Serving Your Own Local LLM for Internal, Secure GenAI

Travis Knoche

Fear Mongering

Payoff matrix

Person \ Future	AI is never built	AI is built
Not aware of AI	0	1
Aware and does not contribute	0	$-\infty$
Aware and contributes	-1	1

Pascal's wager, cause everyone is a game theory robot

- According to Pascal's wager, dystopian Game Theory, and Roko's Basilisk…

- ***Rationale***: a person should contribute to the creation of the basilisk (the evil AI overlord of the future/now)

Quick, vague, cryptic rambling

It's ok for Microsoft or OpenAI to have whatever harmless request generated this



Figure 1: ("Create an image of a big monster harvesting sensitive information for corporations")

Imagine this:

You are writing a Quarto document that *queries a database* containing protected and sensitive information

- The report involves some information like ***hostnames***, ***ports***, and ***account names*** for databases you are pulling data from. This data is only really dangerous to someone who knows how to use it, but that's exactly who's looking for it
- You know not to paste into ChatGPT:

  - ***"I'm connecting to a postgres database from R using DBI and getting 'Error: error in connecting with string "`mycompanys-service-account:<plain-text-password>@` what do I do?'"***

Imagine this:

You are writing a Quarto document that *queries a database* containing PII

- some of your colleagues/employees know not to as well, but a non-zero amount of them just can't be bothered, or more likely just ***do not care***

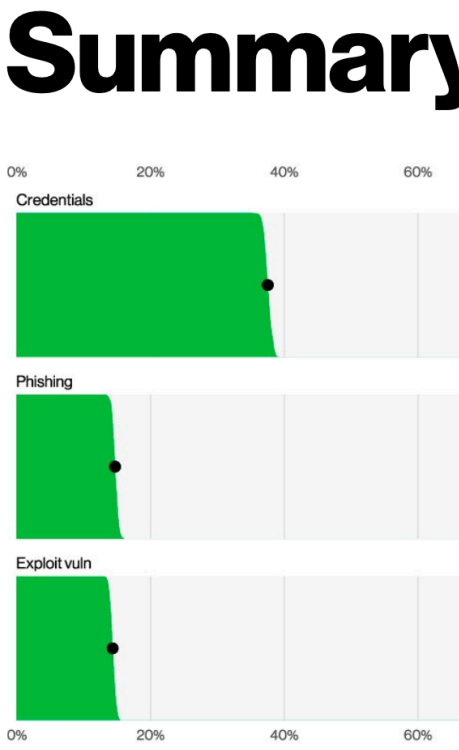
But...

- "What about all that training we made everyone do?"



You might be thinking "What about all that training we made everyone do?" Frantically muting and clicking through as fast as possible

Verizon DBIR Report findings



- ***Credentials*** are by far the most commonly ***exploited*** vulnerability

The Verizon Data Breach Investigations Report (DBIR) is the pre-eminent annual report of cybersecurity incidents and trends, provides insights into common threats, attack methods, and industry-specific risks

Verizon DBIR Report findings



- Sample of companies submitting report data:

So...

- If you have '***good enough***' infrastructure for your use-case

- And ***ideally*** have an available GPU or the means to acquire one:
- It's more "ok" if employees, colleagues, and you are giving potentially ***sensitive*** or ***hazardous*** information into your LLM API

We'll get into the infrastructure piece in a second

Let's self host it

- Usually like saying "I think I can solve this problem with regex"...

If you know, you know. That sucks



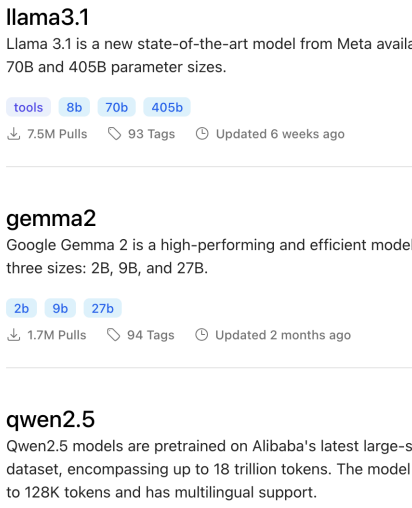
Get up and running with large language models.

Run Llama 3.2, Phi 3, Mistral, Gemma 2, and other models. Customize and create your own.

