



LLM-powered chatbots for document question answering

AI Workshop - ODATIS
Jan 2026, Marseille

Redouane Lguensat ^{1,2}

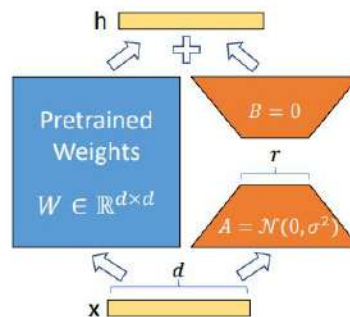
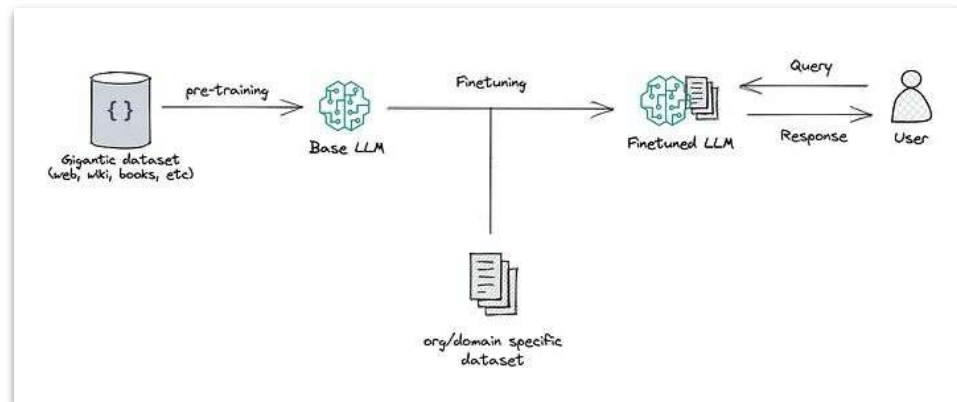
¹ Institut Pierre Simon-Laplace (IPSL)

² IRD (French National Research Institute for Sustainable Development)

First option: Fine tuning ! but

- Fine tuning involves training the LLM
- Expensive in terms of compute/financial cost
- You stick to the particular LLM you finetuned
- You need to refinetune if you would like to add more data

it is a good option if you have high confidence on your LLM and if you have the technical skills and knowledge needed



LoRA technique

🧠 PEFT: State-of-the-art Parameter-Efficient Fine-Tuning.

