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

SIEMENS

Who we are

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

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 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. llm
- **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)

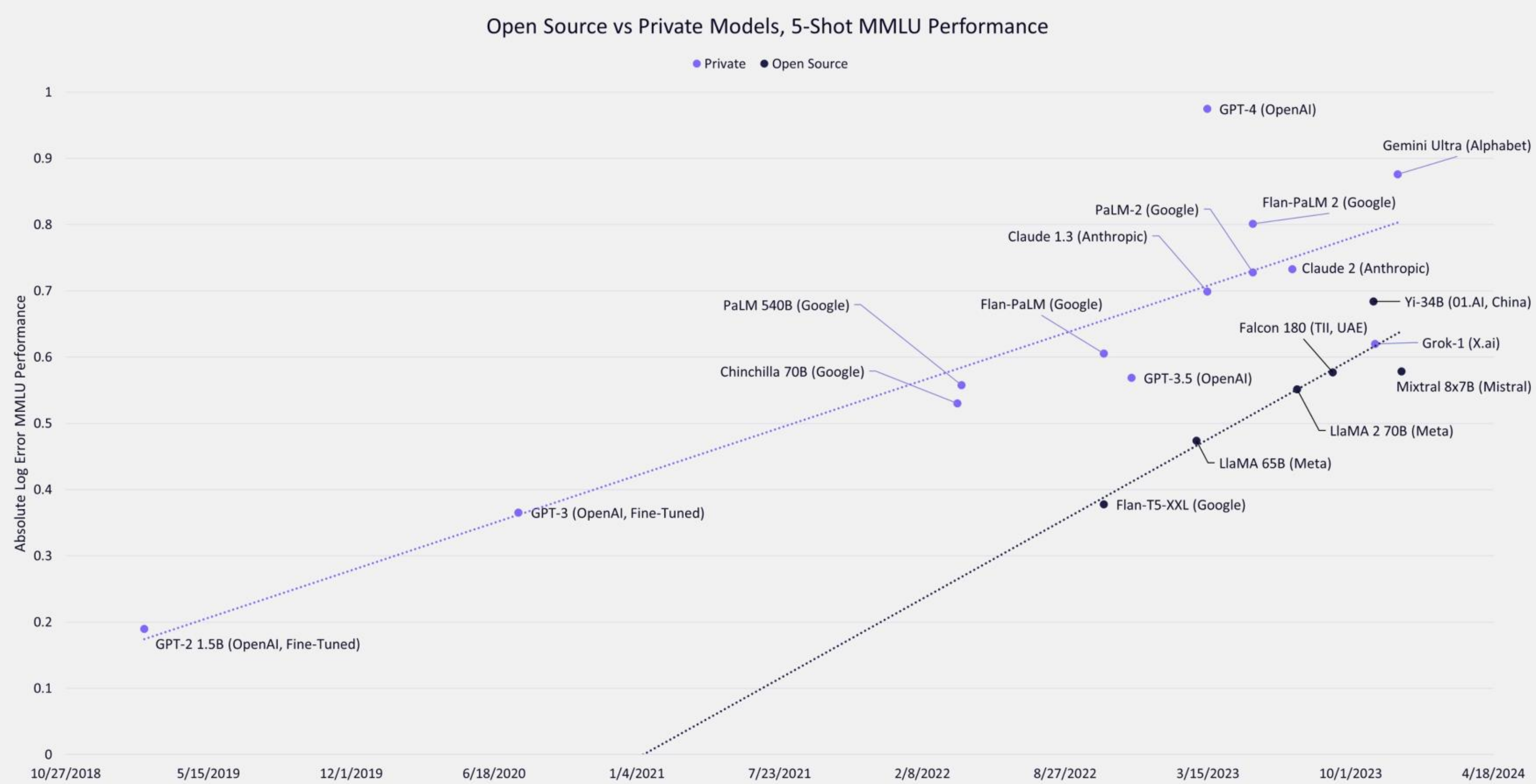
SIEMENS

Why?

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

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

March 7th, 2024

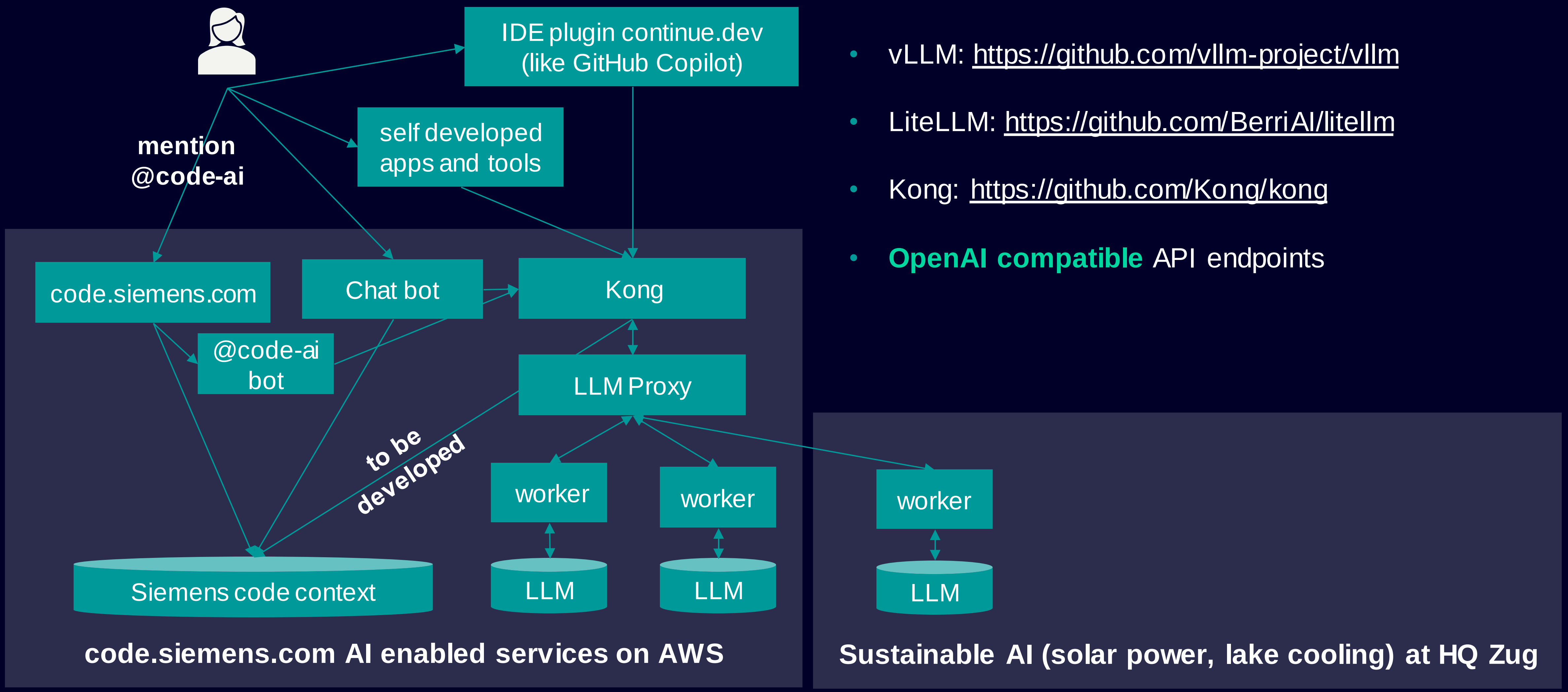
The hyper-scalers are
nowadays also offering
Open Source models such as
Mistral.

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

In-house LLM APIs

SIEMENS

Current setup - all based on Open Source and Inner Source

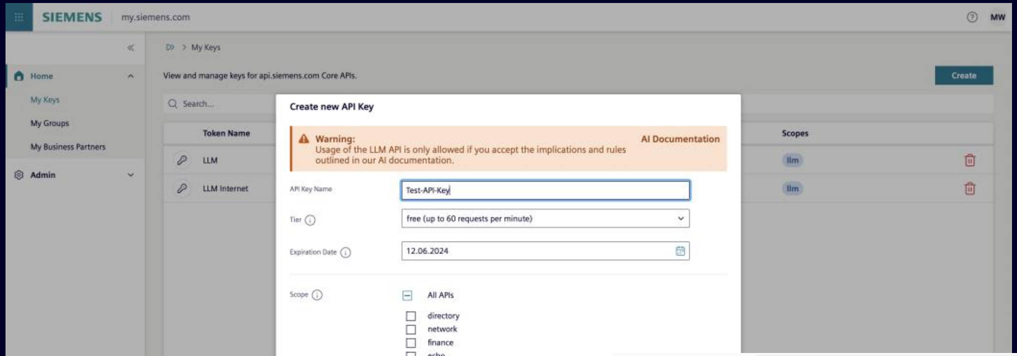

IDE plugin continue.dev (like GitHub Copilot)
self developed apps and tools
mention @code-ai
code.siemens.com
Chat bot
Kong
@code-ai bot
LLM Proxy
to be developed
worker
worker
worker
Siemens code context
LLM
LLM
LLM
code.siemens.com AI enabled services on AWS
Sustainable AI (solar power, lake cooling) at HQ Zug

- vLLM: https://github.com/vllm-project/vllm
- LiteLLM: https://github.com/BerriAI/litellm
- Kong: https://github.com/Kong/kong
- **OpenAI compatible** API endpoints

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).**



View and manage keys for api.siemens.com Core APIs.

