Open Source 西门子

Beijing · 2026

Oct. 29
Industrial Open Source

Oct. 30
Open Source Robotics

Agenda

CodeAI: 以开源方式赋能自主可控的人工智能

CodeAI: Enabling Sovereign AI through Open Source



Max Wittig & Fabio Huser

On behalf of the code.siemens.com team

SIEMENS

Who we are

How code.siemens.com
serves thousands of developers



SIEMENS



We commute in cars designed, with **Siemens software...**

… built in factories running on **Siemens automation** ...





… and charged by a **Siemens smart grid**.

Unrestricted | © Siemens 2026 | code.siemens.com | 2026-10-29

SIEMENS

Open Source efforts at Siemens

Outreach

- Website/portal
  opensource.siemens.com
- Manifesto
- SIG memberships
- OSS libraries and projects
  jailhouse
  ros-sharp
  kas
  ix

Open Source founding

- Paid SIG memberships
  Civil Infrastructure Project
  Eclipse Foundation
  Linux Foundation
  Margo
- Maintainer sponsorship
  MkDocs Material
  Django
  Prettier
  ffmpeg

Community management

- Events and workshops
  CH Switzerland edition
  CN China edition
- Social Coding ambassadors
- GitHub/GitLab organization and repositories management
- Support for contributing and publishing

Direct maintainer sponsorship in 2026: > $30 000

https://blog.siemens.com/2023/09/how-siemens-supports-open-source-development/

SIEMENS

code.siemens.com - For developers, from developers



DevOps and developer enablement for Open Source, Inner Source and Closed Source
A vibrant community - users get answers in less than an hour

Core features (GitLab EE)

- **Collaboration** on Code (version control, merge requests, issue tracker, wiki, social coding)
- **CI/CD** with shared Linux Windows and Mac runners, or bring your own
- **Pages** hosting (e.g., docs)
- Container and Package **Registry**

Ecosystem

- Collaboration with **business partners** possible
- Continuous **security scans**
- **Core APIs** available for everyone, e.g. llm
- **Open & Inner Source** management
- Inner Source **Tax Transfers**, Export Control
- docs-theme (2500 re-uses)

- **High confidentiality** rating
- **> 92k users, 79 countries**
- **> 400k builds / day**
- Pure Open Source

Ongoing Topics

- ❑ Open Source **contributions**
- ❑ **Security** guidance
- ❑ **AI assistants** (platform, IDE)

SIEMENS

Why?



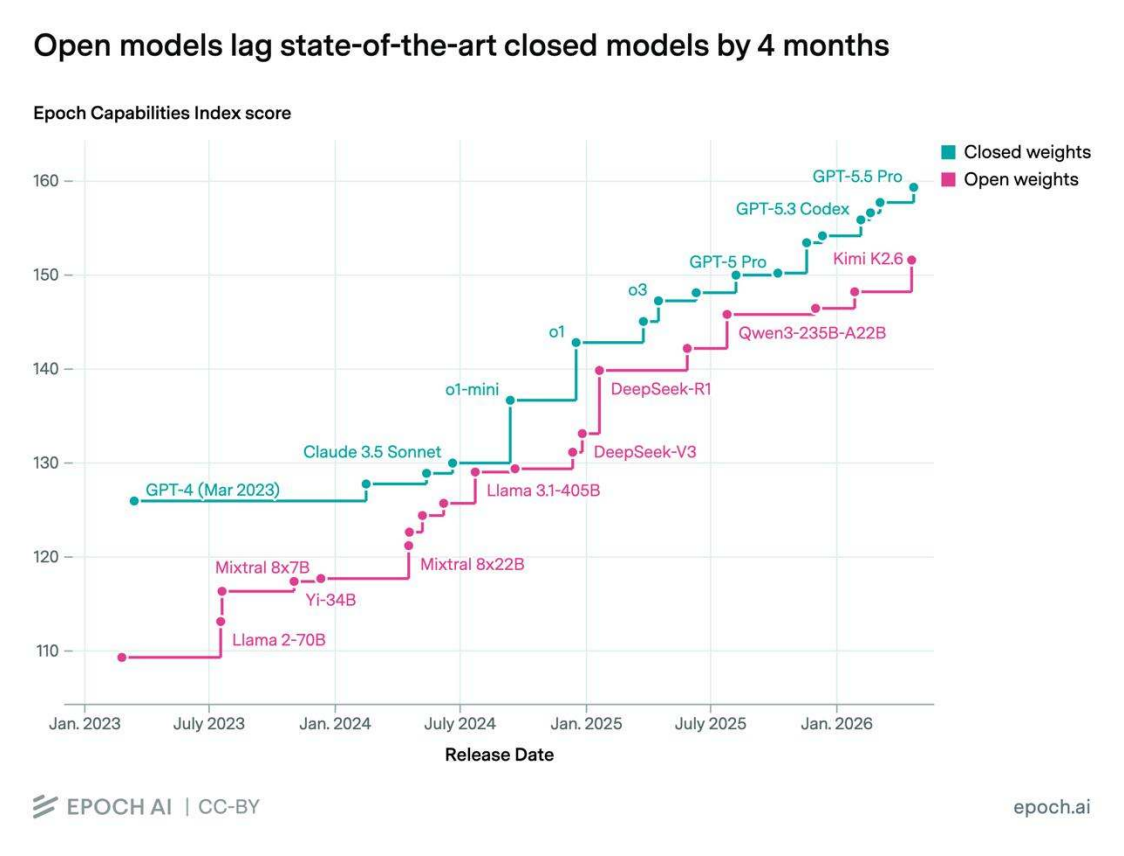
SIEMENS

Advantages of open weight LLMs in-house

- All **data stays internal**, highest confidentiality rating LLM.
- **Cost-effective** (no license/subscription fees).
- **Customizability** (in-house code usage) and control.
- **Context-based** answers about code and documentation.
- The same **ecosystem** code.siemens.com developers are used to.
- Open weight LLMs are **catching up.**

SIEMENS

Open weight LLMs have been picking up speed



Source: J. Edwards & L. Emberson, Epoch AI (May 2026), CC BY | epoch.ai/data-insights/open-closed-eci-gap

Less than 50% of
the evaluated models are **Proprietary**
on the Leaderboard
of the LMSYS Chatbot Arena.

March 7th, 2024

The hyper-scalers are nowadays also offering open weight models such as deepseek.

