# Ola Chat Methods and Tools

**Barclay R. Brown, Ph.D.**
Senior Technical Fellow, AI Research
Applied Research and Technology (ART)
Collins Aerospace

# Agenda

1. What is Open Local AI
2. Architecture / Components
3. Ola Chat Features
4. Ola Chat Applications
5. Installation
6. Loading Models
7. Running Ola Chat
8. Application Notes

# Why “Open Local AI” Now?

- GenAI boosts SE productivity in reqs & test (early field data)
- But: 61% restrict tools; 63% restrict data into public models; bans exist
- Forecast: a majority of LLM workloads move on-prem within ~2 years
- Tension: speed of innovation ↔ confidentiality & export controls
- Goal today: show a practical, replicable, MIT-licensed path

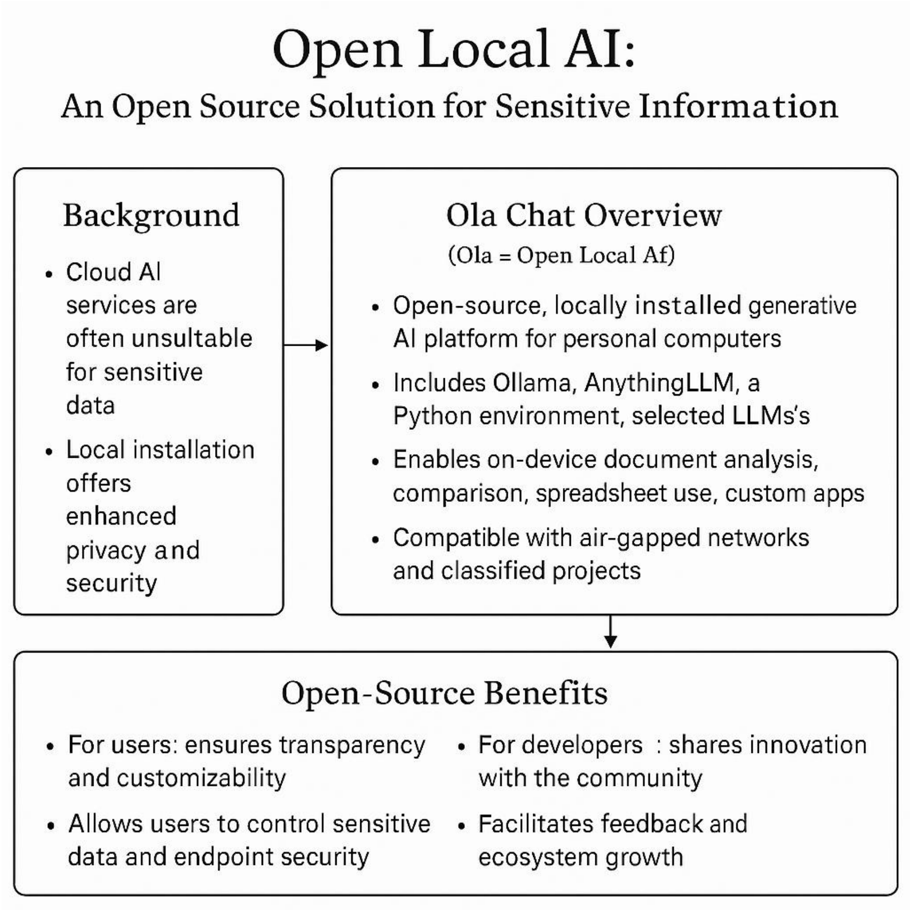
## Design Principles

- **Open**: auditable, modifiable, vendor-neutral
- **Local**: inference + embeddings stay on the workstation
- **Offline-capable**: install → operate with zero internet
- **Fast to adopt**: desktop install in <1 day vs multi-week SaaS approvals
- **Replicable**: published architecture; MIT license

**Is AI for SE more like SAP or more like Excel?**

# Architecture at a Glance

- **Ollama** executes quantized LLMs on CPU/GPU, fully local
- **AnythingLLM** orchestrates chat + retrieval; local vector indexes
- **Windows installer**: unzip → run; works on air-gapped networks
- **Scales down & up**: 8–20B models on dev boxes; heavier models on shared workstations
- **Model notes**: new small models (e.g., Phi-4 Reasoning / Mini) deliver strong utility even on low-/no-GPU machines
- **New OpenAI open source models** (GPT-OSS-20 and GPT-OSS-120) bring previously frontier model performance



## Open Local AI:

An Open Source Solution for Sensitive Information

### Background

- Cloud AI services are often unsultable for sensitive data
- Local installation offers enhanced privacy and security

### Ola Chat Overview

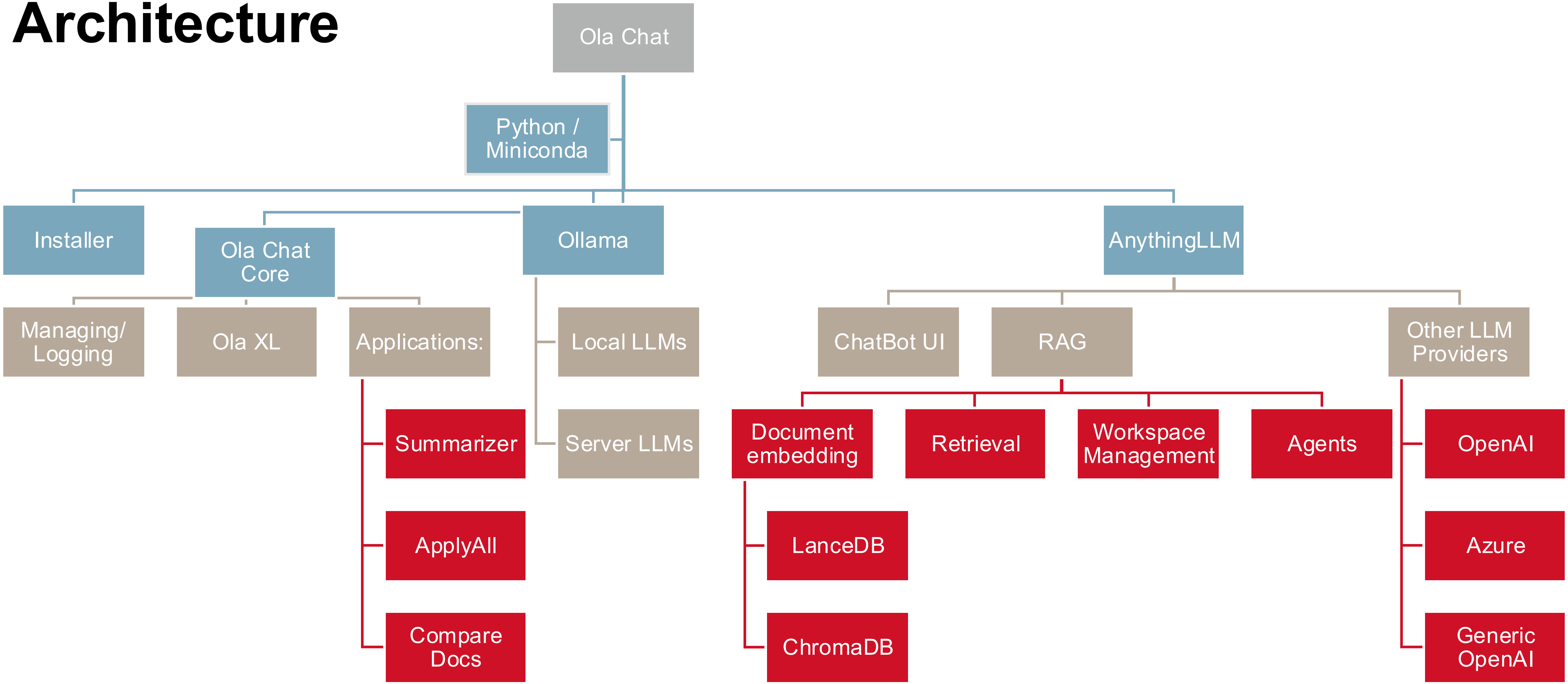
(Ola = Open Local Af)

- Open-source, locally installed generative AI platform for personal computers
- Includes Ollama, AnythingLLM, a Python environment, selected LLMs's
- Enables on-device document analysis, comparison, spreadsheet use, custom apps
- Compatible with air-gapped networks and classified projects

### Open-Source Benefits

- For users: ensures transparency and customizability
- Allows users to control sensitive data and endpoint security
- For developers : shares innovation with the community
- Facilitates feedback and ecosystem growth

# Ola Chat Architecture



# Deployment Approach

### The Case for Open Local AI

- Faster approvals (local install), immediate productivity lift
- Data sovereignty: nothing leaves; easier audits
- Extensible: add domain agents, pipelines, adapters
- Proven components; MIT license → fork & tailor
- **Call to action**: pilot on a dev workstation; identify 1–2 high-value SE use cases; plan a shared-workstation build for bigger models
- Deploy hosted solution to scale

### Choose your adventure…

“We can do what they did”

OR

“We can use their stuff (and build on it)”

# Open Local AI (Ola) Chat

## What is Ola Chat?

- Open-Source based generative AI chatbot interface and application platform
- Ready to use functionality, similar to ChatGPT
- Application platform for developers
- Open-source components plus about 12k lines custom code
- Main features:
  - Chatbot using local, server or cloud models
  - End-user RAG with vector store/retrieval
  - Ola XL spreadsheet interface
  - Summarizer, document comparison, other applications
- Approved as open-source release by RTX