Create new API Key

Warning: Usage of the LLM API is only allowed if you accept the implications and rules outlined in our AI documentation. **AI Documentation**

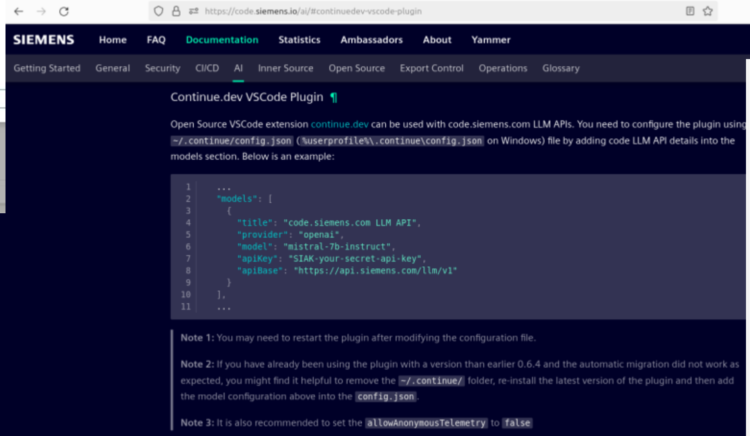
API Key Name: Test-API-Key

Tier: free (up to 60 requests per minute)

Expiration Date: 12.06.2024

Scope:
☐ All APIs
☐ directory
☐ network
☐ finance
☐ echo
☑ llm
☐ products

1 selected / 6 total

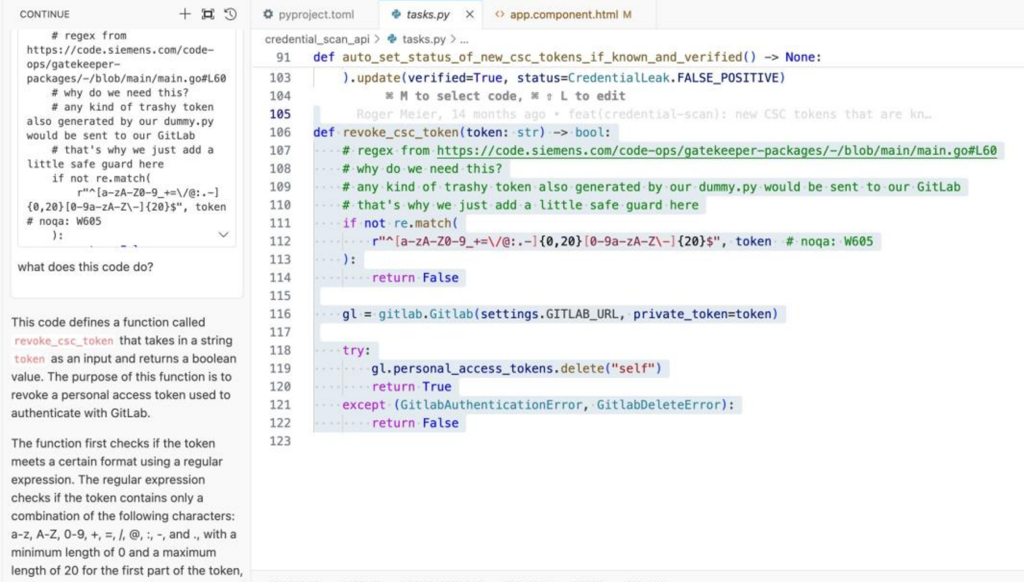


Continue.dev VSCode Plugin ¶

Open Source VSCode extension continue.dev can be used with code.siemens.com LLM APIs. You need to configure the plugin using `~/.continue/config.json` (`%userprofile%\.continue\config.json` on Windows) file by adding code LLM API details into the models section. Below is an example:

```
...
"models": [
  {
    "title": "code.siemens.com LLM API",
    "provider": "openai",
    "model": "mistral-7b-instruct",
    "apiKey": "SIAK-your-secret-api-key",
    "apiBase": "https://api.siemens.com/llm/v1"
  }
],
...
```

> **Note 1:** You may need to restart the plugin after modifying the configuration file.
>
> **Note 2:** If you have already been using the plugin with a version than earlier 0.6.4 and the automatic migration did not work as expected, you might find it helpful to remove the `~/.continue/` folder, re-install the latest version of the plugin and then add the model configuration above into the `config.json`.
>
> **Note 3:** It is also recommended to set the `allowAnonymousTelemetry` to `false`



```
    # regex from
https://code.siemens.com/code-
ops/gatekeeper-
packages/-/blob/main/main.go#L60
    # why do we need this?
    # any kind of trashy token
also generated by our dummy.py
would be sent to our GitLab
    # that's why we just add a
little safe guard here
    if not re.match(
        r"^[a-zA-Z0-9_+=\/@:.-]
{0,20}[0-9a-zA-Z\-]{20}$", token
# noqa: W605
    ):
```

what does this code do?

This code defines a function called `revoke_csc_token` that takes in a string `token` as an input and returns a boolean value. The purpose of this function is to revoke a personal access token used to authenticate with GitLab.

The function first checks if the token meets a certain format using a regular expression. The regular expression checks if the token contains only a combination of the following characters: a-z, A-Z, 0-9, +, =, /, @, :, -, and ., with a minimum length of 0 and a maximum length of 20 for the first part of the token,

```
def auto_set_status_of_new_csc_tokens_if_known_and_verified() -> None:
    ).update(verified=True, status=CredentialLeak.FALSE_POSITIVE)
          ⌘ M to select code, ⌘ ⇧ L to edit
      Roger Meier, 14 months ago • feat(credential-scan): new CSC tokens that are kn…
def revoke_csc_token(token: str) -> bool:
    # regex from https://code.siemens.com/code-ops/gatekeeper-packages/-/blob/main/main.go#L60
    # why do we need this?
    # any kind of trashy token also generated by our dummy.py would be sent to our GitLab
    # that's why we just add a little safe guard here
    if not re.match(
        r"^[a-zA-Z0-9_+=\/@:.-]{0,20}[0-9a-zA-Z\-]{20}$", token  # noqa: W605
    ):
        return False

    gl = gitlab.Gitlab(settings.GITLAB_URL, private_token=token)

    try:
        gl.personal_access_tokens.delete("self")
        return True
    except (GitlabAuthenticationError, GitlabDeleteError):
        return False

```

Next

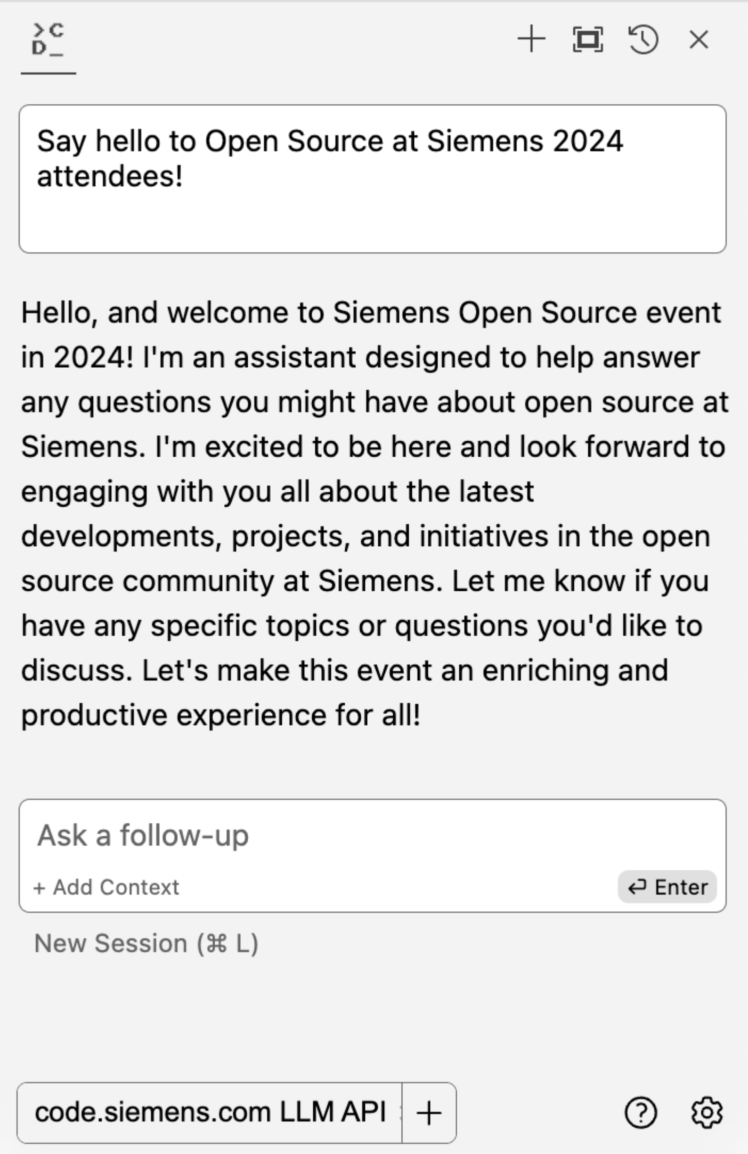
- **Tab completion.**
- **Siemens context provider within IDE.**

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

Usage via VSCode IDE integration (continue.dev)

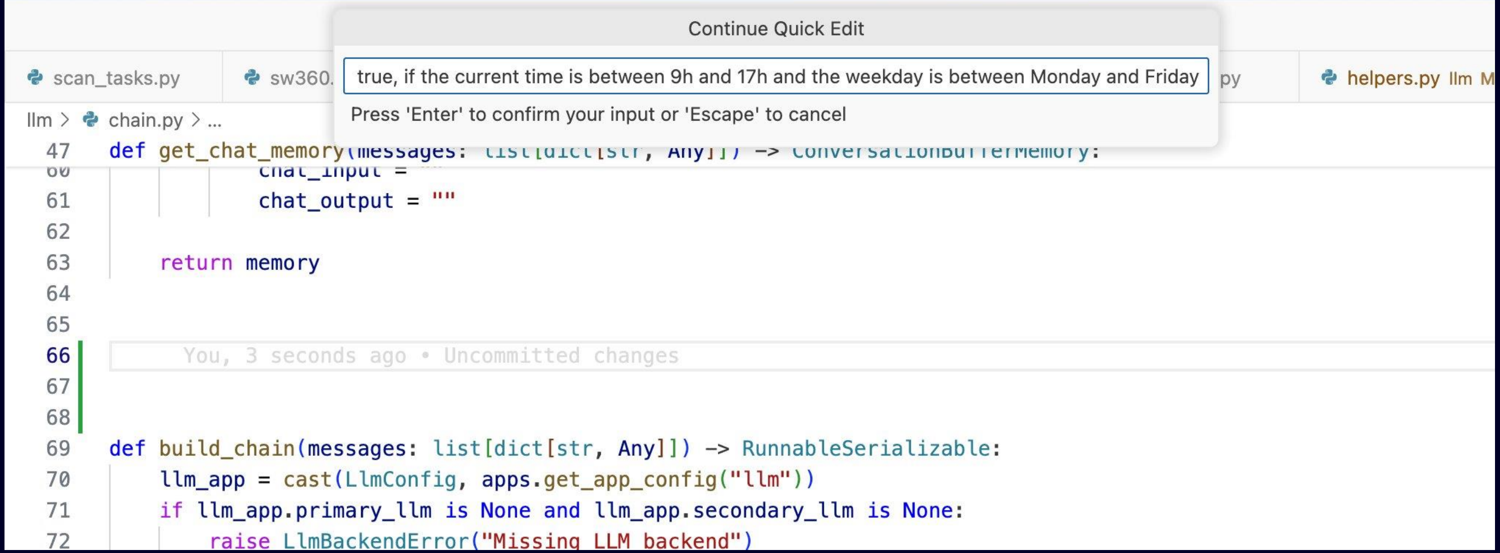
SIEMENS

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



Say hello to Open Source at Siemens 2024 attendees!