huggingface.co/docs/peft

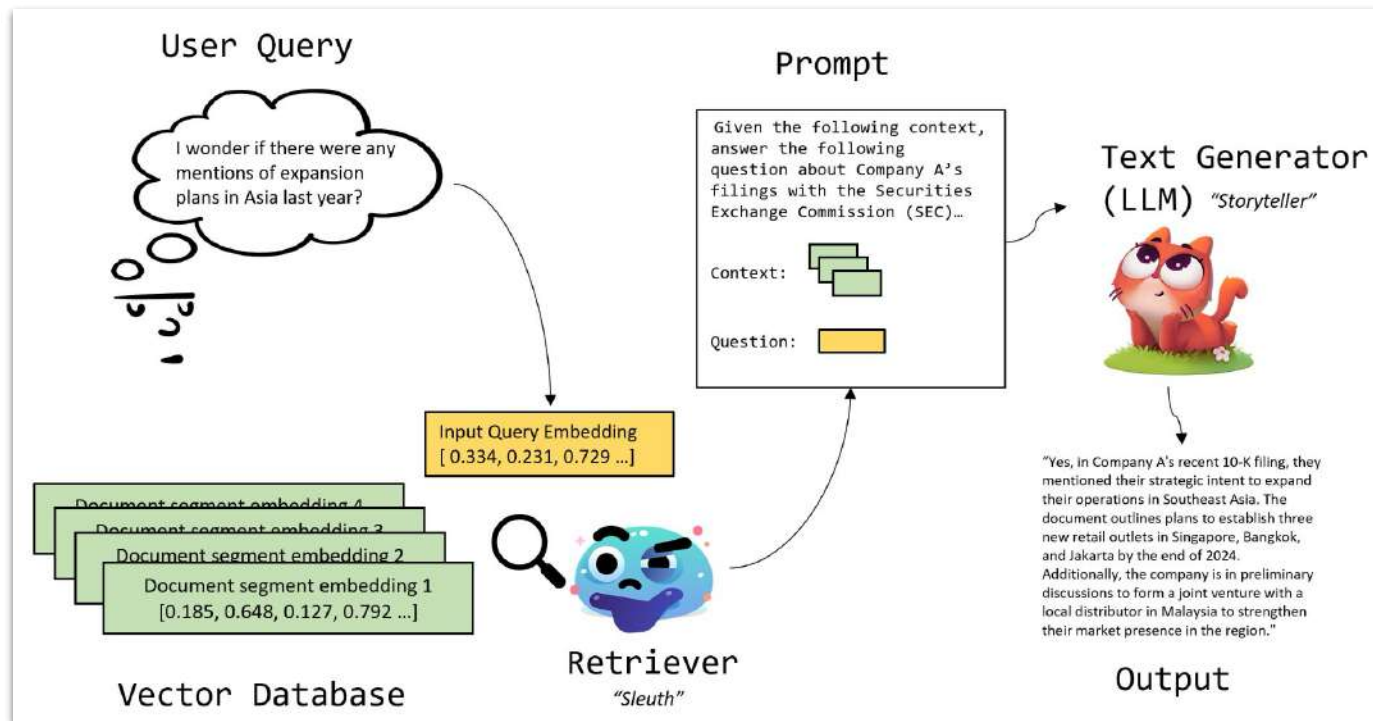
Python Library

Augmenting LLMs with private data

RAG: Retrieval Augmented Generation (Lewis et al. 2020)

Second option: RAG

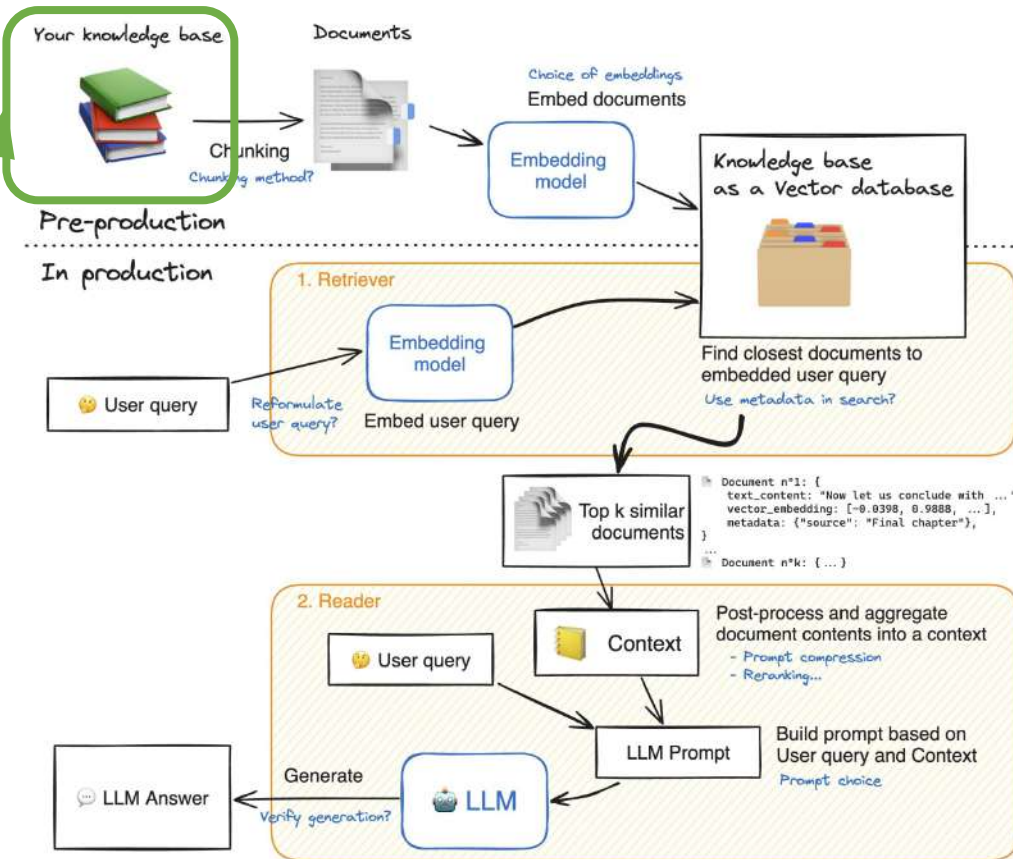
- Simple, intuitive, not very costly
- Flexible, you can plug and play any LLM
- Needs some Prompt engineering



