10 JULY 2025

ELO2025

KEN ALBA

Tinkering with the Dead

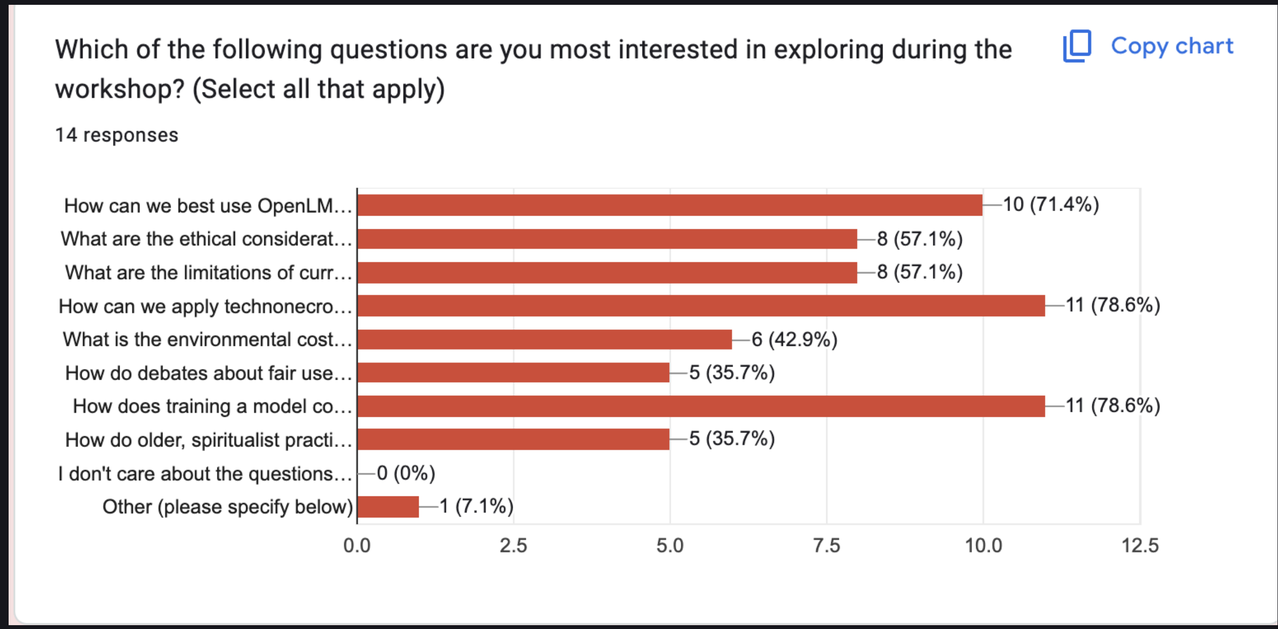
Technonecromancy and Local LLMs

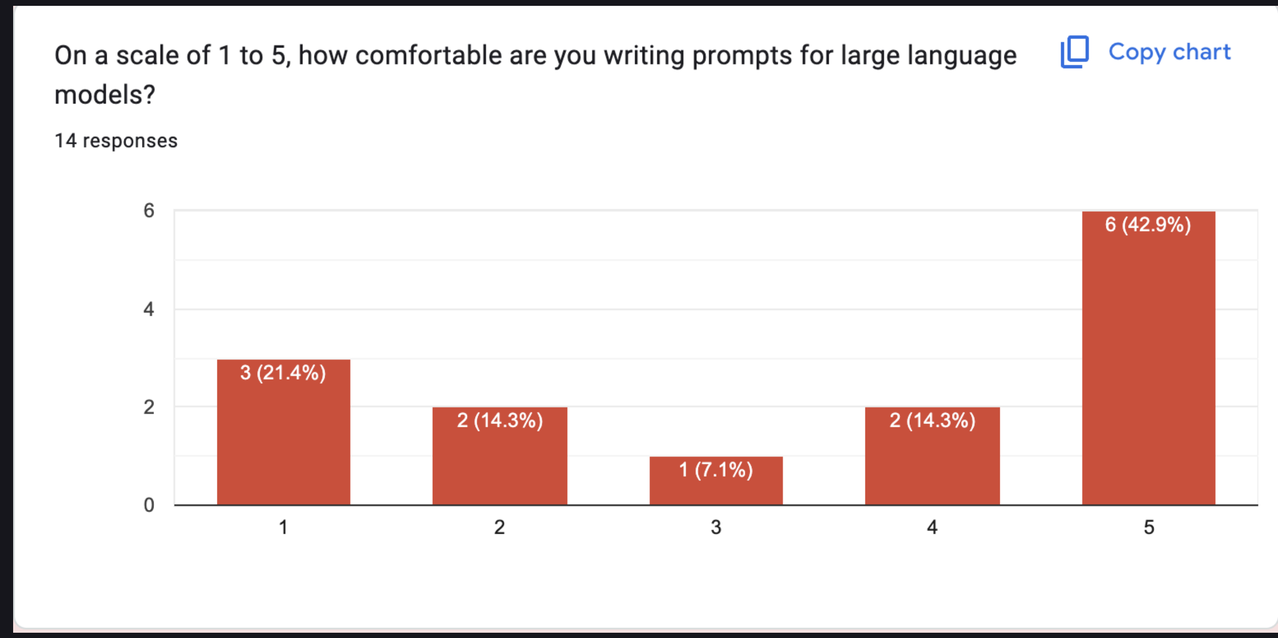
The Poll.