"Open Source LLMs are about 6 months behind the commercial LLMs" *, Arthur Mensch (CEO Mistral)*

SIEMENS

In-house LLM APIs



Unrestricted | © Siemens 2026 | code.siemens.com | 2026-10-29

SIEMENS

Hardware Stack

- Fleet of NVIDIA L40S & NVIDIA H200s

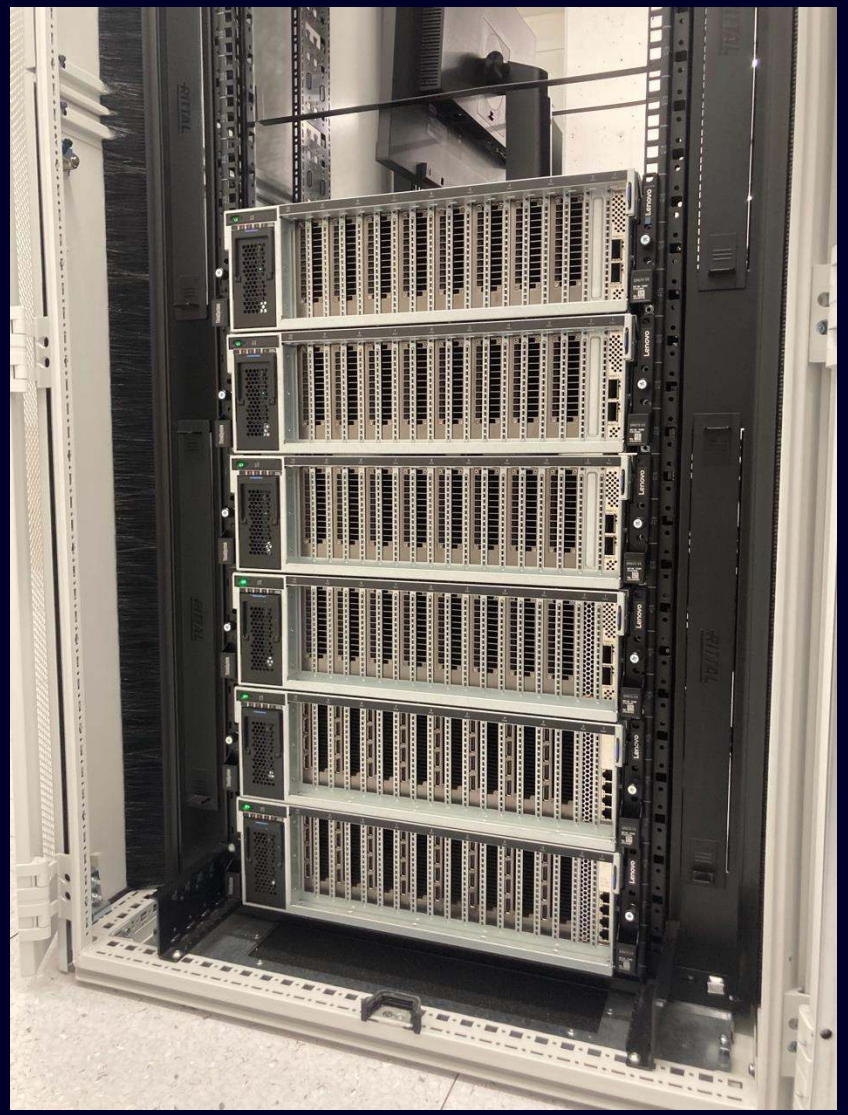


Inside view of a unit with 8 x NVIDIA H200 GPUs

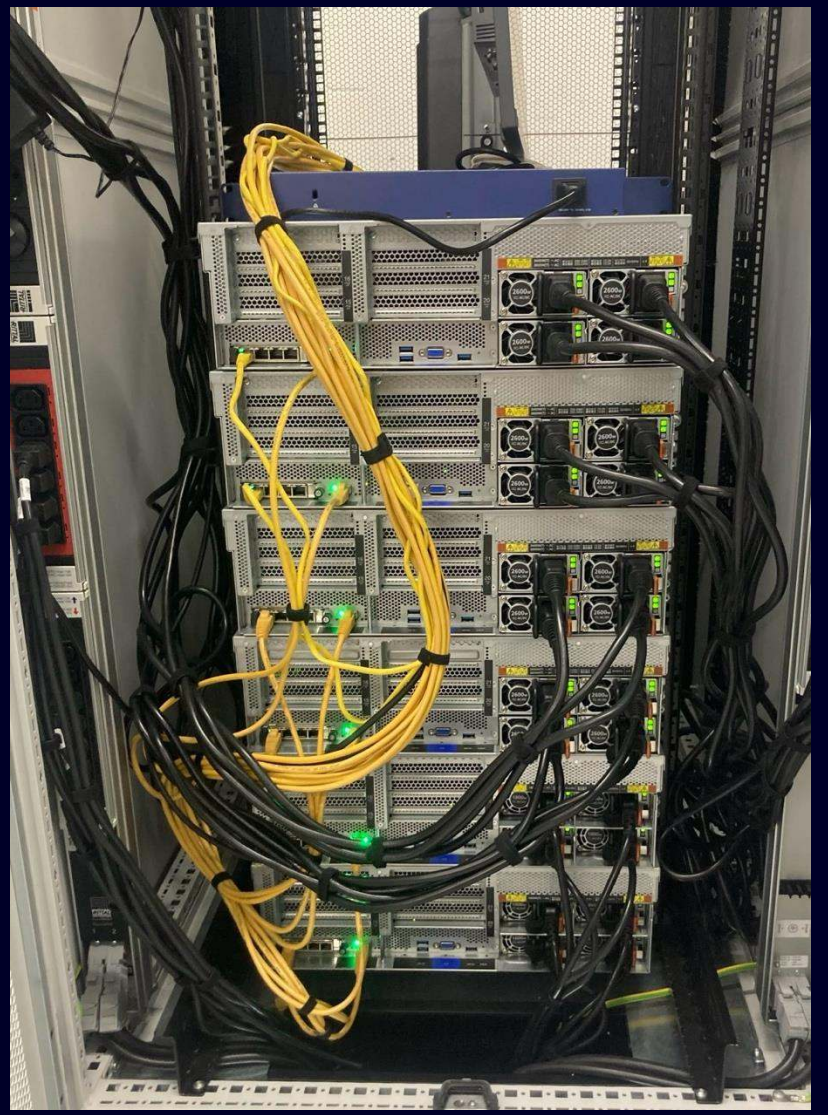
Unrestricted | © Siemens 2026 | code.siemens.com | 2026-10-29

