one or more user queries, participants will evaluate the accuracy and relevance of answers generated by the RAG system in Open WebUI, and report on whether the retrieved context aligns with the queries, achieving at least 80% correct identification as measured by an instructor-provided answer key.

INTRODUCTION



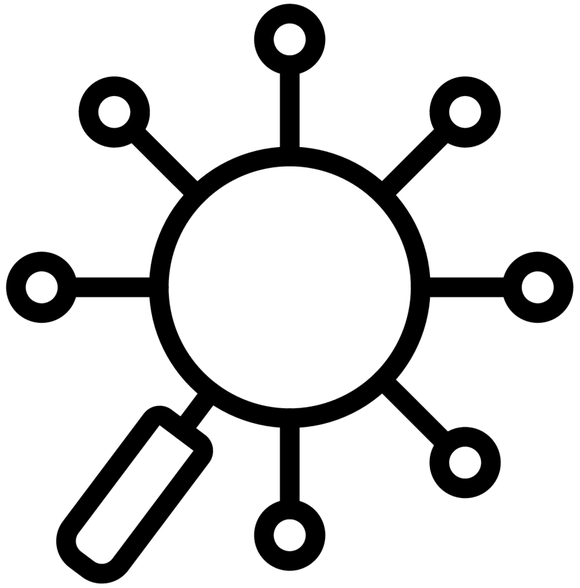
- Before we talk about RAG, we need to understand how AI has changed over time.
- The “AI” discussed in the 1950s was rule-based — systems that followed fixed instructions and could not learn or create anything new.
- Today’s AI is Generative AI, which can produce new content such as text, images, audio, and video.
- Generative AI works because of Large Language Models (LLMs) — the “brain” that understands language, reasons, and generates responses.
- Popular LLMs include GPT-5.1 (OpenAI), Gemini 2.5 Pro, Claude, Copilot, and many open-source models like Llama.
- Modern AI is no longer about simple automation; it is about intelligent content generation.



WHAT IS RAG ?

Retrieval-Augmented Generation (RAG) is a technique that combines a Large Language Model with your own knowledge base so the AI can give answers that are accurate, updated, and grounded in your data. Instead of relying only on the model's memory, RAG retrieves relevant documents, embeds them, and feeds the context to the LLM, allowing the AI to answer questions using verified information from your files, reports, SOPs, and intelligence archives — reducing hallucination and turning AI into a reliable, domain-specific assistant.

WHY YOU NEED RAG ? WIIFM



Knowledge Continuity
RAG preserves operational knowledge so it stays in the unit even when officers retire, transfer, or change roles.

Rapid Intelligence Retrieval
Officers can quickly access past reports, case notes, and SOPs through an AI chatbot instead of manual searching.

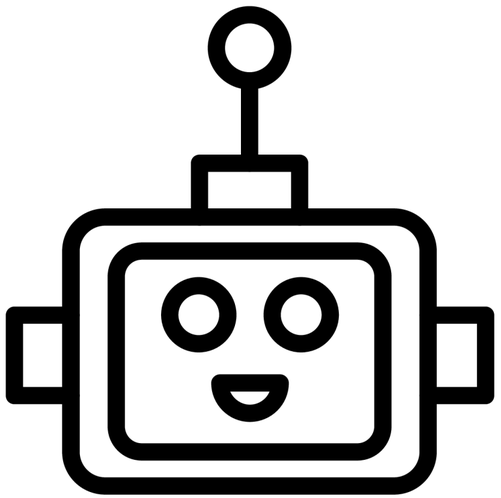
Better Decision-Making
LLM + RAG provides accurate, up-to-date insights, helping officers make faster and more confident operational decisions.

Enhanced Intelligence Analysis
AI assists in extracting patterns and insights from large datasets such as
OSINT, digital forensics, chat logs, and PDFs.

Higher Productivity & Efficiency
Automation of routine tasks (summaries, drafting, documentation) frees officers to focus on core investigative and analytical work.

