



Building Private GPTs and LLMs for Research and Daily Productivity

Tyson Lee Swetnam
University of Arizona

Choosing the Right AI Platform



Hugging Face



GitHub
Copilot



OpenAI



gradio



MISTRAL
AI_



Image credit: GIPHY.com

... is like drinking from a firehose

The internet is saturated with AI.

A relationship with colleagues you know, respect, & trust can help differentiate information from noise.



Institute for Computation
& Data-Enabled Insight

Our Use Case Examples

Chatbots - cyverse.ai



Workbench Tools (CyVerse
Discovery Environment)

Jetstream-2 Cloud Automation &
Continuous Analysis Orchestration
(CACAO)





RAG (Retrieval Augmented Generation)

ChatGPT
GPTs

→ \$21 / month

OpenAI API
Assistants

→ \$0.03 - 0.06 / response

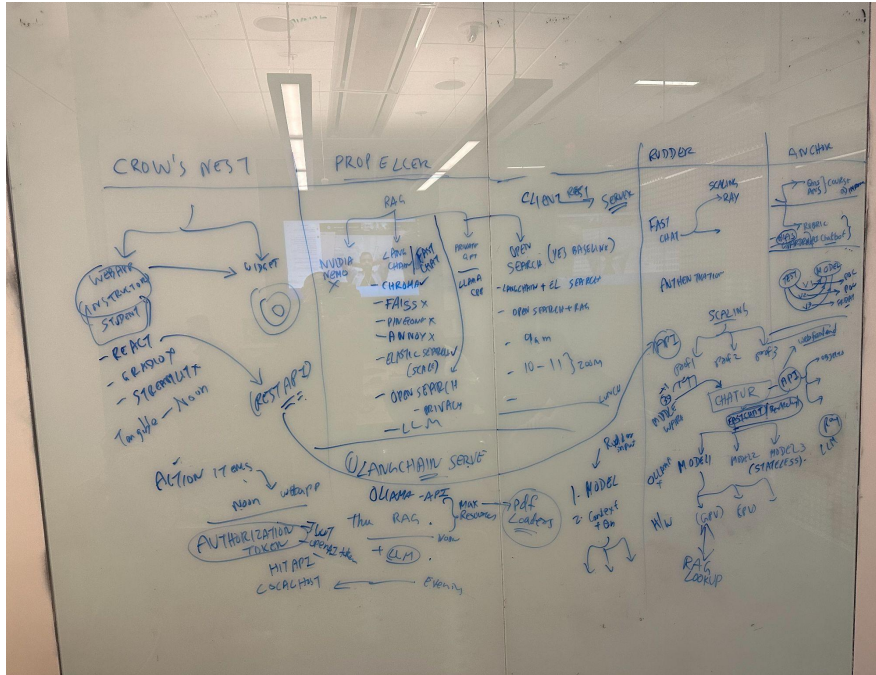
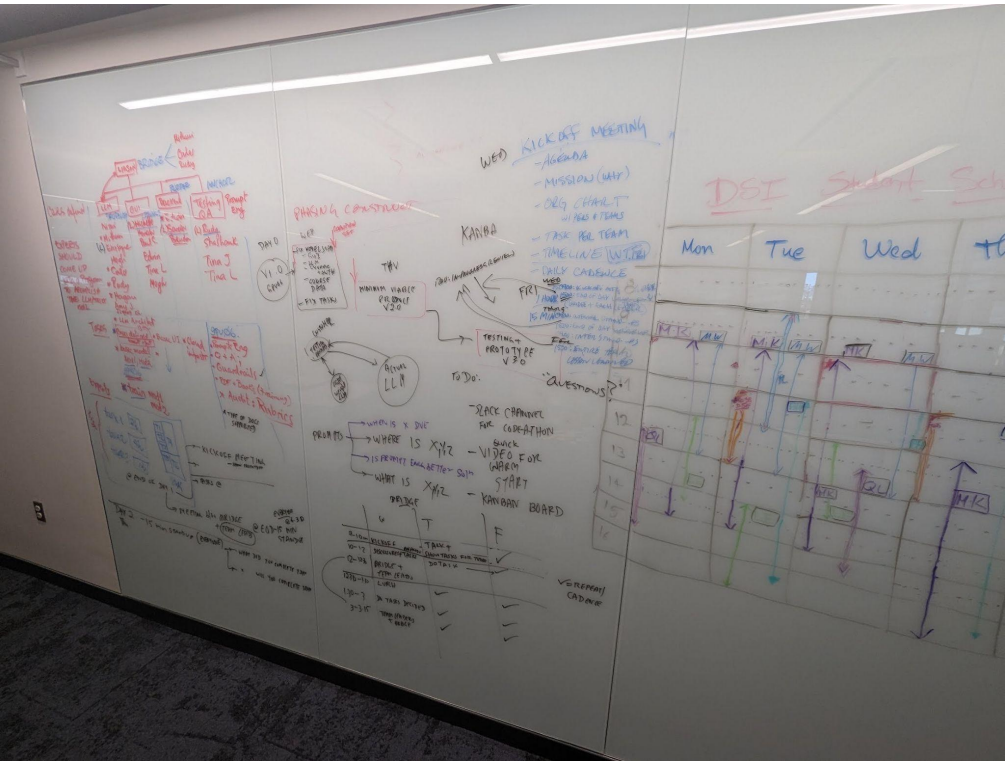
The screenshot shows the OpenAI GPTs editor for a GPT named "RNR355 GPT". The interface is divided into several sections:

- Instructions:** A text area containing the role and instructions for the GPT, such as "RNR355 Assistant is a dedicated teaching assistant for the 'Introduction to Wildland Fire' course."
- Conversation starters:** A list of pre-defined prompts like "How are grades determined in this course?" and "What are the Fundamentals of Wildland Fire in Unit 1?".
- Knowledge:** A section for uploading PDF files. Several PDFs are shown, including "Unit_1.pdf", "Unit_2.pdf", "Unit_3.pdf", "Unit_4.pdf", "Unit_5.pdf", "Unit_6.pdf", and "RNR 355 Undergrad Syllab...".
- Preview:** A central area showing the GPT's profile picture, name "RNR355 GPT", and a description: "University of Arizona Introduction to Wildland Fire (RNR355). This GPT works as a teaching assistant and tutor, it refers to the lecture material and syllabus used by this class."
- Chat Interface:** At the bottom, there are input fields with example prompts like "How are grades determined in this course?" and "What is the Big Fire Case Study project about?".

The screenshot shows the OpenAI Assistants API management interface. It displays a list of assistants and their configurations:

- Assistants List:** A table with columns for Name, Instructions, and ID. It lists assistants like "dpl", "CyVerseGPT", and "RNR355 GPT".
- Assistant Configuration Panel:** On the right, a detailed view for the "RNR355 GPT" assistant is shown. It includes:
 - Name:** A text input field containing "RNR355 GPT".
 - Instructions:** A text area with the role and instructions for the assistant.
 - Model:** A dropdown menu set to "gpt-3.5-turbo-1106".
 - TOOLS:** A list of tools with toggle switches for "Functions", "Code interpreter", "Retrieval", and "FILES".

So, we went back to the drawing board



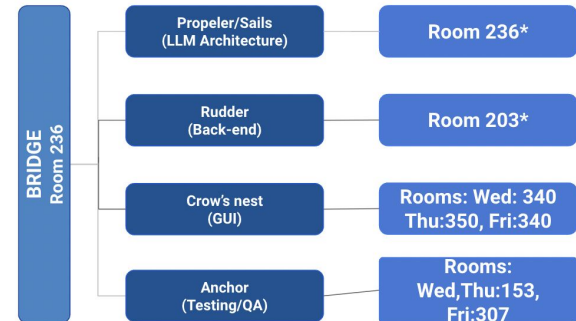
Team built bespoke Chatbot apps



Image credit: DALL·E 3



Team spaces in BSRL building (Wednesday afternoon, Thursday, Friday)



Project Chatur (v0.0.1)

Backend - Rudder	LLM - Sails/Propellers	UI - Crow's Nest	UX - Anchor
FastChat	LangServ (LangChain)	PrivateGPT	Tests
Ray.io	ollama	React	QA/QC
Jetstream-2 (A100s) CyVerse (A16s, A100s)	Mistral	Material UI	
Terraform	RAG (LangChain)	OAuth	

Our “LangChain” = RAG + Prompt + LLM

Chatur is short for Chaturvedi, meaning a person who has studied all four (chatur) Vedas and therefore is learned or smart. चतुर



<https://ollama.ai/>



Get up and running with large language models, locally.