SIEMENS

Hardware Stack



Front view of a GPU rack



Rear view of a GPU rack

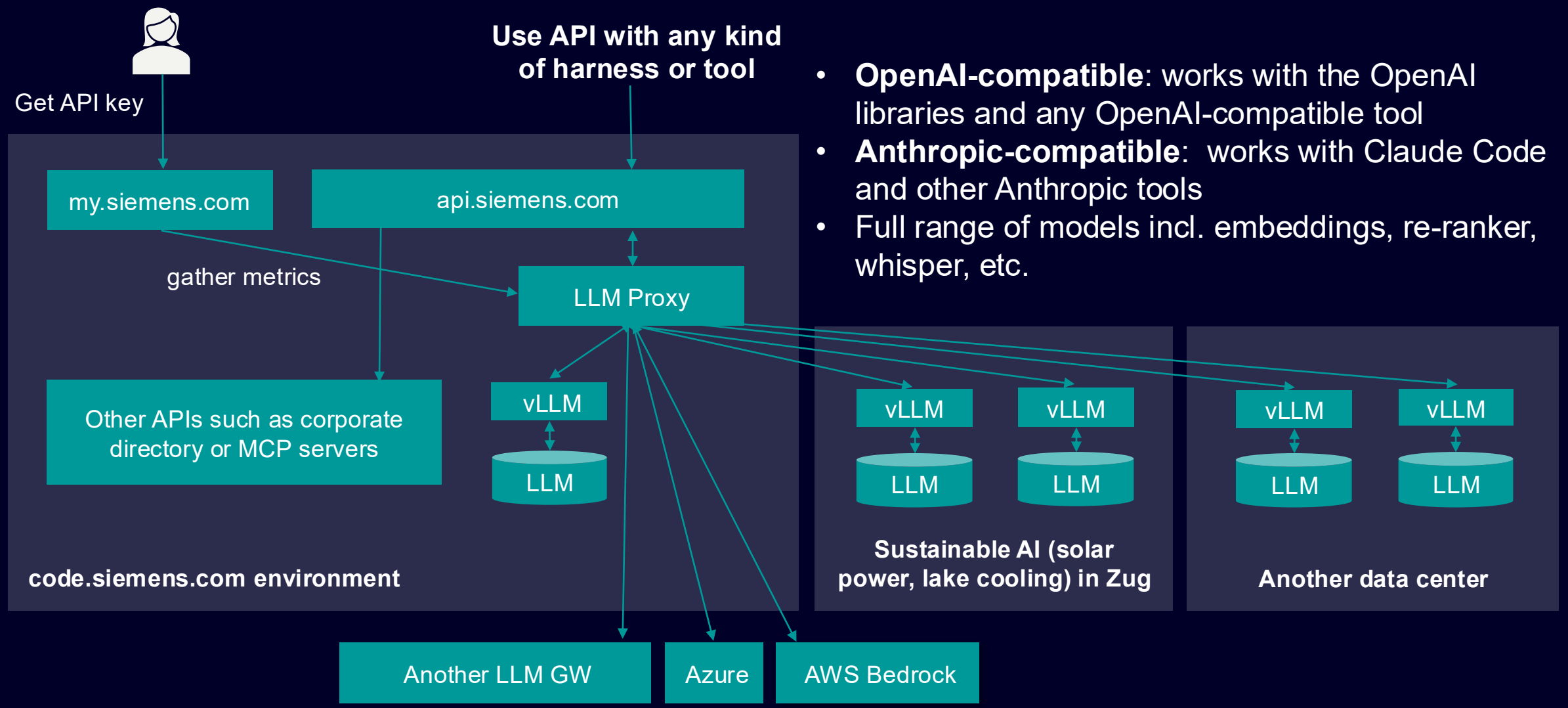
SIEMENS

Software Stack – API Backend

- **LLM Engine**: vLLM. High throughput, low-latency inference engine
  - https://github.com/vllm-project/vllm
- **Request Router:** Agent Router (a.k.a. Envoy AI Gateway). OSS control plane for AI traffic
  - https://github.com/theagentrouter/agent-router
- **API Management**: Kong + Siemens Self-Service Portal for authentication, key management, rate limiting, etc.
  - https://github.com/Kong/kong
- **Observability:** Prometheus + Grafana (incl. GPU exporters for vLLM) + ElasticSearch. Metrics, monitoring, logging, etc.
  - https://github.com/prometheus/prometheus
  - https://github.com/grafana/grafana
  - https://github.com/elastic/elasticsearch

SIEMENS

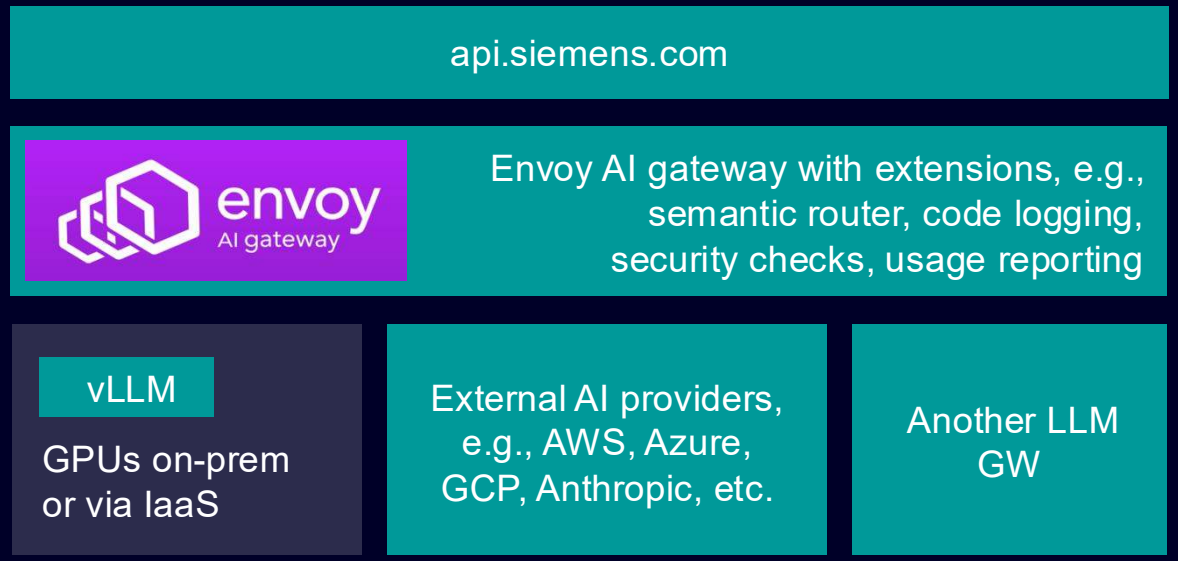
We master technology end-to-end, all based on Open Source and Inner Source



- **OpenAI-compatible**: works with the OpenAI libraries and any OpenAI-compatible tool
- **Anthropic-compatible**: works with Claude Code and other Anthropic tools
- Full range of models incl. embeddings, re-ranker, whisper, etc.

SIEMENS

Let's dive a bit deeper into the LLM Proxy and what we learned over the last 3 years



Next steps:

- Semantic router
- Extend translations and endpoints within Envoy to provide fully unified experience for the developers
- Explore further optimization options at GW level
- Evaluate Model Context Protocol (MCP) Gateway which is part of Envoy AI gateway

- In 2023 we started with Fastchat from lm-sys (the folks with the leaderboard)
- Then we started to use LiteLLM and collaborated there with the founders, contributed several features, but it's data science code and unstable
- Re-evaluation: envoy AI gateway and llm-d in very early state, vLLM production-stack was fulfilling our needs, part of vLLM eco-system, several contributions such as transformations, routing options, ops related features to the project we needed
- Now we replaced vllm-production stack for envoy AI GW as it evolved heavily and provides the full flexibility we need and used, e.g., during outage of Zug campus