AI TOOLS (LOCAL PC / LOCAL SERVER INSTALLATION)

- **GPT4All –** simple runtime for desktops/laptops, supports many open-source models.
- **Ollama –** user-friendly tool to run LLMs locally with one-line commands.
- **LM Studio –** desktop app for running, testing, and comparing models.
- **Open WebUI –** modern interface for LLMs with built-in RAG, agents, extensions.
- **AnythingLLM –** full-featured RAG suite; upload files and chat with your data.
- **PrivateGPT –** run private Q&A and document analysis offline.

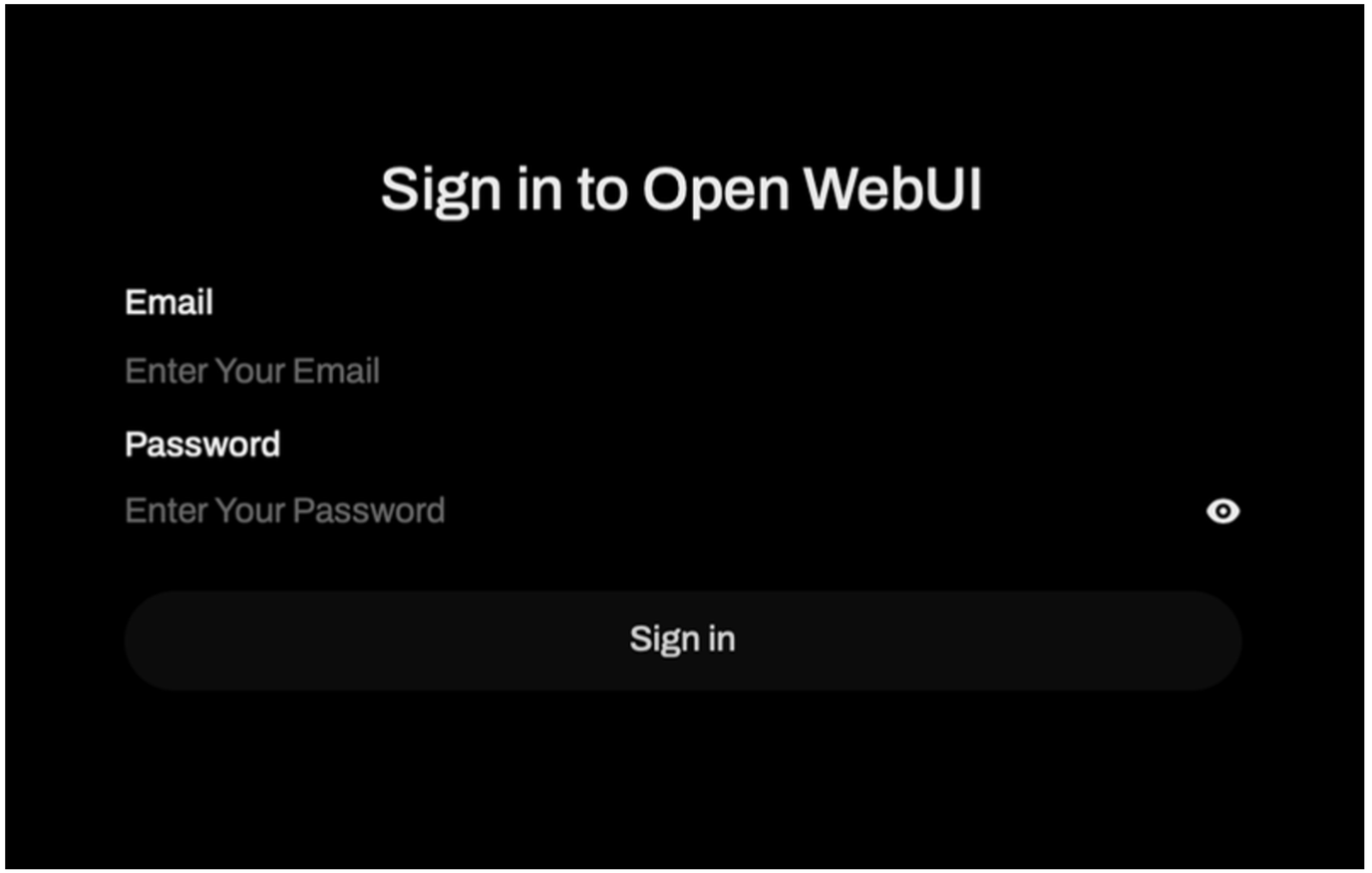


THE DEMO

Login - demo@matnet.my

Passwd - interpol2025

THE DEMO



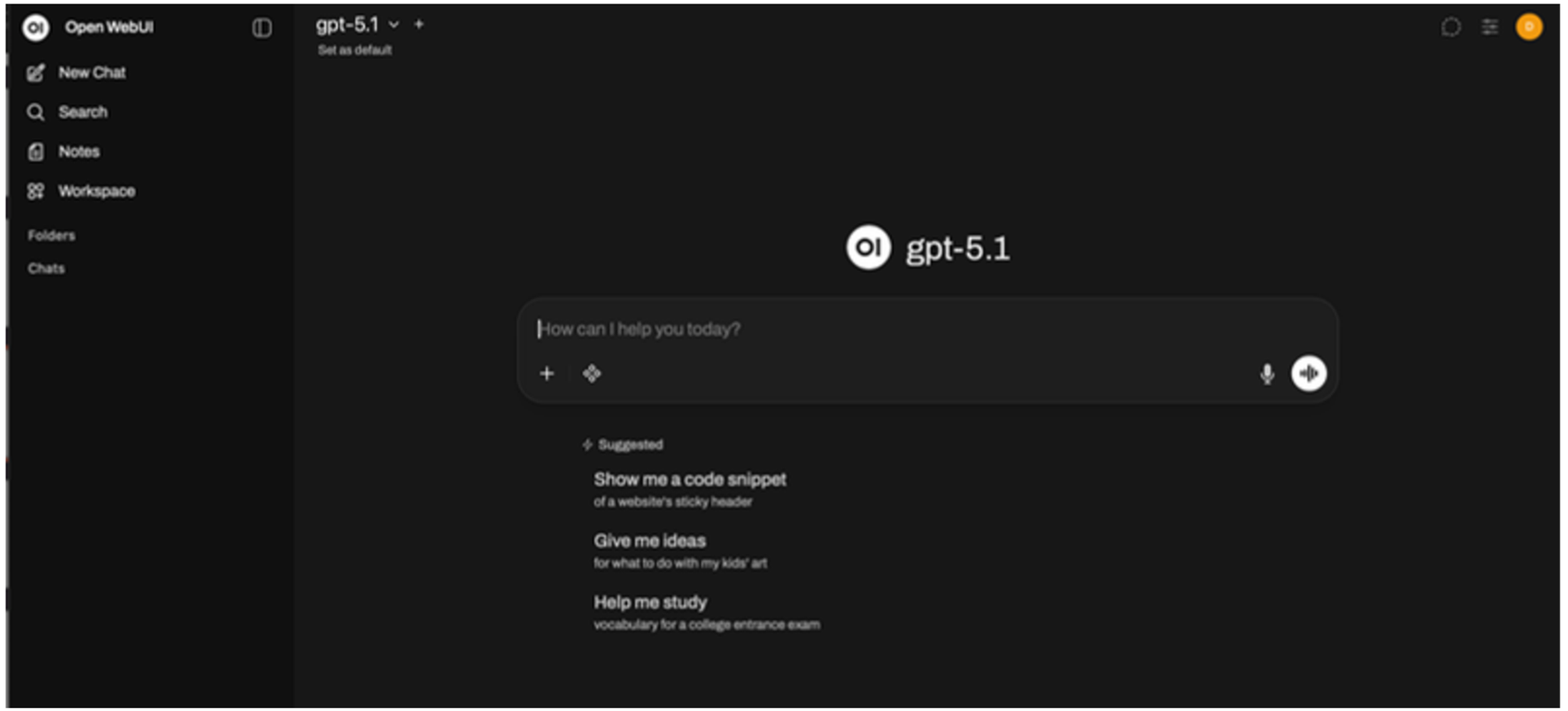
Login - demo@matnet.my

Passwd - interpol2025

QUICK SURVEY.