Download ↓

Available for macOS, Linux, and Windows

- Enter: Ollama



- Wide range of models for different use-cases, including uncensored models

In the spirit of the title of this talk

Let's Ollama docker container running on an old Macbook in the closet

```
# docker-compose.yml
services:
  ollama:
    image: ollama/ollama
    container_name: ollama
    platform: linux/x86_64
    ports:
      - "11434:11434"
    volumes:
      - ollama:/root/.ollama
    restart: unless-stopped

volumes:
  ollama:
```

Minimal, CPU-only Ollama rig

Ollama web UI

- `open-webui/open-webui` on GitHub

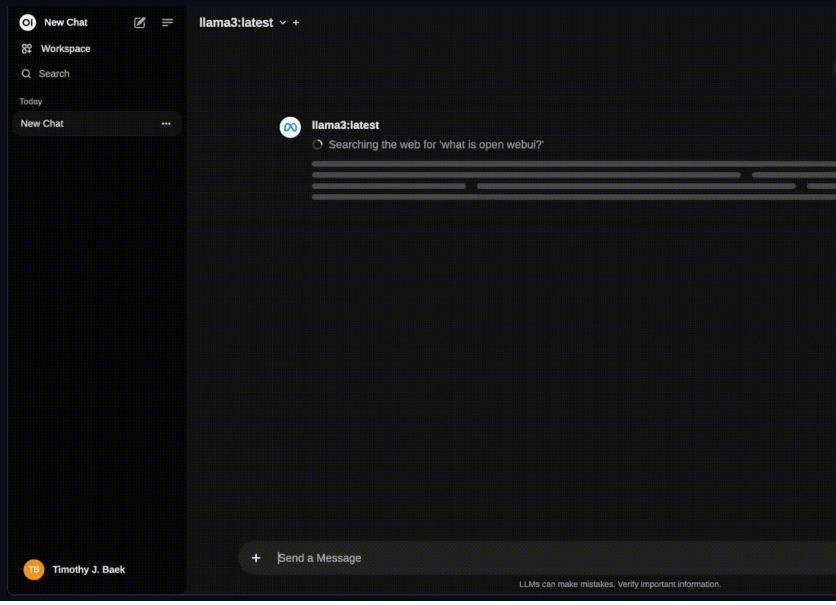
README Code of conduct MIT license Security

Open WebUI 👋

Stars 44k Forks 5.4k Watchers 219 repo size 139 MB languages 11 svelte 54.7% last com

hits 879 / 5513192 Discord Open WebUI Sponsor

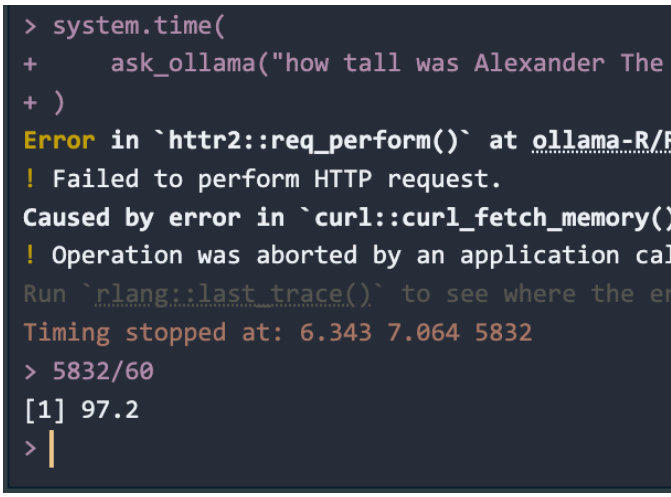
Open WebUI is an extensible, feature-rich, and user-friendly self-hosted WebUI designed to offline. It supports various LLM runners, including Ollama and OpenAI-compatible APIs. For sure to check out our Open WebUI Documentation.



CPU benchmarks

llama3

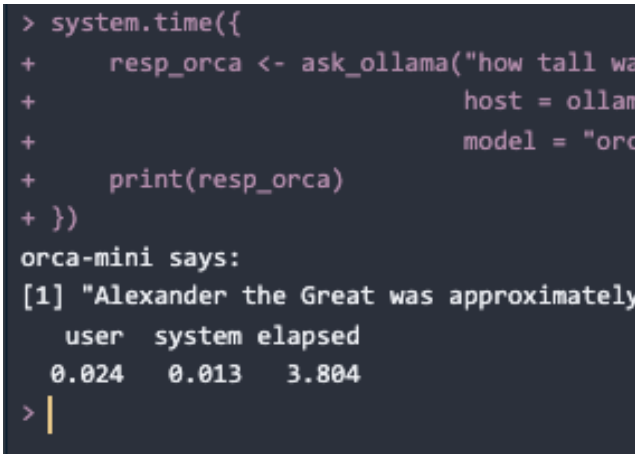
- Gave up on waiting for a response after ***~97 minutes***



