



Konferencija Srce DEI

Arhitektura umjetne inteligencije temeljena na alatima otvorenog koda

Miro Mačković

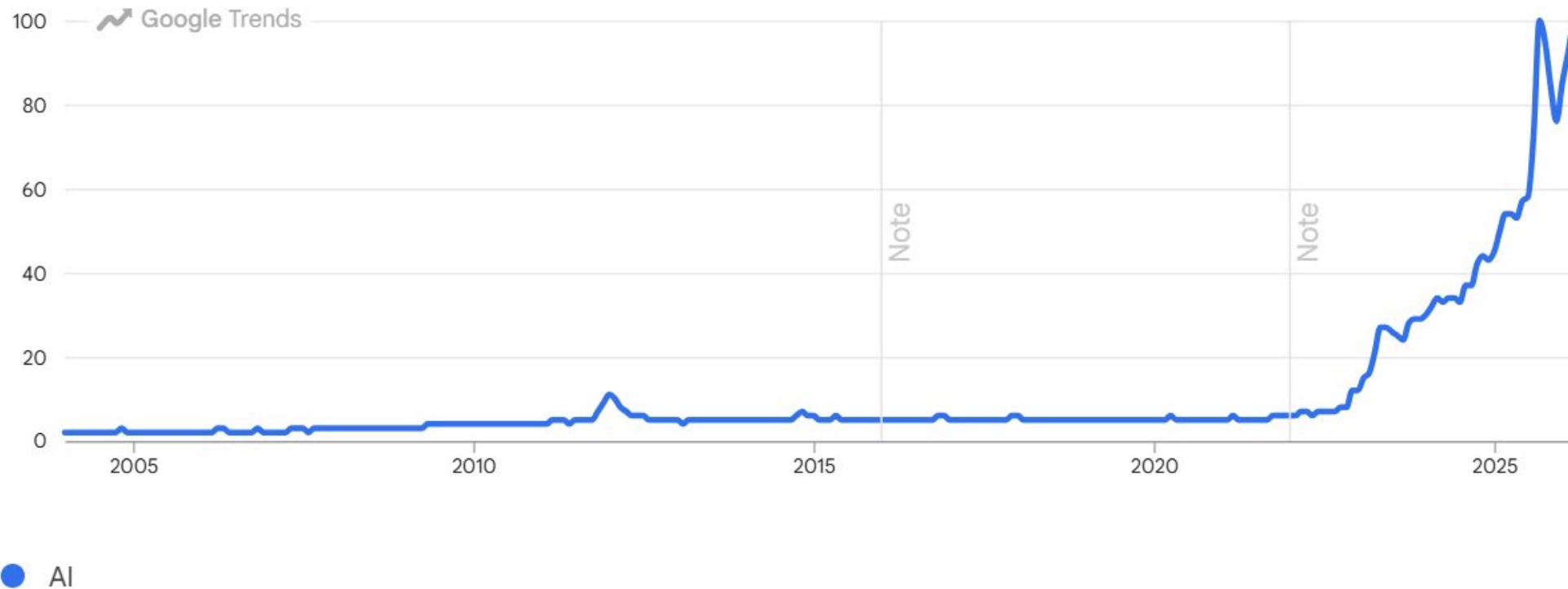
Sveučilište u Zagrebu
Sveučilišni računski centar

