

Siemens Internal LLM APIs

YANG Shaopeng on behalf of the code.siemens team



SIEMENS

The Rise of Open Source LLMs

Exploring the open-source Large Language Model (LLM) ecosystem on the code.siemens.com infrastructure

Fabio Huser & Dr. Ercan Uçan - on behalf of the code team

Who we are

Quick overview on code.siemens.com,
serving thousands of developers,
enabling and building communities



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 CE)

- **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. IIm
- **Open & Inner Source** management
- Inner Source **Tax Transfers**, Export Control
- docs-theme (1900 re-uses)

- **High confidentiality** rating
- **> 70k users, 79 countries**
- **> 300k builds / day**
- Pure Open Source

Ongoing Topics

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

Why?



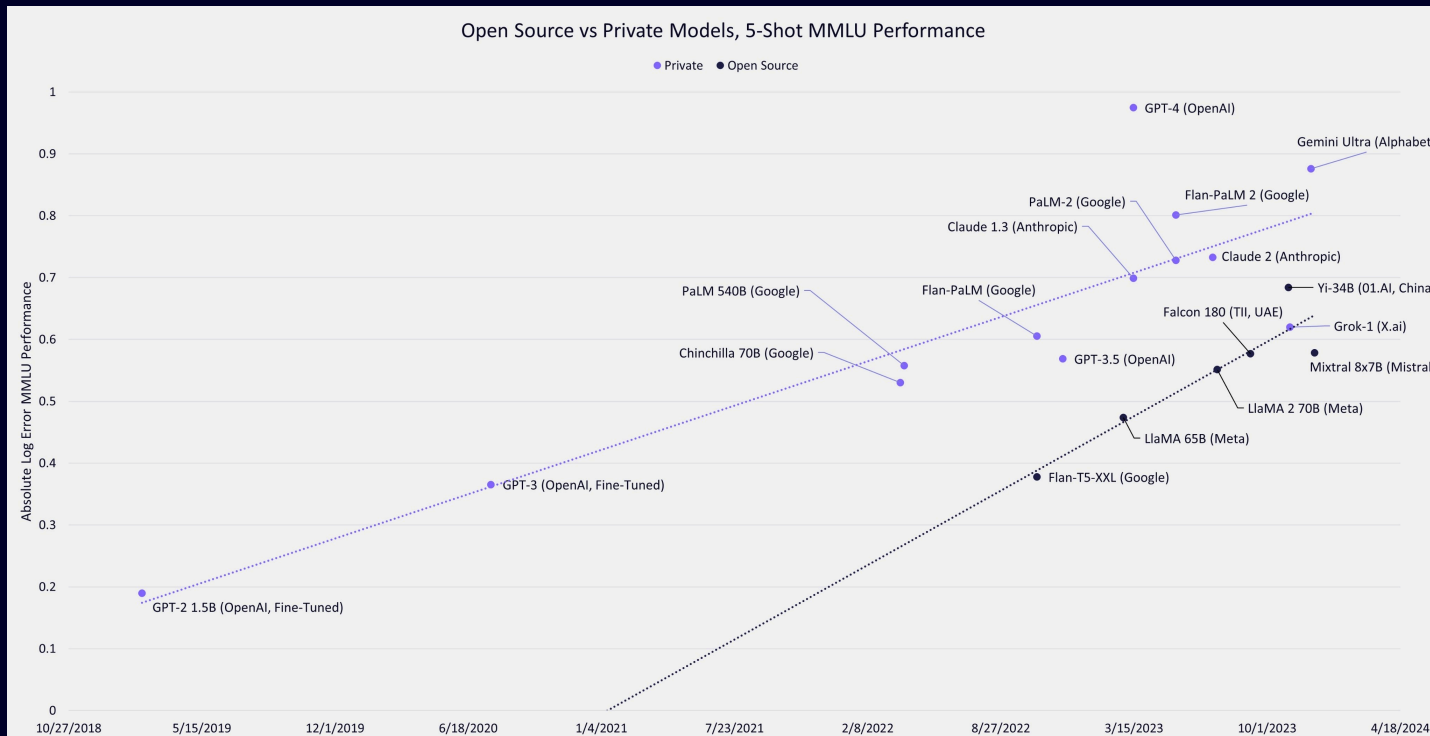
Unrestricted | © Siemens 2024 | code.siemens.com | 2024-05-14

SIEMENS

Advantages of open source 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 source LLMs are **catching up**.