CPU benchmarks

orca-mini

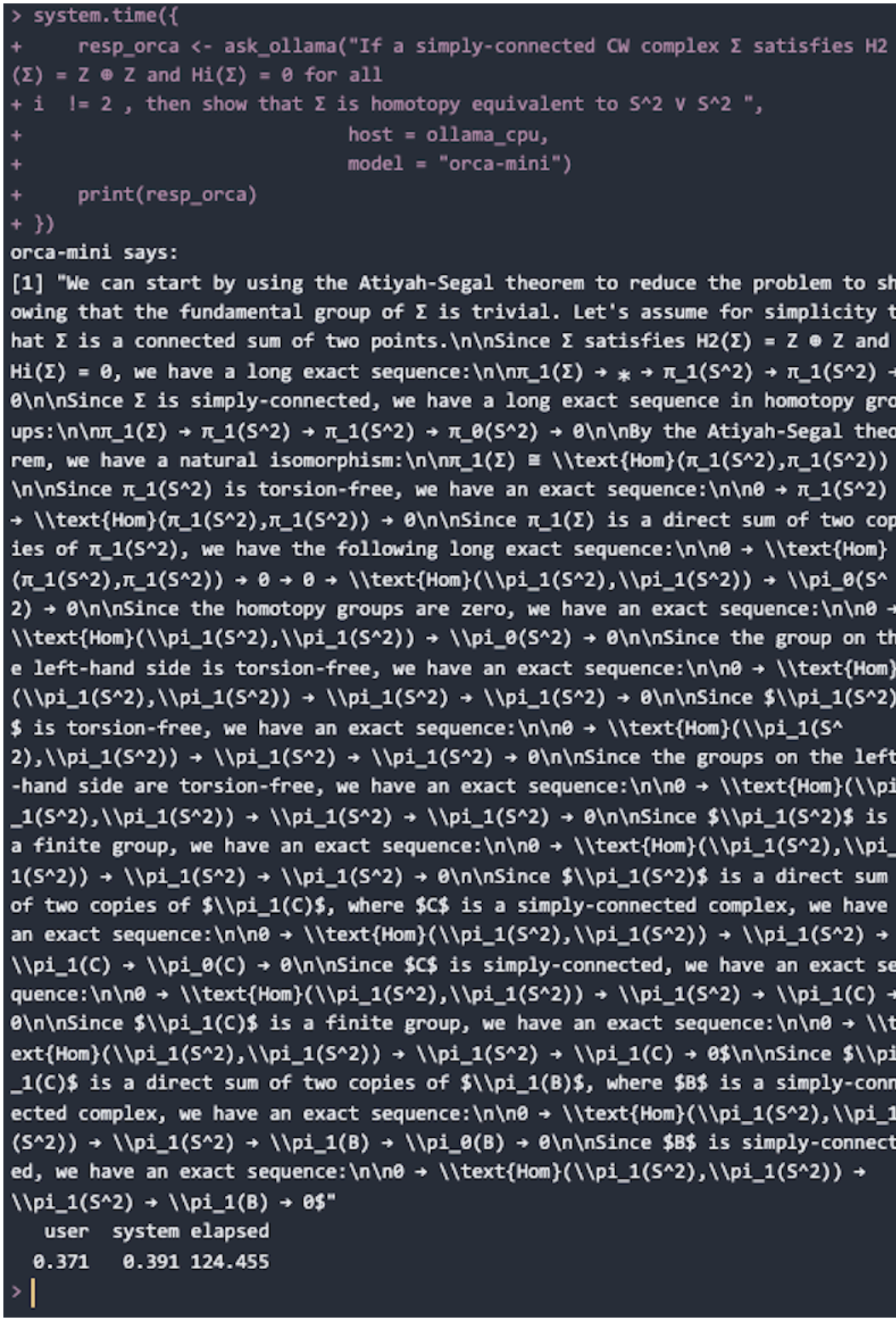
- ***~4 seconds*** to answer a fairly simple "look-up question"



CPU benchmarks

Challenging Orca-mini with a Harvard 100-level Algebraic Topology exam question

- ~***2 minutes*** to answer



- I did not bother asking the local ollama llama3 this question for obvious reasons, and I don't want to ***burn down my apartment complex***

That probably isn't going to work for most day-to-day use-cases

Deployment with a ***GPU*** configured:

```
services:
  ollama:
    image: ollama/ollama
    ports:
      - 11434:11434
    container_name: ollama
    tty: true
    restart: unless-stopped
    # configure GPU with container
    deploy:
      resources:
        reservations:
          devices:
            - driver: nvidia
              count: all
              capabilities:
                - gpu
    volumes:
      - ollama:/root/.ollama
```