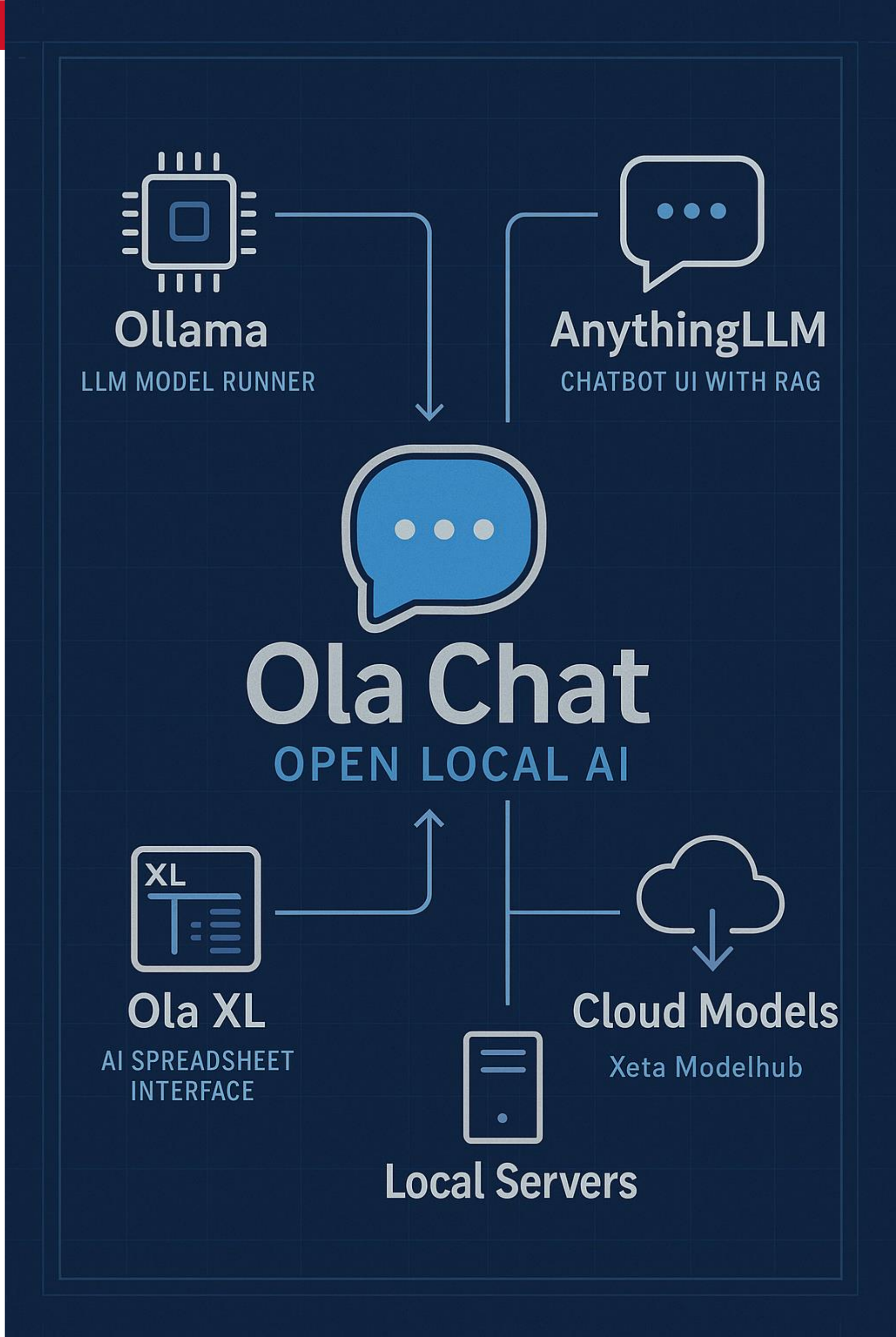
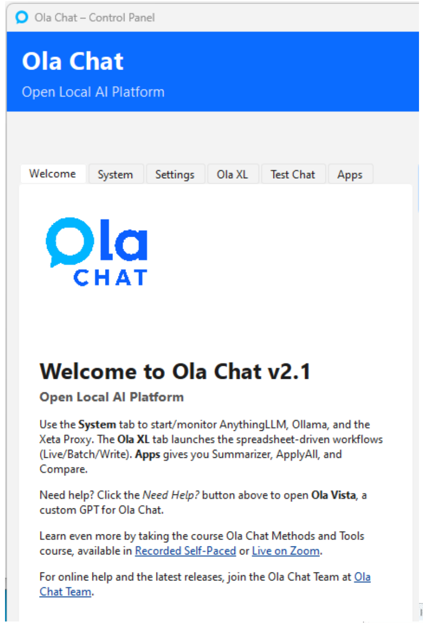


*Image generated by author*

# Ola CHAT Open Local AI Chat

## Current Progress

- User Community: 1000 across RTX with 38 Beta Testers
- Showing time savings and creative applications in systems engineering
- Educational / Outreach
    - Generative AI Applications course (~1000 completions so far)
    - GATTACA Workshops monthly
    - Ola Chat office hours weekly
    - Ola Chat Methods and Tools course (first offering July 2025)



### Welcome to Ola Chat v2.1

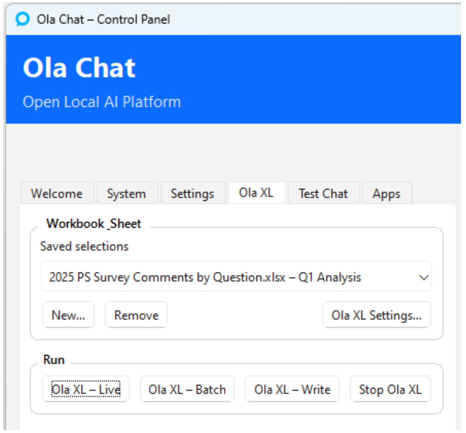
**Open Local AI Platform**

Use the **System** tab to start/monitor AnythingLLM, Ollama, and the Xeta Proxy. The **Ola XL** tab launches the spreadsheet-driven workflows (Live/Batch/Write). **Apps** gives you Summarizer, ApplyAll, and Compare.

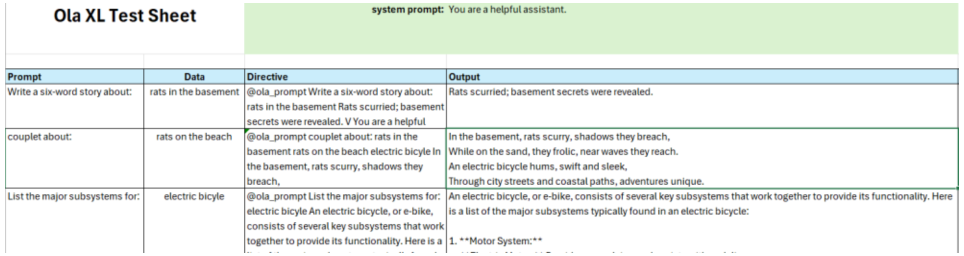
Need help? Click the *Need Help?* button above to open **Ola Vista**, a custom GPT for Ola Chat.

Learn even more by taking the course Ola Chat Methods and Tools course, available in Recorded Self-Paced or Live on Zoom.

For online help and the latest releases, join the Ola Chat Team at Ola Chat Team.



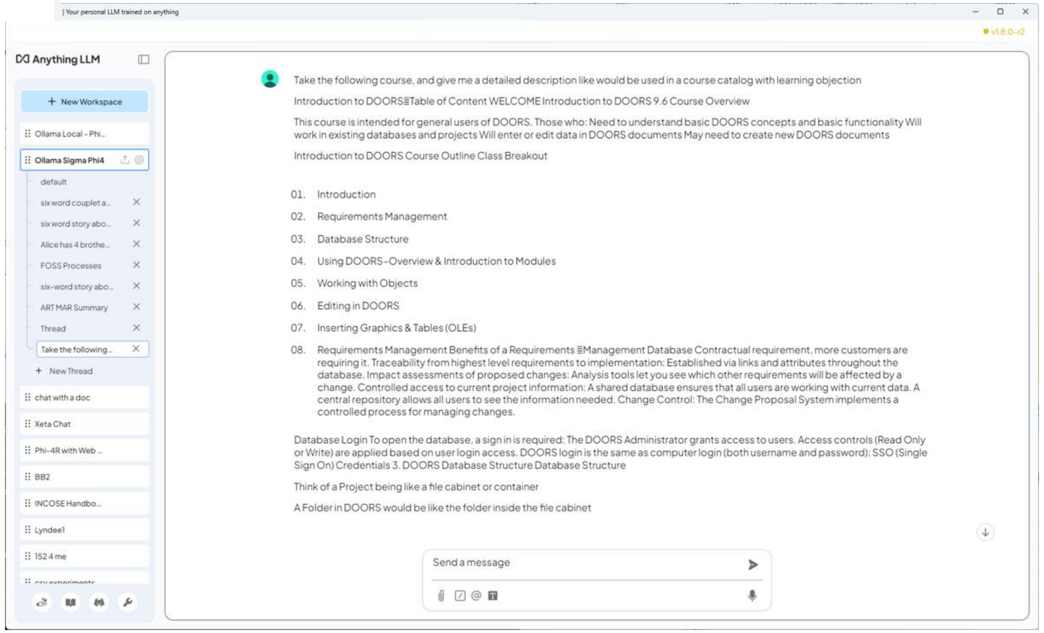
Saved selections: 2025 PS Survey Comments by Question.xlsx – Q1 Analysis



**Ola XL Test Sheet**

**system prompt:** You are a helpful assistant.

| Prompt | Data | Directive | Output |
|---|---|---|---|
| Write a six-word story about: | rats in the basement | @ola_prompt Write a six-word story about: rats in the basement Rats scurried; basement secrets were revealed. V You are a helpful | Rats scurried; basement secrets were revealed. |
| couplet about: | rats on the beach | @ola_prompt couplet about: rats in the basement rats on the beach electric bicyle In the basement, rats scurry, shadows they breach, | In the basement, rats scurry, shadows they breach,<br>While on the sand, they frolic, near waves they reach.<br>An electric bicycle hums, swift and sleek,<br>Through city streets and coastal paths, adventures unique. |
| List the major subsystems for: | electric bicyle | @ola_prompt List the major subsystems for: electric bicyle An electric bicycle, or e-bike, consists of several key subsystems that work together to provide its functionality. Here is a | An electric bicycle, or e-bike, consists of several key subsystems that work together to provide its functionality. Here is a list of the major subsystems typically found in an electric bicycle:<br><br>1. **Motor System:** |



Take the following course, and give me a detailed description like would be used in a course catalog with learning objection

Introduction to DOORS≡Table of Content WELCOME Introduction to DOORS 9.6 Course Overview

This course is intended for general users of DOORS. Those who: Need to understand basic DOORS concepts and basic functionality Will work in existing databases and projects Will enter or edit data in DOORS documents May need to create new DOORS documents

Introduction to DOORS Course Outline Class Breakout

01. Introduction
02. Requirements Management
03. Database Structure
04. Using DOORS - Overview & Introduction to Modules
05. Working with Objects
06. Editing in DOORS
07. Inserting Graphics & Tables (OLEs)
08. Requirements Management Benefits of a Requirements ≡Management Database Contractual requirement, more customers are requiring it. Traceability from highest level requirements to implementation: Established via links and attributes throughout the database. Impact assessments of proposed changes: Analysis tools let you see which other requirements will be affected by a change. Controlled access to current project information: A shared database ensures that all users are working with current data. A central repository allows all users to see the information needed. Change Control: The Change Proposal System implements a controlled process for managing changes.

Database Login To open the database, a sign in is required: The DOORS Administrator grants access to users. Access controls (Read Only or Write) are applied based on user login access. DOORS login is the same as computer login (both username and password): SSO (Single Sign On) Credentials 3. DOORS Database Structure Database Structure