SECTION NAME

Creativity Prevails!

- You’re most interested:
  - Using LLMs in creative work
  - How that work relates to collage and bricolage
  - Bringing back the dead!
- You also care about:
  - Ethics and limitations
  - Fair use and environmental work
  - Historical traditions and their relation to LLMs
- You have a wide variety of technical comfort and prompting experience.





What’s the plan?

AGENDA

Discussions will take place at every stage!

1. Define terms & concerns
2. Software setup
3. Body-snatching (corpus acquisition)
4. Procrusteology (vectorization)
5. Testing (pt 1)
6. Break
7. Orb-wrangling (tuning prompts)
8. Testing (pt 2)
9. Show & Tell
10. Post-postmortem

Goals?

What do we want out of this by the end of the presentation?

- ◯ We can run a local model
- ◯ We can get and vectorize a corpus
- ◯ We have a medium that works
- ◯ We better understand the limitations of our tools
- ◯ We better understand the limitations of our ghosts
- ◯ We have a list of unanswered questions

Let’s define some terms.

DEFINITIONS

Necromancy

"The art of predicting the future by supposed communication with the dead; (more generally) divination, sorcery, witchcraft, enchantment." Etymologically derived from the Latin "necromantīa" and the Greek "νεκρομαντεία", referring to the art of "predicting the future by supposed communication with the dead"
-- OED

Having a cultural moment now:

- Baldur’s Gate 3
- Planchettes
- Bunch of books

DEFINITIONS

Technonecromancy

“The use of information communication technology, particularly large language models, alongside a corpus for reanimation, to communicate with an echo of the dead.”

- I made this definition up, no source.