Srce DEI 2026



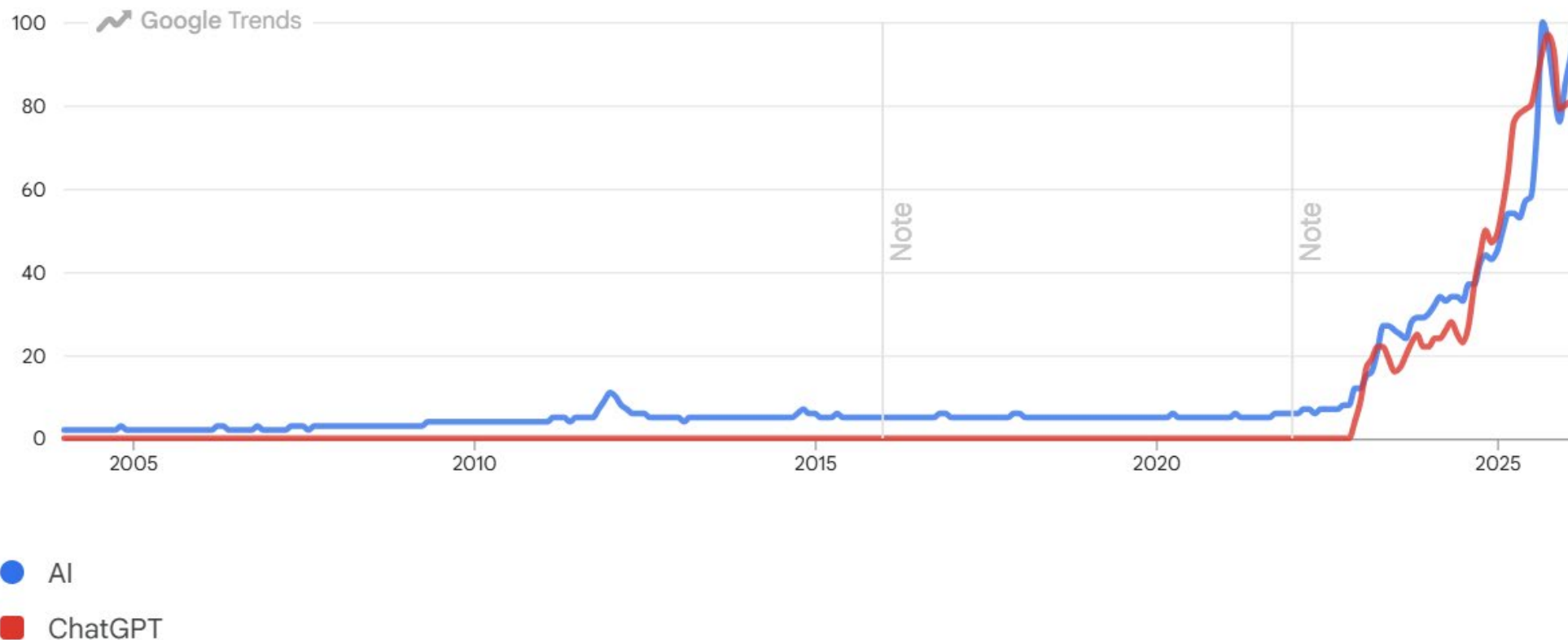


Umjetna inteligencija





Umjetna inteligencija





ChatGPT

- Attention Is All You Need – Google, 2017.
- Generative pre-trained transformer – OpenAI, 2018.
- ChatGPT, November 30, 2022



ChatGPT

ChatGPT

Log in Sign up for free

New chat

Search chats

Images

Apps

Deep research

See plans and pricing

Settings

Help

Get responses tailored to you

Log in to get answers based on saved chats, plus create images and upload files.

Log in

What's on the agenda today?

+ Ask anything Voice

By messaging ChatGPT, an AI chatbot, you agree to our [Terms](#) and have read our [Privacy Policy](#). See [Cookie Preferences](#).



ChatGPT (Claude, Le Chat, Gemini...)

- GPU
- LLM
 - > 1T parametara
- API
 - stateless (Chat Completions)
 - stateful (Responses)
- UI
 - web sučelje, perzistentnost



Srce

- GPU
- LLM
 - open source model, open-weight model, open model
 - 8B parametara, 20B, 120B (quantized)
 - vLLM
- API
 - LiteLLM
 - stateless (Chat Completions)
 - stateful (Responses)
- UI
 - Open WebUI
- RAG
 - Onyx



vLLM

- vLLM v0.1.0
 - 20. lipnja 2023. Sky Computing Lab at UC Berkeley
 - vLLM < PyTorch < Linux Foundation
- PagedAttention algoritam
- Continuous batching of incoming requests
- Chunked prefill
- Prefix caching





vLLM

```
docker run --runtime nvidia --gpus all \  
    -v \  
    ~/.cache/huggingface:/root/.cache/huggingface \  
    --env "HF_TOKEN=$HF_TOKEN" \  
    -p 8000:8000 \  
    --ipc=host \  
    vllm/vllm-openai:latest \  
    --model Qwen/Qwen3-0.6B
```



vLLM

```
curl http://localhost:8000/v1/chat/completions \
  -H "Content-Type: application/json" \
  -d '{
    "model": "Qwen/Qwen2.5-1.5B-Instruct",
    "messages": [
      {"role": "system", "content": "You are
a helpful assistant."},
      {"role": "user", "content": "Who won
the world series in 2020?"}
    ]
  }'
```