HOW DO YOU FEEL ABOUT THE OPEN WEBUI INTERFACE? IS IT USER-FRIENDLY OR DOES IT FEEL COMPLICATED?

THE DEMO

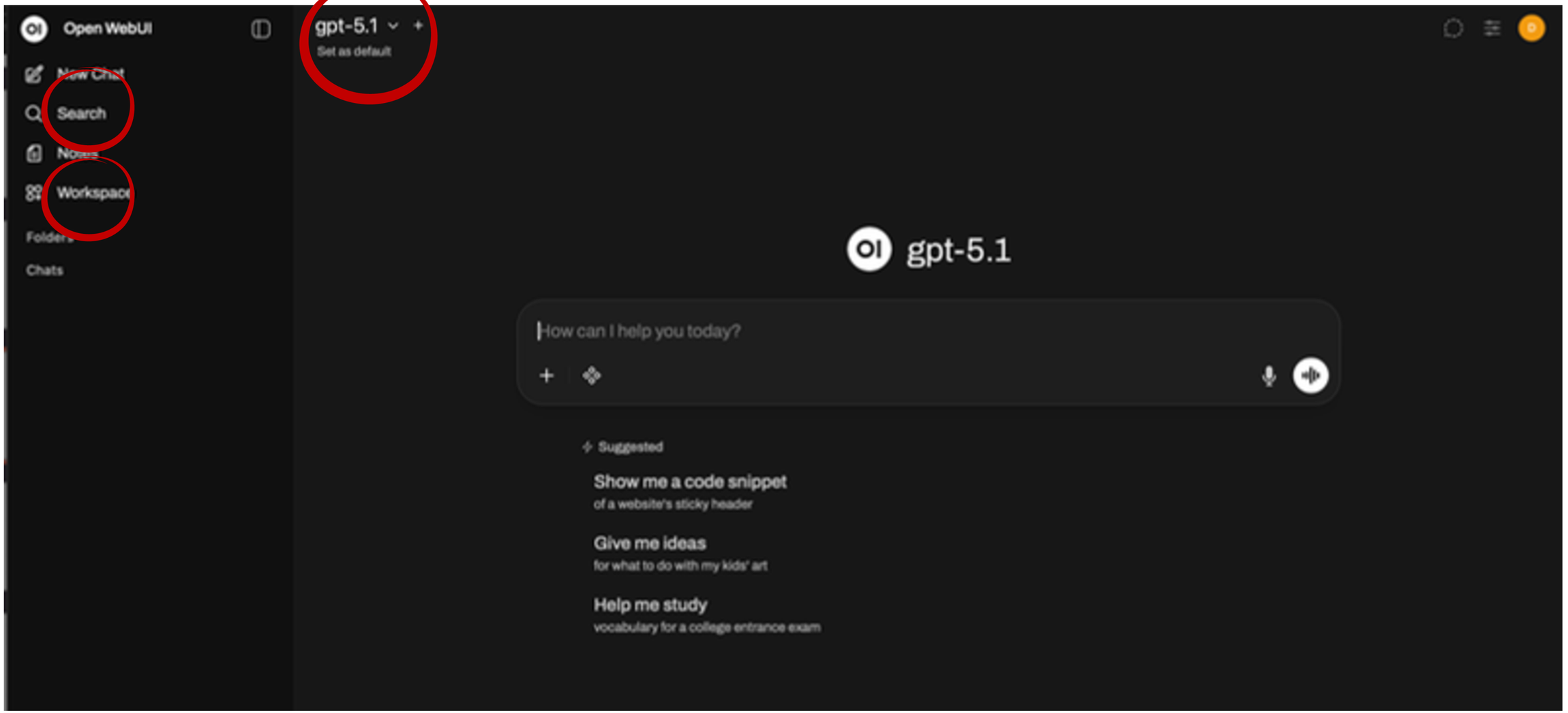


Login - demo@matnet.my

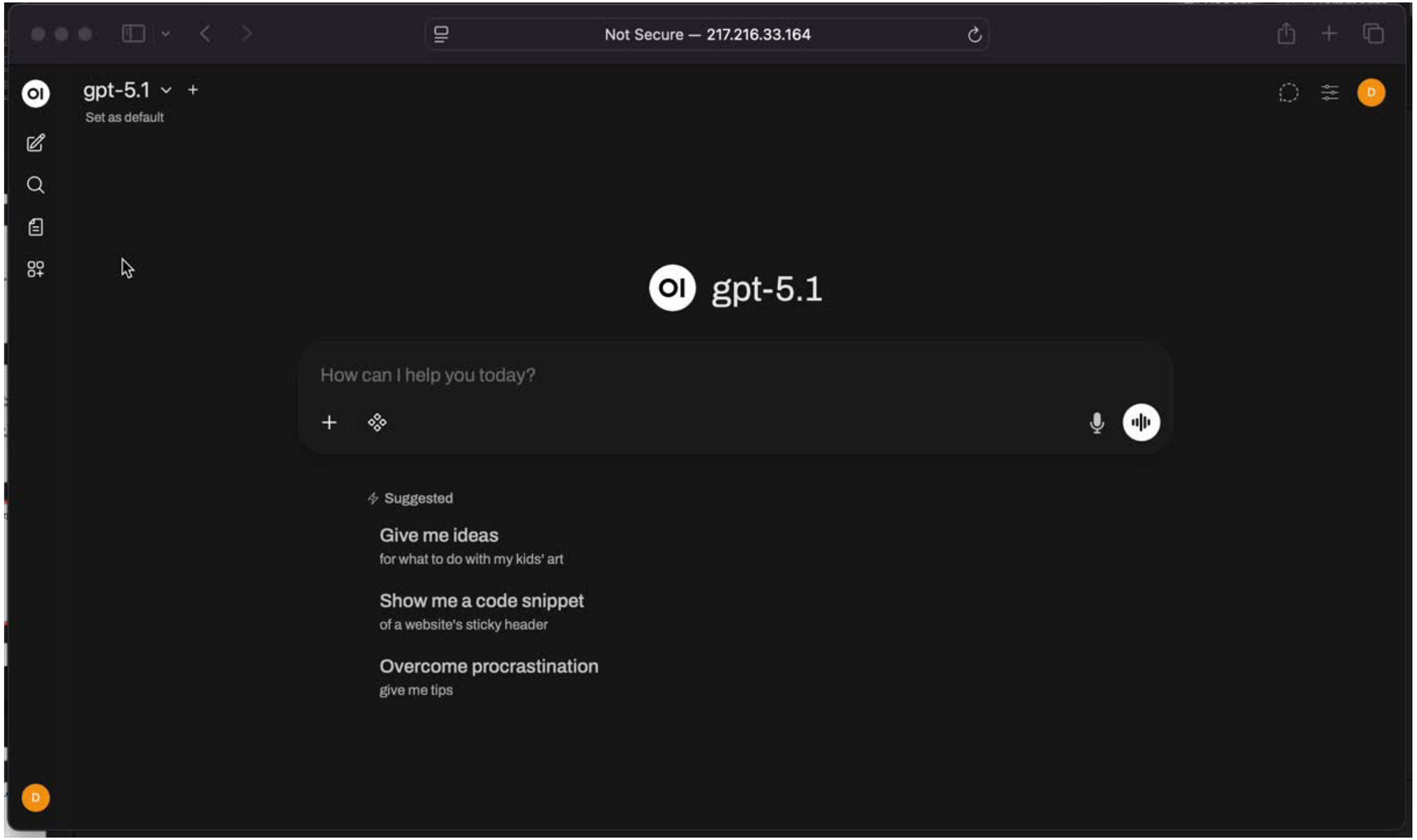