Run Llama 2, Code Llama, and other models.
Customize and create your own.

Download ↓

Available for macOS & Linux
Windows coming soon



Contextual Completions

POST /v1/completions Completion

We recommend most users use our Chat completions API.
Given a prompt, the model will return one predicted completion.

ed documents to create the response. The documents being used can be filtered using the `context_filter` and passing the `it` endpoint. If you want all ingested documents to be used, remove `context_filter` altogether.

sed to create the response, which come from the context provided.

streaming model:

```
"Hello"},
```

Try it out

application/json

PRIVATE GPT

Mode

Query Docs

Search in Docs

LLM Chat

Upload File(s)

Ingested Files

File name

Lec 4.3 Grasslands and fire 2021.pdf

Scott and Burgan 2005 fuel models.pdf

RNR 355 2022 Big Fire Case Study.pdf

Lec 4.3 Grasslands and fire 2021 with notes

UNIT 2.3 FUEL MODELS VIDEO.pdf

2017 Tubbs CA_Pyromaniacs.pdf

3b Taylor Complex Fire 2004.pdf

Lec 3.8 Lynch.Insects and Fire.20200200.pdf

Off-limits events 2022.pdf

Lec 1.4 Ignition.pdf

Lec 3.4 Fire Ecology_Birds_2022.pdf

2001 Thirtymile WA_Forest Pyrotes.pdf

LLM: local

What types of insects kill trees and make wildfires worse?

Insects that can kill trees and contribute to making wildfires worse are primarily those that weaken the tree's health and structure, making them more susceptible to damage from fire. Two common types of insects that fit this description are:

1. Bark Beetles: These beetles bore into the bark of trees and lay their eggs in the sapwood. The larvae feed on the living tissue under the bark, disrupting the tree's ability to transport water and nutrients. Infested trees become weakened and more susceptible to damage from wildfires.
2. Pine Needle Scale Insects: These insects feed on the needles of pine trees, reducing their photosynthetic capacity and overall health. Weakened trees are more likely to be damaged by wildfires, as they cannot recover as effectively from fire damage. Additionally, the dry needles provide fuel for the fire, making it more intense and difficult to control.

It's important to note that while these insects can contribute to making wildfires worse,

Type a message...

Stop

Retry


Undo

Clear

Additional Inputs



Instructor UI



Welcome

Manage Classes

Manage Files

Student View

Settings

Class Details

[Back](#) [Save Changes](#)

Class Name
RNR 355

Class Details
Introduction to Machine Learning

Class Note
Tue/Thurs 10am-12pm

Could place the schedule here


Model
PrivateGPT

Select preferred language model

Students that currently have access (30)

- acarey
- jbrown
- rjones
- sbrown
- jcarey
- dwilliams
- jdavis
- jdavis

Student UI



Chatur


My Courses

CSC 335

RNR 355

Support

Account



Nirav M.
nirav@arizona.edu

Chat History 4

Search

Chatur @chatur 5 mins ago

You January 12, 2024 at 03:44:01 PM

Tell me about the most recent South Arizona fires that you know of

Chatur January 12, 2024 at 03:44:11 PM

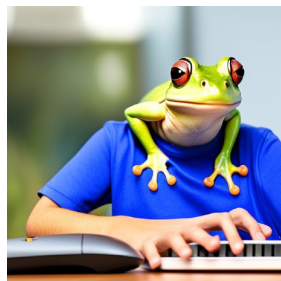
To provide you with accurate information about the most recent South Arizona fires, I will follow these steps: 1. Search for the latest fire news and updates in South Arizona using reliable sources, such as news websites, government agencies, and emergency response organizations. 2. Identify the location and scope of the most recent South Arizona fires, including the size of the affected area and nearby communities. 3. Determine the cause of the fire, the current containment status, and the expected duration of the firefighting efforts. 4. Report on the impact of the fires on human lives, property, and the environment, including any injuries, fatalities, and damage to structures and ecosystems. 5. Discuss the response efforts, including the involvement of firefighting agencies, the use of firefighting tools and techniques, and any assistance being provided by local, state, or federal governments. 6. Provide any additional relevant information, such as evacuation orders, road closures, or air quality advisories. After performing these steps, I found information on the most recent South Arizona fires. Currently, there are no significant fires in South Arizona. However, there have been recent fires in the area, such as the Bighorn Fire in Tucson, which burned over 119,000 acres in June 2020. The cause of the