Think of a Project being like a file cabinet or container

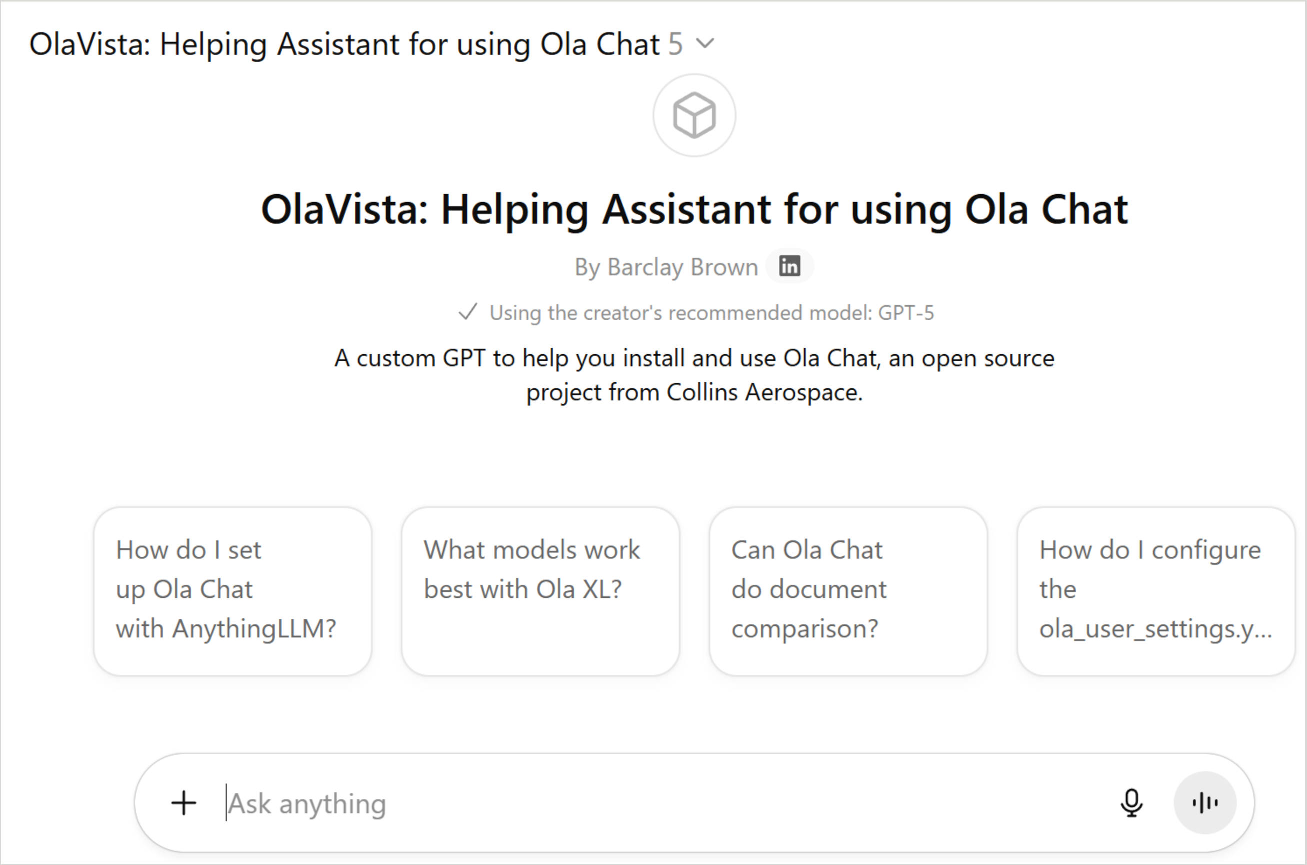
A Folder in DOORS would be like the folder inside the file cabinet

# Meet OlaVista: Ola Chat Assistant

To access:

- ChatGPT.com
  - Explore Custom GPTs
    - Look for OlaVista
  - Direct link: https://chatgpt.com/g/g-682f17b82b708191950afd867b5495e2-ola-chat

- OlaVista runs on an **external** AI, therefore, no RTX data. The subject matter is Ola Chat, which is approved as open source (release pending)
- It’s fine to ask it about Ola Chat, AnythingLLM, Ollama, Ola XL, etc. but don’t include company data in the question



OlaVista: Helping Assistant for using Ola Chat 5

## OlaVista: Helping Assistant for using Ola Chat

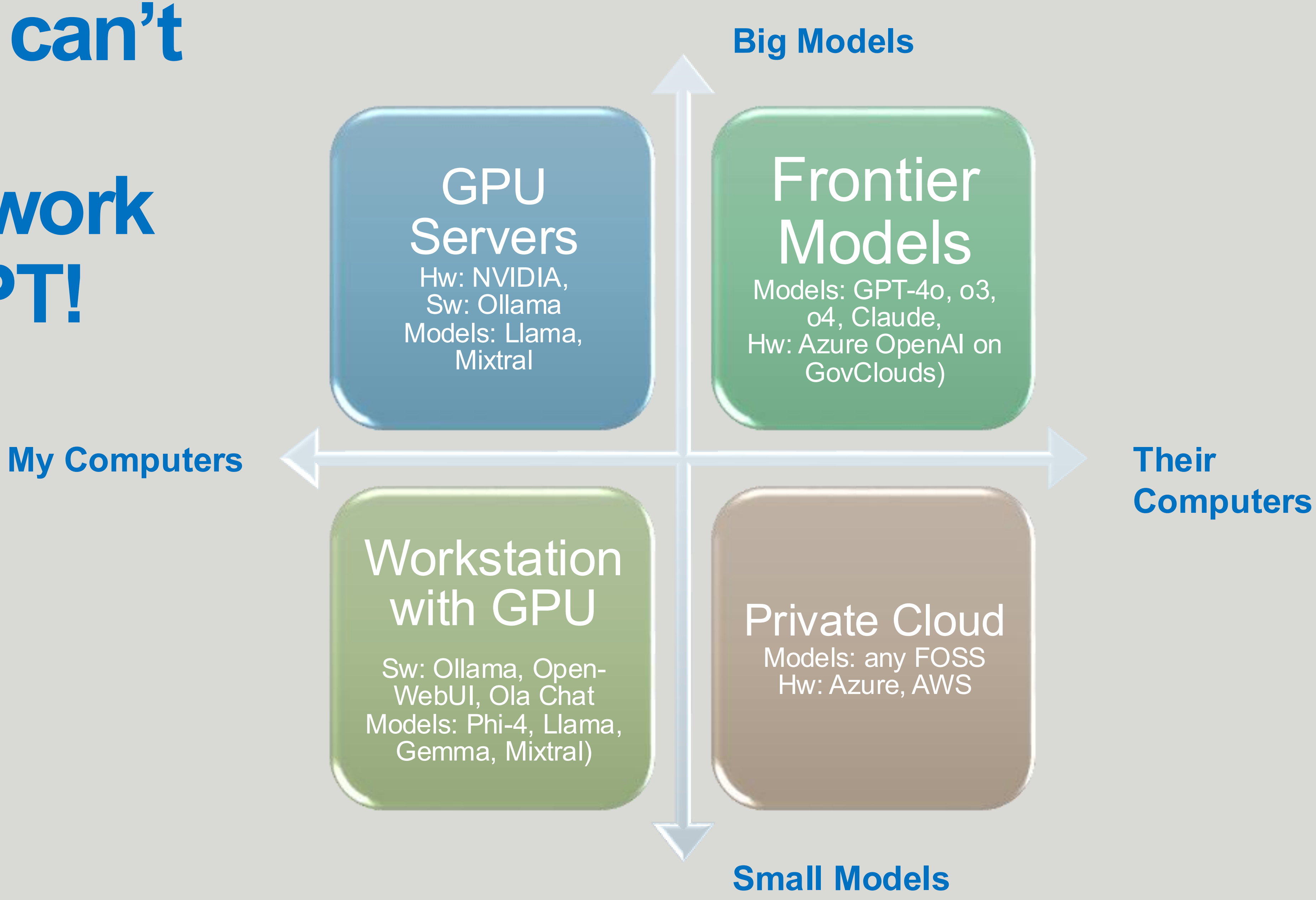
By Barclay Brown

✓ Using the creator's recommended model: GPT-5

A custom GPT to help you install and use Ola Chat, an open source project from Collins Aerospace.

How do I set up Ola Chat with AnythingLLM?
What models work best with Ola XL?
Can Ola Chat do document comparison?
How do I configure the ola_user_settings.y...
+
Ask anything

# But, wait, I can’t just do my company work on ChatGPT!



Big Models

GPU Servers
Hw: NVIDIA,
Sw: Ollama
Models: Llama, Mixtral

Frontier Models
Models: GPT-4o, o3, o4, Claude,
Hw: Azure OpenAI on GovClouds)

My Computers

Their Computers

Workstation with GPU
Sw: Ollama, Open-WebUI, Ola Chat
Models: Phi-4, Llama, Gemma, Mixtral)

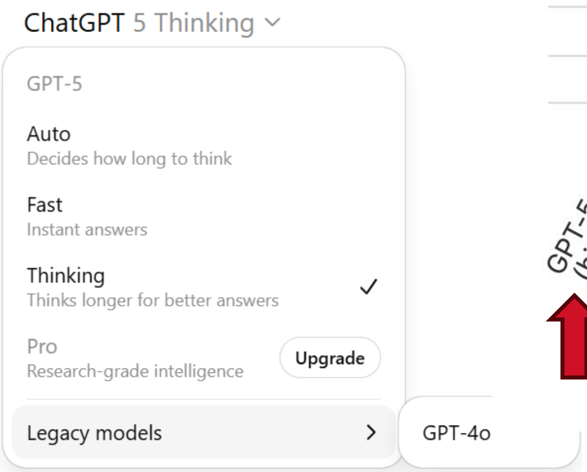
Private Cloud
Models: any FOSS
Hw: Azure, AWS

Small Models

# Models, Models… Everywhere

## GPT-5

- New single model with multiple versions
- Performance/ intelligence varies a lot
- Be sure to use GPT-5 Thinking if you want the actual best



ChatGPT 5 Thinking ⌄

GPT-5

Auto
Decides how long to think

Fast
Instant answers

Thinking ✓
Thinks longer for better answers

Pro
Research-grade intelligence
Upgrade

Legacy models ›

GPT-4o

## Artificial Analysis Intelligence Index

Artificial Analysis Intelligence Index v3.0 incorporates 10 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME 2025, IFBench, AA-LCR, Terminal-Bench Hard, $\tau^2$-Bench Telecom