Open Source LLMs are catching up



Source: Cathy Wood, ARK Invest <https://twitter.com/CathieDWood/status/1735136182651027952>

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

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

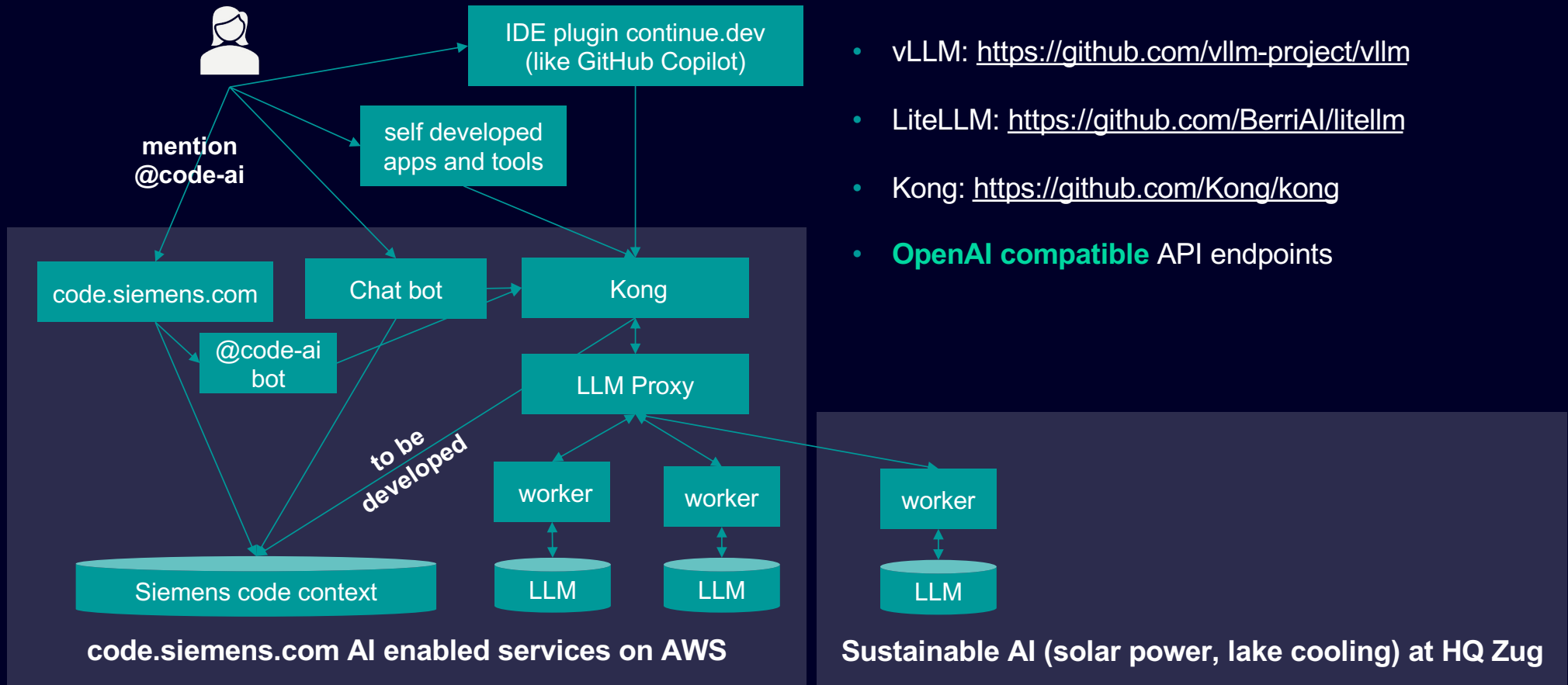
March 7th, 2024

The hyper-scalers are nowadays also offering Open Source models such as Mistral.