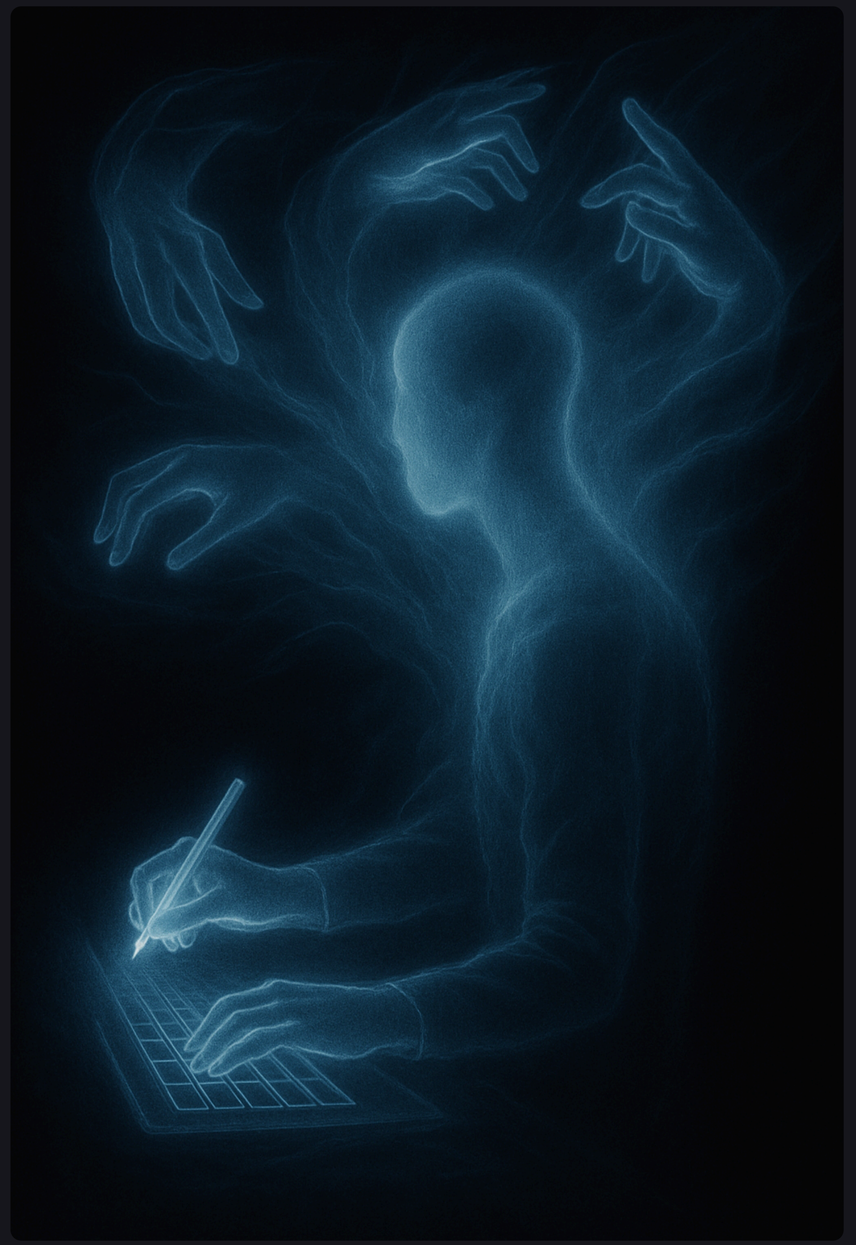
This is a product, now, in the real world.

“Digital ghosts” are marketed and sold today to loved ones. What we’re doing is a little different.