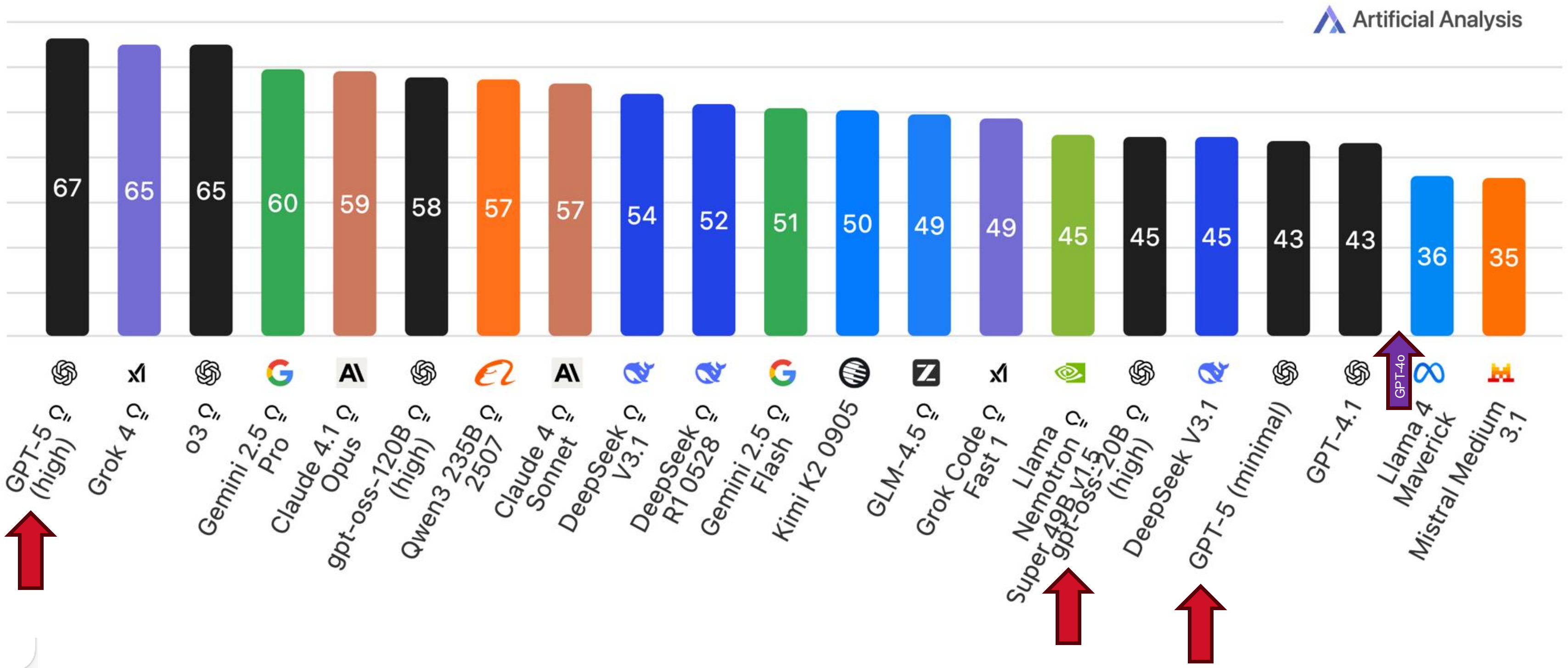


*Diagram used with permission of the creator*

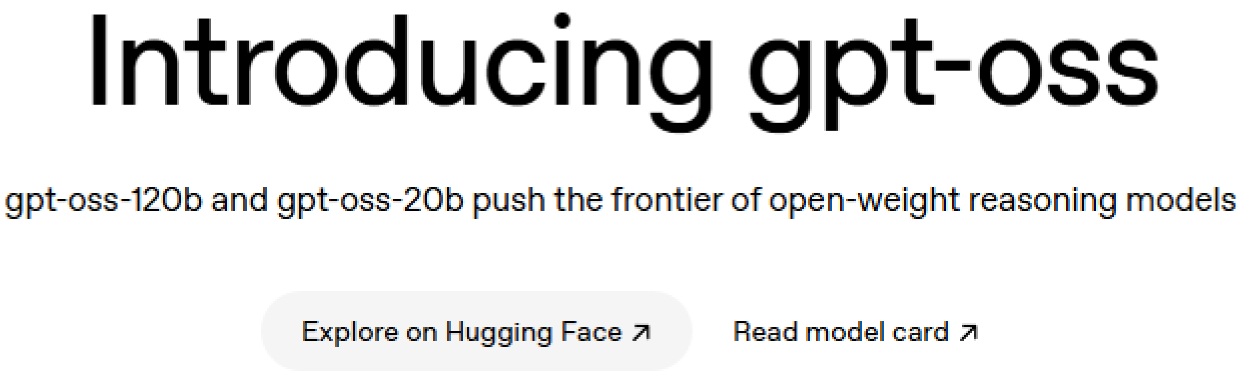
See also post by Ethan Mollick: https://www.linkedin.com/posts/emollick_the-issue-with-gpt-5-in-a-nutshell-is-that-activity-7359977503651729408-zgY0?utm_source=share&utm_medium=member_desktop&rcm=ACoAAAKu9PEBBWCYdSzIWNcmQVg0EcIQDRdJ-5k

# In other news, GPT-OSS

*Another reason to go local!*

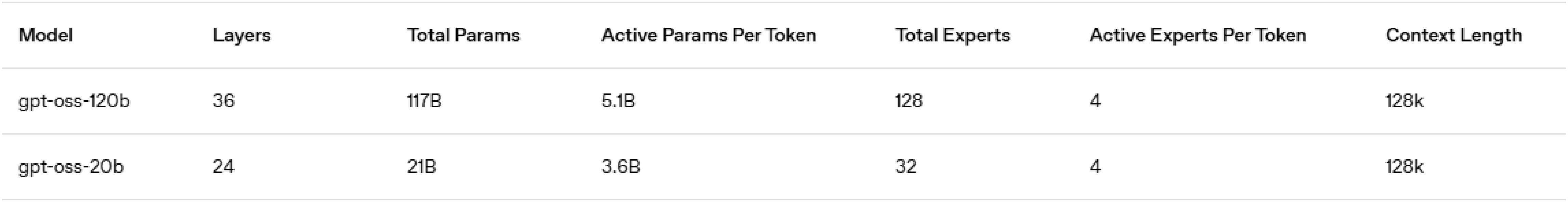
**GPT-OSS-20, GPT-OSS-120**

- First open source models from OpenAI since GPT-2
- Apache 2 open source license
- GPT-OSS-20 <> GPT-o3-mini
- GPT-OSS-120 <> GPT-04-mini
- Very good for a lot of uses



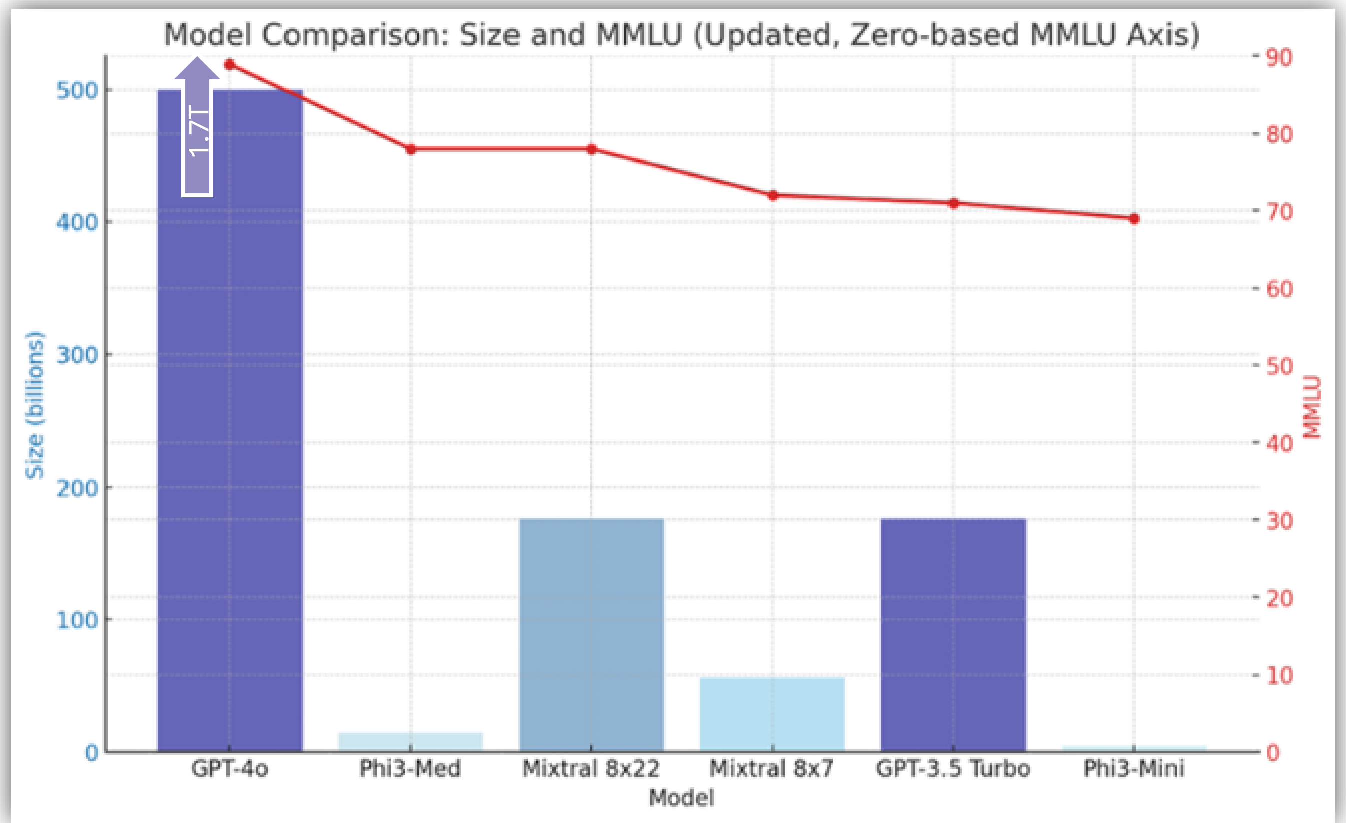
## Introducing gpt-oss

gpt-oss-120b and gpt-oss-20b push the frontier of open-weight reasoning models



| Model | Layers | Total Params | Active Params Per Token | Total Experts | Active Experts Per Token | Context Length |
|---|---|---|---|---|---|---|
| gpt-oss-120b | 36 | 117B | 5.1B | 128 | 4 | 128k |
| gpt-oss-20b | 24 | 21B | 3.6B | 32 | 4 | 128k |