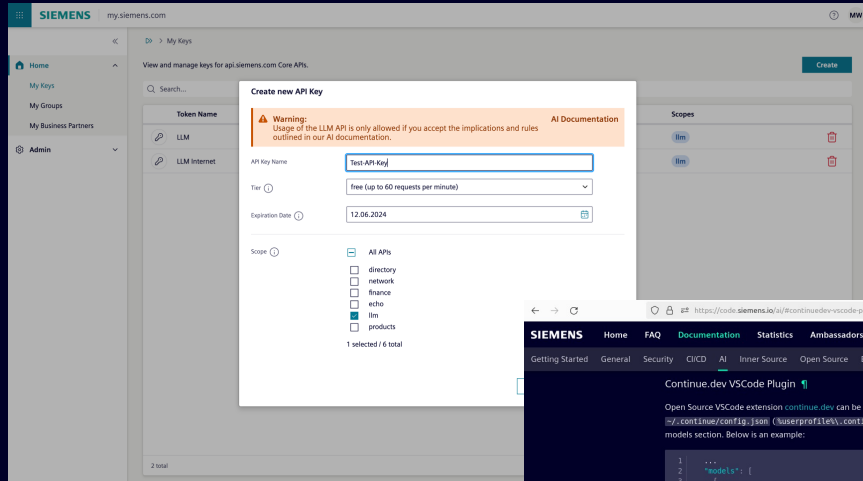


In-house LLM APIs

Current setup - all based on Open Source and Inner Source



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

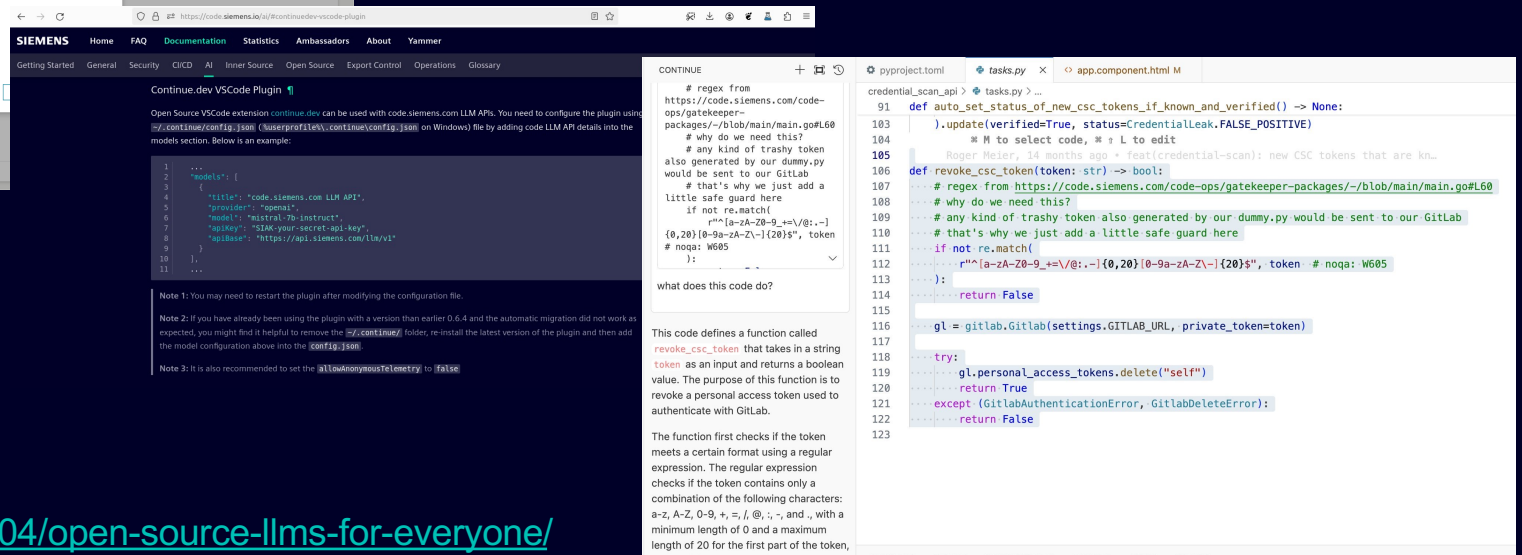


- Users get started in less than 5 minutes.
- All data stays within Siemens.
- Open Source IDE plugin continue.dev (Apache-2.0).

Next

- Tab completion.
- Siemens context provider within IDE.

<https://blog.siemens.com/2024/04/open-source-llms-for-everyone/>





Usage via VSCode IDE integration (continue.dev)

Continue.dev – open-source Visual Studio Code (VS Code) integration

Say hello to Open Source at Siemens 2024 attendees!

Hello, and welcome to Siemens Open Source event in 2024! I'm an assistant designed to help answer any questions you might have about open source at Siemens. I'm excited to be here and look forward to engaging with you all about the latest developments, projects, and initiatives in the open source community at Siemens. Let me know if you have any specific topics or questions you'd like to discuss. Let's make this event an enriching and productive experience for all!

Ask a follow-up

+ Add Context

New Session (⌘ L)