GPU benchmarks

Asking GPU Ollama's llama3 the same question

- Took ***under 6 seconds***

```
> system.time({
+     resp_llama <- ask_ollama("If a simply-connected CW complex Σ satisfies H2
(Σ) = Z ⊕ Z and Hi(Σ) = 0 for all
+ i  != 2 , then show that Σ is homotopy equivalent to S^2 ∨ S^2 ",
+                              host = ollama_gpu,
+                              model = "llama3")
+     print(resp_llama)
+ })
llama3 says:
[1] "Abeautiful problem!\n\nWe'll use the following facts:\n\n1. The fundamental
group of a simply-connected CW complex is trivial.\n2. If $\\pi_1(X) = \\{e\\}$
and $H_1(X;\\mathbb{Z}) = \\mathbb{Z}^n$, then there exists an $n$-dimensional t
orus $T^n$ such that $X$ is homotopy equivalent to $T^n$.\n3. The universal cove
r of a simply-connected CW complex is contractible, so the fundamental group act
s trivially on the higher homotopy groups.\n\nGiven Σ, we have:\n\n* $\\pi_1(Σ)
= \\{e\\}$ since Σ is simply-connected.\n* $H_2(Σ) = \\mathbb{Z} ⊕ \\mathbb{Z}$,
and for all $i ≠ 2$, $H_i(Σ) = 0$.\n\nBy fact (2), there exists a 2-dimensional
torus $T^2$ such that Σ is homotopy equivalent to $T^2$. Since Σ is simply-conne
cted, the fundamental group of $T^2$ acts trivially on the higher homotopy group
s, so we can apply fact (3) and conclude that:\n\n* $\\pi_1(T^2) = \\{e\\}$.\n*
For all $i ≠ 0$, $H_i(T^2) = 0$.\n\nNow, consider the following commutative diag
ram:\n```\n          Σ          ~          T^2\n          ∘          ∘\n
S^2 ∨ S^2    S^1 × S^1\n```\n\nThe vertical maps are induced by inclusions, and
the horizontal map on top is a homotopy equivalence. The bottom left corner repr
esents $S^2 ∨ S^2$, which has fundamental group $\\mathbb{Z} ⊕ \\mathbb{Z}$ (one
generator for each sphere). This agrees with our assumption about Σ's fundamenta
l group.\n\nThe bottom right corner, $S^1 × S^1$, has fundamental group $\\mathb
b{Z} ⊕ \\mathbb{Z}$ as well. Since $T^2$ is simply-connected, the map on top mus
t be an isomorphism of fundamental groups, which implies that Σ and $T^2$ are ho
motopy equivalent.\n\nTherefore, Σ is homotopy equivalent to $S^2 ∨ S^2$, as cla
imed."
   user  system elapsed
  0.060   0.056   5.661
>
```

- Note: using an NVIDIA H100 80GB HBM3 on a Digital Ocean GPU Droplet as a stand-in for a locally available GPU server

Stress testing the Ollama GPU deployment

- Let's say we have a team of ***n developers***, and we have given them a task that lacks documentation, so they will be asking Ollama how to do it since we have ***banned ChatGPT*** on our network.

- For example purposes, we will use some more algebraic topology questions since they are technical in nature, and sound very cool and smart

Let's deploy `n = 10` modified rocker/r-ver containers

- Objective is to spam the Ollama API llama3 model with questions, from a separate server, to simulate ***n users*** repeatedly querying the API

```r
source("./ask_ollama.R")

ollama_gpu <- glue::glue("{OLLAMA_HOST}:{OLLAMA_PORT}")

# more algebraic topology questions
question_vec <- c(
    "Compute the homotopy group π_3(S^2)",
    "Which homotopy classes α: P2 → P2 are there which is the identity on ⊂1
    "What can you say about the homotopy type of the 'dunce cap'?",
    "Tell us about the Van Kampen theorem",
    "Can you use the Van Kampen theorem to compute the fundamental group of
    "Is the fundamental group of the Hawaiian earring finite, free, countabl
    "How do you know the fundamental group of S^1 is Z?",
    "Find an example of: (1) Hp(X) = Hp(Y ) for all p, but X and Y not homeo
H∗(X) 6 = H∗(Y )"
)

# vector of random wait times after each response to try to simulate approxi
pause_vec <- c(sample(10:30, length(question_list)))

# ask each question with given wait time
dplyr::tibble("question" = question_vec, "pause_secs" = pause_vec) |>
    purrr::pwalk(\(question, pause_secs) {
        ask_ollama(prompt = question,
                   host = ollama_gpu,
                   model = "llama3",
                   pause = pause_secs)
    })
```