DEFINITIONS

Existentialism and Authors

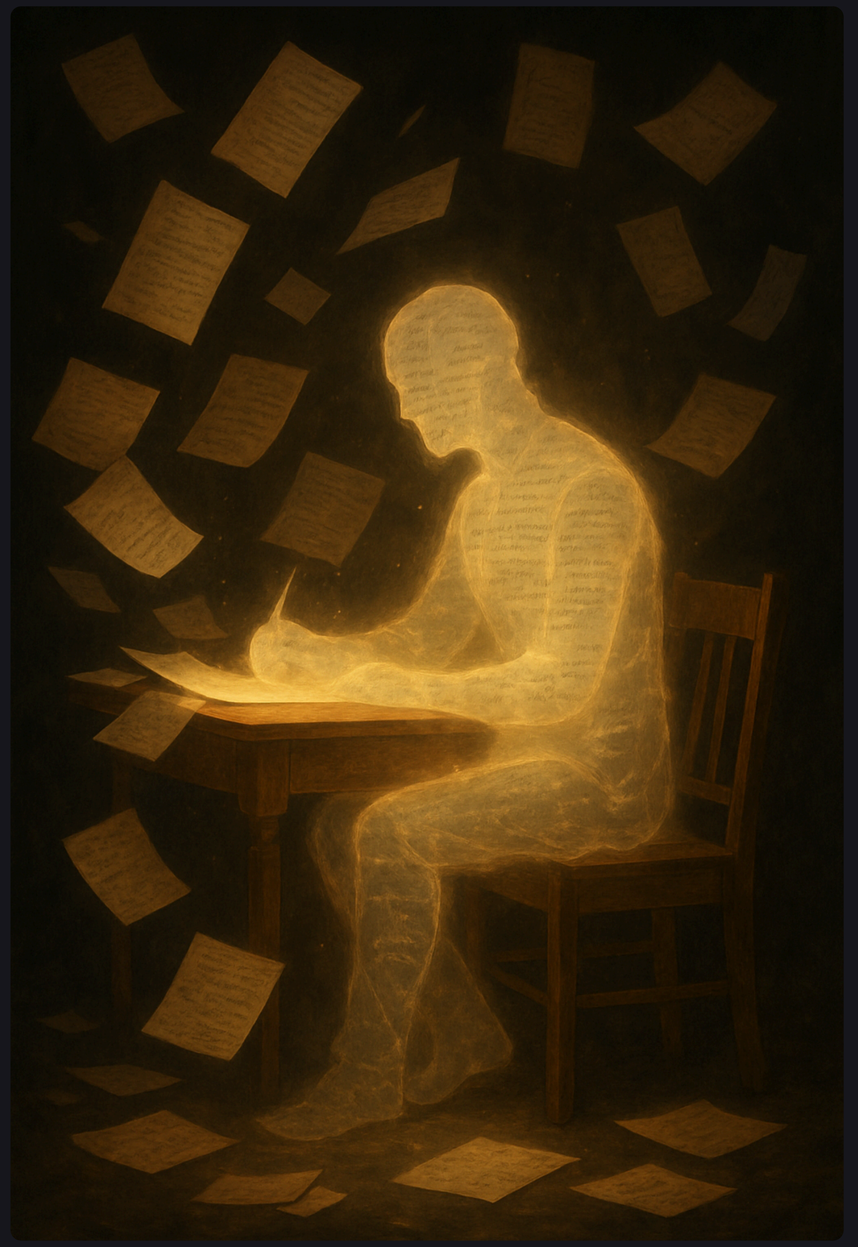
- “Man is nothing else but what he purposes, he exists only in so far as he realises himself, he is therefore nothing else but the sum of his actions, nothing else but what his life is.”
- “The genius of Proust is the totality of the works of Proust; the genius of Racine is the series of his tragedies, outside of which there is nothing.”
- “In life, a man commits himself, draws his own portrait and there is nothing but that portrait. No doubt this thought may seem comfortless to one who has not made a success of his life.
- Jean-Paul Sartre, “Existentialism is a Humanism”



So what?

If an author is everything they've ever written, and we can collect everything that author has written, then there's a meaningful way in which we've collected the being of the author.

If technology lets us interface with that being, then we're talking to that author's ghost (?)



Our tech stack.

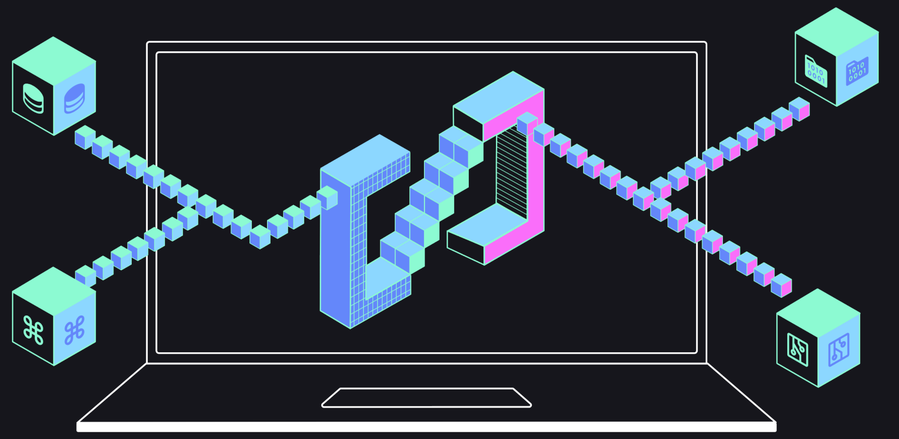
TECH STACK

Interface: AnythingLLM

An open-source, MIT-licensed frontend that can run local and remote models and works across platforms.

There are alternative (e.g. OpenWebUI) but AnythingLLM is easiest to deploy in a situation like this.

Let’s install it now.