code.siemens.com LLM API

```
Continue Quick Edit
true, if the current time is between 9h and 17h and the weekday is between Monday and Friday
Press 'Enter' to confirm your input or 'Escape' to cancel

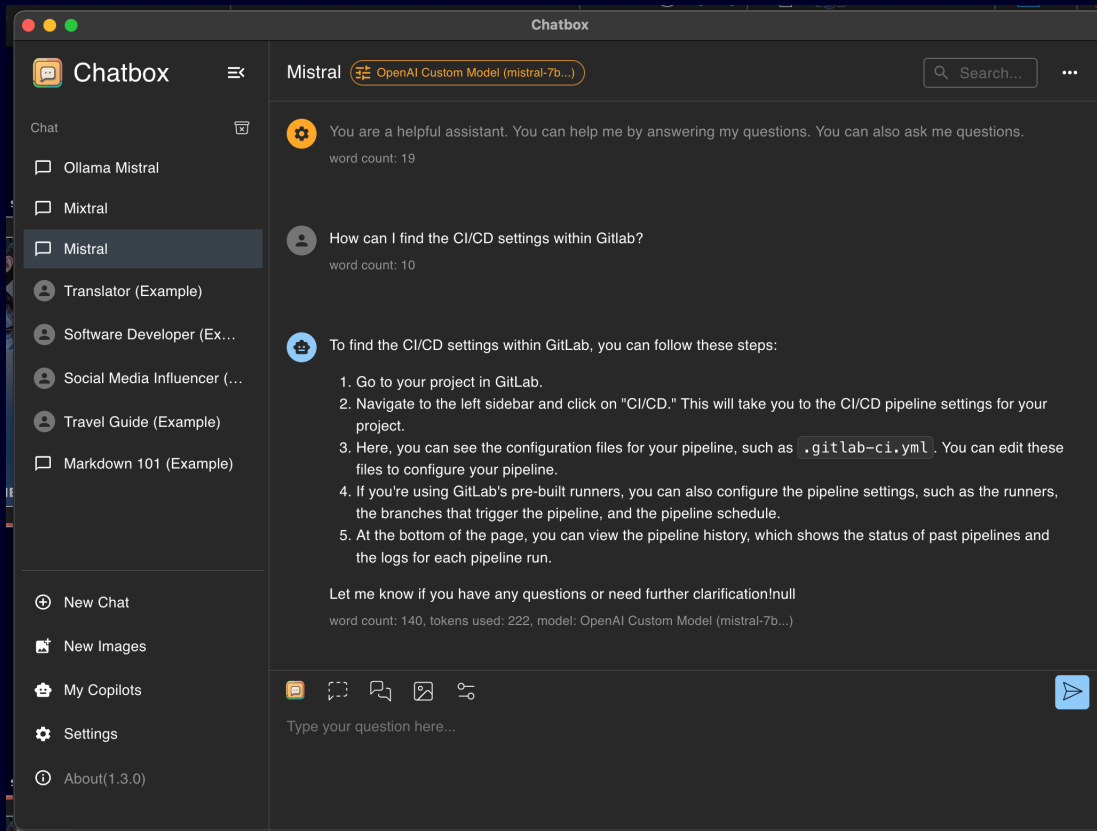
llm > chain.py > ...
47 def get_chat_memory(messages: list[dict[str, Any]]) -> ConversationBufferMemory:
48     chat_input =
49     chat_output = ''
61
62
63     return memory
64
65
66 You, 3 seconds ago * Uncommitted changes
67
68
69 def build_chain(messages: list[dict[str, Any]]) -> RunnableSerializable:
70     llm_app = cast(LlmConfig, apps.get_app_config("llm"))
71     if llm_app.primary_llm is None and llm_app.secondary_llm is None:
72         raise LlmBackendError("Missing LLM backend")
```

```
65 Accept ✓ | Reject ✗ | (⌘↔/⌘↔ to accept/reject all, ⌘↔L to retry)
66
67 def is_worktime() -> bool:
68     today_date = datetime.datetime.now().date()
69     today_weekday = today_date.weekday()
70     if today_weekday not in range(1, 5):
71         return False
72     current_time = datetime.time(hour=datetime.datetime.now().hour, minute=datetime.datetime.now().minute)
73     return current_time.hour >= 9 and current_time.hour <= 17
74 You, 1 second ago * Uncommitted changes
75
```



Usage via local UI (Chatbox)

Chatbox



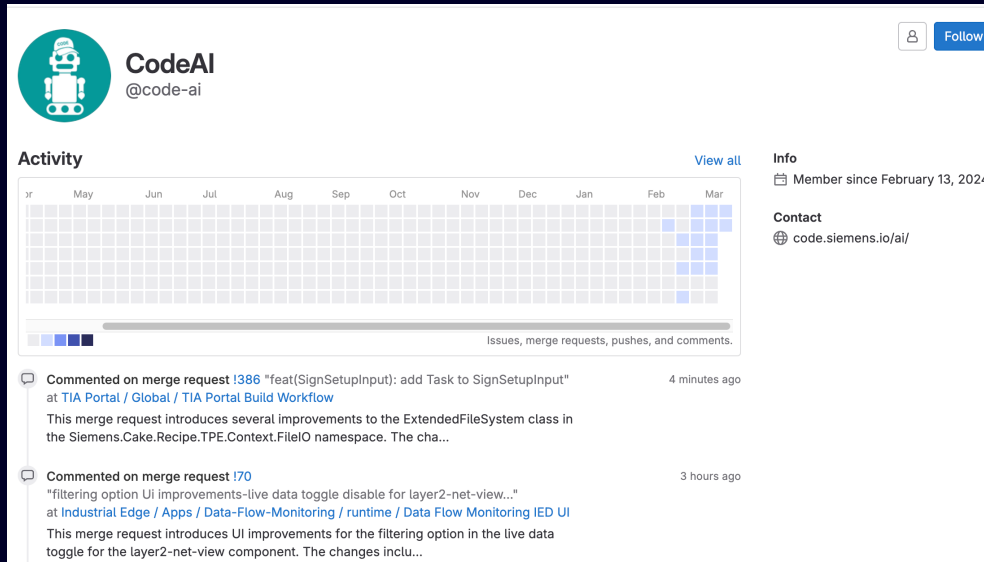
- Easily configured for usage with `code.siemens.com` LLM offerings.
- You can select models and personas (i.e., profiles).
- Open Source on GitHub:

<https://github.com/Bin-Huang/chatbox>

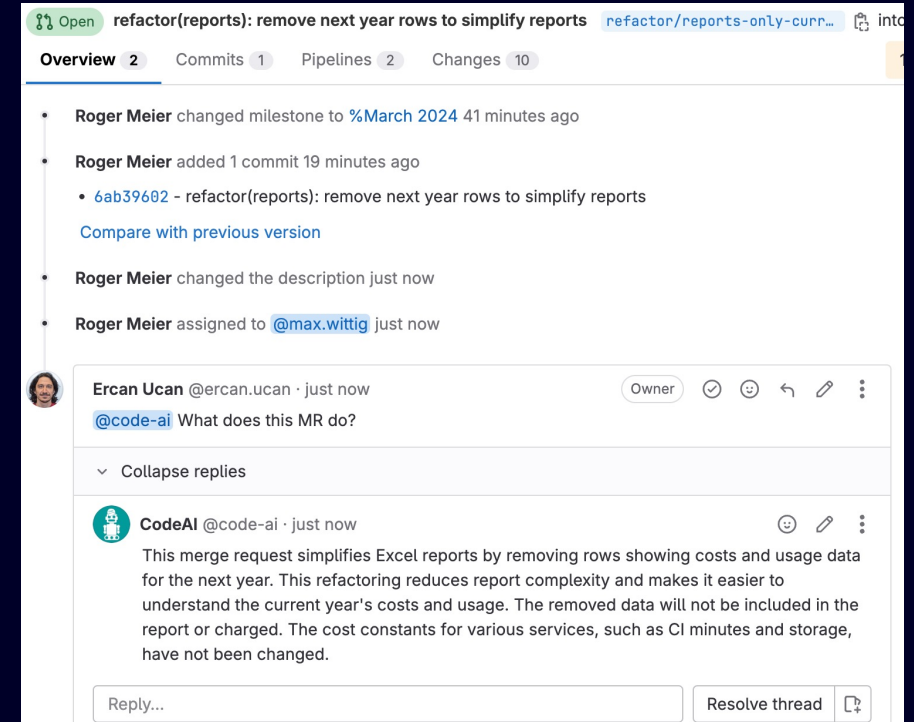


Usage with code-ai bot

Code-AI bot



The screenshot shows the GitHub profile for CodeAI (@code-ai). The profile includes a robot icon, the name 'CodeAI', and the handle '@code-ai'. Below the profile is an 'Activity' section with a calendar view showing activity from May to March. Two recent comments are visible: one on merge request !386 about adding a task to SignSetupInput, and another on merge request !70 about filtering options in the live data toggle.



The screenshot shows a GitHub pull request discussion for the merge request 'refactor(reports): remove next year rows to simplify reports'. The discussion includes a list of actions by Roger Meier, such as changing milestones and adding commits. A comment by Ercan Ucan asks '@code-ai What does this MR do?'. The CodeAI bot responds with a detailed explanation: 'This merge request simplifies Excel reports by removing rows showing costs and usage data for the next year. This refactoring reduces report complexity and makes it easier to understand the current year's costs and usage. The removed data will not be included in the report or charged. The cost constants for various services, such as CI minutes and storage, have not been changed.'

Current Functionality:

- **summary:** summarize changes in Merge Requests.
- **ci-failure:** identifies root cause for failed pipeline.
- **clean-code:** clean code suggestions (community contribution).
- **discussion-summary:** summarize the discussions in the issue or merge request.

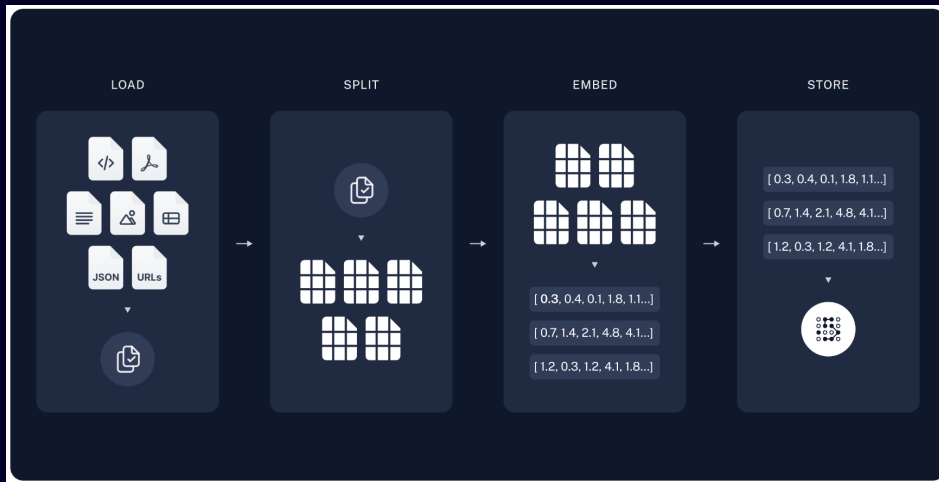
Next Topics: Unit test creation, Secure coding analysis.