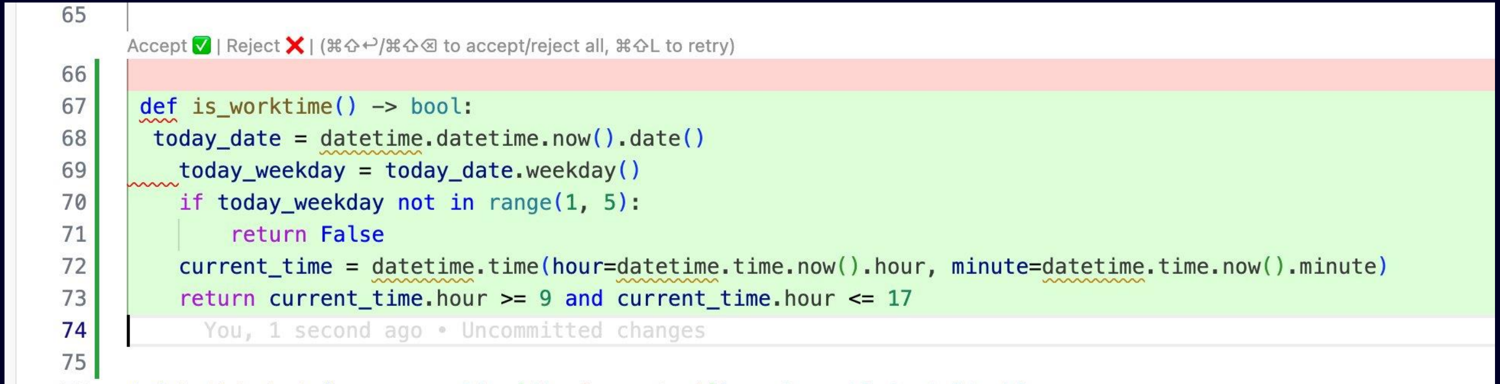
Hello, and welcome to Siemens Open Source event in 2024! I'm an assistant designed to help answer any questions you might have about open source at Siemens. I'm excited to be here and look forward to engaging with you all about the latest developments, projects, and initiatives in the open source community 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!



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

```python
def get_chat_memory(messages: list[dict[str, Any]]) -> ConversationBufferMemory:
            chat_input = ""
            chat_output = ""

    return memory


      You, 3 seconds ago • Uncommitted changes


def build_chain(messages: list[dict[str, Any]]) -> RunnableSerializable:
    llm_app = cast(LlmConfig, apps.get_app_config("llm"))
    if llm_app.primary_llm is None and llm_app.secondary_llm is None:
        raise LlmBackendError("Missing LLM backend")
```

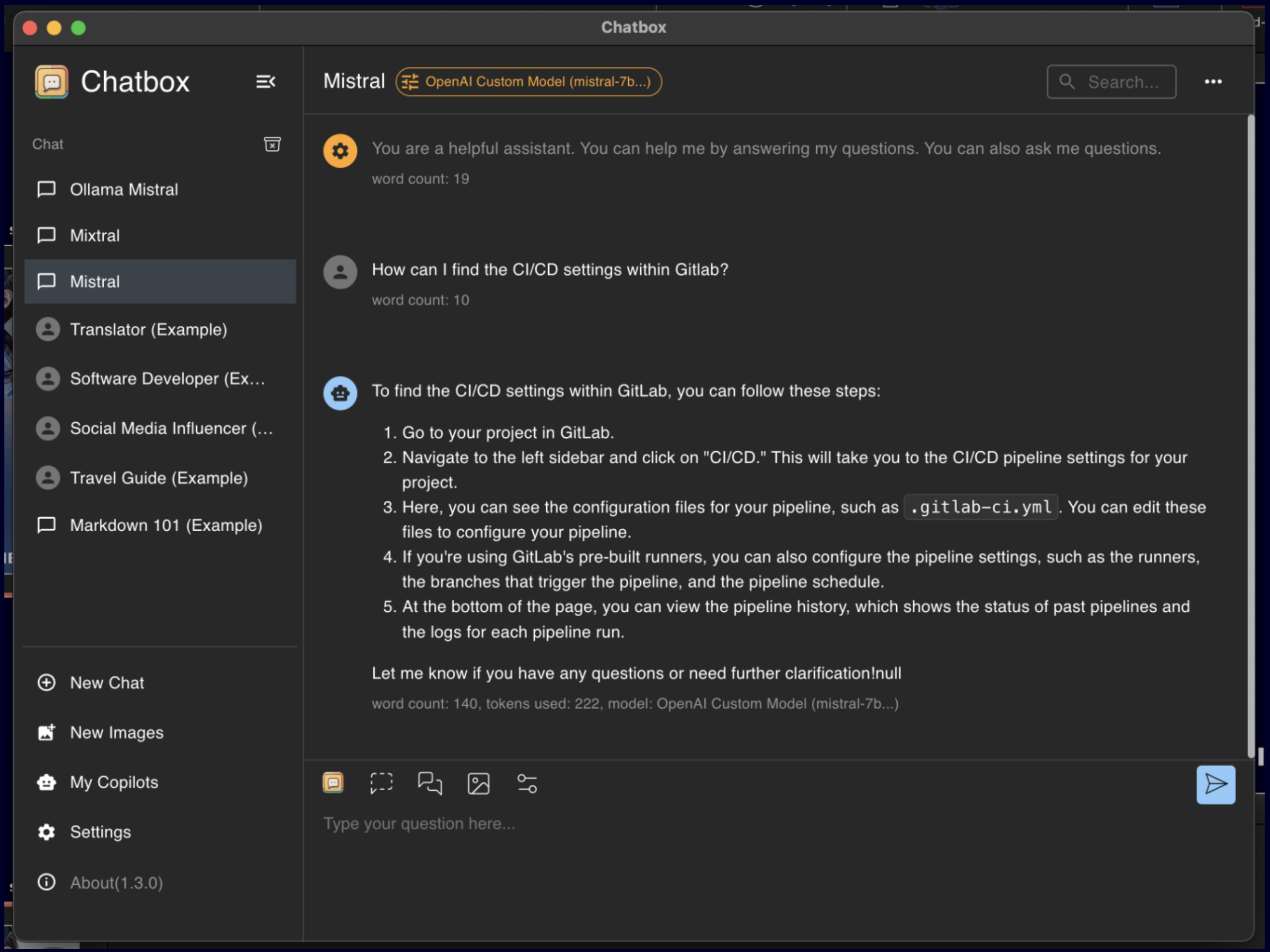


Accept ✅ | Reject ❌ | (⌘⇧↩/⌘⇧⌫ to accept/reject all, ⌘⇧L to retry)

```python

def is_worktime() -> bool:
 today_date = datetime.datetime.now().date()
   today_weekday = today_date.weekday()
   if today_weekday not in range(1, 5):
      return False
   current_time = datetime.time(hour=datetime.time.now().hour, minute=datetime.time.now().minute)
   return current_time.hour >= 9 and current_time.hour <= 17
      You, 1 second ago • Uncommitted changes

```

Usage via local UI (Chatbox)

Chatbox



You are a helpful assistant. You can help me by answering my questions. You can also ask me questions.

word count: 19

How can I find the CI/CD settings within Gitlab?

word count: 10

To find the CI/CD settings within GitLab, you can follow these steps:

1. Go to your project in GitLab.
2. Navigate to the left sidebar and click on "CI/CD." This will take you to the CI/CD pipeline settings for your project.
3. Here, you can see the configuration files for your pipeline, such as `.gitlab-ci.yml`. You can edit these files to configure your pipeline.
4. If you're using GitLab's pre-built runners, you can also configure the pipeline settings, such as the runners, the branches that trigger the pipeline, and the pipeline schedule.
5. At the bottom of the page, you can view the pipeline history, which shows the status of past pipelines and the logs for each pipeline run.