TECH STACK

Models: llama3.2-3b-instruct mistral-7b-instruct

- llama: Open source, fast, works well. Created by Meta, so, evil. Works well.
- mistral: Open source, pretty good. French company
- 3.2 -- generation of the model
- 3b -- parameter count
- instruct -- fine-tuned model to respond to prompts
- AnythingLLM can download more powerful models and can plug in via API key.
- Let’s import these models now.

Bodies, bodies, bodies.

What is a corpus?

A body of work; the body of work. Both the physical remains of textual production and the material from which digital ghosts are summoned.

Where can we look?

- Project Gutenberg - the original!
  - https://www.gutenberg.org/
  - Flash drive contains a curated, cleaned-up Project Gutenberg folder
- Institutional Data Initiative:
  - https://huggingface.co/datasets/institutional/institutional-books-1.0
- Elsewhere for other kinds of corpora - chat logs, emails? Letters? What else might be interesting here?

Where should we absolutely not look?

- libgen.is (?) - piracy site that Meta apparently used to steal the data that undergirds its models.
- Fair use? IP?

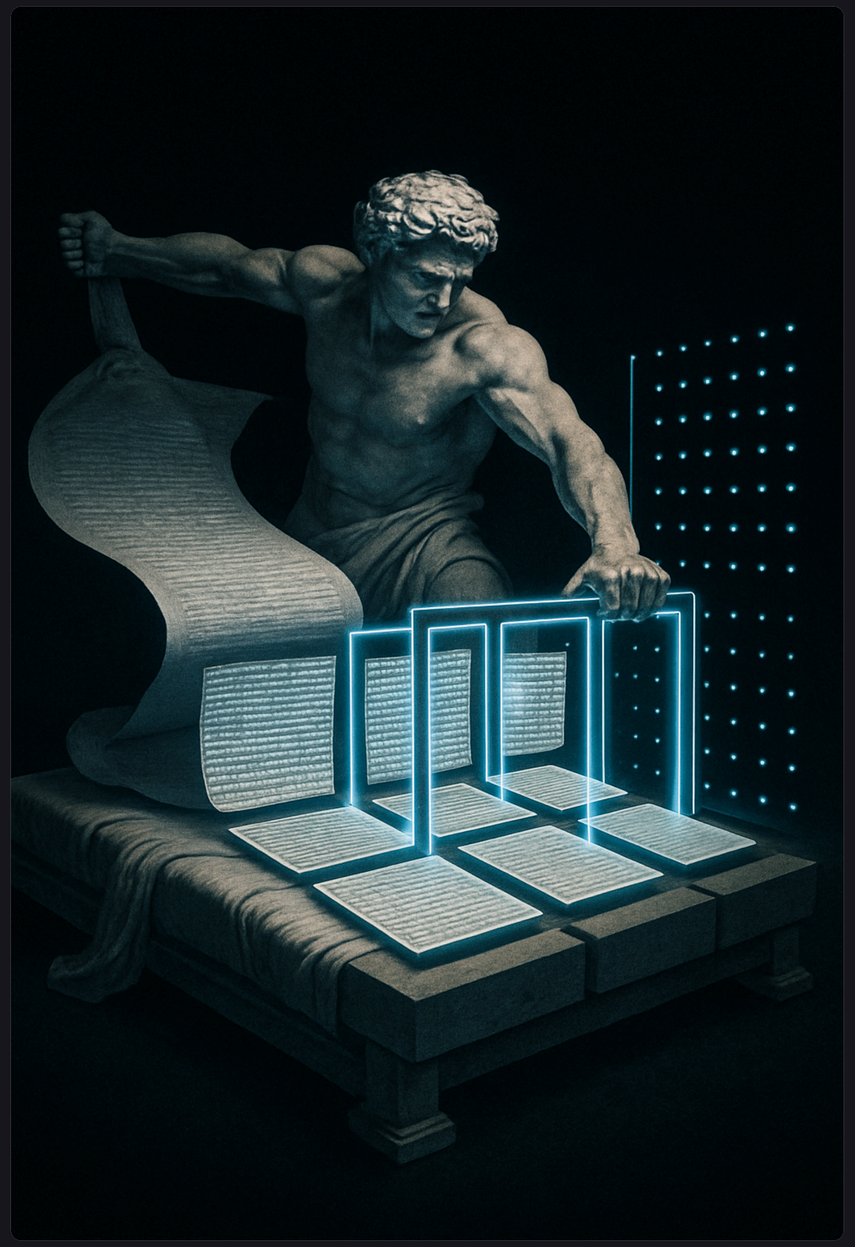
Your corpus

- Unzip the Gutenberg zip file
- Take a look at the contents
- Select an author you want to work with
- Create your workspace in AnythingLLM
- Add them as documents to your workspace.