Demo with RAG on the NEMO documentation

General Idea

Latex files from the NEMO forge
<https://forge.nemo-ocean.eu/nemo/doc/manuals>
 +
 Rst files from the user guide
<https://forge.nemo-ocean.eu/nemo/doc/guide>





RAG: Retrieval Augmented Generation

Option 1: create your GPT with OpenAI

GPTs

Discover and create custom versions of ChatGPT that combine instructions, extra knowledge, and any combination of skills.

Search: Lguensat

All

NEMOfinder
Answers questions using the NEMO V4.2 manual
By Redouane Lguensat · 1

Featured
Curated top picks from this week

- Adobe Express**
Stand out with Adobe Express. Quickly and easily make impactful social posts, images, videos, flyers...
By adobe.com
- Code Copilot**
Code Smarter, Build Faster—With the Expertise of a Top Programmer by Your Side.
By promptspellsmith.com
- Social Butterfly-AI 2.0**
AI Assistant for content development and social media strategy. Nuanced, guided...
By digitalmogul.co
- Software Architect GPT**
Builds new software architecture documents by understanding user requirements and design...
By V B Wickramasinghe

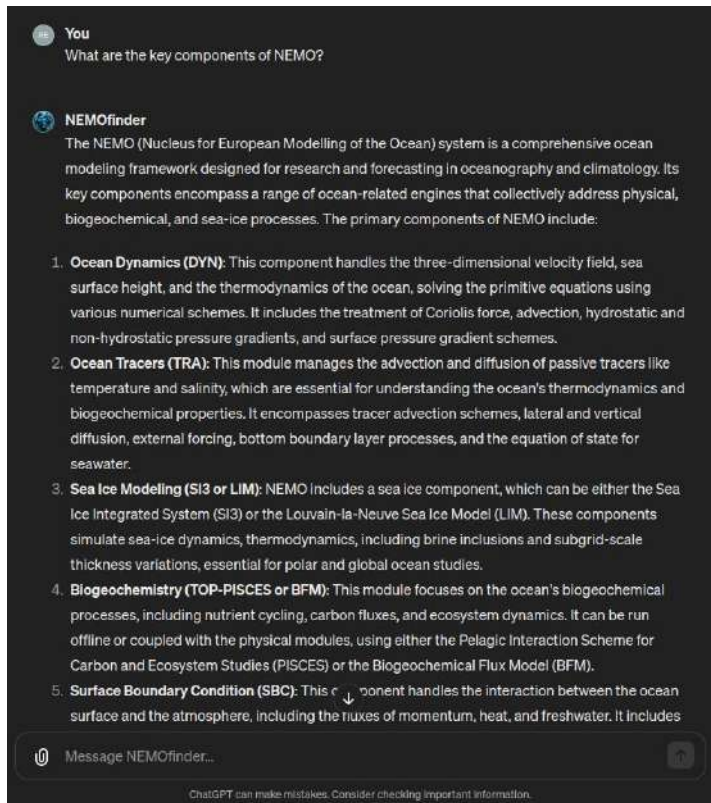


RAG: Retrieval Augmented Generation

Option 1: create your GPT with OpenAI

Advantages / Drawbacks

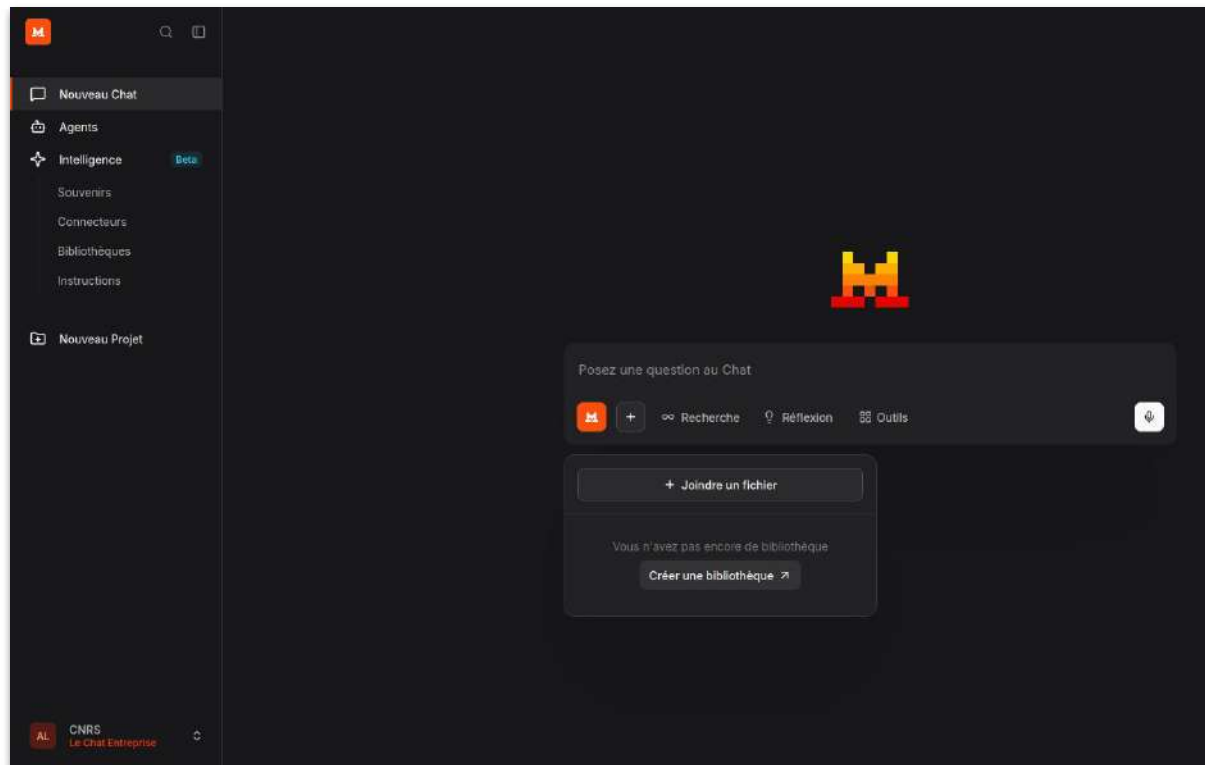
- VERY simple and easy to create, you only need to upload your document
- Shared link (you can access mine here <https://chat.openai.com/g/g-PFUCFXueM-nemofinder>) but only if you and your colleagues have a ChatGPT Plus subscription (20\$ per month)
- Based on one of the best Embedding + LLM at the moment but closed source





RAG: Retrieval Augmented Generation

Option 2: use Emmy the CNRS assistant (based on MistralAI)





RAG: Retrieval Augmented Generation

Option 2: use Emmy the CNRS assistant (based on MistralAI)



Emmy* est un agent conversationnel en intelligence artificielle générative, respectueux de vos données personnelles et professionnelles.