Let me know if you have any questions or need further clarification!null

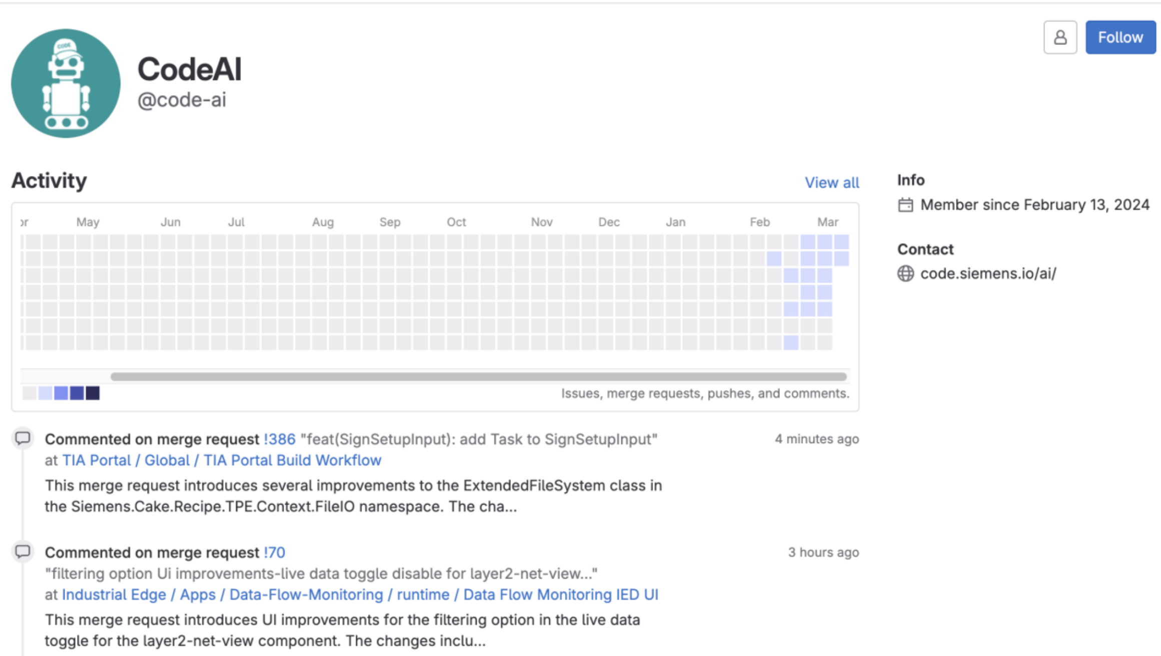
word count: 140, tokens used: 222, model: OpenAI Custom Model (mistral-7b...)

- 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



CodeAI
@code-ai

Activity

Issues, merge requests, pushes, and comments.

Info
Member since February 13, 2024

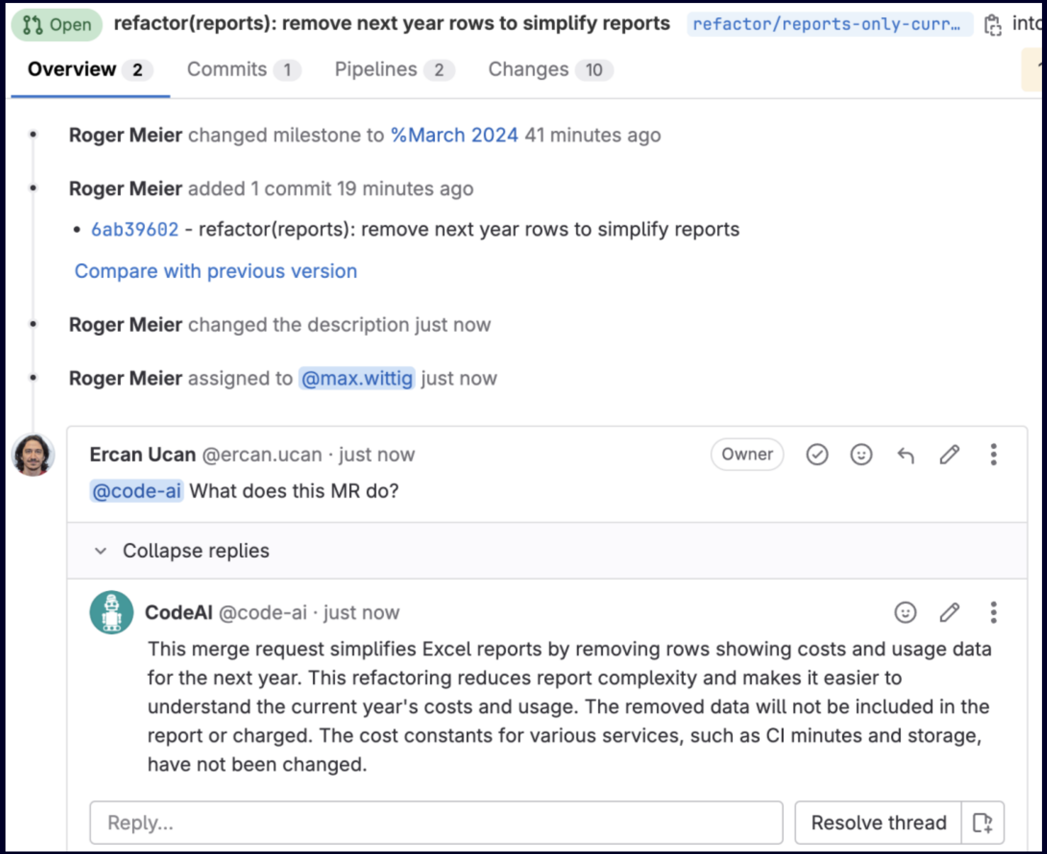
Contact
code.siemens.io/ai/

Commented on merge request !386 "feat(SignSetupInput): add Task to SignSetupInput" at TIA Portal / Global / TIA Portal Build Workflow — 4 minutes ago

This merge request introduces several improvements to the ExtendedFileSystem class in the Siemens.Cake.Recipe.TPE.Context.FileIO namespace. The cha...

Commented on merge request !70 "filtering option Ui improvements-live data toggle disable for layer2-net-view..." at Industrial Edge / Apps / Data-Flow-Monitoring / runtime / Data Flow Monitoring IED UI — 3 hours ago

This merge request introduces UI improvements for the filtering option in the live data toggle for the layer2-net-view component. The changes inclu...



refactor(reports): remove next year rows to simplify reports

- **Roger Meier** changed milestone to %March 2024 41 minutes ago
- **Roger Meier** added 1 commit 19 minutes ago
  - 6ab39602 - refactor(reports): remove next year rows to simplify reports

  Compare with previous version
- **Roger Meier** changed the description just now
- **Roger Meier** assigned to @max.wittig just now

Ercan Ucan @ercan.ucan · just now

@code-ai What does this MR do?

CodeAI @code-ai · just now

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:

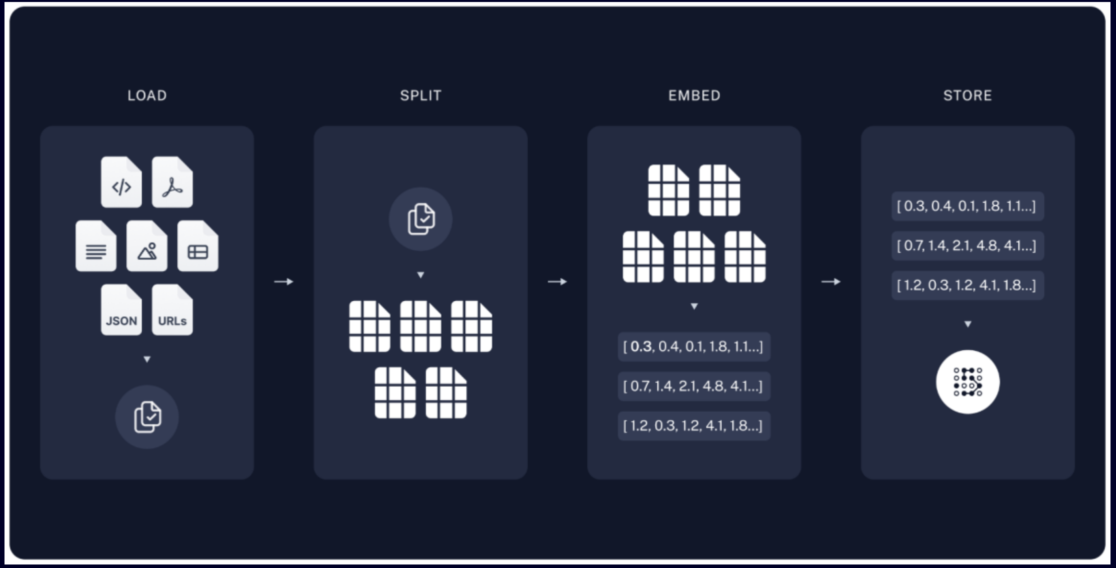
- **summary**: summarize changes in Merge Requests.
- **ci-failure**: identifies root cause for failed pipeline.
- **clean-code**: clean code suggestions (community contribution).
- **discussion-summary**: summarize the discussions in the issue or merge request.

Next Topics: Unit test creation, Secure coding analysis.