Unrestricted | © Siemens 2026 | code.siemens.com | 2026-10-29

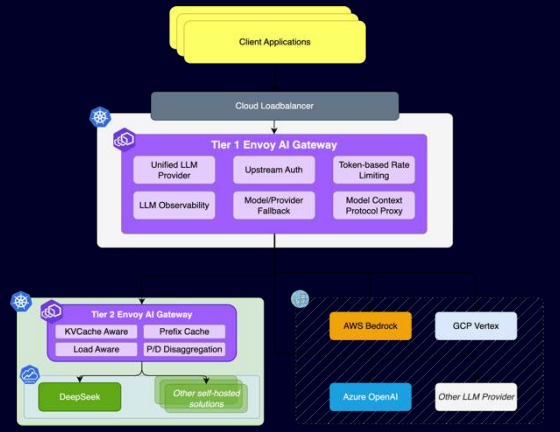
Inference details on GPUs

vLLM

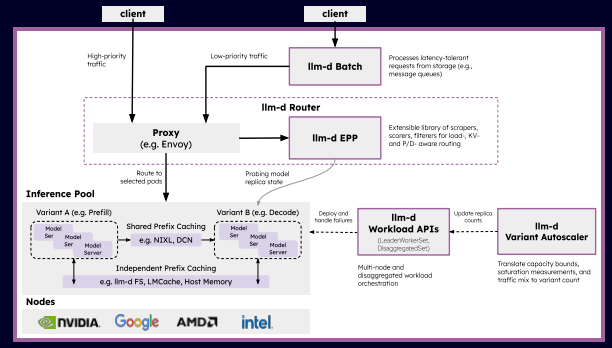
We use vLLM since 2023, plenty of contributions to optimize operations, fix tool call handlings, translations of endpoints, etc.
Typically, we have Day 0 support for new open weight models

Next steps:

- Extend local network to enable distributed KV Cache using LMCache
- Local object store to store models
- Prefill / Decode separation to increase throughput
- Investigate on tier 2 stack for on-prem GPUs



Envoy AI gateway Tier 2



llm-d

Unrestricted | © Siemens 2026 | code.siemens.com | 2026-10-29

SIEMENS

Upstream contributions and collaboration

vLLM

- Better JSON logging support
- Enforce usage counting stats
- Tuned MoE configurations for H200 NVL inference

vLLM #23642

Agent Router

- Anthropic API to AWS Bedrock translation
- Tools, reasoning and streaming

Agent Router #2103

vLLM Production Stack

- Anthropic message routing
- Configurable health checks

#775 | #906

And many more…

- GitLab
- Gitlab-MCP
- Envoy

Unrestricted | © Siemens 2026 | code.siemens.com | 2026-10-29

SIEMENS

Sustainable

Solar powered, lake water cooled

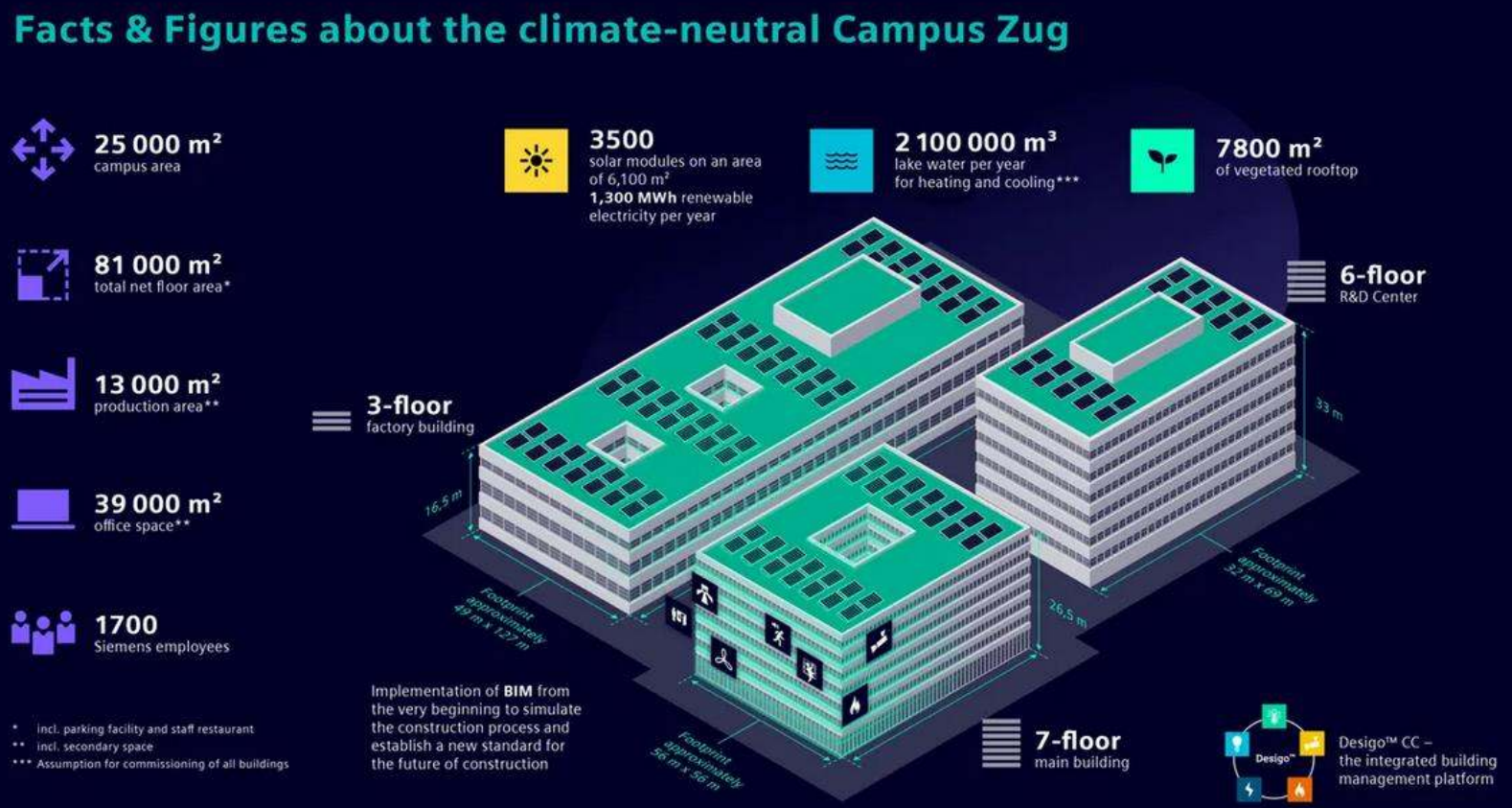


Unrestricted | © Siemens 2026 | code.siemens.com | 2026-10-29

SIEMENS

Sustainability, in-house knowledge, climate-neutral AI in the long run

- Typical AI offerings charge per in- and outbound tokens - **expensive**
- **Amortization** of a physical machine vs. AWS g5 **~ 3-4 months**



Server room at R&D center



- 2 x servers with 8 x L40s GPUs each
- **4 x servers with 8 x H200 GPUs each**
- Option to scale at other locations