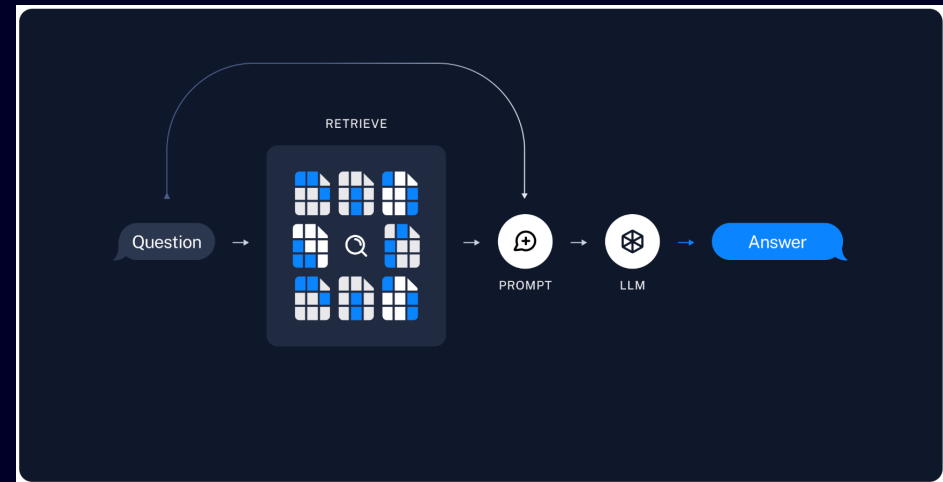
Usage via RAG with context

RAG: Retrieval Augment Generation with context (code.siemens.com)

Ingest



Retrieve



LangChain: https://python.langchain.com/docs/use_cases/question_answering/

- Ingested with code.siemens.com documentation, various inner source project code samples.

RAG: Chat bot with Siemens context

SIEMENS Code-Apps

Welcome to Code-Apps, Ercan
Get insights into all your code.siemens.com projects.

give me a code example using si-wizard

Here's an example of using the `si-wizard` component from Siemens Simple Icons Library in an Angular project:

```
import { Component } from '@angular/core';

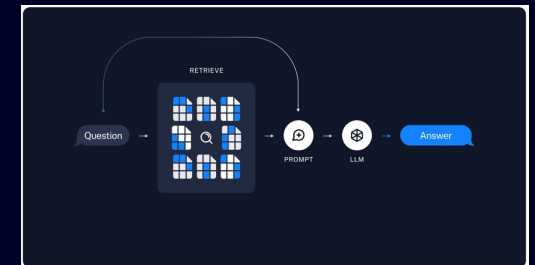
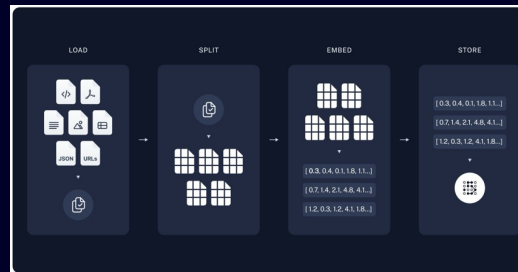
@Component({
  selector: 'app-wizard-example',
  template: `
    <si-wizard>
      <si-wizard-step heading="Step 1" (next)="logEvent('Step 1 finished!')">
        <si-empty-state icon="element-occupied" heading="Occupied" />
      </si-wizard-step>
      <si-wizard-step
        heading="Step 2"
        (next)="logEvent('Step 2 finished!')"
        (back)="logEvent('Back to Step 1!')"
      >
        <si-empty-state icon="element-home" heading="Home" />
      </si-wizard-step>
      <!-- Add more steps as needed -->
    </si-wizard>
  `
})
export class WizardExampleComponent {
  logEvent(event: string) {
    console.log(event);
  }
}
```

This example defines a component called `WizardExampleComponent` that uses the `si-wizard` and `si-wizard-step` components. Each `si-wizard-step` has a heading and event handlers for the `next` and `back` buttons. In this example, the `logEvent` method is used to log a message to the console when a step is finished. You can replace this method with your own logic as needed.

Send a message (e.g., 'ci with nodejs', 'si-wizard code example')

Code-Apps is the supplementary service for all code.siemens.com related topics.

Retrieval Augmented Generation (RAG)



Initial implementation in a hackathon

Now in production and within testing phase

Upcoming features:

- Projects on code.siemens.com use it with their context.
- Selection of context based on permissions.

Legal Aspects

Copyright, The Stack



How can we avoid code from a dataset ending up within our code?

The Stack
6 TB of permissive code data

Dataset Collection
GH Archive → Query → GH clone → Row dataset (127 Ml repos, 82.8 Mln, 103 TB of data) → near-deduplication → License filtering → 69 TB of data

Licensing + Governance
Row dataset: MIT, Apache 2.0, BSD-3-Clause, etc.
Permissive: MIT, Apache 2.0, BSD-3-Clause, etc.
Opt-out: If users would like to exclude their code from the corpus, we have an opt-out mechanism. Visit: <https://www.bigcode-project.org/about/about-the-stack/>

Permissive license distribution of licenses used to filter the dataset

MIT (2.7%)	Apache-2.0 (18.7%)	BSD-3-Clause (5.9%)	Unlicense (2.0%)
CC-0 (1.5%)	BSD-2-Clause (1.2%)	CC-BY-4.0 (1.1%)	CC-BY-3.0 (0.7%)
0BSD (0.4%)	MIT-0 (0.3%)	MIT (0.2%)	Other (1.6%) (2.2%)

Programming Languages
Bar chart showing lines of code for various languages: Python, JavaScript, Java, C++, etc.