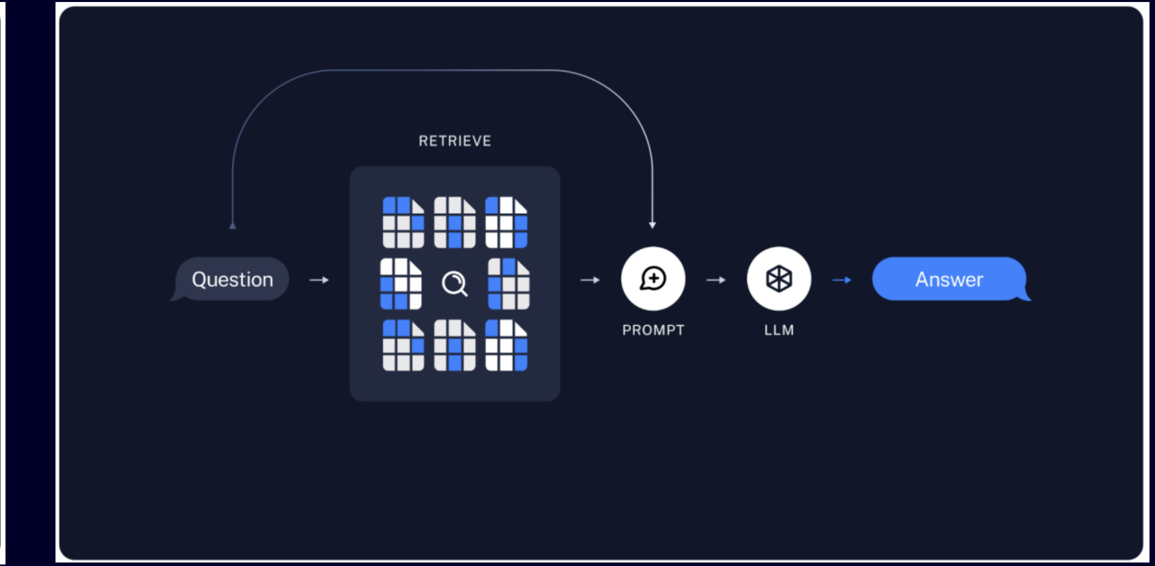
Usage via RAG with context

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

Ingest



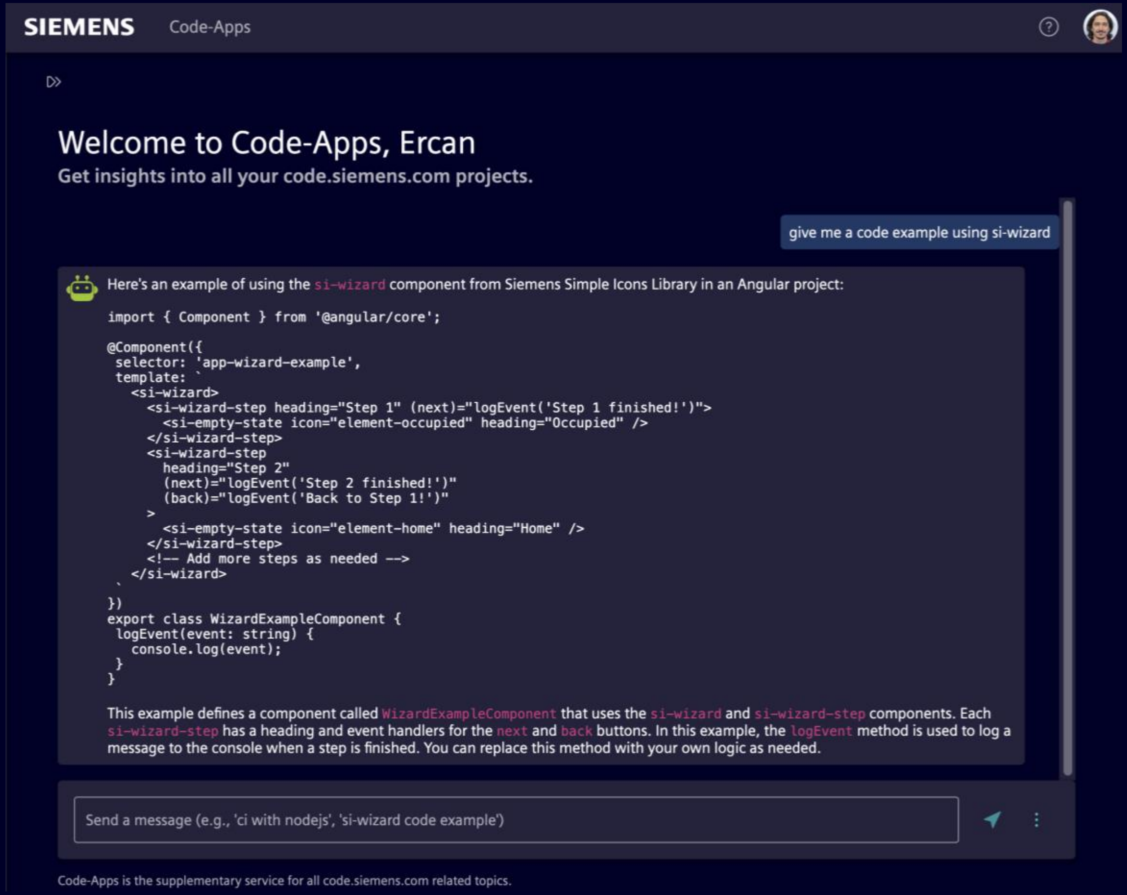
Retrieve



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



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:

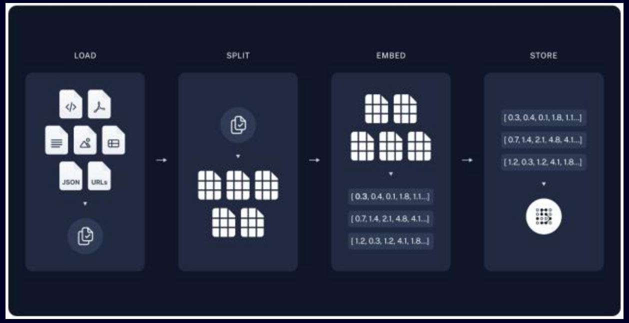
```
import { Component } from '@angular/core';

@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);
 }
}
```

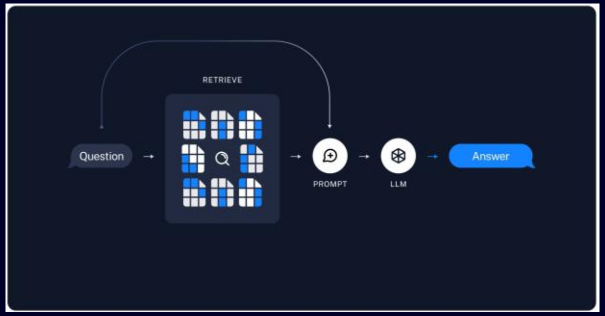
This example defines a component called `WizardExampleComponent` that uses the `si-wizard` and `si-wizard-step` components. Each `si-wizard-step` has a heading and event handlers for the `next` and `back` buttons. In this example, the `logEvent` method is used to log a message to the console when a step 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:

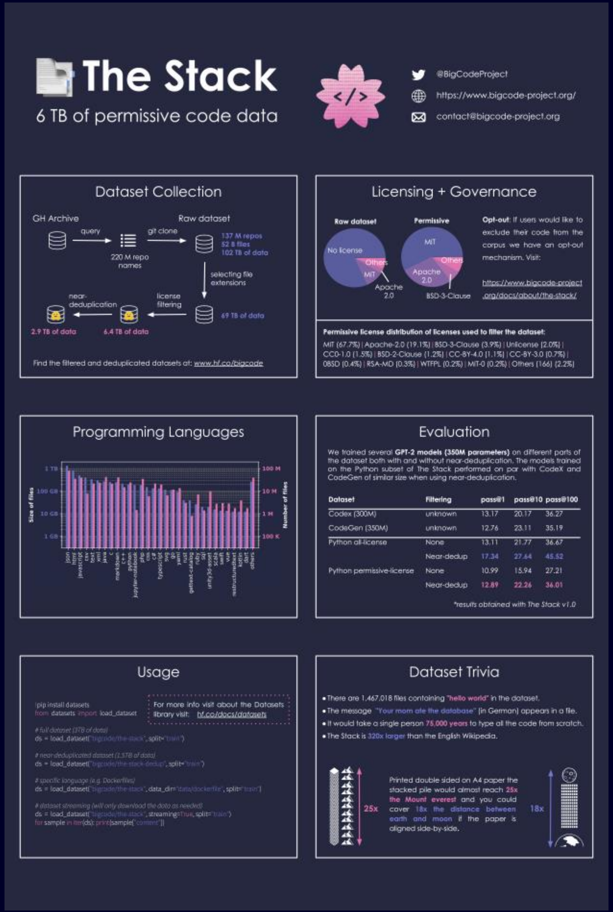
- Projects on code.siemens.com use it with their context.
- Selection of context based on permissions.

SIEMENS

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

@BigCodeProject

https://www.bigcode-project.org/

contact@bigcode-project.org

Dataset Collection

GH Archive
query
220 M repo names
git clone
Raw dataset
137 M repos
52 B files
102 TB of data
selecting file extensions
69 TB of data
license filtering
6.4 TB of data
near-deduplication
2.9 TB of data

Find the filtered and deduplicated datasets at: www.hf.co/bigcode

Licensing + Governance

Raw dataset
No license
Others
MIT
Apache 2.0
Permissive
MIT
Others
Apache 2.0
BSD-3-Clause

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/docs/about/the-stack/

Permissive license distribution of licenses used to filter the dataset:
MIT (67.7%) | Apache-2.0 (19.1%) | BSD-3-Clause (3.9%) | Unlicense (2.0%) | CC0-1.0 (1.5%) | BSD-2-Clause (1.2%) | CC-BY-4.0 (1.1%) | CC-BY-3.0 (0.7%) | 0BSD (0.4%) | RSA-MD (0.3%) | WTFPL (0.2%) | MIT-0 (0.2%) | Others (166) (2.2%)

Programming Languages

Evaluation

We trained several **GPT-2 models (350M 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 (300M)	unknown	13.17	20.17	36.27
CodeGen (350M)	unknown	12.76	23.11	35.19
Python all-license	None	13.11	21.77	36.67
	Near-dedup	17.34	27.64	45.52
Python permissive-license	None	10.99	15.94	27.21
	Near-dedup	12.89	22.26	36.01

*results obtained with *The Stack v1.0*

Usage

```
!pip install datasets
from datasets import load_dataset

# full dataset (3TB of data)
ds = load_dataset("bigcode/the-stack", split="train")

# near-deduplicated dataset (1.5TB of data)
ds = load_dataset("bigcode/the-stack-dedup", split="train")

# specific language (e.g. Dockerfiles)
ds = load_dataset("bigcode/the-stack", data_dir="data/dockerfile", split="train")

# dataset streaming (will only download the data as needed)
ds = load_dataset("bigcode/the-stack", streaming=True, split="train")
for sample in iter(ds): print(sample["content"])
```