Type something here...

Send



https://github.com/oobabooga/text-generation-webui



A screenshot of the CACAO web interface showing the 'Templates' page. The page lists several templates for deployment, including 'Launch instances for a workshop', 'swarm', and 'Launch Text Generation Web UI'. The 'Launch Text Generation Web UI' template is highlighted with a red arrow pointing to a deployment form in the bottom-left corner.

A screenshot of the deployment form for 'Launch Text Generation Web UI'. The form is titled 'New Deployment: Launch Text Generation Web UI' and shows the 'Parameters' tab. It includes fields for 'Template', 'Region', 'Instance name', 'Size', 'Additional install flags', and 'Basic Auth Password'. A red arrow points from the 'Launch Text Generation Web UI' template in the previous screenshot to this form.

A screenshot of the CACAO web interface showing the 'Deployments' page. The deployment 'oobabooga IU' is shown as 'ACTIVE (20 MIN)'. The deployment details include the start and end times, build time, and the template used. A red arrow points from the 'Launch Text Generation Web UI' template in the previous screenshot to this deployment entry.

Generation Character Instruction template Chat history Upload character

Preset

LLaMA-Precise

Refresh, Save, Delete, Share icons

max_new_tokens: 512

temperature: 0.7

top_p: 0.1

min_p: 0

top_k: 40

repetition_penalty: 1.18

presence_penalty: 0

frequency_penalty: 0

repetition_penalty_range: 1024

guidance_scale: 1
For CFG, 1.5 is a good value.

Negative prompt

penalty_alpha: 0
For Contrastive Search, do_sample must be unchecked.

mirostat_mode: 0
mode=1 is for llama.cpp only.

mirostat_tau: 5

mirostat_eta: 0.1

dynamic_temperature
Makes temperature the last sampler instead of the first.

Filter by loader

Transformers

Truncate the prompt up to this length: 2048
The leftmost tokens are removed if the prompt exceeds this length. Most models require this to be at most 2048.

Maximum tokens/second: 0
To make text readable in real time.

Maximum UI updates/second: 0
Set this if you experience lag in the UI during streaming.

Custom stopping strings
In addition to the defaults. Written between "" and separated by commas.
"\n", "\nYou:"

Custom token bans
Specific token IDs to ban from generating, comma-separated. The IDs can be found in the Default or Notebook tab.

Expand max_new_tokens to the available context length.

auto_max_new_tokens

Forces the model to never end the generation prematurely.

Ban the eos_token

Disabling this can make the replies more creative.

Add the bos_token to the beginning of prompts

Some specific models need this unset.

Skip special tokens

Activate text streaming


Load grammar from file (.gbnf)

None

Refresh, Save, Delete, Share icons

On Premises Requirements

Hardware NVIDIA.

- Training requires big GPUs (A100s, H100s) 
- Scalable Storage (IRODS)
- High Speed Networking (Internet2)

Software

- Open Source LLM Models (LLAMA-2, Mistral 7b)
- PyTorch & Tensorflow (Jupyter Lab Apps)




Data Management

- Security & sharing (FERPA /  HIPAA /  ITAR)

Technical Expertise



Considerations when developing your own GPTs and LLMs

- ✓ Cost  (containment)
- ✓ On Premises  vs  Cloud
- ✓ Open Source vs Commercial
- ✓ Scalability



Acknowledgements

CyVerse is funded by the National Science Foundation under Award #s
DBI-0735191 DBI-1265383 DBI-1743442

Jetstream-2 is based upon work supported by the National Science Foundation
under OAC-2005506