Passwd - interpol2025

QUICK QUIZ

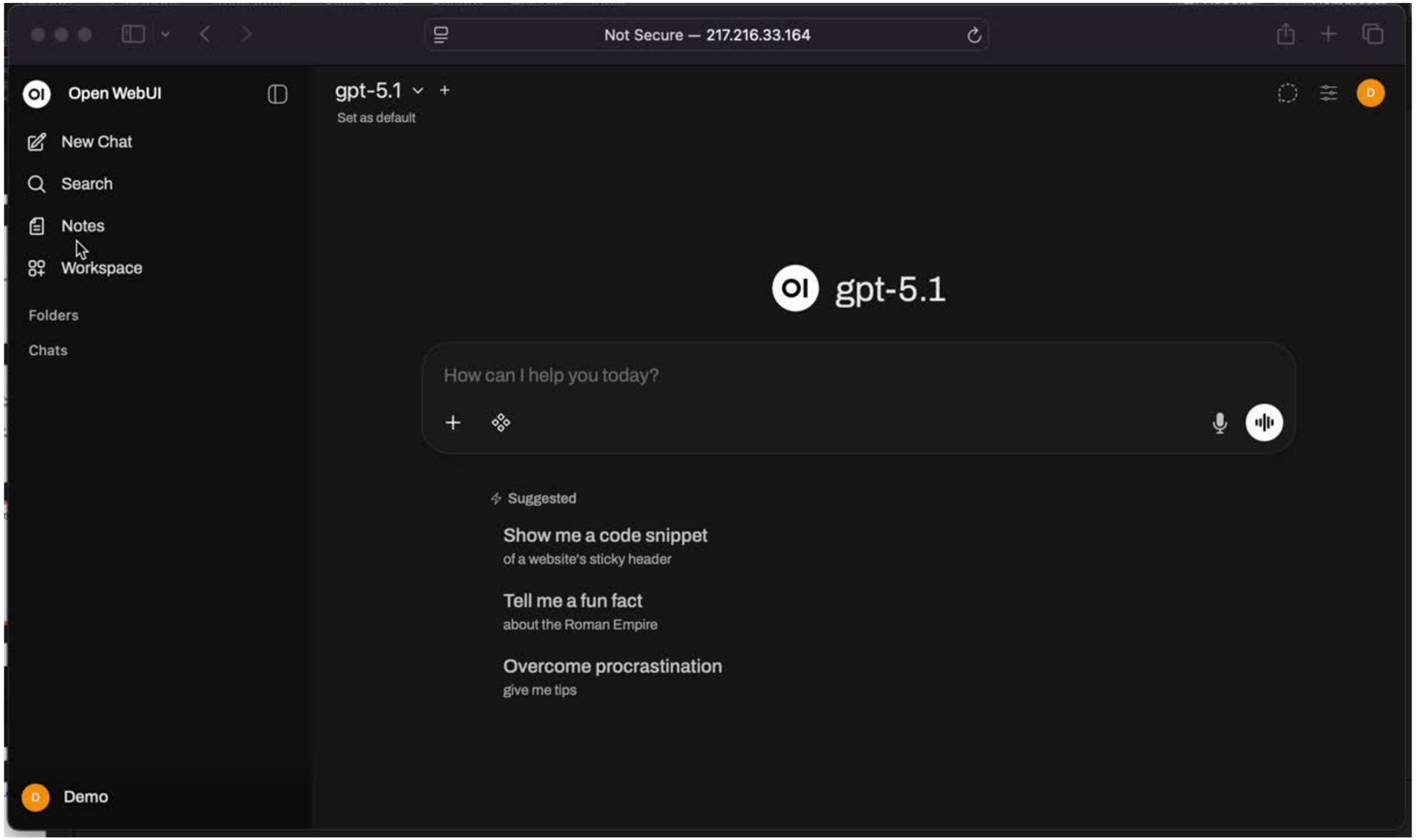
List 3 basic functions of OpenWebUI ? (LO1)