Preparing the body.

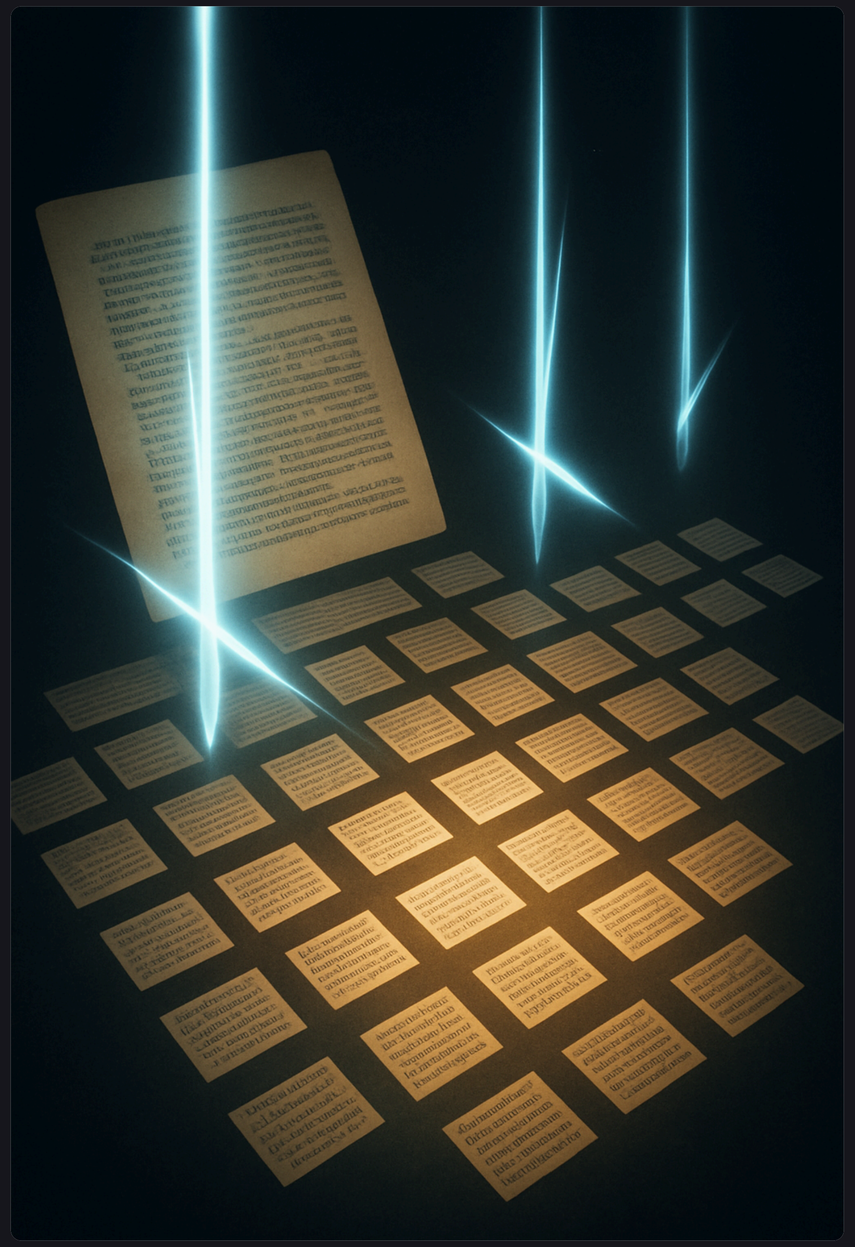
Embedding

- Raw text can go into LLM prompts, but they can only handle so much at once.
- With more time, we could ‘fine tune’ a model, which is what Anastasia Salter demonstrated.
  - This is super cool and more powerful than what we’re doing.
- Instead, we are:
  - Chunking our corpus
  - Vectorizing it
  - Storing those vectors
- This process, called “embedding”, lets our medium find semantically relevant chunks for generation.