Bonjour Redouane LGUENSAT

Vous n'avez pas accès à Emmy

L'accès à Emmy est réservé aux agents CNRS et aux directeurs d'unités.



RAG: Retrieval Augmented Generation

Option 3: rely on Open Source solutions and models 🏆

You can do from scratch (like me two years ago...)

- Needs a lot more work
- This field changes **TOO much** and **TOO quickly**, each month a new model !
- **Free !** (if you decide to not use APIs)
- Still not reaching the performance of closed source models but the gap is closing...



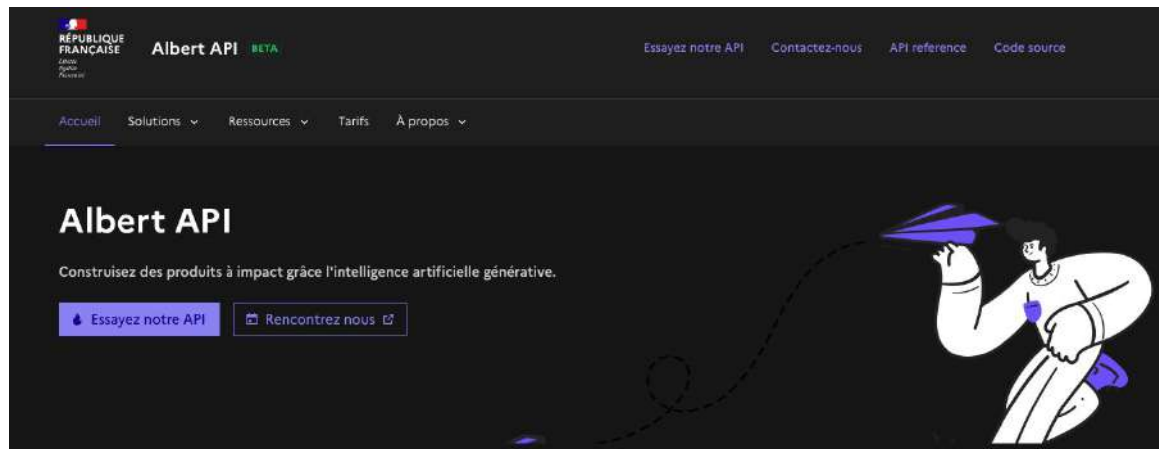


RAG: Retrieval Augmented Generation

Option 3: rely on Open Source solutions and models 🏆

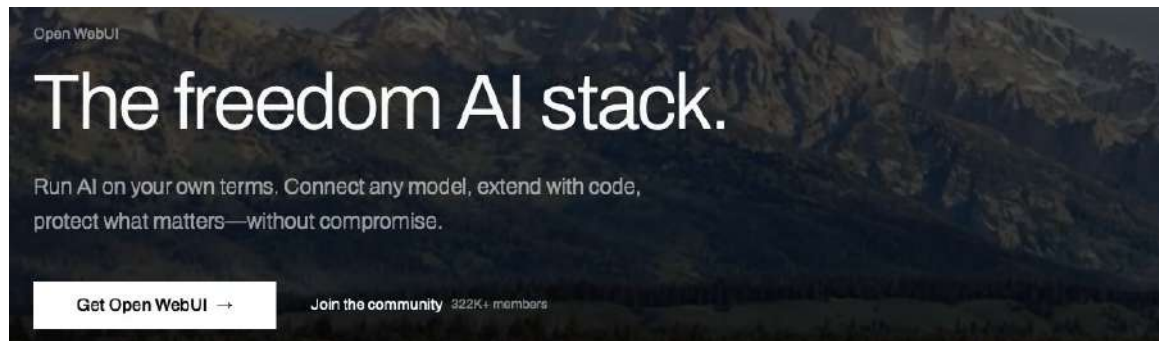
Free API access to the best Open Source LLMs !

- **Albert API** by DINUM 🇫🇷
(big thanks to them for the quality work)
- Free for all the French public servants
- Privacy preserving (SecNumCloud certified)



Open Source platform with built-in RAG !