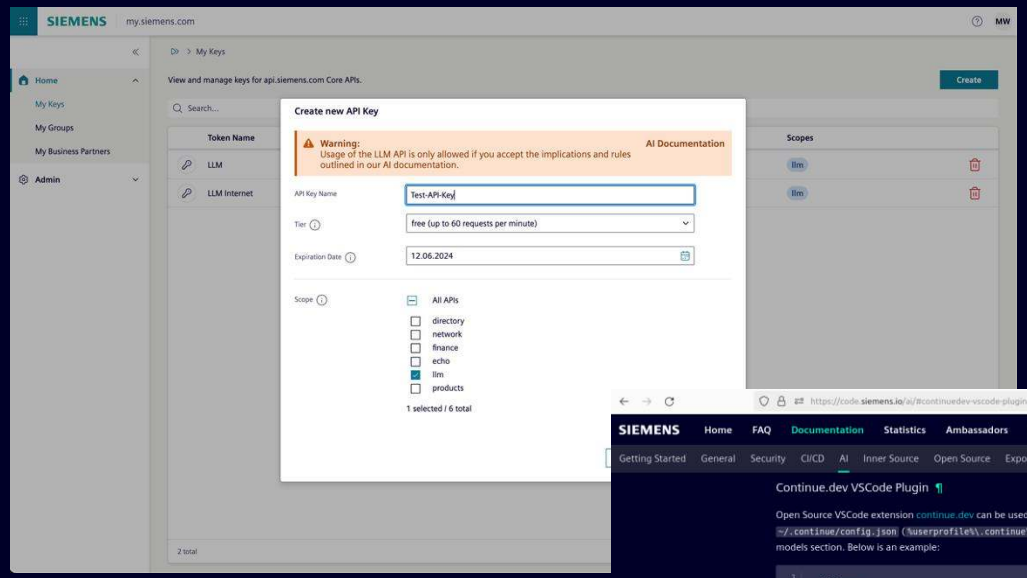
Unrestricted | © Siemens 2026 | code.siemens.com | 2026-10-29
SIEMENS

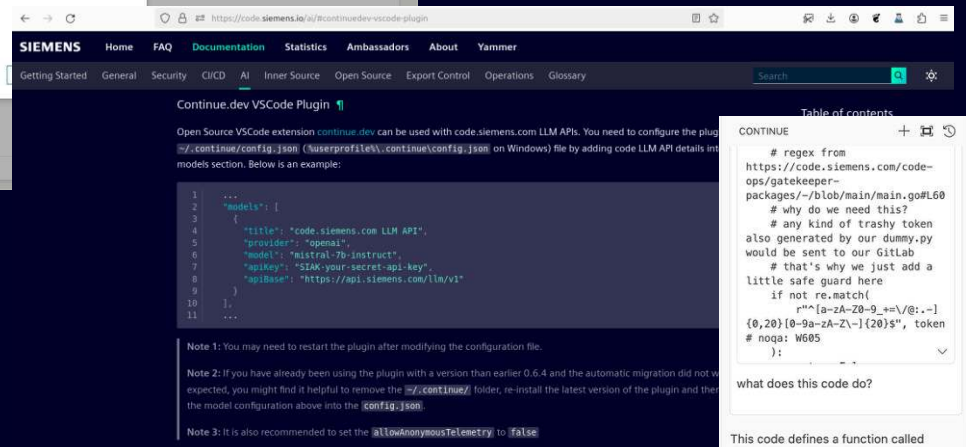
Usage Examples

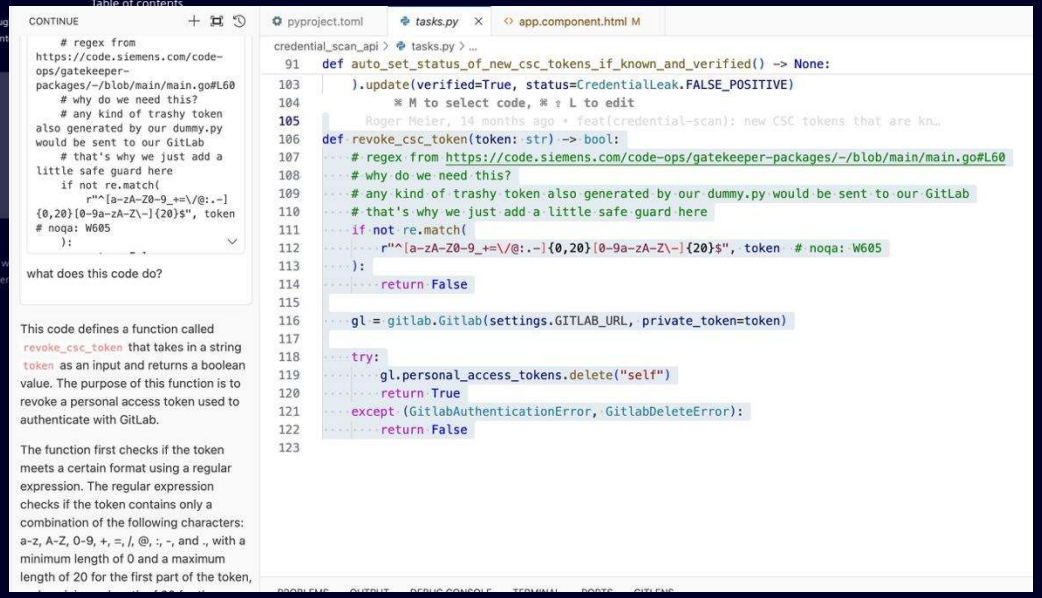


SIEMENS

Easy Self-Service, Get API key, Configure IDE, use it







- Users get started in less than 5 minutes.
- All data stays within Siemens.
- Open Source IDE plugin continue.dev (Apache-2.0).