# Can an LLM really run on my PC?

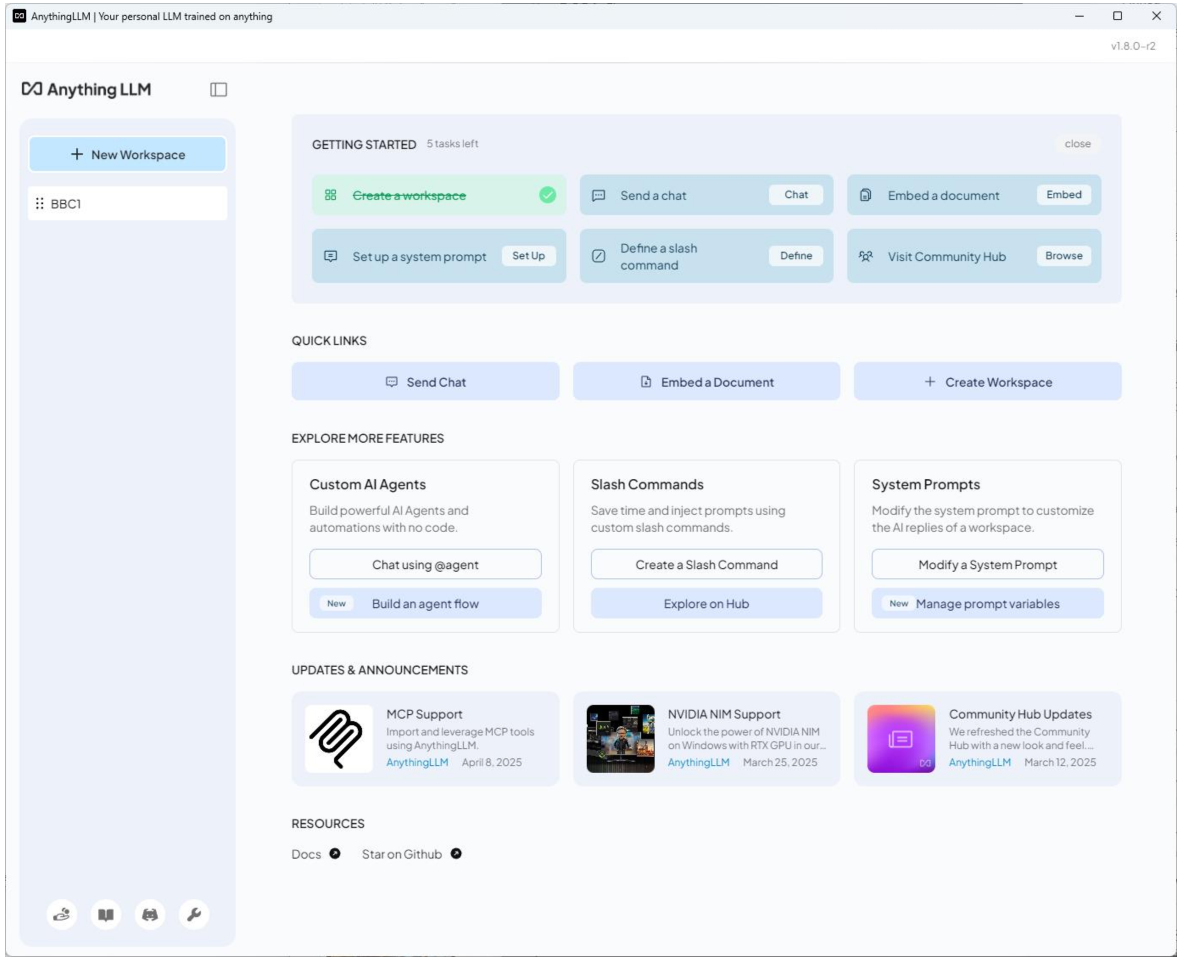


| Model | MMLU-5 | MMLU-Pro |
|---|---|---|
| • GPT-5 | 91.3 | **87** |
| • Grok-4 | 86.6 | **87** |
| • GPT-OSS-120 | **90.0** | |
| • GPT-o3 | **92.3** | 85.6 |
| • Gemini 2.5 Pro Exp | | 84.1 |
| • Claude 3.7 Sonnet | | 82.7 |
| • GPT-5 mini | | 82.5 |
| • GPT-5 nano | | 77.9 |
| • GPT-4o | **88.7** | 72.6 |
| • GPT-o3-mini-high | 87.0 | |
| • Llama 3.1 (405B) | 85.4 | |
| • GPT-OSS-20 | **85.3** | **67.1** |
| • GPT-o4-mini | 82 | 63 |
| • Phi 4 Reasoning Plus | | 76 |
| • Phi 4 (14B) | 84.8 | 71.5 |
| • Llama 3.3 (70B) | 82.0 | |
| • Mixtral 8x22 (176B) | 77.8 | |
| • Llama 3.2 Vision (11B) | 68.4 | |
| • Mixtral 8x7B (56B total) | 70.6 | |
| • Phi-4 Mini 128k (3.8B) | 67.3 | 52.8 |
| • Mistral 7B | 60.1 | |

**GPT-OSS-20 <> GPT-o3-mini / GPT-OSS-120 <> GPT-o4-mini**

# Ola Chat and AnythingLLM

- AnythingLLM as UI
  - MIT License
  - Python integration via API
  - Flexible model access (many model providers)
- Ola Chat Installation Flexibility
  - Ola Chat Setup – windows installer
  - Can also install AnythingLLM and Ollama independently
  - Add Ola Chat source code in Python, with supporting libraries: Ola Chat is not compiled



AnythingLLM | Your personal LLM trained on anything

v1.8.0-r2

Anything LLM

GETTING STARTED 5 tasks left

~~Create a workspace~~

Send a chat

Embed a document

Set up a system prompt

Define a slash command

Visit Community Hub

QUICK LINKS

EXPLORE MORE FEATURES

Custom AI Agents

Build powerful AI Agents and automations with no code.

Slash Commands

Save time and inject prompts using custom slash commands.

System Prompts

Modify the system prompt to customize the AI replies of a workspace.

UPDATES & ANNOUNCEMENTS

MCP Support

Import and leverage MCP tools using AnythingLLM.

AnythingLLM April 8, 2025

NVIDIA NIM Support

Unlock the power of NVIDIA NIM on Windows with RTX GPU in our...

AnythingLLM March 25, 2025

Community Hub Updates

We refreshed the Community Hub with a new look and feel....

AnythingLLM March 12, 2025

RESOURCES

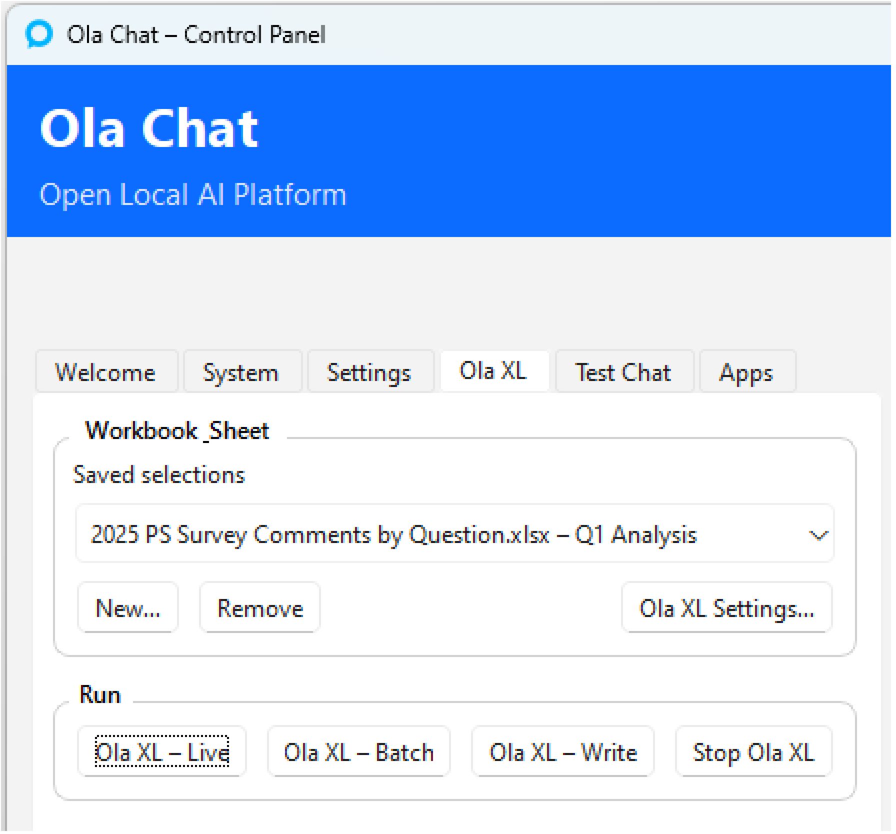
Docs

Star on Github

**Be sure to follow your BU FOSS processes**

# Ola XL – Spreadsheet AI

- Allows entry of prompts, data and directives in ordinary Excel spreadsheets
- Results returned to spreadsheet or written to Word document
- Quick start: @ola_prompt
  - “ “ is the separator for the TEXTJOIN; TRUE means skip empty cells
  - @ola_prompt is the directive to Ola Chat
  - A4 is the prompt
  - B4 is the data to be sent to the LLM with the prompt
  - D4 is the output cell
  - “V” means spill any multi-part data vertically (not used)
  - $D$1 is the system prompt (optional)



Ola Chat – Control Panel

Ola Chat

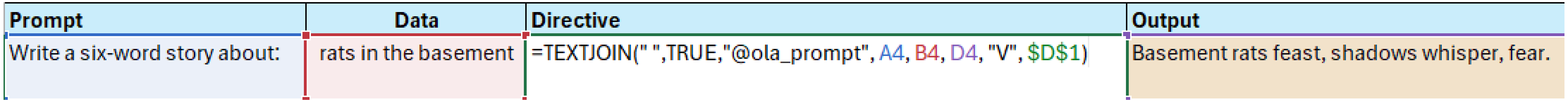
Open Local AI Platform

Workbook_Sheet

Saved selections

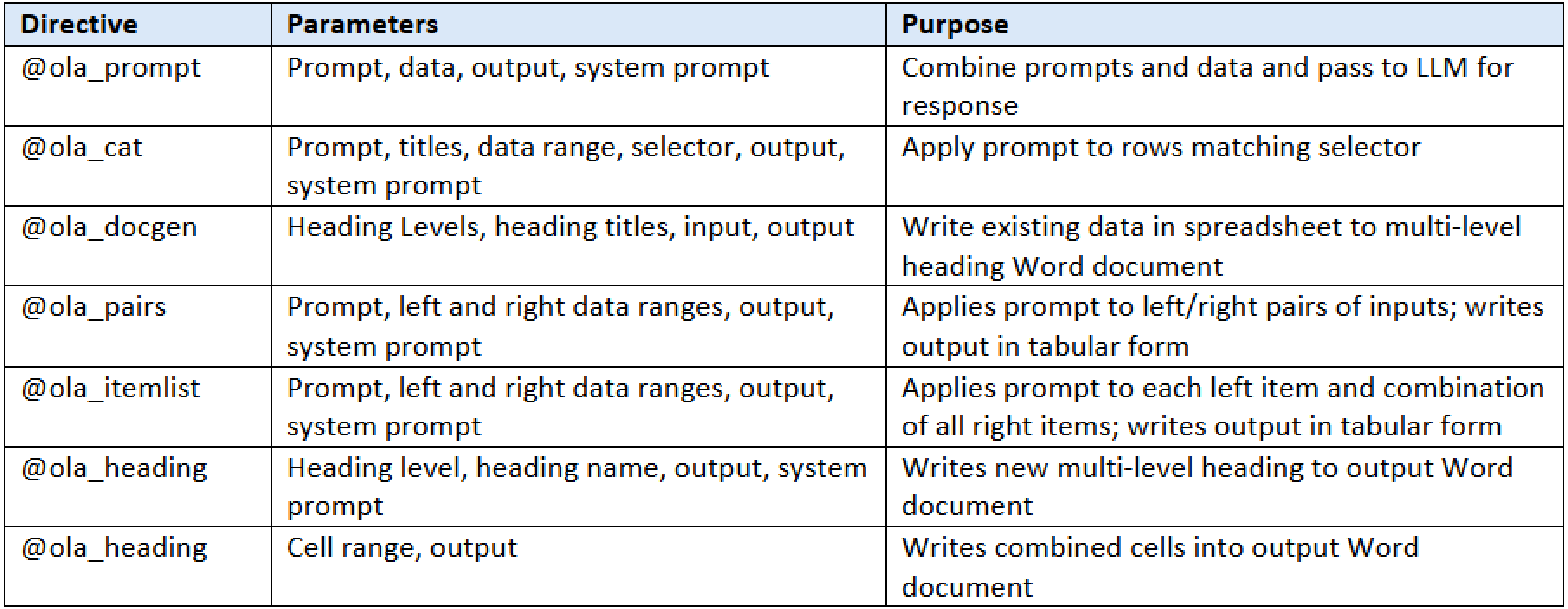
2025 PS Survey Comments by Question.xlsx – Q1 Analysis

Run



| Prompt | Data | Directive | Output |
|---|---|---|---|
| Write a six-word story about: | rats in the basement | =TEXTJOIN(" ",TRUE,"@ola_prompt", A4, B4, D4, "V", $D$1) | Basement rats feast, shadows whisper, fear. |

# Additional Ola XL directives (functions)



| Directive | Parameters | Purpose |
|---|---|---|
| @ola_prompt | Prompt, data, output, system prompt | Combine prompts and data and pass to LLM for response |
| @ola_cat | Prompt, titles, data range, selector, output, system prompt | Apply prompt to rows matching selector |
| @ola_docgen | Heading Levels, heading titles, input, output | Write existing data in spreadsheet to multi-level heading Word document |
| @ola_pairs | Prompt, left and right data ranges, output, system prompt | Applies prompt to left/right pairs of inputs; writes output in tabular form |
| @ola_itemlist | Prompt, left and right data ranges, output, system prompt | Applies prompt to each left item and combination of all right items; writes output in tabular form |
| @ola_heading | Heading level, heading name, output, system prompt | Writes new multi-level heading to output Word document |
| @ola_heading | Cell range, output | Writes combined cells into output Word document |