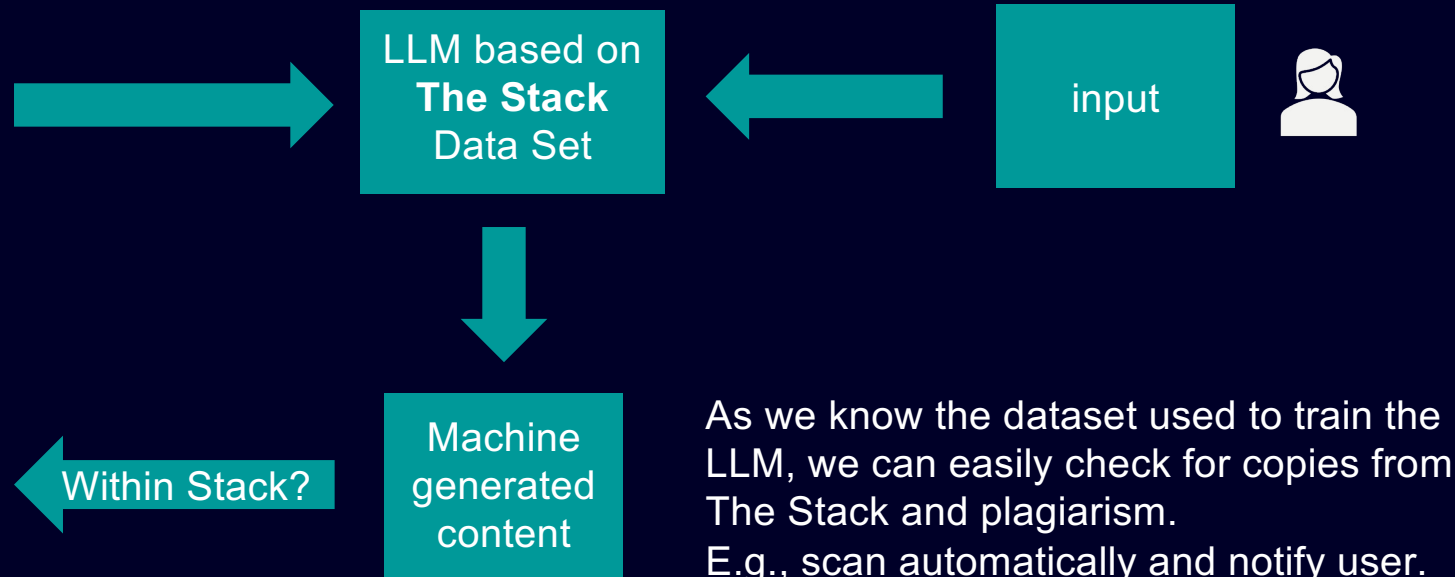
Evaluation
We trained several GPT-2 models (50M parameters) on different parts of the dataset both with and without near-deduplication. The models trained on the Python subset of The Stack performed on par with CodeX and CodeGen of similar size when using near-deduplication.

Dataset	Filtering	pass@1	pass@10	pass@100
CodeX (50M)	unknown	13.17	20.17	32.29
CodeGen (50M)	unknown	12.76	23.11	35.19
Python-all-license	None	13.11	21.77	34.27
	Near-dedup	17.34	27.44	43.52
Python-permissive-license	None	10.97	16.64	27.29
	Near-dedup	12.89	22.24	34.91

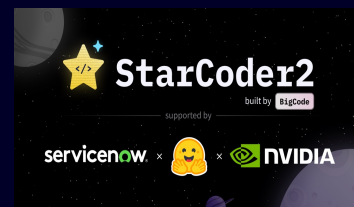
Usage
pip install dataset
from datasets import load_dataset
Full dataset (17B of data)
ds = load_dataset("bigcode/the-stack")
Python dataset (1.8B of data)
ds = load_dataset("bigcode/the-stack", "python")
Python dataset (1.8B of data)
ds = load_dataset("bigcode/the-stack", "python", split="train")
Python dataset (1.8B of data)
ds = load_dataset("bigcode/the-stack", "python", split="train", streaming=True, batch_size=1024)

Dataset Trivia
• There are 1,407,018 files containing "hello world" in the dataset.
• The message "four men ate the database" (in German) appears in a file.
• It would take a single person 78,000 years to type all of the code from scratch.
• The Stack is 20x larger than the English Wikipedia.
Printed codes sized on A4 paper the stacked pile would amount more like the Mount Everest and you could cover the distance between TBx with said moon if the paper is aligned side-by-side.

The Stack v2 (67.5 TB) was released February 28th



As we know the dataset used to train the LLM, we can easily check for copies from The Stack and plagiarism. E.g., scan automatically and notify user.