LiteLLM

- OpenAI-compatible API
- Rate Limiting: Per-user and per-model
- Cost Tracking: Usage tracking and budget management
- Authentication: API key management and user authentication



LiteLLM

config.yaml:

model_list:

- model_name: gpt-oss-20b
 litellm_params:
 model: gpt-oss-20b
 api_base: http://vllm:8000
 api_key: API_KEY

general_settings:

master_key: sk-1234567

database_url: "postgresql://llmproxy:dbpassword9090@db:5432/litellm"



OpenAI compatible API

The screenshot displays the Swagger UI for the LiteLLM API. The browser title is "LiteLLM API - Swagger UI". The page header includes the API name "LiteLLM API" with version "1.82.3" and "OAS 3.1". Below the header, there are several links: "Proxy Server to call 100+ LLMs in the OpenAI format", "LiteLLM Admin Panel on /ui", "LiteLLM Model Cost Map", "LiteLLM Model Hub", and "LiteLLM Chat UI". An "Authorize" button is visible on the right side of the header.

The main content area is titled "model management" and lists the following endpoints:

- GET /models Model List
- GET /v1/models Model List
- GET /models/{model_id} Model Info
- GET /v1/models/{model_id} Model Info
- GET /v1/model/info Model Info V1
- GET /model/info Model Info V1
- GET /model_group/info Model Group Info
- GET /public/model_hub Public Model Hub
- GET /public/model_hub/info Public Model Hub Info
- GET /public/litellm_model_cost_map Get Litellm Model Cost Map
- PATCH /model/{model_id}/update Patch Model
- POST /model/delete Delete Model
- POST /model/new Add New Model
- POST /model/update Update Model
- POST /model_group/make_public Update Public Model Groups
- POST /model_hub/update_useful_links Update Useful Links
- POST /access_group/new Create Model Group



OpenAI compatible API

- IDE extenzije
- Agenti
- Aplikacije

The screenshot shows the Kilo Code IDE interface. The main editor displays a TypeScript file named `page.tsx` with the following code:

```
src > app > page.tsx > ...
1  "use client";
2
3  import { useState, useEffect, useRef, useCallback, useMemo } from "react";
4
5  interface Task {
6    id: string;
7    text: string;
8    completed: boolean;
9  }
10
11 type Filter = "all" | "active" | "completed";
12
13 // Swipe threshold in pixels to trigger delete
14 const SWIPE_THRESHOLD = 100;
15
16 function TaskItem({
17   task,
18   onToggle,
19   onDelete,
20 }): {
21   task: Task;
22   onToggle: (id: string) => void;
23   onDelete: (id: string) => void;
24 } {
25   const [swipeX, setSwipeX] = useState(0);
26   const [isSwiping, setIsSwiping] = useState(false);
27   const startXRef = useRef(0);
28   const currentXRef = useRef(0);
29   const itemRef = useRef<HTMLLIElement>(null);
30
31   const handleTouchStart = useCallback((e: React.TouchEvent) => {
32     startXRef.current = e.touches[0].clientX;
33     currentXRef.current = 0;
```

The sidebar on the left contains a list of agents and a recent task:

- Code**: The default agent. Executes tools based on configur...
- Plan**: Plan mode. Disallows all edit tools.
- Debug**: Diagnose and fix software issues with systematic de...
- Orchestrator**: Coordinate complex tasks, delegating to specializ...
- Ask**: Get answers and explanations without making chang...