# Ola XL Modes

### Live Mode

- Normal, default mode
- Ola XL opens the designated spreadsheet and performs the Ola XL directives on the live open sheet, writing results in real time
- Only performs directives with no output present
- Saves sheet periodically

### Batch Mode

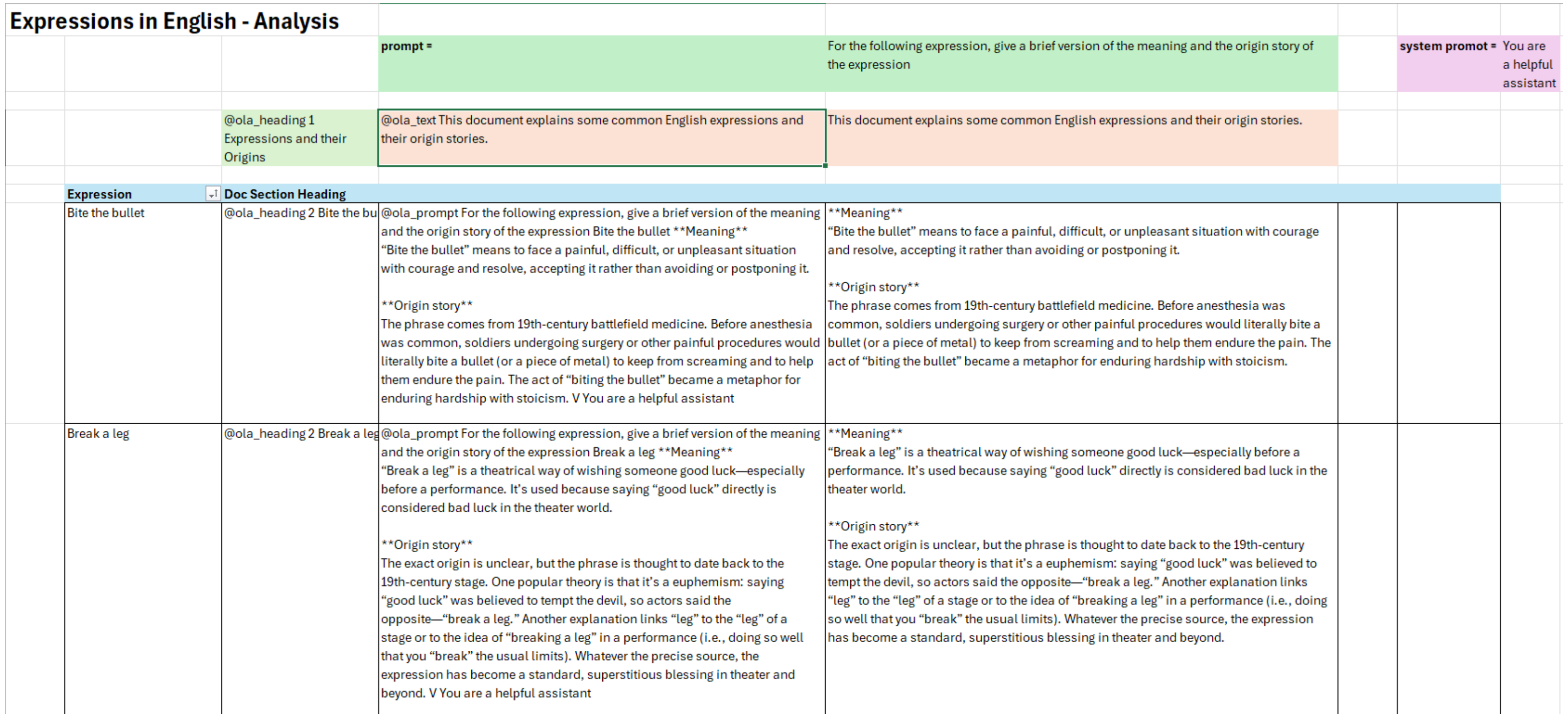
- Same as live mode but operates without opening Excel sheet

### Write Mode

- Writes all output to Word document
- Output to Excel sheet optional
- Allows for intelligence complex document generation

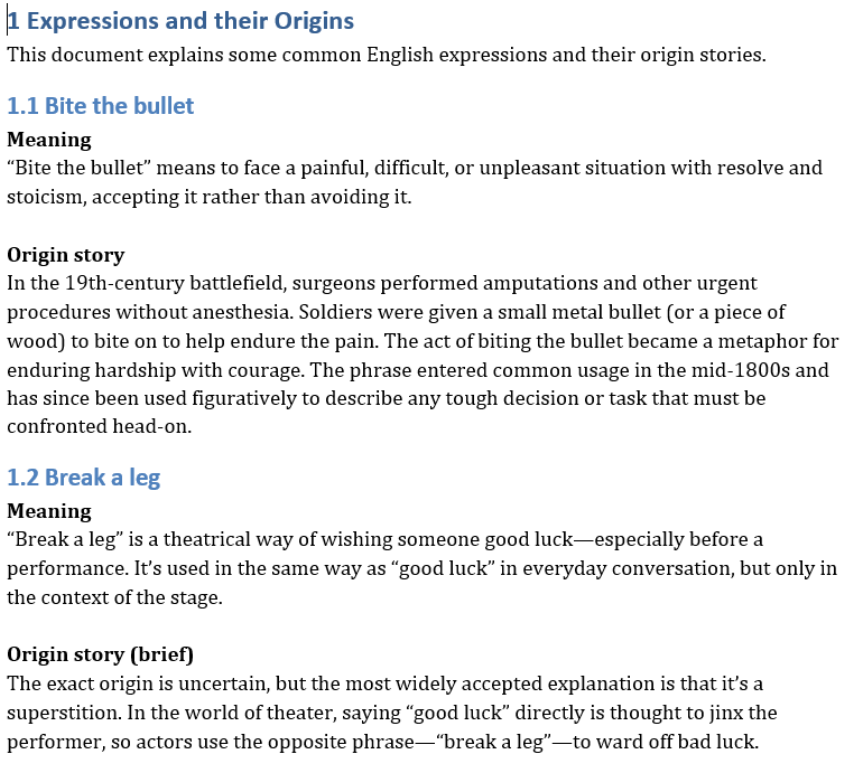
# Ola XL Write Mode - Example

## Analyze List of Items



Expressions in English - Analysis

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | **prompt =** | For the following expression, give a brief version of the meaning and the origin story of the expression | | **system promot =** | You are a helpful assistant |
| | | @ola_heading 1 Expressions and their Origins | @ola_text This document explains some common English expressions and their origin stories. | This document explains some common English expressions and their origin stories. | | | |
| | **Expression** | **Doc Section Heading** | | | | | |
| | Bite the bullet | @ola_heading 2 Bite the bu | @ola_prompt For the following expression, give a brief version of the meaning and the origin story of the expression Bite the bullet **Meaning**<br>“Bite the bullet” means to face a painful, difficult, or unpleasant situation with courage and resolve, accepting it rather than avoiding or postponing it.<br><br>**Origin story**<br>The phrase comes from 19th-century battlefield medicine. Before anesthesia was common, soldiers undergoing surgery or other painful procedures would literally bite a bullet (or a piece of metal) to keep from screaming and to help them endure the pain. The act of “biting the bullet” became a metaphor for enduring hardship with stoicism. V You are a helpful assistant | **Meaning**<br>“Bite the bullet” means to face a painful, difficult, or unpleasant situation with courage and resolve, accepting it rather than avoiding or postponing it.<br><br>**Origin story**<br>The phrase comes from 19th-century battlefield medicine. Before anesthesia was common, soldiers undergoing surgery or other painful procedures would literally bite a bullet (or a piece of metal) to keep from screaming and to help them endure the pain. The act of “biting the bullet” became a metaphor for enduring hardship with stoicism. | | | |
| | Break a leg | @ola_heading 2 Break a leg | @ola_prompt For the following expression, give a brief version of the meaning and the origin story of the expression Break a leg **Meaning**<br>“Break a leg” is a theatrical way of wishing someone good luck—especially before a performance. It’s used because saying “good luck” directly is considered bad luck in the theater world.<br><br>**Origin story**<br>The exact origin is unclear, but the phrase is thought to date back to the 19th-century stage. One popular theory is that it’s a euphemism: saying “good luck” was believed to tempt the devil, so actors said the opposite—“break a leg.” Another explanation links “leg” to the “leg” of a stage or to the idea of “breaking a leg” in a performance (i.e., doing so well that you “break” the usual limits). Whatever the precise source, the expression has become a standard, superstitious blessing in theater and beyond. V You are a helpful assistant | **Meaning**<br>“Break a leg” is a theatrical way of wishing someone good luck—especially before a performance. It’s used because saying “good luck” directly is considered bad luck in the theater world.<br><br>**Origin story**<br>The exact origin is unclear, but the phrase is thought to date back to the 19th-century stage. One popular theory is that it’s a euphemism: saying “good luck” was believed to tempt the devil, so actors said the opposite—“break a leg.” Another explanation links “leg” to the “leg” of a stage or to the idea of “breaking a leg” in a performance (i.e., doing so well that you “break” the usual limits). Whatever the precise source, the expression has become a standard, superstitious blessing in theater and beyond. | | | |



### 1 Expressions and their Origins

This document explains some common English expressions and their origin stories.

### 1.1 Bite the bullet

**Meaning**

“Bite the bullet” means to face a painful, difficult, or unpleasant situation with resolve and stoicism, accepting it rather than avoiding it.

**Origin story**

In the 19th-century battlefield, surgeons performed amputations and other urgent procedures without anesthesia. Soldiers were given a small metal bullet (or a piece of wood) to bite on to help endure the pain. The act of biting the bullet became a metaphor for enduring hardship with courage. The phrase entered common usage in the mid-1800s and has since been used figuratively to describe any tough decision or task that must be confronted head-on.

### 1.2 Break a leg

**Meaning**

“Break a leg” is a theatrical way of wishing someone good luck—especially before a performance. It’s used in the same way as “good luck” in everyday conversation, but only in the context of the stage.

**Origin story (brief)**

The exact origin is uncertain, but the most widely accepted explanation is that it’s a superstition. In the world of theater, saying “good luck” directly is thought to jinx the performer, so actors use the opposite phrase—“break a leg”—to ward off bad luck.