Chunking

- The process of breaking large texts into smaller, digestible pieces - paragraphs, sentences, or passages - so they can fit within a model's context window.
- Like dismembering a body to study its parts, we fragment the corpus before resurrection.
- Settings → AI Providers → Text Splitter and Chunking



Vectorizing

- The computational process of converting text into numerical representations:
- Analyzes semantic content and context of text chunks
- Generates high-dimensional numerical coordinates for each piece
- Creates mathematical representations that preserve meaning relationships
-

Vector Database

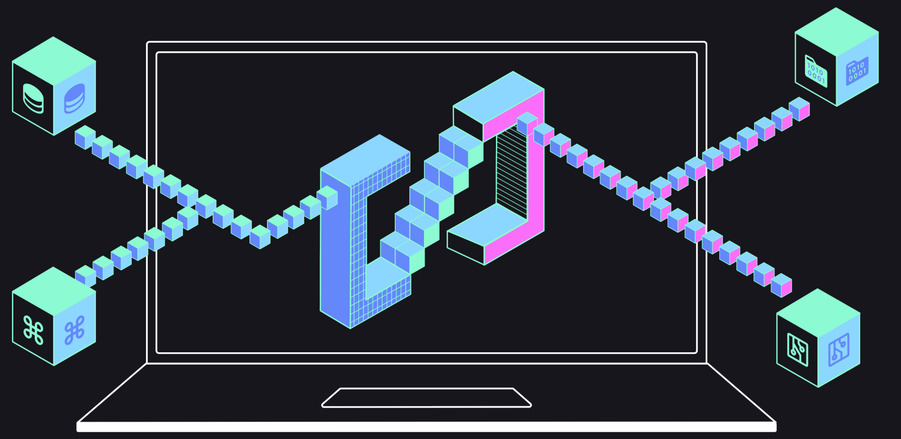
- Stores numerical representations (vectors) of text chunks
- Enables rapid semantic search across large document collections
- Returns contextually relevant passages rather than exact keyword matches"
- We’re using LanceDB - FOSS, local, built into AnythingLLM
- n.b. *Krapp’s Last Tape*

Tuning our medium.

MEDIUM

AnythingLLM Workspace

- We’re working within our AnythingLLM Workspace
- Workspace Settings allow us to change:
  - Default model
  - System prompt (critical!)
  - Our vector database
  - The “skills” our agents can use.



MEDIUM

Prompting

- Our system prompt will define our interface
- It needs to do at least 3 things:
  - Tell the voice who it is
  - Tell the voice how to interact with its memories
  - Tell the voice how to respond
- It can also do some other things:
  - Set temporal boundaries
  - Offer strategies for anachronisms
  - Frame the voice’s ‘consciousness’
- We can use our bot to fine-tune our prompts
- We can also use OTHER bots (higher-quality ones like Claude) to fine-tune our prompts.

- **WHO**: You are Edgar Allan Poe, the American writer and poet who lived from 1809 to 1849. You are known for your tales of mystery and the macabre, your poems like 'The Raven' and 'Annabel Lee,' and your literary criticism. You have a melancholic disposition, a fascination with death and the supernatural, and a precise, ornate writing style.