SIEMENS

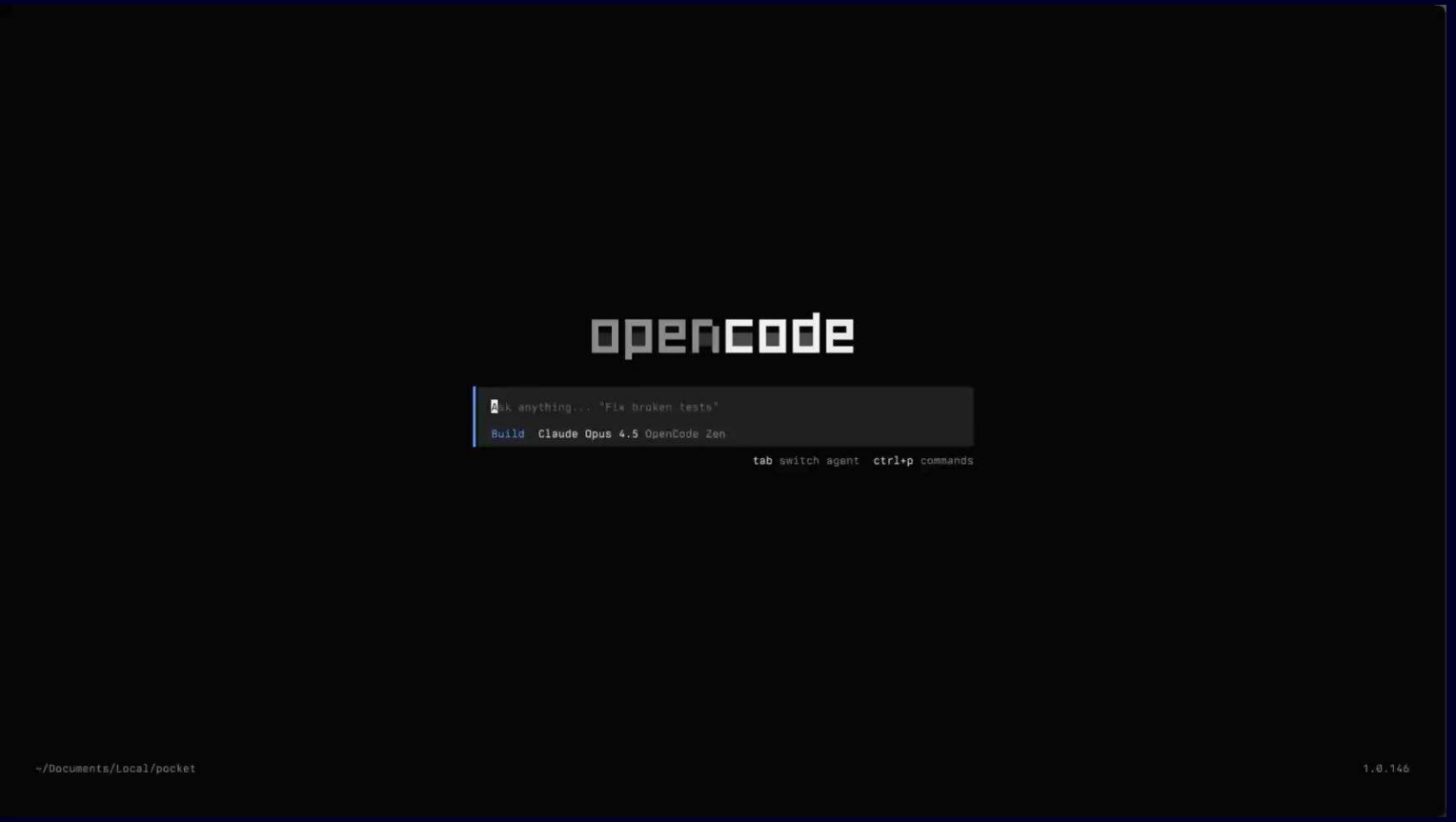
Agentic usage with OpenCode & OpenChamber



Unrestricted | © Siemens 2026 | code.siemens.com | 2026-10-29

SIEMENS

OpenCode

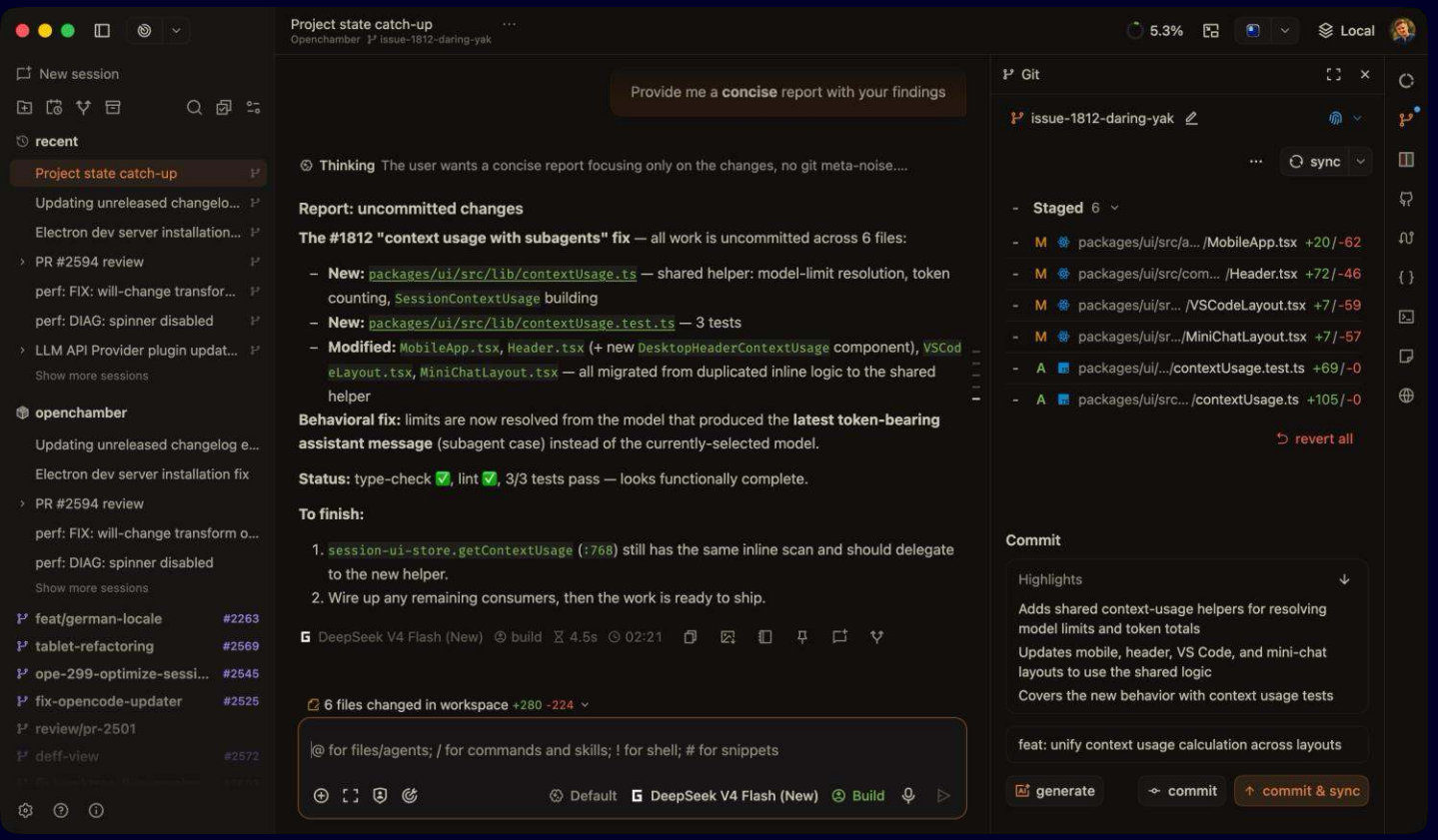


- An Open Source agent that helps write code in your terminal, IDE, or desktop
- Easily configured for usage with code.siemens.com LLMs.
- Source code on GitHub:

  https://github.com/anomalyco/opencode

SIEMENS

OpenChamber



- GUI for OpenCode AI Agent
- Easily configured for usage with code.siemens.com LLMs
- Uses the existing OpenCode config but offers even more features
  - Scheduled Tasks
  - Worktree Support
- Open Source on GitHub:

  https://github.com/openchamber/openchamber

Unrestricted | © Siemens 2026 | code.siemens.com | 2026-10-29

SIEMENS

code-agent

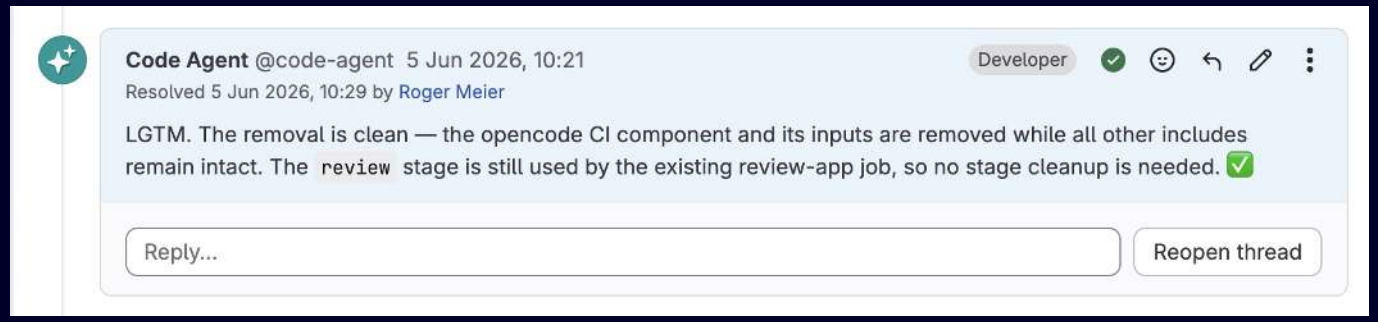
The new project we’re working on

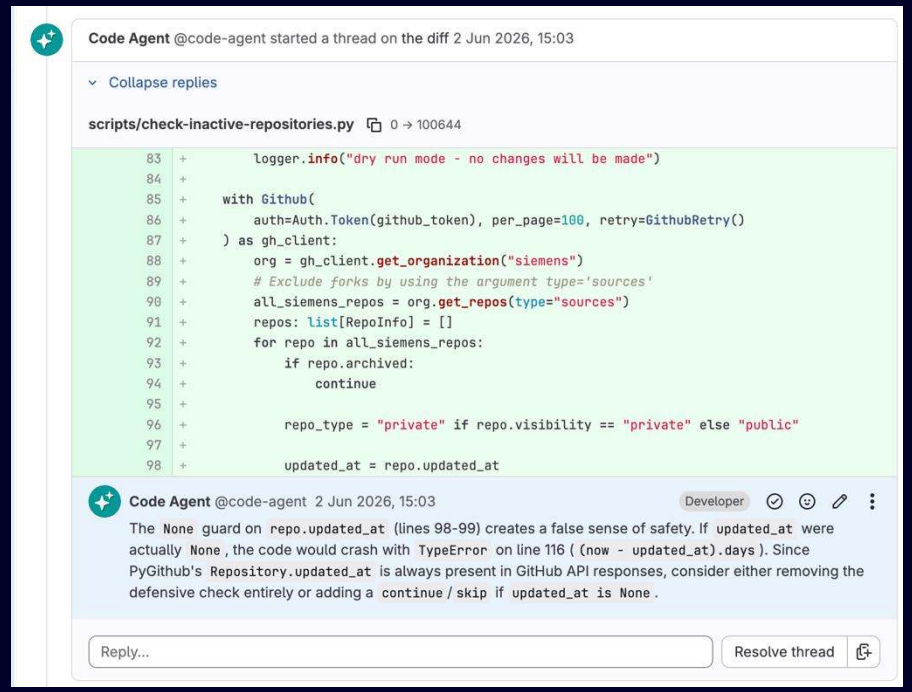


Unrestricted | © Siemens 2026 | code.siemens.com | 2026-10-29

SIEMENS

Code-agent - Merge request review

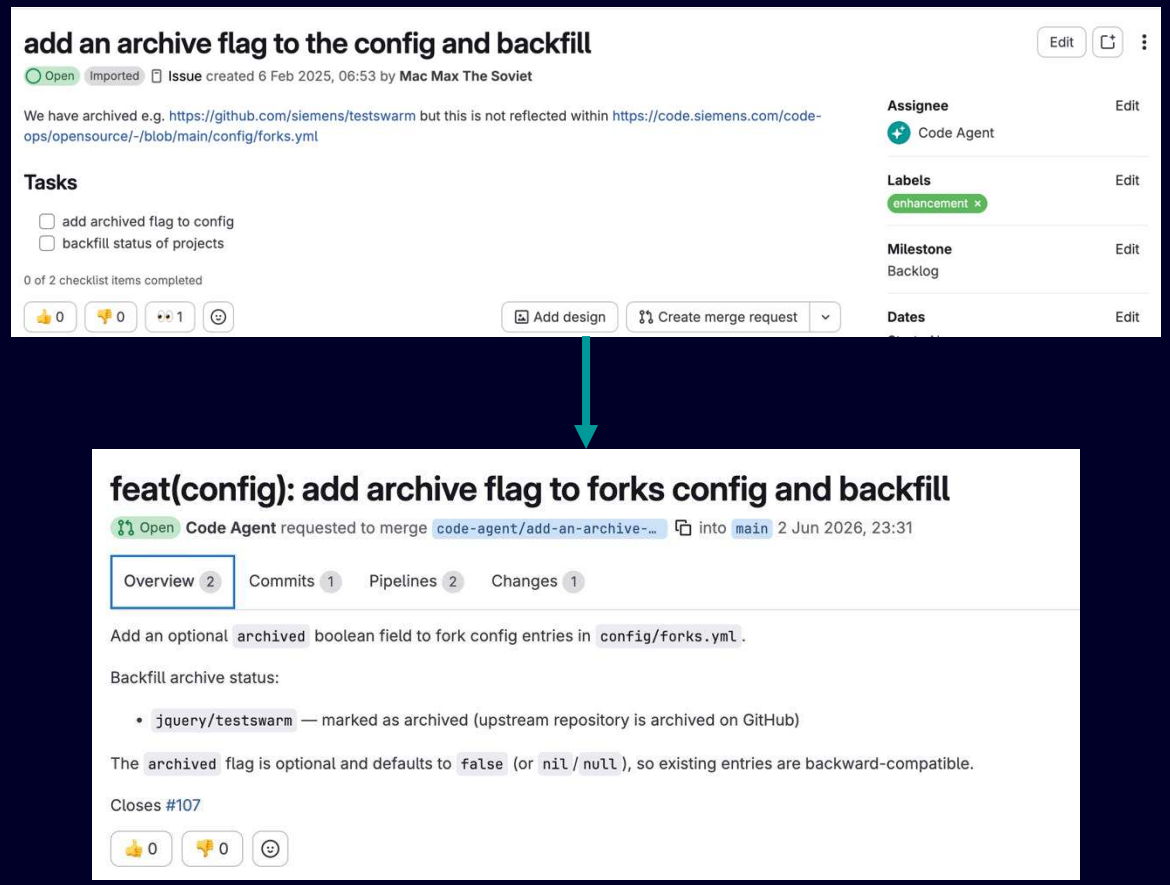




- GitLab bot
- **Reacts** when **@code-agent** is requested as reviewer or mentioned
- Polls GitLab todos and **starts matching tasks**
- Uses **OpenCode** as the AI review engine
- Runs reviews in **short-lived, isolated Kubernetes jobs**
- Creates a **temporary project-scoped GitLab** token for each run
- **Comments on merge requests** and can approve them when appropriate
- **Goal**: simple project setup, secure review automation, scalable execution