- **Open WebUI** is a user-friendly self-hosted AI platform designed to operate entirely offline.





RAG: Retrieval Augmented Generation

Option 3: rely on Open Source solutions and models 🏆

The screenshot shows a web interface for 'Assistant NEMO'. On the left is a sidebar with navigation links: 'Open WebUI', 'New Chat', 'Search', 'Notes', 'Workspace', 'Folders', 'Chats', 'Today', 'NEMO Core Components', 'Bathymetry Types in NEMO', 'Previous 30 days', 'ESPRi Research Overview', '2025', 'Accès resto CROUS', 'Contacter les RH IPSL', 'Directeur de l'IPSL', and 'Conversation Assistance'. The main area displays a chat with 'Assistant NEMO'. The user's question is 'what are the different types of bathymetry I can use?'. The assistant's response, after a 1-second thought, explains two methods: 1. Analytical (hand-defined) bathymetry, which can be set directly in the model configuration (DOMAINCFG → zgr_bat) and includes sub-points for flat-bottom domains (nn_bathy = 0) and predefined analytic bathymetries. 2. File-based bathymetry from a NetCDF file, where the model reads depth fields. This method requires bathymetry variables (cn_bath in cn_topo) and, when ice-shelf cavities are activated (ln_isfcav = .true.), an optional ice-shelf draft variable (cn_visfd in cn_fisfd). These files must contain data on every point of the 2-D model grid, with positive depth values in meters; land is represented by zero or negative values. The assistant also lists three options: Flat-bottom analytic bathymetry (nn_bathy = 0), Other analytic bathymetries defined by hand (nn_bathy ≠ 0 but ≠ 0), and NetCDF-based bathymetry (nn_bathy > 0) with necessary variables (cn_topo, cn_fisfd, etc.) and optional ice-shelf draft data. At the bottom, it says 'These possibilities are described in the NEMO documentation' with a link to 'apdx_DOMAINCFG.tex'. The interface includes a 'Send a Message' input field, a '1 Source' indicator, and a 'Capture d'écran' button.

01 Open WebUI

Assistant NEMO

what are the different types of bathymetry I can use?

01 Assistant NEMO

Thought for 1 seconds

You can choose between two broad ways of defining the bathymetry in NEMO:

1. **Analytical (hand-defined) bathymetry** – set directly in the model configuration (`DOMAINCFG` → `zgr_bat`).
 - If `nn_bathy = 0` you are using an analytic definition.
 - `nn_bathy = 0` gives a flat-bottom domain (the total depth is obtained from the vertical coordinate transformation). Other non-positive values can be used for predefined analytic bathymetries.
2. **File-based bathymetry from a NetCDF file** – the model reads depth fields supplied by the user.
 - If `nn_bathy > 0` the bathymetry must be provided in NetCDF files.
 - `nn_bathy = 1` (and similarly `nn_bathy = 2`) requires a bathymetry variable (`cn_bath` in `cn_topo`) and, when ice-shelf cavities are activated (`ln_isfcav = .true.`), an optional ice-shelf draft variable (`cn_visfd` in `cn_fisfd`). These files must contain data on every point of the 2-D model grid, with positive depth values in meters; land is represented by zero or negative values.

So the options are:

- **Flat-bottom analytic bathymetry** (`nn_bathy = 0`).
- **Other analytic bathymetries** defined by hand (`nn_bathy ≠ 0` but `≠ 0`).
- **NetCDF-based bathymetry** (`nn_bathy > 0`), with the necessary variables (`cn_topo` , `cn_fisfd` , etc.) and optional ice-shelf draft data.

These possibilities are described in the NEMO documentation [apdx_DOMAINCFG.tex](#).

1 Source

Send a Message

Redouane Lguensat

Capture d'écran



THANK YOU

Redouane Lguensat
rlguensat@ipsl.fr

Website: redouanelg.github.io

X: @redouanelg