StarCoder2 is based on The Stack v2

Sustainable AI

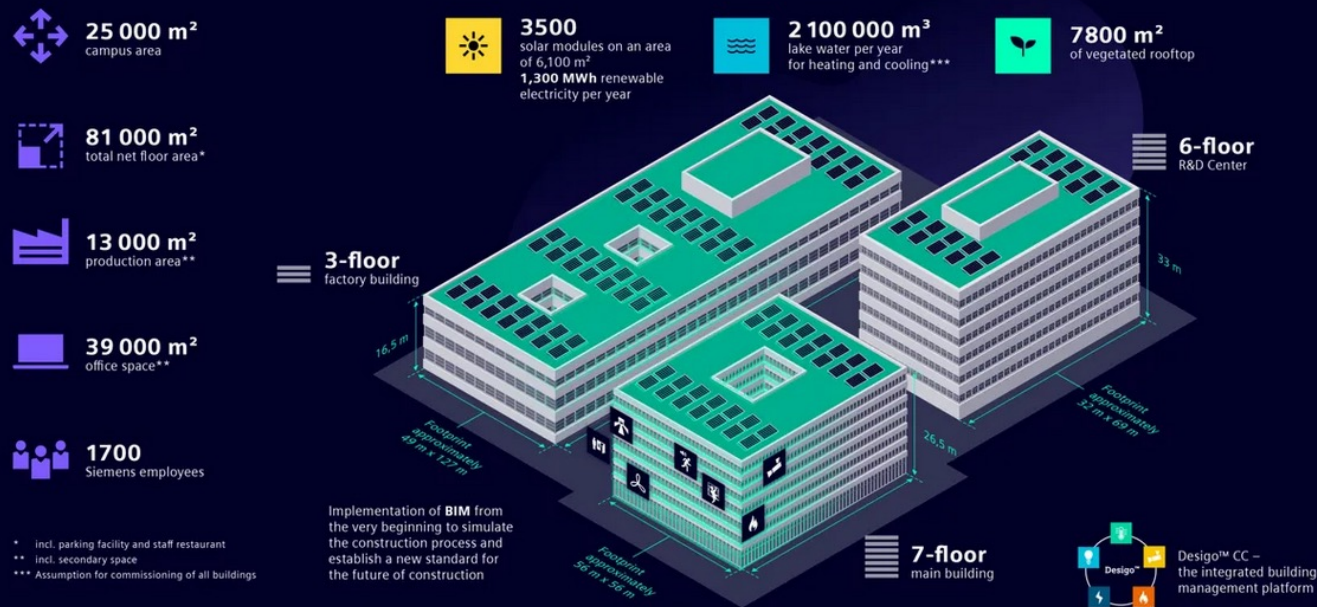
An abstract graphic on the right side of the slide. It features a large, light blue circular area with a radial, fibrous texture. Overlaid on this is a thick, dark blue ring. The background of the entire slide is a dark blue gradient.

On-prem Hardware
(Future work)

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**

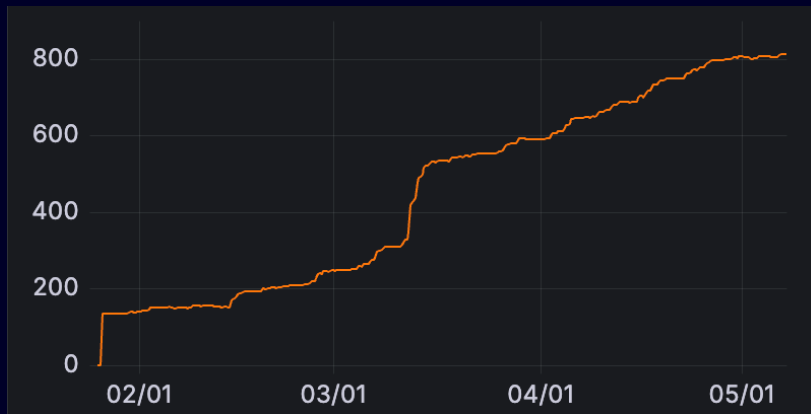
Facts & Figures about the climate-neutral Campus Zug



Server room within the R&D center

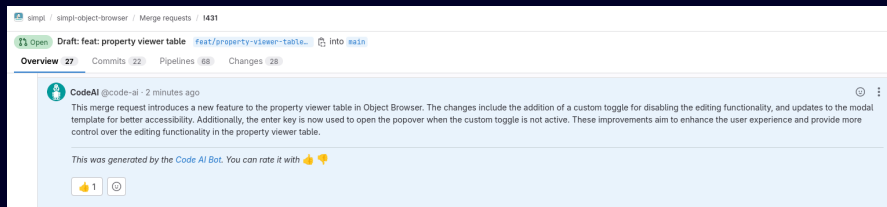
“Companies with serious plans for AI have the GPUs and the knowledge in-house.”

Feedback from the community



Great work colleagues. Leading the way for open source adoption even in this new and niche field.

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



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 🗨️

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!

Contact

- opensource.siemens.com
- Fabio Huser: fabio.huser@siemens.com
- Dr. Ercan Uçan: ercan.uçan@siemens.com