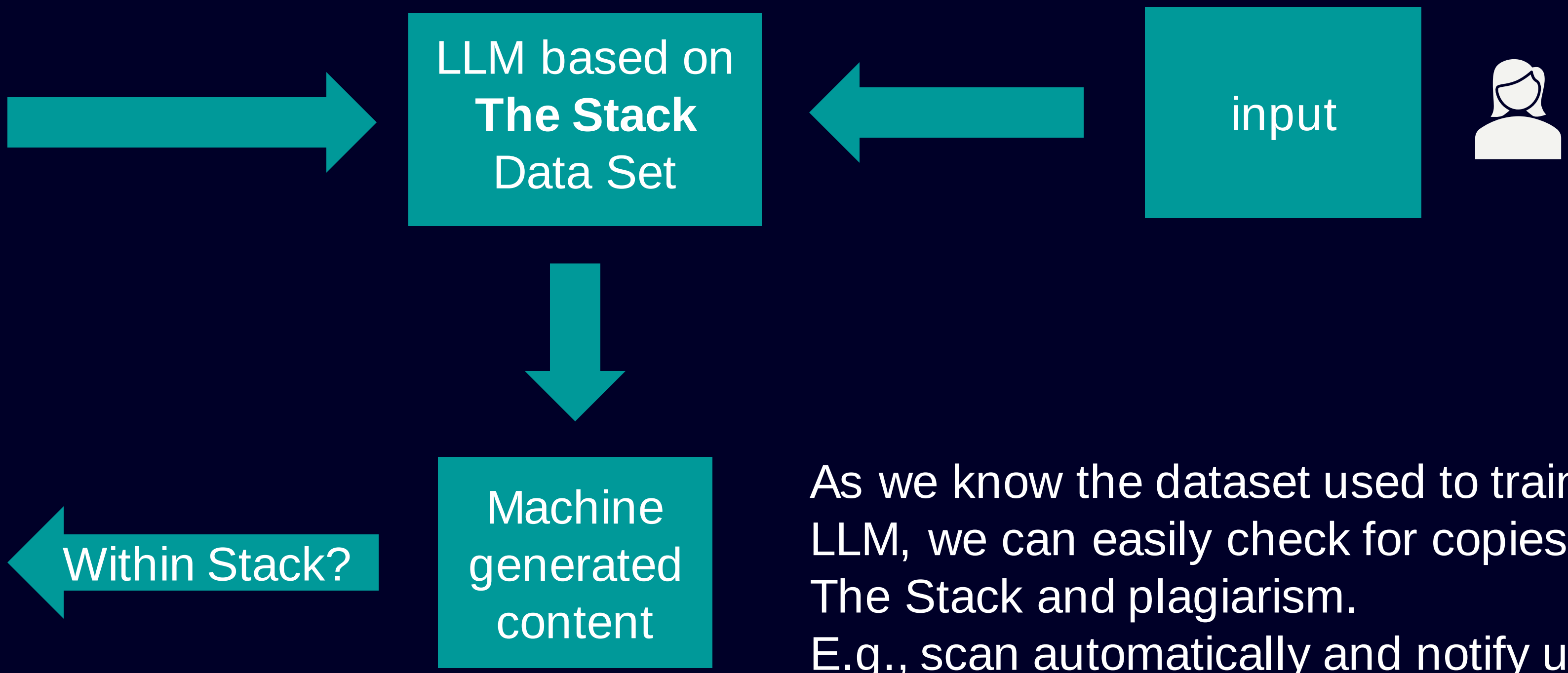
For more info visit about the Datasets library visit: hf.co/docs/datasets

Dataset Trivia

- There are 1,467,018 files containing "hello world" in the dataset.
- The message "Your mom ate the database" (in German) appears in a file.
- It would take a single person 75,000 years to type all the code from scratch.
- The Stack is 320x larger than the English Wikipedia.

Printed double sided on A4 paper the stacked pile would almost reach 25x the Mount everest and you could cover 18x the distance between earth and moon if the paper is aligned side-by-side.

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



LLM based on **The Stack** Data Set

input

Machine generated content

Within Stack?

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.

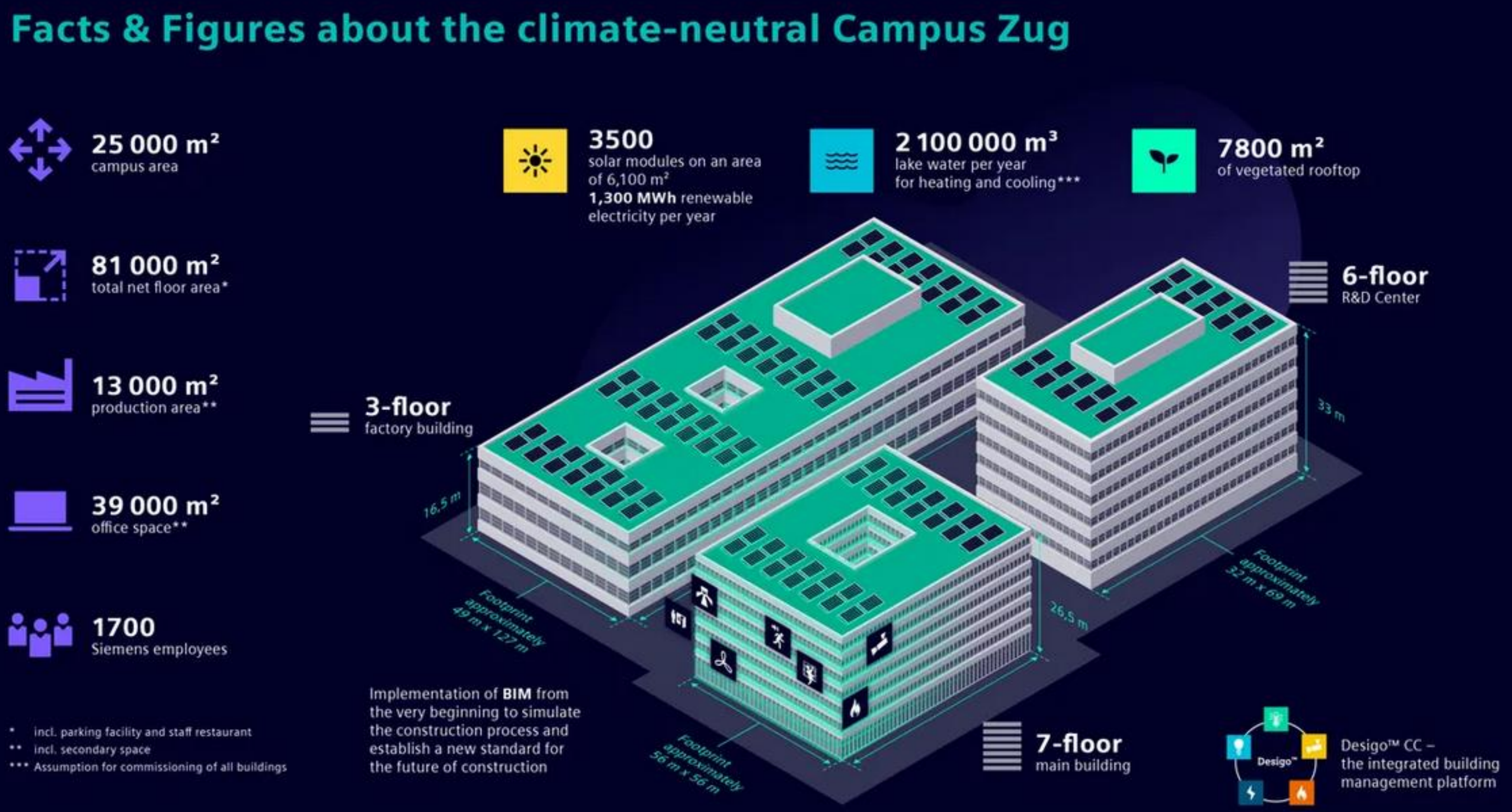
StarCoder2 is based on The Stack v2

Sustainable AI

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

25 000 m^2 campus area

81 000 m^2 total net floor area*

13 000 m^2 production area**

39 000 m^2 office space**

1700 Siemens employees

3500 solar modules on an area of 6,100 m^2 **1,300 MWh** renewable electricity per year

2 100 000 m^3 lake water per year for heating and cooling***

7800 m^2 of vegetated rooftop

3-floor factory building

6-floor R&D Center

7-floor main building

Implementation of **BIM** from the very beginning to simulate the construction process and establish a new standard for the future of construction

Desigo™ CC – the integrated building management platform

16,5 m
Footprint approximately 49 m x 127 m
33 m
Footprint approximately 32 m x 69 m
26,5 m
Footprint approximately 56 m x 56 m
Desigo™

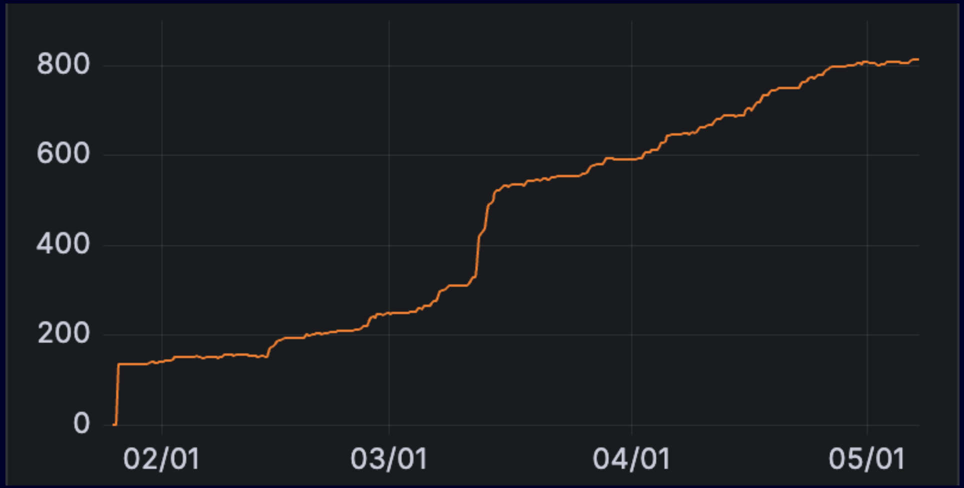
* incl. parking facility and staff restaurant
** incl. secondary space
*** Assumption for commissioning of all buildings

Server room within the R&D center

> *"Companies with serious plans for AI have the GPUs and the knowledge in-house."*