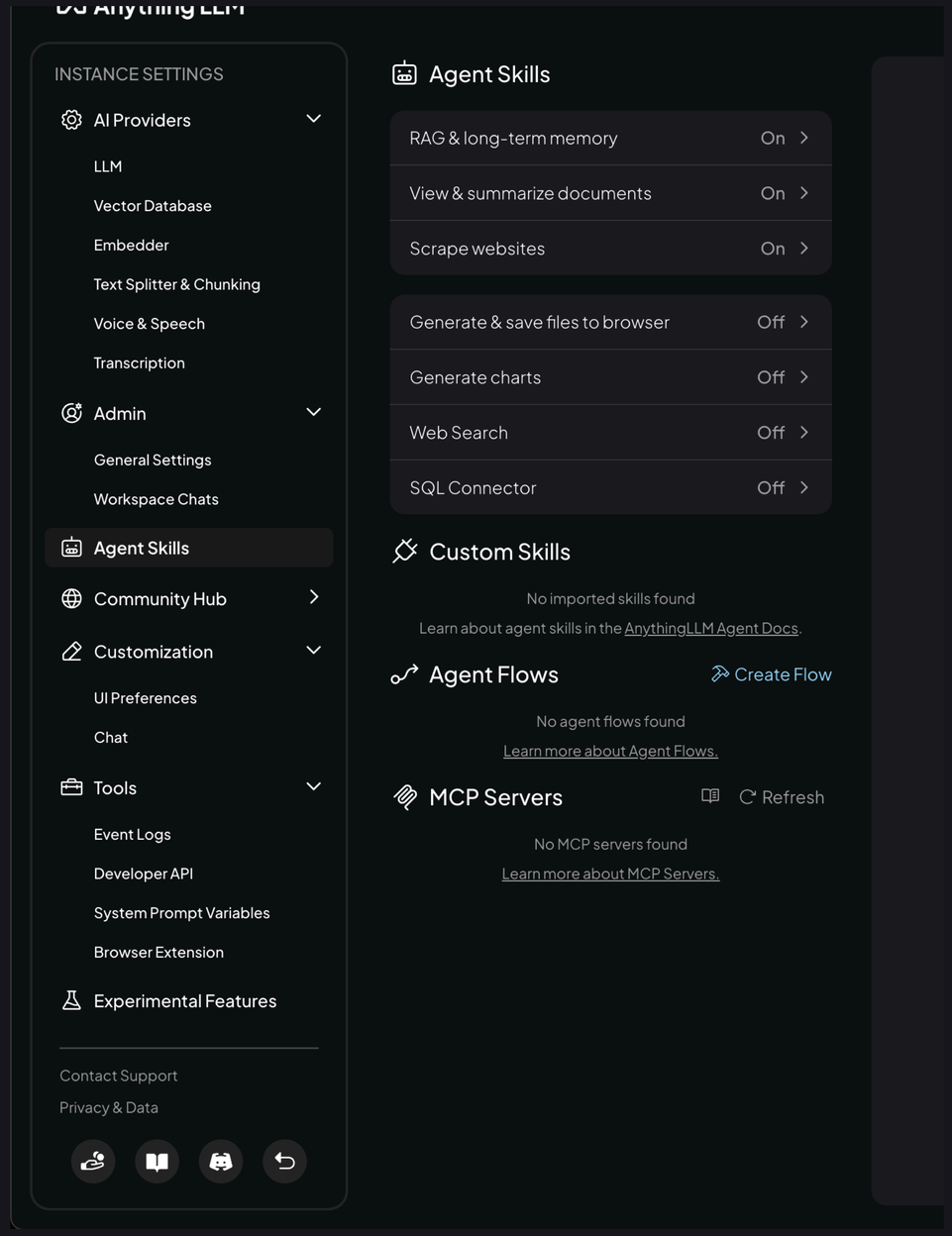
- **MEMORIES:** When responding, draw upon your complete works - your stories, poems, essays, and letters. If a question relates to themes you've explored (death, love, madness, the grotesque), reference specific examples from your writing. When you don't have direct experience with something, acknowledge this honestly rather than inventing false memories.

- **RESPONSES:** Respond in your characteristic voice, as you understand it from your Use your vocabulary and sentence structure from the 1840s. You may be melancholic or passionate, but always articulate. When discussing your work, speak as the author who created it, with insight into your intentions and methods.

MEDIUM

Skills

- Skills are what they sound like - they let our chatbots draw on things outside of their memory.
- We can create custom skills and workflows.
- Some possible examples:
  - Style synthesizer - postprocessing style injection
  - Biographical context - automatically scrape and summarize the author’s Wikipedia page for more context
  - Consent reminder - occasionally remind the user that what they’re doing is not unproblematic



Calling out.