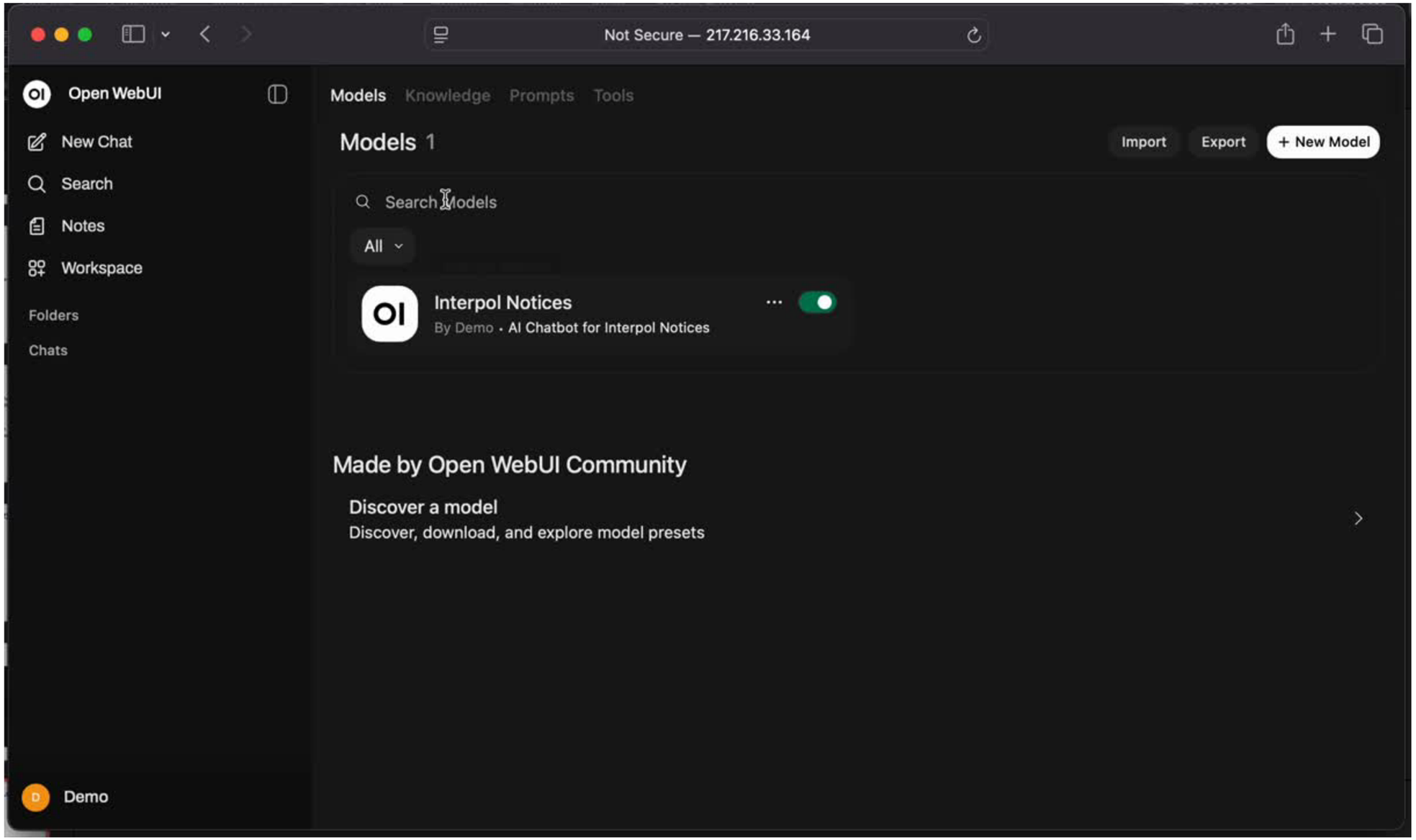
CREATE KNOWLEDGE BASE (LO2)