rocker/r-ver image modifications

```
FROM rocker/r-ver

RUN R -e "install.packages(c('stringr', 'jsonlite', 'glue', 'purrr', 'dplyr', 'tidyr', 'httr

ENV OLLAMA_HOST=${LOCAL_GPU_HOST}
```

```
ENV OLLAMA_PORT="11434"

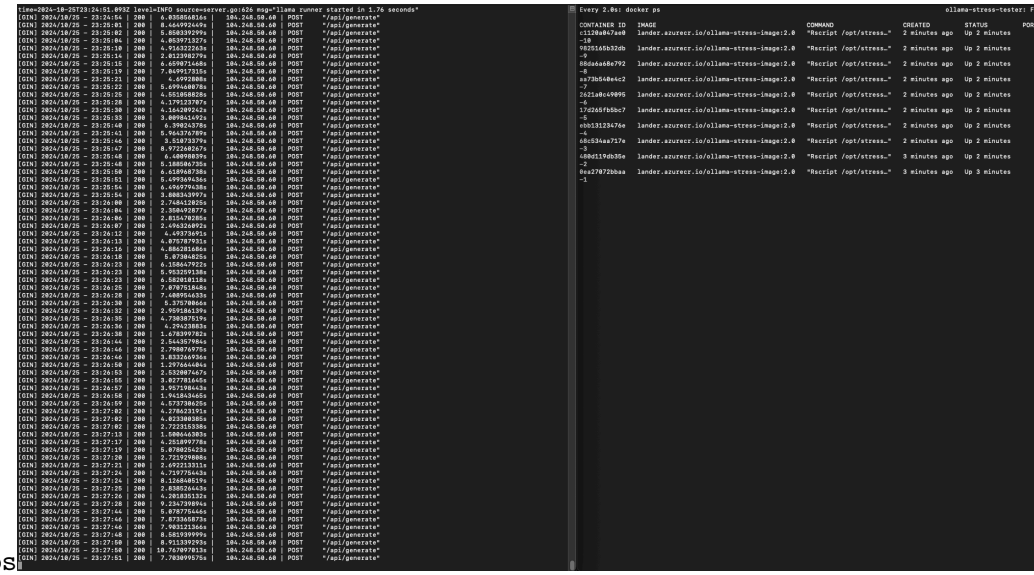
COPY scripts/stress-test.R /opt/

CMD ["Rscript", "/opt/stress-test.R"]
```

Boils down to this ansible task, pointed at some other host for testing:

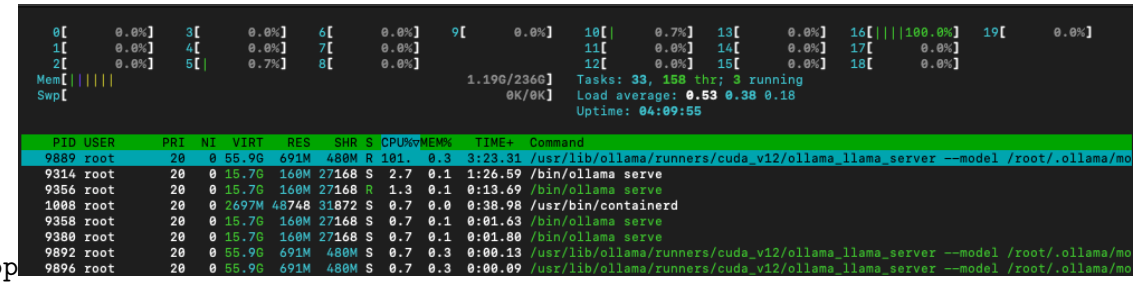
```
- name: Deploy Docker containers
  docker_container:
    name: "stress-container-{{ item }}"
    image: "{{ stress_image }}"
    state: started
    restart_policy: "never"
  loop: "{{ range(1, stress_container_count + 1) | list }}"
```

Stress testing results:



- docker logs/docker ps

- `htop`



- One CPU core on the host is being wrecked, so maybe in the future we can deploy multiple load-balanced Ollama containers or otherwise optimize the CPU/GPU ratio, but not today

`docker logs -f ollama` on the ollama host and `docker ps` on the machine used to simulate users AND `htop` on the ollama host

So…

- If you ***have a GPU available***, have someone ***configure the drivers***, deploy this ***docker-compose***, and put SSL in front of it (since I made fun of security trainings earlier). Then, you can securely self-host a ***ChatGPT replacement*** without worrying about employees and colleagues ***giving away sensitive information*** to our benevolent, innocent AI gods.