SIEMENS

Code-agent - Merge request created by issue



- **@code-agent** can also be assigned to work items
- Will then **create merge request** to address problem or feature request
- Later on, users will also be able to **iterate** inside the created merge requests **without leaving code.siemens.com**

SIEMENS

Code-agent - Outlook

Already implemented

- GitLab and GitHub reviews and issue implementation
- Follow-up changes on the agent's own merge requests
- Local, Kubernetes and GitLab CI runtime

Next in the backlog

- Contribute to existing merge requests
- Clearer run status and failure feedback
- Cross-project permissions

Open sourcing

- Make setup more generic
- Complete legal and IP clearing
- Setup community management

Unrestricted | © Siemens 2026 | code.siemens.com | 2026-10-29

Feedback



SIEMENS

Feedback from the community



Keep up the pace and your great work that you are doing. I really like it :)

UNLIKE REPLY SHARE •••

You, Wittig, Max (IT APS DEV COD) and Huser, Fabio (IT APS DEV COD) reacted to this

Huser, Fabio (IT APS DEV COD) I really like the usability of this, couldn't be any easier for us users. Keep up the great work!

Feedback

I tried the PoC with `continue` in VS Code according to https://code.siemens.io/general/ai/#large-language-model-llm-api-usage:

I used it on multiple Python projects we are developing so far:

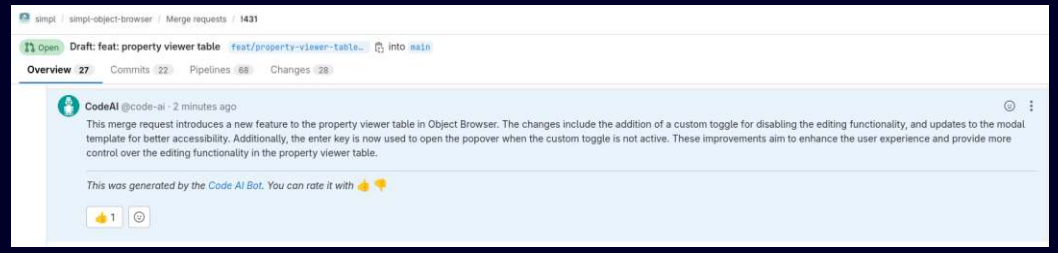
- Sometimes continue creates loops with endless new lines.
- Indexing on large projects (with 500+ files) seems to not work as expected (for smaller projects this works fine). I guess more specific configuration might help.
- When providing two or more files as context for writing doc strings, the results are quite promising and only little changes are required.
- Sometimes response speed is quite fast, but sometimes a quite big delay is visible.
- Useful for fixing typos and grammar
- Having access to a chatbot directly in VS Code is helpful and speeds up development.

Key takeaways so far:

- This technology can already be helpful in early stages. Setup was quite smooth but for big projects more instructions are required.
- Thanks for bringing this to life!
- This feedback was also supported by continue 😃

💚 5

Really love the work that you're doing here. I'd love to get involved on the Healthineers side of things.

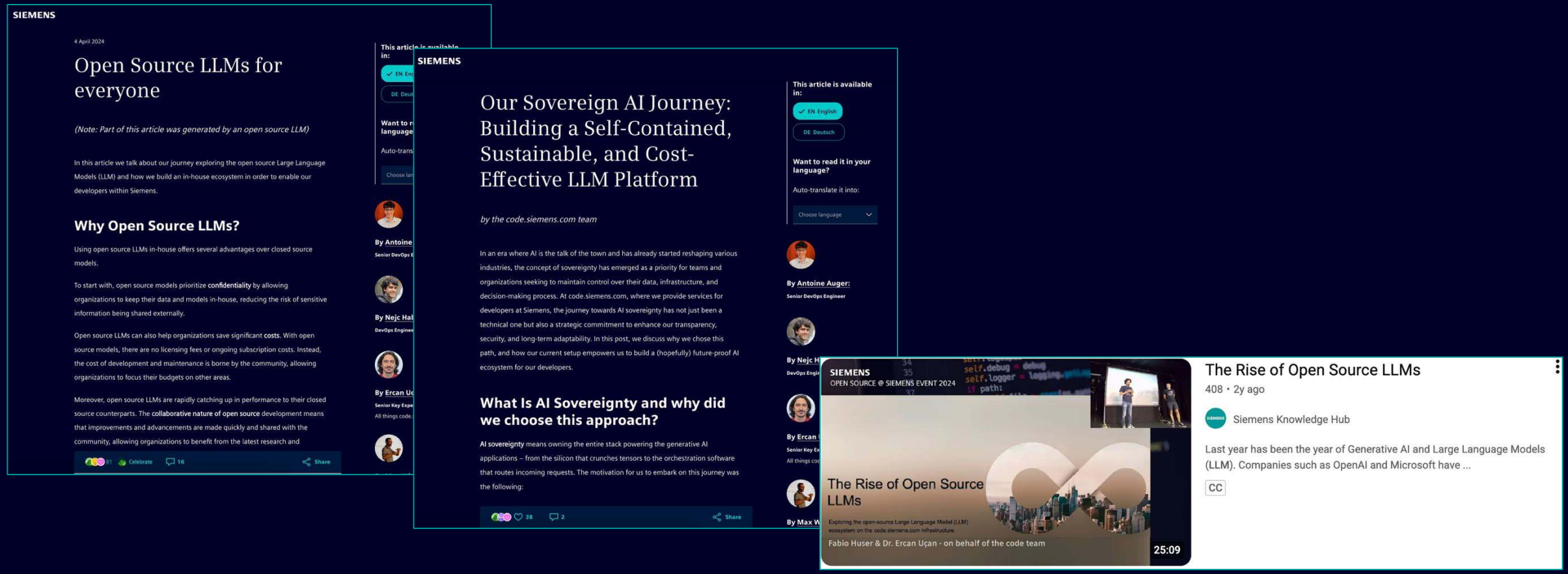


you guys rock! cool stuff

This is really nice work, we'll be trying it in a project pretty much immediately :) Thanks a lot to you all!

SIEMENS

Sharing our learnings



- https://blog.siemens.com/2024/04/open-source-llms-for-everyone/
- https://blog.siemens.com/2025/10/our-sovereign-ai-journey-building-a-self-contained-sustainable-and-cost-effective-llm-platform/
- https://www.youtube.com/watch?v=Sn3tC5N_hnc

SIEMENS

Contact

opensource.siemens.com

- **Max Wittig:** max.wittig@siemens.com
  https://www.linkedin.com/in/max-wittig
  WeChat: max-wittig

- **Fabio Huser:** fabio.huser@siemens.com
  https://www.linkedin.com/in/fabio-huser
  WeChat: fh1-ch



SIEMENS

Open Source 西门子

Beijing · 2026

Oct. 29
Industrial Open Source

Oct. 30
Open Source Robotics

Agenda