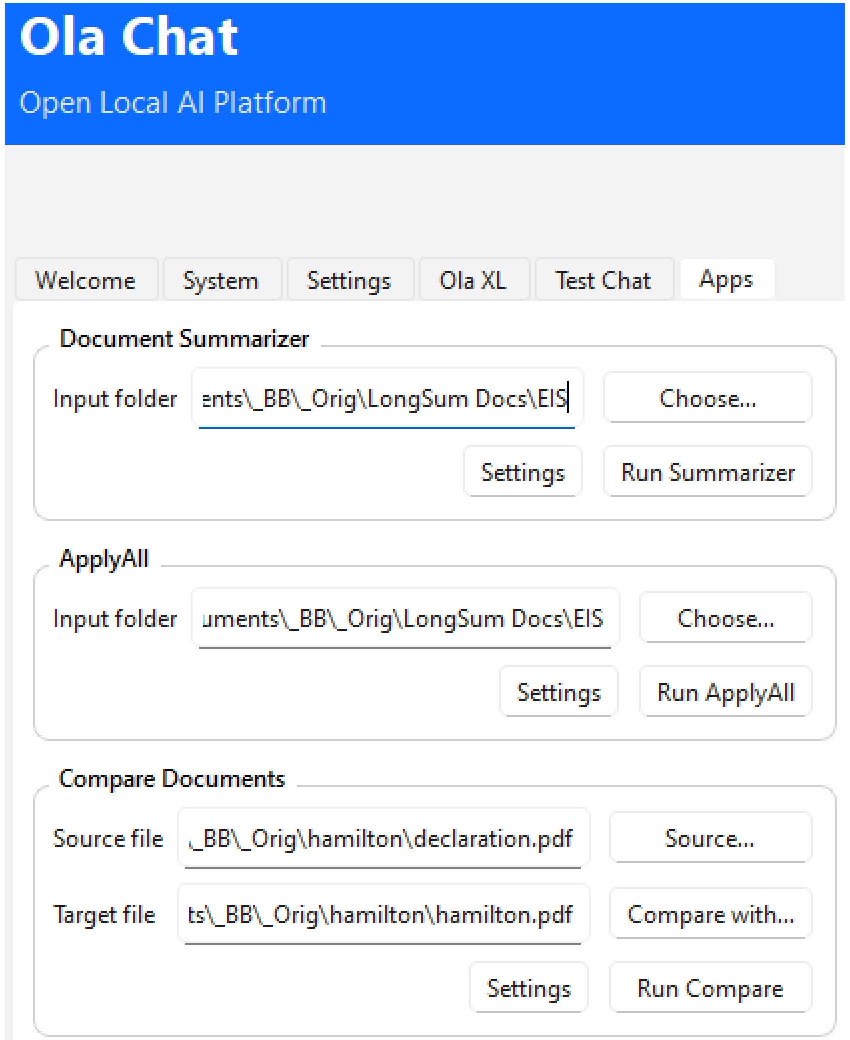
**Combines @ola_prompt, @ola_text, @ola_heading directives**

Collins Aerospace
An RTX Business

# Other Ola Chat Applications

**Built-In Features (end-user)**

- Document Summarizer – Summarizes long documents
- Apply All – iteratively applies a prompt to a set of files in a folder
- Compare – compares two documents by identification of concepts and iterative comparison
- Most or even all of this can now be done using Ola XL (but new applications can be added)



Document Summarizer

Input folder ents_BB_Orig\LongSum Docs\EIS

ApplyAll

Input folder uments_BB_Orig\LongSum Docs\EIS

Compare Documents

Source file _BB_Orig\hamilton\declaration.pdf

Target file ts_BB_Orig\hamilton\hamilton.pdf

**Applications put advanced AI workflows within reach of non-coders**

# Installer Variants

- Four options available
  - OlaChatSetup_2-x-x_update: Full installer without Ollama/AnythingLLM/models (for UI updates)
  - OlaChatSetup_2-x-x_nomodels: Full installer without models (if models pre-installed)
  - OlaChatSetup_2-x-x_patch: Patch installer, smaller: skips core libs/env
  - OlaChatSetup_2-x-x_full: Complete installer: includes Ollama, AnythingLLM, Phi-4 models

# Loading Models

- Methods to load into Ollama
  - Full installer includes Phi-4 Mini & embedding model
  - Command line: `ollama pull <model>`
  - Ola Chat 'lmz' menu option for model ZIP
  - Manual .gguf or safetensors via HuggingFace

# Ola Chat Summarizer

- Documents smaller than context can simply be pasted
- There is also a summarizer built into AnythingLLM
- Summarizes multiple long documents using portioning if necessary
  - Creates portions (stored as txt files)
  - Summarizes each portion (also stored)
  - Combines summaries (concatenate, then summarize using prompt)

```
> sum
2025-07-24 09:03:29,575 - OlaChat Log - INFO - Starting Ola
Document Summarizer with mode: summarize

======================================================
Starting Ola Document Summarizer with mode: summarize
======================================================

(Input folder is set in ola_user_settings.yaml as
summarize_input_folder.
Results are stored in new folders under the input folder.)

Input folder: C:\Users\e21131490\Documents\_BB\_Orig\LongSum
Docs\EIS
Summarize from doc: 1 and to doc: 0 (0 indicates all)
Using modelset: Ollama, and model: phi4:14b-q8_0
Using prompt: Please summarize the text provided,
retaining...
Ollama server is: http://192.168.999.999:11434

Proceed with summarization, mode: summarize? (y/n) [n]
```

# Ola Chat Compare Docs

- Documents are compared using novel algorithm:
  1. Split Doc 1 into propositions
  2. Compare each proposition to Doc 2
  3. Categorize:

     Nothing in Doc 2 relates to the prop

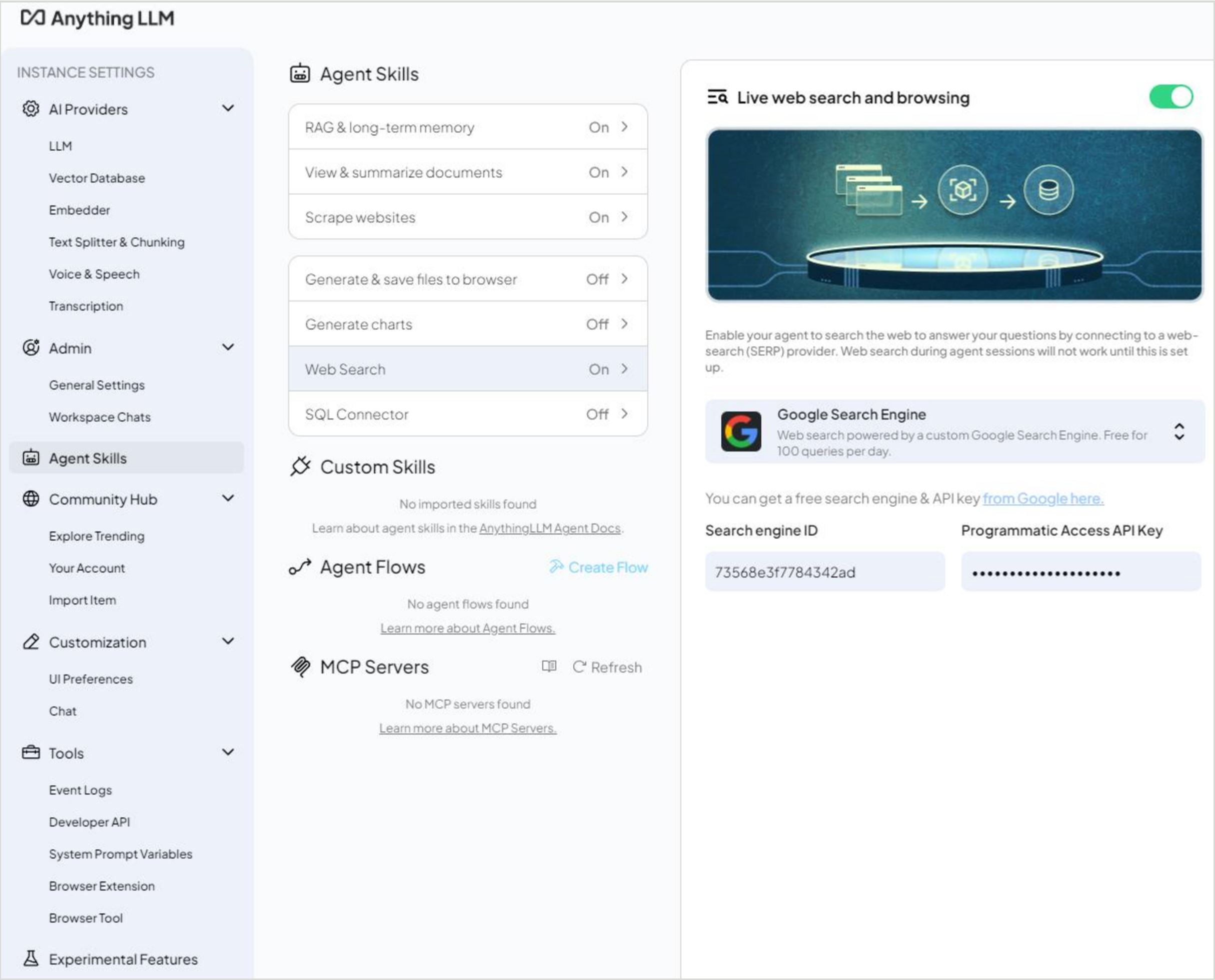
     Doc 2 is consistent with the prop

     Doc 2 contradicts the prop

```
> sum
2025-07-24 09:03:29,575 - OlaChat Log - INFO - Starting Ola
Document Summarizer with mode: summarize

=======================================================
Starting Ola Document Summarizer with mode: summarize
=======================================================

(Input folder is set in ola_user_settings.yaml as
summarize_input_folder.
Results are stored in new folders under the input folder.)

Input folder: C:\Users\e21131490\Documents\_BB\_Orig\LongSum
Docs\EIS
Summarize from doc: 1 and to doc: 0 (0 indicates all)
Using modelset: Ollama, and model: phi4:14b-q8_0
Using prompt: Please summarize the text provided,
retaining...
Ollama server is: http://192.168.999.999:11434

Proceed with summarization, mode: summarize? (y/n) [n]
```

# Ola Chat with Web Searching

- Must use a search engine, e.g. Google
  - So, no company info can be sent in a prompt that will result in a Google Search– CAREFUL!
- Open your workspace Agent Configuration in the AnythingLLM settings
- Under Agent Skills, confirm Web-Browsing is enabled (default “Default” tool)
- In the Search Provider dropdown, choose Google
- Note: Google Search allows up to 100 free searches/day by default AnythingLLM
- May not work with some VPNs in place (need more testing)



Agent Skills

| | |
|---|---|
| RAG & long-term memory | On |
| View & summarize documents | On |
| Scrape websites | On |
| Generate & save files to browser | Off |
| Generate charts | Off |
| Web Search | On |
| SQL Connector | Off |

Custom Skills

No imported skills found

Learn about agent skills in the AnythingLLM Agent Docs.

Agent Flows — Create Flow

No agent flows found

Learn more about Agent Flows.

MCP Servers — Refresh

No MCP servers found

Learn more about MCP Servers.

Live web search and browsing

Enable your agent to search the web to answer your questions by connecting to a web-search (SERP) provider. Web search during agent sessions will not work until this is set up.

Google Search Engine

Web search powered by a custom Google Search Engine. Free for 100 queries per day.

You can get a free search engine & API key from Google here.