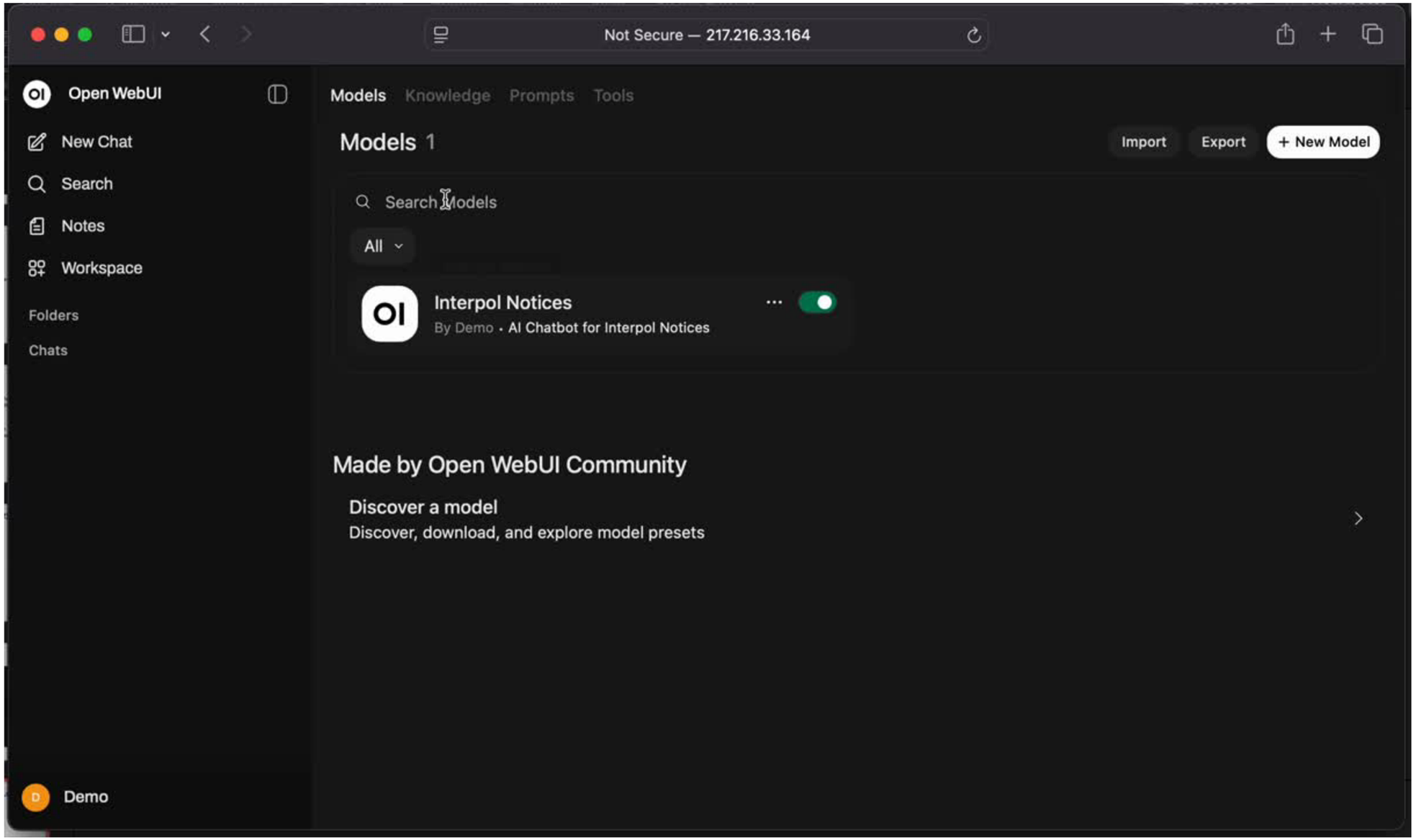
CREATE MODEL (LO2)



UPLOADING DOCS AND VERIFY (LO3)



EVALUATE CHATBOT ASNWER (LO4)



SUMMARY

- Today’s AI is powered by LLMs and becomes truly useful when combined with our own knowledge through RAG.
- RAG solves the problem of lost expertise, fragmented documents, and inconsistent information in daily operations.
- Tools like Open WebUI allow us to build secure, private, and domain-specific AI assistants using our internal data.
- With proper setup collections, embeddings, vector databases, and queries officers can retrieve accurate, up-to-date intelligence instantly.
- Integrating RAG into our workflow improves decision-making, productivity, and knowledge continuity across the organization.
- As Suzy AI showed at the start, today's AI is intelligent; RAG ensures that intelligence is also accurate and relevant to our mission

THANK YOU FOR YOUR ATTENTION

PHONE
+6019-3846574

WEBSITE
www.matnet.my

EMAIL
norzaidi@rmp.gov.my

ADDRESS
Kuala Lumpur