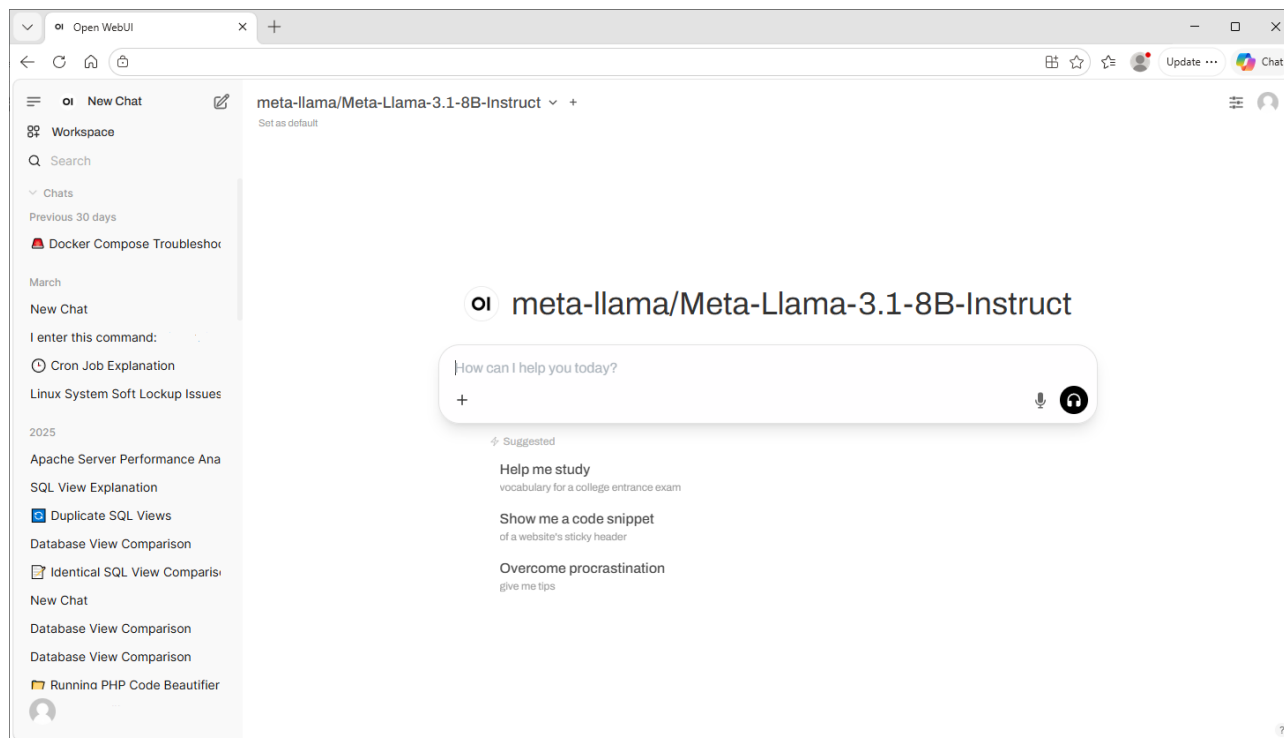
Recent task: Refactor src/page.tsx for efficiency just now

At the bottom, the status bar shows: Ln 1, Col 1 Spaces: 2 UTF-8 LF {} TypeScript JSX Autocomplete



Open WebUI

```
docker run -d -p 3000:8080 --add-  
host=host.docker.internal:host-gateway -v open-  
webui:/app/backend/data --name open-webui --restart  
always ghcr.io/open-webui/open-webui:main
```