Search engine ID: 73568e3f7784342ad

Programmatic Access API Key: ••••••••••••••••••••

# Extension/Customization Possibilities

### In Ola Chat Python Codebase

- GUI Customization
- Module addition
- New Applications
- Custom vector DB loading

### In AnythingLLM

•**Custom Agent Skills (NodeJS plugins)**
Write a folder containing
•plugin.json (defines skill metadata & parameters)
•handler.js (implements the skill's logic in Node.js)
and drop it into plugins/agent-skills in your STORAGE_DIR
•**Custom Data Connectors (plugins)**
Build "data source" plugins to pull in documents or records from proprietary APIs, databases or file systems
•**Custom Authentication (OAuth plugins)**
Implement OAuth or other auth flows so end-users can log in via external identity providers
•**Custom Model Provider**
Complete flexibility for modifying the user interaction flow
•**AnythingLLM REST API / SDK**
Use its HTTP endpoints or official JavaScript/Python SDK to embed chat, RAG or management features into your own applications
•**Embedded Chat Widgets**
Drop prebuilt React/JS widgets into any web page to provide in-page AI chat powered by your AnythingLLM instance

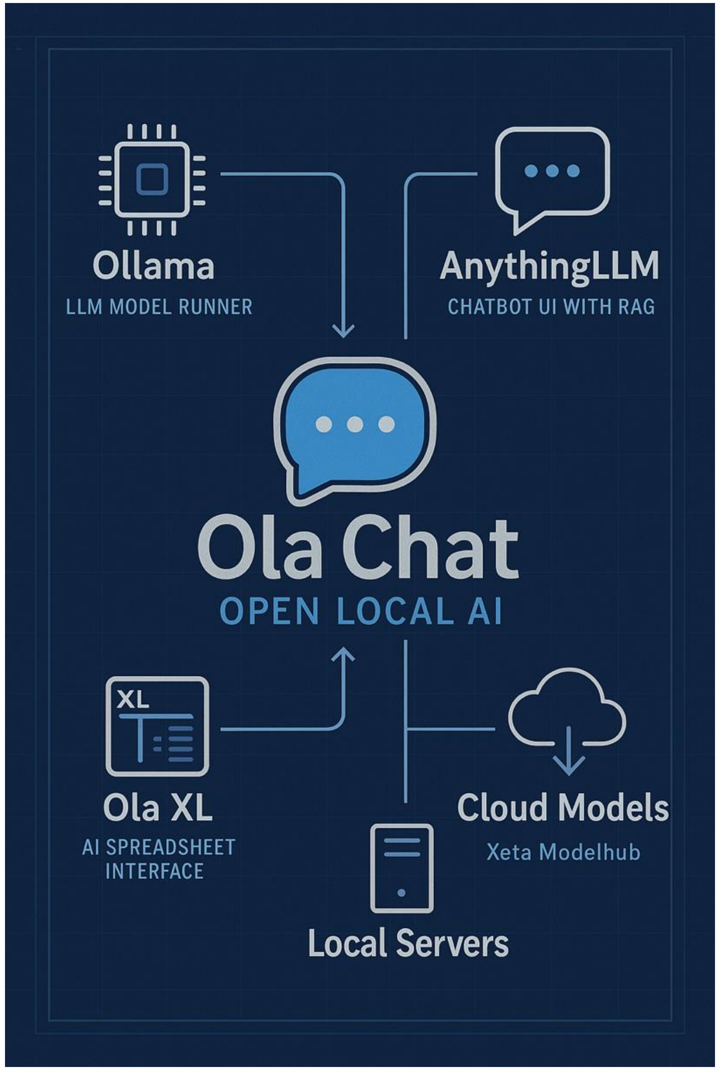


*Image generated by author*

# BACKUP

# Building Block Architecture

- Ollama
  - Open source “engine” that runs models on your local PC (or a local server)
  - Approved open source across RTX
- Anything LLM
  - UI for Ollama and other hosted models
- Models
  - Recommend Phi-4 mini and Phi-4
  - More GPU available? Try  or Mixtral 8x7B
  - Phi-4 mini available in Ola Chat model zips or Ollama Pull
- Hobson
  - AI assistant via email
  - Prompts and attached documents
- Ola Chat
  - Chatbot interface for Ollama
  - Saved chats, file upload, custom modes
  - Prompt lists
- RTX “Innersource” – internal open source
- Distribution via MS Team (for now)

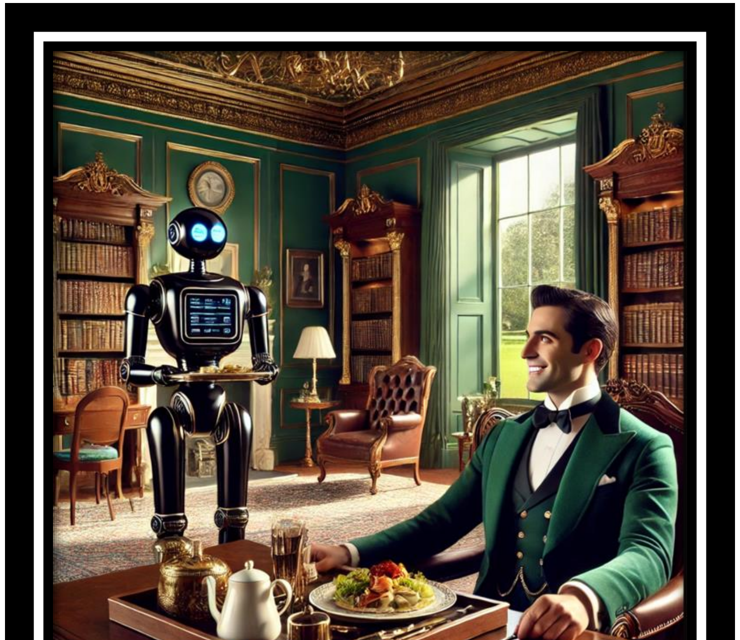
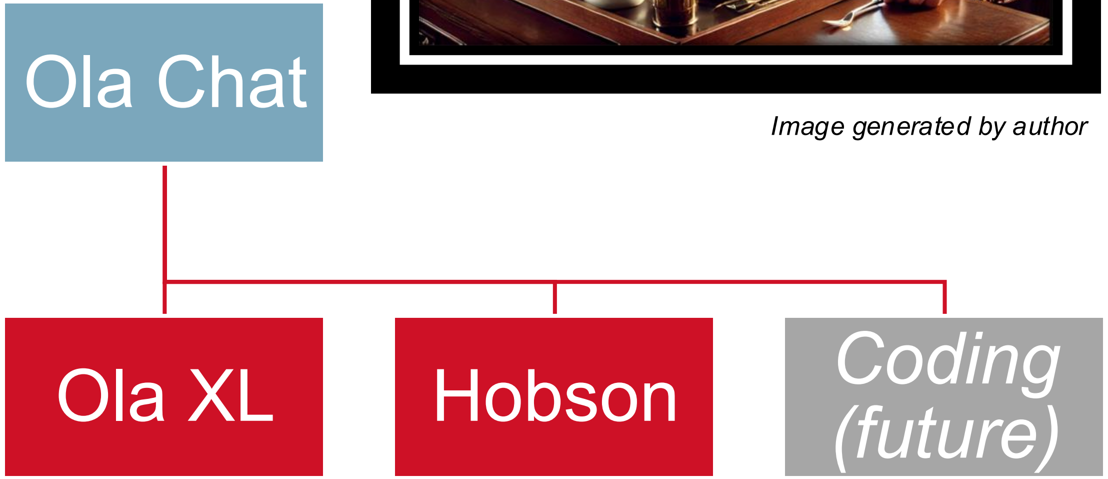
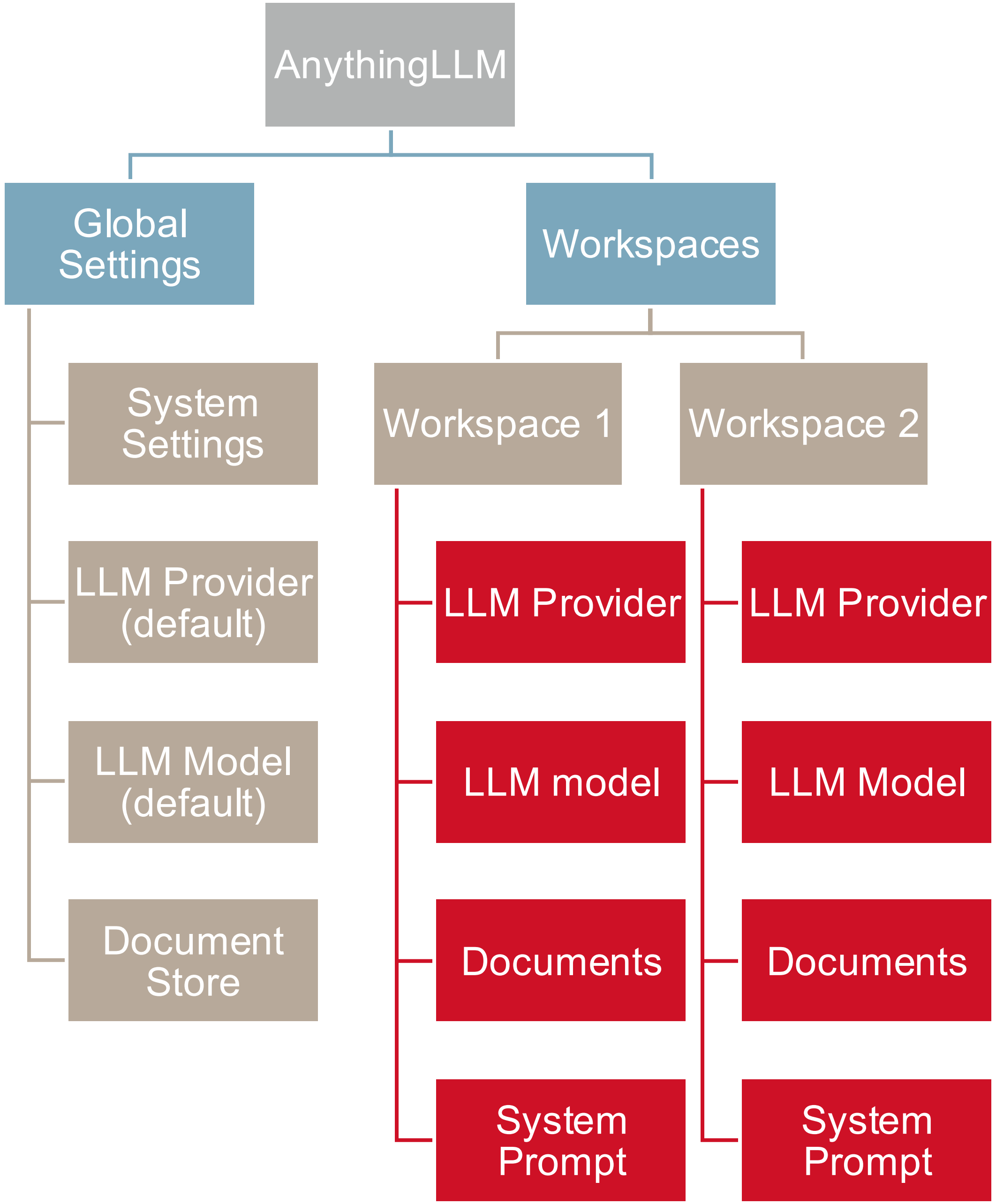


*Image generated by author*



**Join Ola Chat / Hobson team using code: 4zsxdx7**

# AnythingLLM Settings



# Installation Requirements

- Prepare the environment
  - Install to user-level folder (avoid Program Files)
  - Install AnythingLLM for current user only
  - GPU users: ensure NVIDIA CUDA drivers (e.g., v12.5)
  - Verify with nvidia-smi command

# Installation Steps

- Step-by-step process
  - Download OlaChatSetup_x-x-x_zip.txt and rename to .zip
  - Unzip folder and run OlaChatSetup.exe
  - Unblock executables if Windows Defender blocks
  - Choose install for current user
  - Skip Ollama/AnythingLLM download if pre-installed

# Running Ola Chat

- Quick start and detailed
  - Start via Windows Start menu or start_ola_chat.bat
  - User-level install typically in AppData\Local\OlaChat
  - Batch script reads registry, activates conda env, runs ola_chat
  - Logs and settings in ola_chat_logs and settings folders
  - Optional Xeta Proxy server start via menu