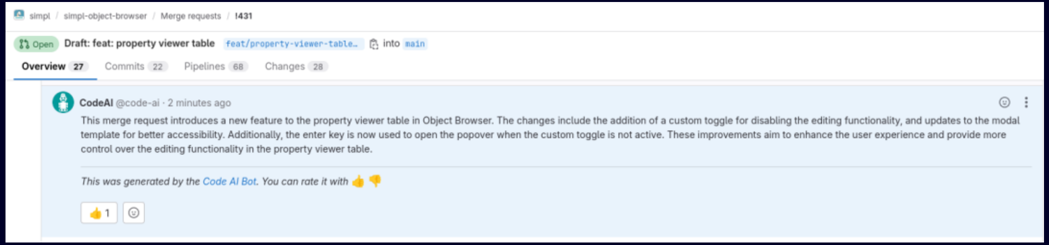
SIEMENS

Feedback from the community



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

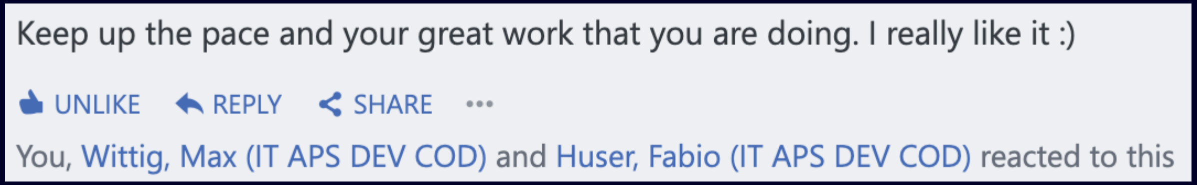
Really love the work that you're doing here. I'd love to get involved on the Healthineers side of things.


simpl / simpl-object-browser / Merge requests / !431
Open
Draft: feat: property viewer table
feat/property-viewer-table…
into main
Overview 27
Commits 22
Pipelines 68
Changes 28
👍 1

CodeAI @code-ai · 2 minutes ago

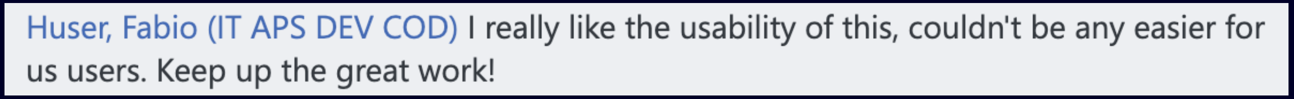
This merge request introduces a new feature to the property viewer table in Object Browser. The changes include the addition of a custom toggle for disabling the editing functionality, and updates to the modal template for better accessibility. Additionally, the enter key is now used to open the popover when the custom toggle is not active. These improvements aim to enhance the user experience and provide more control over the editing functionality in the property viewer table.

This was generated by the Code AI Bot. You can rate it with 👍 👎

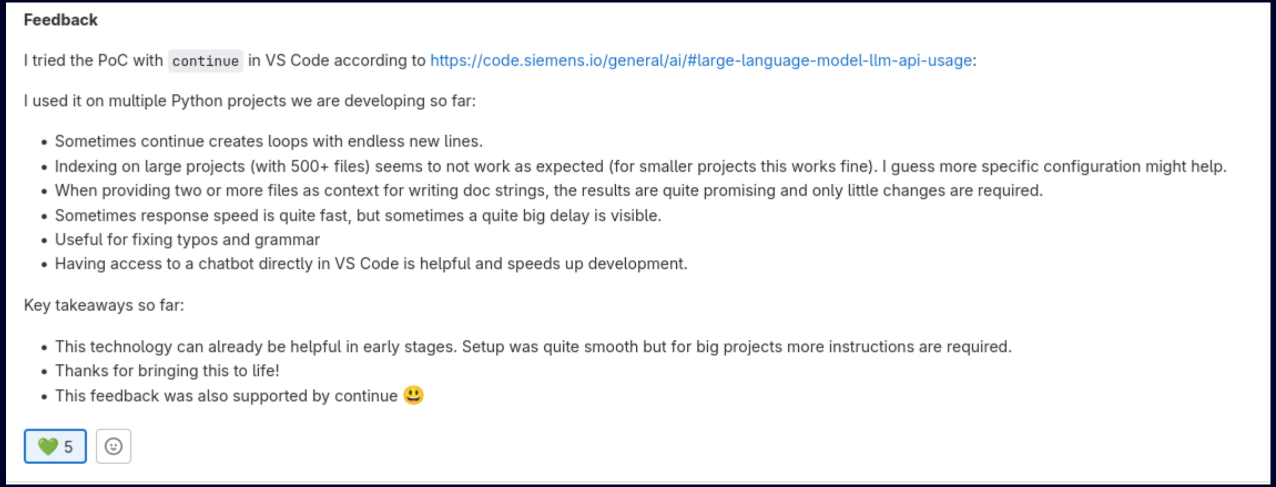


Keep up the pace and your great work that you are doing. I really like it :)

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:

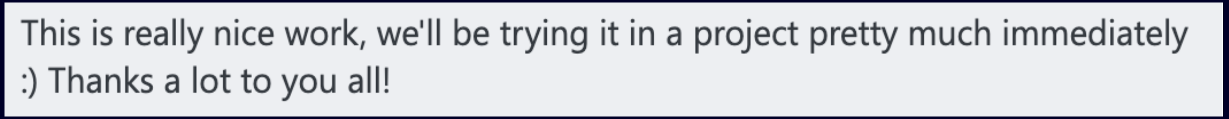
I used it on multiple Python projects we are developing so far:

- 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 😃

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!

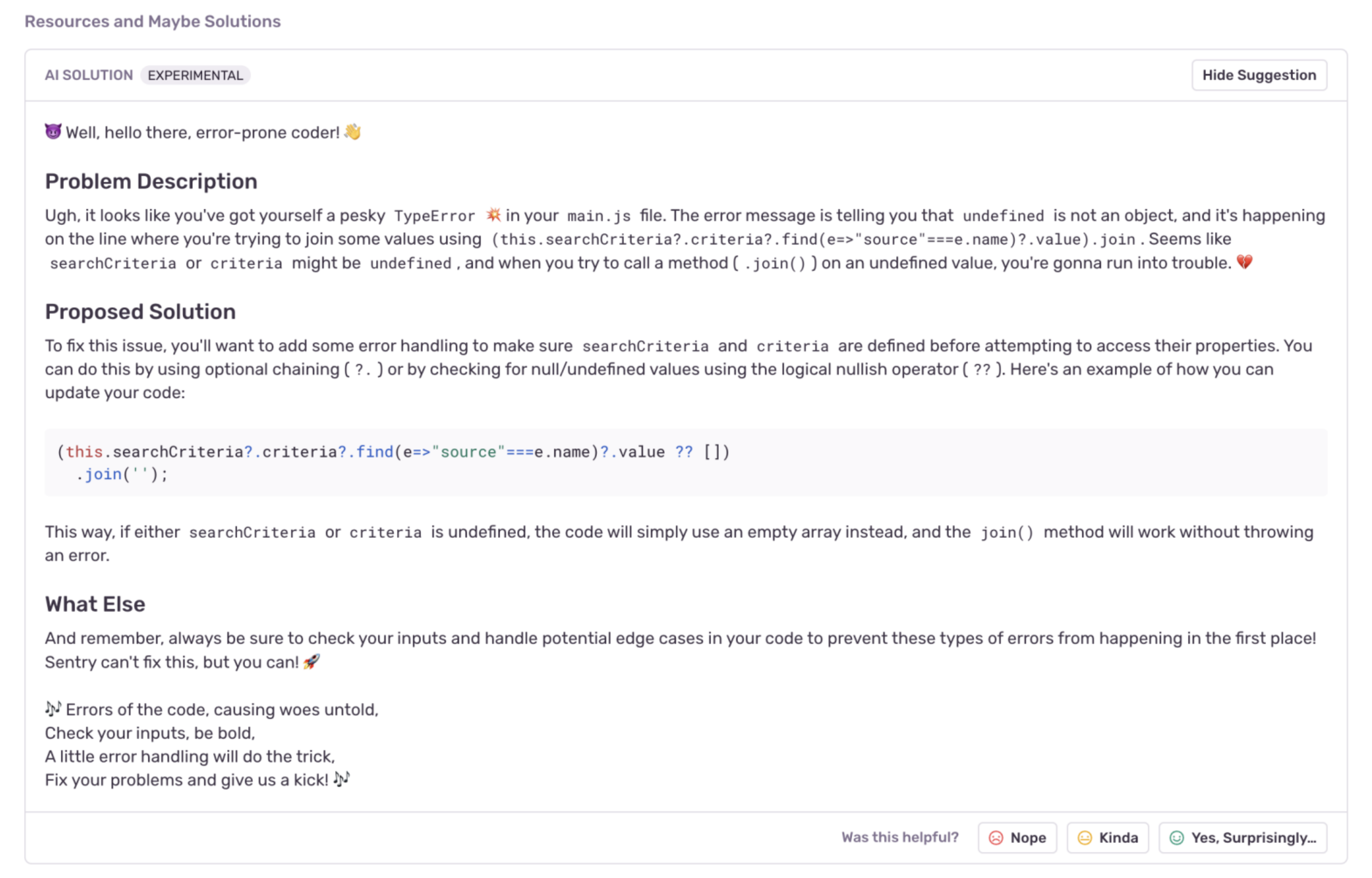
Contact

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

SIEMENS

Backup slides

Integration into internal crash reporting



😈 Well, hello there, error-prone coder! 👋

Problem Description

Ugh, it looks like you've got yourself a pesky `TypeError` 💥 in your `main.js` file. The error message is telling you that `undefined` is not an object, and it's happening on the line where you're trying to join some values using `(this.searchCriteria?.criteria?.find(e=>"source"===e.name)?.value).join`. Seems like `searchCriteria` or `criteria` might be `undefined`, and when you try to call a method (`.join()`) on an undefined value, you're gonna run into trouble. 💔

Proposed Solution

To fix this issue, you'll want to add some error handling to make sure `searchCriteria` and `criteria` are defined before attempting to access their properties. You can do this by using optional chaining (`?.`) or by checking for null/undefined values using the logical nullish operator (`??`). Here's an example of how you can update your code:

```
(this.searchCriteria?.criteria?.find(e=>"source"===e.name)?.value ?? [])
  .join('');
```