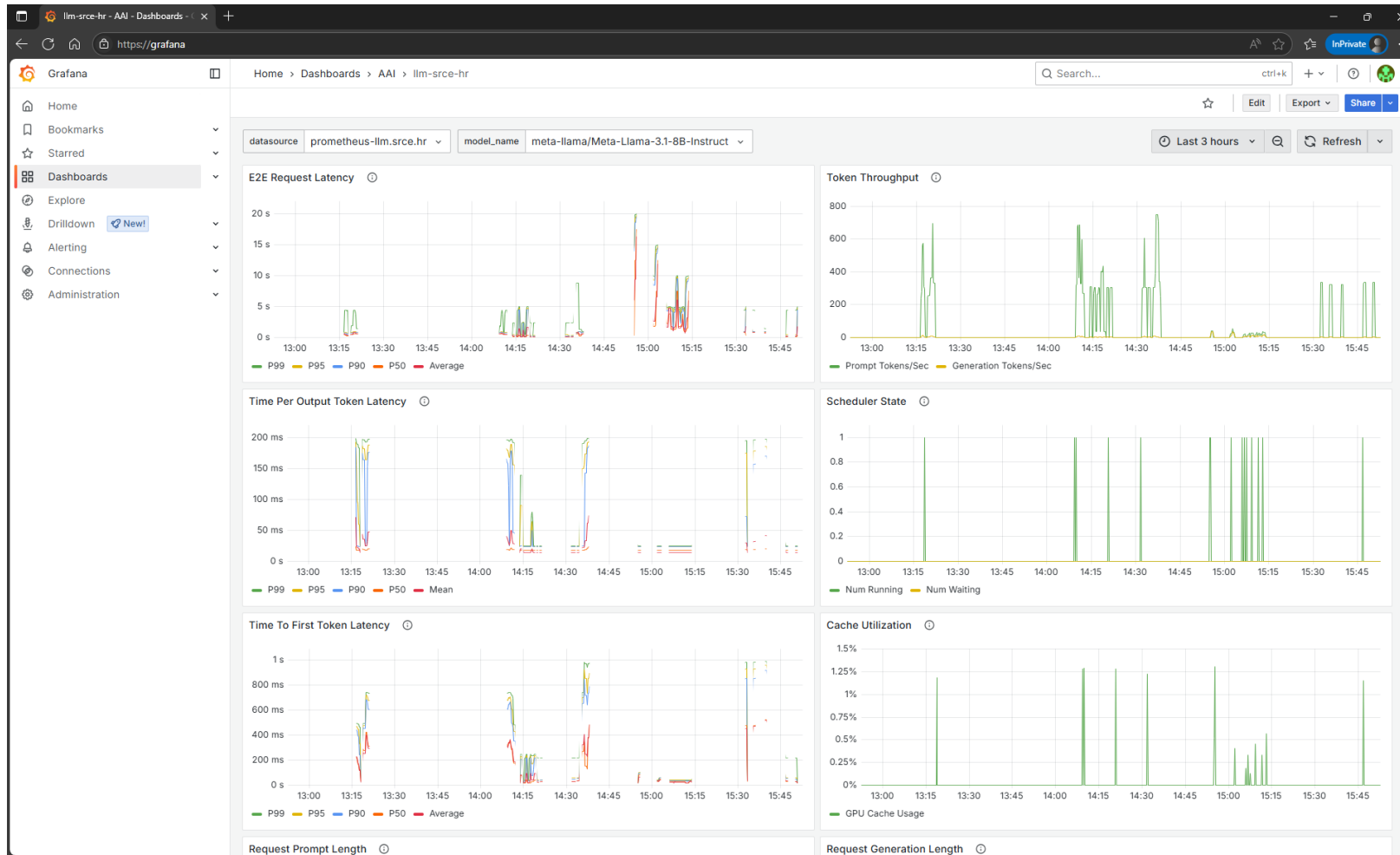
Onyx stack

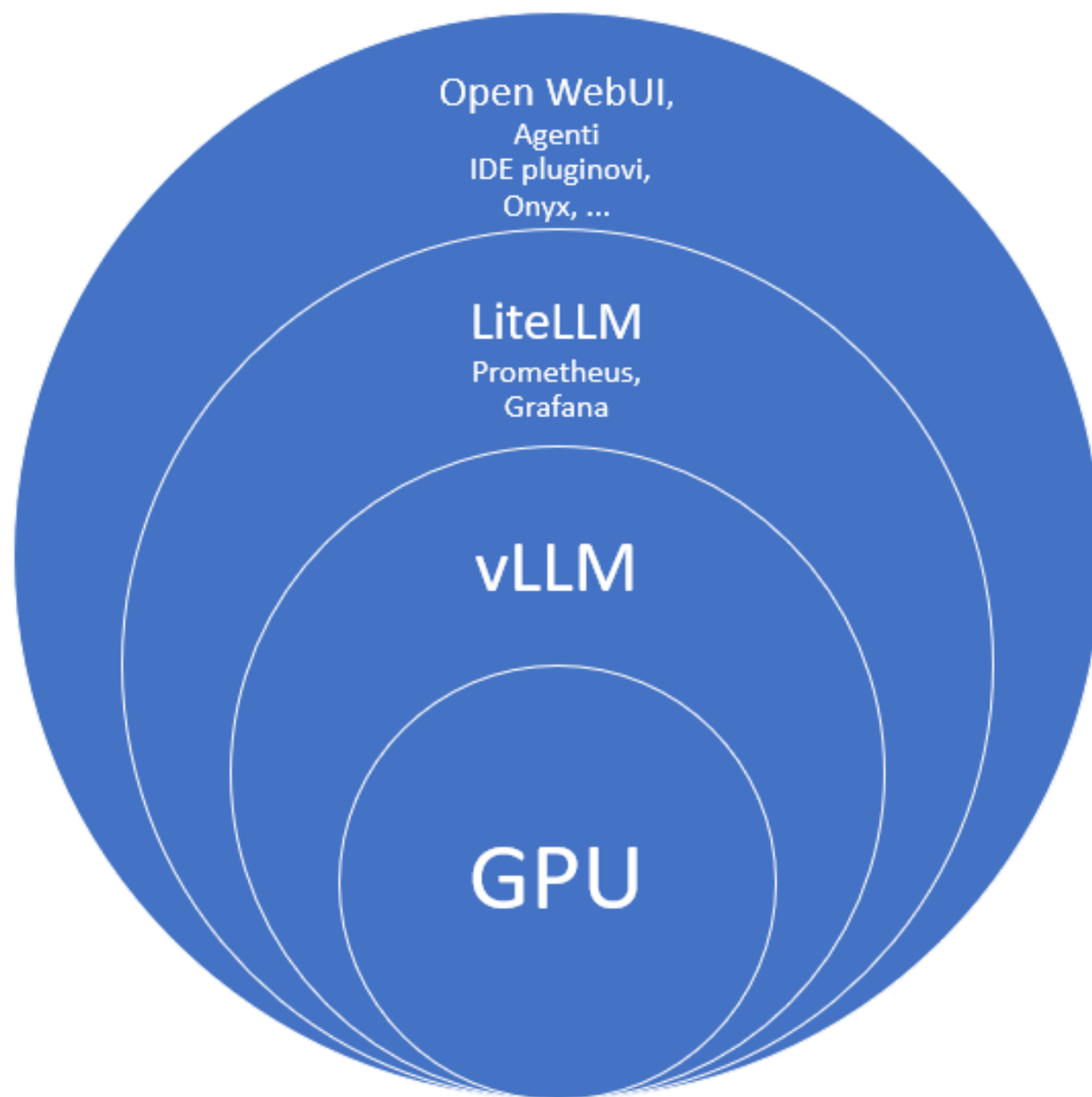
- onyx-nginx-1
- onyx-web_server-1
- onyx-background-1
- onyx-api_server-1
- onyx-relational_db-1
- onyx-inference_model_server-1
- onyx-index-1
- onyx-indexing_model_server-1
- onyx-cache-1

The screenshot displays the Onyx AI platform interface. On the left, a sidebar shows navigation options: 'New Session', 'Projects', 'Agents' (Sales Assistant, HR Policy, More Agents), and 'Sessions' (Summarize Most Recent 3..., Onyx AI Intro, Onyx AI Use Cases Overvie..., POC Documents, Onyx AI Latest News). The main chat area shows a search for 'Onyx' with a result titled 'What is Onyx?'. The result text describes Onyx as an open-source enterprise AI platform. Below the text are two bullet points: '40+ plug-and-play connectors' and 'LLM-agnostic architecture'. A follow-up question 'What are the main features' is shown with a response: 'Considering Onyx's features... The user asked earlier for the most highlighted features of Onyx. I'll give a concise list of its key features. Since Onyx seems like an internal product, I might not need to browse the web for references.' At the bottom, there's an input field for follow-up questions and a 'Deep Research' button. On the right, an 'All Sources' panel lists search results, including 'Onyx: Open Source AI Platform', 'AI agent startup Onyx raised a larger seed round with this 9-slid...', 'Onyx Followup', 'Joachim in #general', and 'Onyx Overview'.



Prometheus i Grafana







Hvala!



Ovo djelo je dano na korištenje pod licencom Creative Commons *Imenovanje* 4.0 međunarodna.

Srce politikom otvorenog pristupa široj javnosti osigurava dostupnost i korištenje svih rezultata rada Srca, a prvenstveno obrazovnih i stručnih informacija i sadržaja nastalih djelovanjem i radom Srca.

www.srce.unizg.hr

creativecommons.org/licenses/by/4.0/deed

www.srce.unizg.hr/otvoreni-pristup