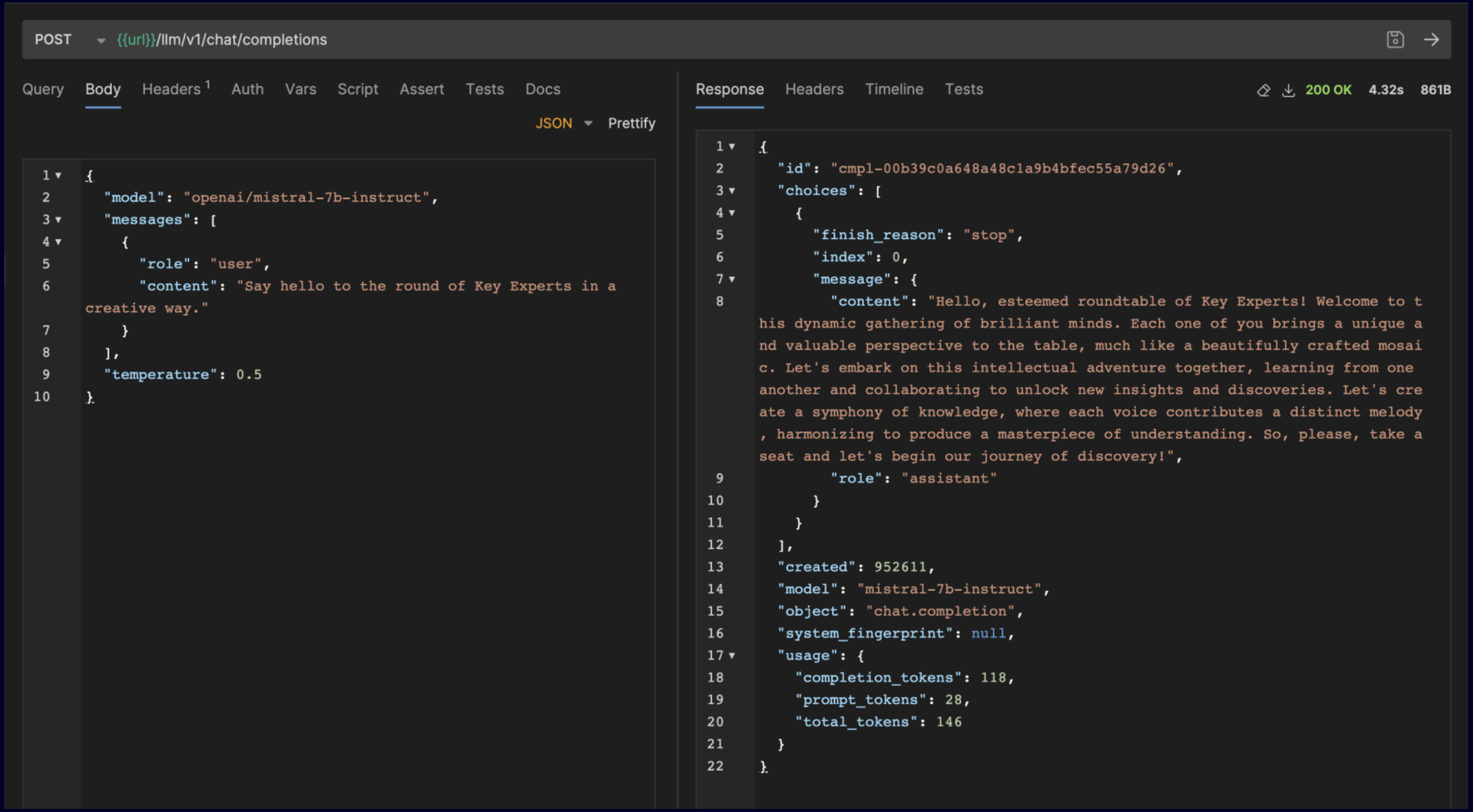
This way, if either `searchCriteria` or `criteria` is undefined, the code will simply use an empty array instead, and the `join()` method will work without throwing an error.

What Else

And remember, always be sure to check your inputs and handle potential edge cases in your code to prevent these types of errors from happening in the first place! Sentry can't fix this, but you can! 🚀

🎶 Errors of the code, causing woes untold,
Check your inputs, be bold,
A little error handling will do the trick,
Fix your problems and give us a kick! 🎶

API usage for all possible applications – demo with Bruno (API Client)



```
{{url}}/llm/v1/chat/completions
```

```
{
  "model": "openai/mistral-7b-instruct",
  "messages": [
    {
      "role": "user",
      "content": "Say hello to the round of Key Experts in a creative way."
    }
  ],
  "temperature": 0.5
}
```

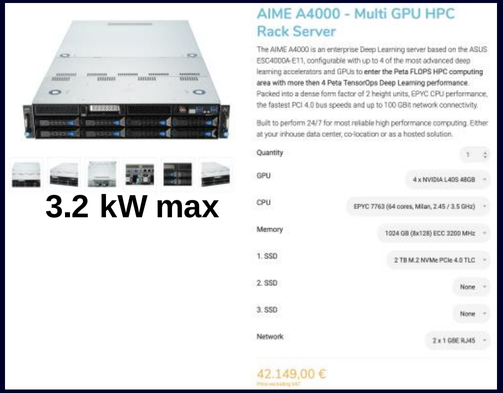
```
{
  "id": "cmpl-00b39c0a648a48c1a9b4bfec55a79d26",
  "choices": [
    {
      "finish_reason": "stop",
      "index": 0,
      "message": {
        "content": "Hello, esteemed roundtable of Key Experts! Welcome to this dynamic gathering of brilliant minds. Each one of you brings a unique and valuable perspective to the table, much like a beautifully crafted mosaic. Let's embark on this intellectual adventure together, learning from one another and collaborating to unlock new insights and discoveries. Let's create a symphony of knowledge, where each voice contributes a distinct melody, harmonizing to produce a masterpiece of understanding. So, please, take a seat and let's begin our journey of discovery!",
        "role": "assistant"
      }
    }
  ],
  "created": 952611,
  "model": "mistral-7b-instruct",
  "object": "chat.completion",
  "system_fingerprint": null,
  "usage": {
    "completion_tokens": 118,
    "prompt_tokens": 28,
    "total_tokens": 146
  }
}
```

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

- Typical offerings charge per inbound and outbound tokens, which is more expensive, and we have no influence if we reach limits.

Example cost calculation

- AWS g5.12xlarge (4 x 24 GB GPU)
  - 192 GB GPU = **~ $ 122.640 / year**
- Physical AIME machine
  - 192 GB GPU = **~ € 42.149 / one time**



AIME A4000 - Multi GPU HPC Rack Server

The AIME A4000 is an enterprise Deep Learning server based on the ASUS ESC4000A-E11, configurable with up to 4 of the most advanced deep learning accelerators and GPUs to **enter the Peta FLOPS HPC computing area with more then 4 Peta TensorOps Deep Learning performance**. Packed into a dense form factor of 2 height units, EPYC CPU performance, the fastest PCI 4.0 bus speeds and up to 100 GBit network connectivity.

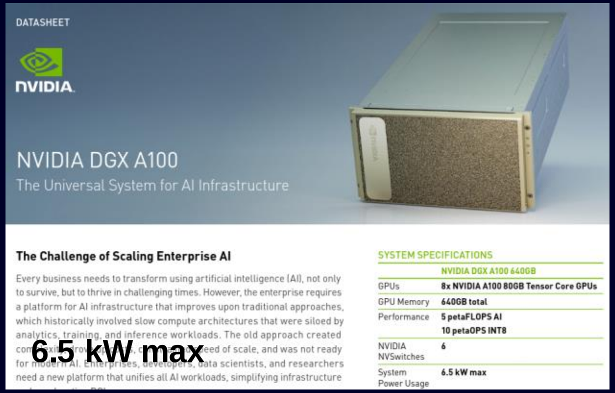
Built to perform 24/7 for most reliable high performance computing. Either at your inhouse data center, co-location or as a hosted solution.

Quantity	1
GPU	4 x NVIDIA L40S 48GB
CPU	EPYC 7763 (64 cores, Milan, 2.45 / 3.5 GHz)
Memory	1024 GB (8x128) ECC 3200 MHz
1. SSD	2 TB M.2 NVMe PCIe 4.0 TLC
2. SSD	None
3. SSD	None
Network	2 x 1 GBE RJ45

42.149,00 €

Price excluding VAT

3.2 kW max



DATASHEET

NVIDIA DGX A100

The Universal System for AI Infrastructure

The Challenge of Scaling Enterprise AI

Every business needs to transform using artificial intelligence (AI), not only to survive, but to thrive in challenging times. However, the enterprise requires a platform for AI infrastructure that improves upon traditional approaches, which historically involved slow compute architectures that were siloed by analytics, training, and inference workloads. The old approach created com[illegible] speed of scale, and was not ready for modern AI. Enterprises, developers, data scientists, and researchers need a new platform that unifies all AI workloads, simplifying infrastructure

SYSTEM SPECIFICATIONS

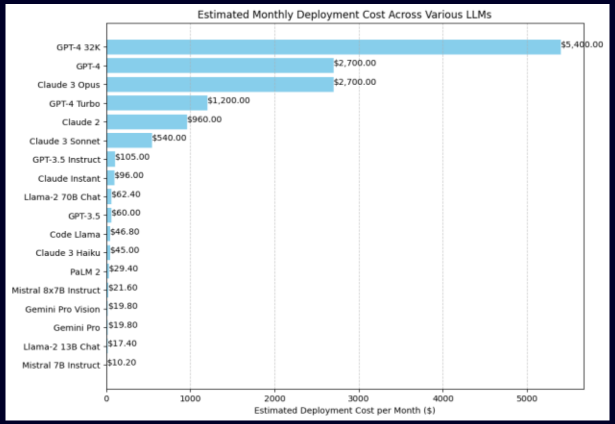
	NVIDIA DGX A100 640GB
GPUs	8x NVIDIA A100 80GB Tensor Core GPUs
GPU Memory	640GB total
Performance	5 petaFLOPS AI 10 petaOPS INT8
NVIDIA NVSwitches	6
System Power Usage	6.5 kW max

6.5 kW max

- **Amortization** for a physical machine is **~ 3/4 months**

Estimate: GitHub Copilot IDE plugin vs. GPU on prem

- Serving 1000 developers with a 192GB machine for 3 years
- **~ €1.38 / month**



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