

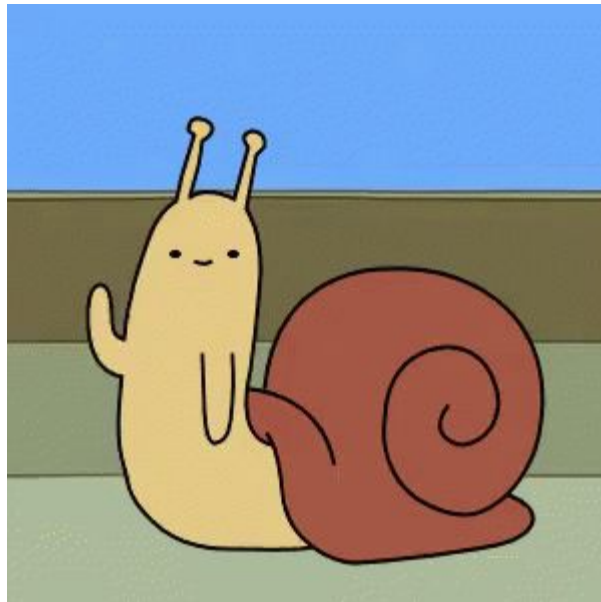
Clickhouse + Helicone

Clickhouse enables next gen AI Observability

👋 Hi I'm Justin

- Co-founder of Helicone
- Platform/infra Engineer

Ex: Sisu Data, Apple and Intel

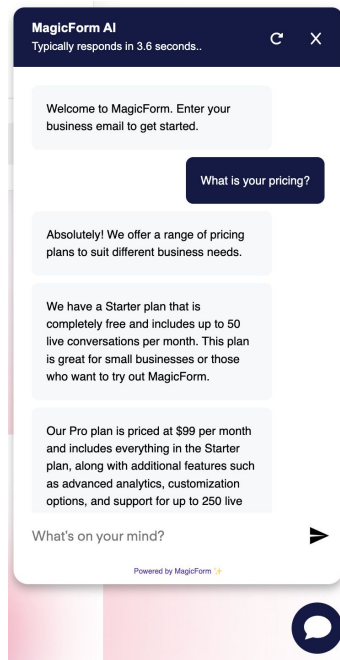
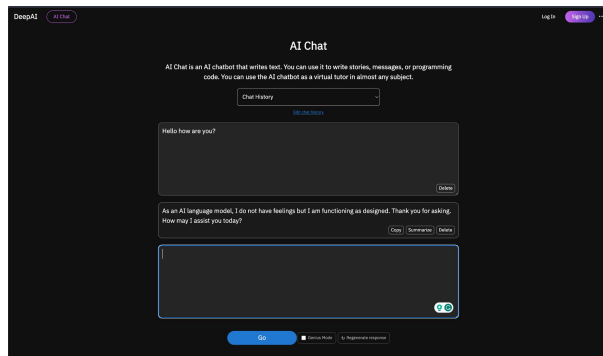
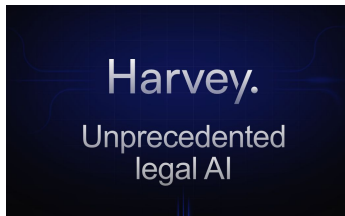


Everyone is using LLMs



It's all over the place!

- Chatbots
- Agents
- Business
- Entertainment
- News
- EVERYWHERE



Helicone

Open source observability and monitoring for LLMs

The screenshot displays the GitHub repository for Helicone, a public project. The repository has 597 branches, 0 tags, 6 forks, and 587 stars. The commit history shows recent updates, including a retry on 522 (#686) and a response copy logging (#482). The README section provides an overview of the project, stating it is an open-source observability platform for LLMs. It lists features such as logging requests to OpenAI, caching, cost and latency tracking, and a playground for iterating on prompts. The repository also includes a list of contributors, deployments, and languages used in the project.

Helicone

Visit Us [Helicone.ai](#) View Documentation [Docs](#) Join our community [Discord](#) Follow @Helicone.ai

Open-source observability platform for LLMs

Helicone is an open-source observability platform for Language Learning Models (LLMs). It offers the following features:

- Logs all of your requests to OpenAI in a user-friendly UI
- Caching, custom rate limits, and retries
- Track costs and latencies by users and custom properties
- Every log is a playground: iterate on prompts and chat conversations in a UI
- Share results and collaborate with your friends or teammates

Contributors 18

Deployments 500

- Preview 13 hours ago
- Production 13 hours ago

Languages

- TypeScript 93.2%
- Python 4.1%
- PL/pgSQL 1.4%
- Shell 1.1%
- JavaScript 0.1%
- CSS 0.1%

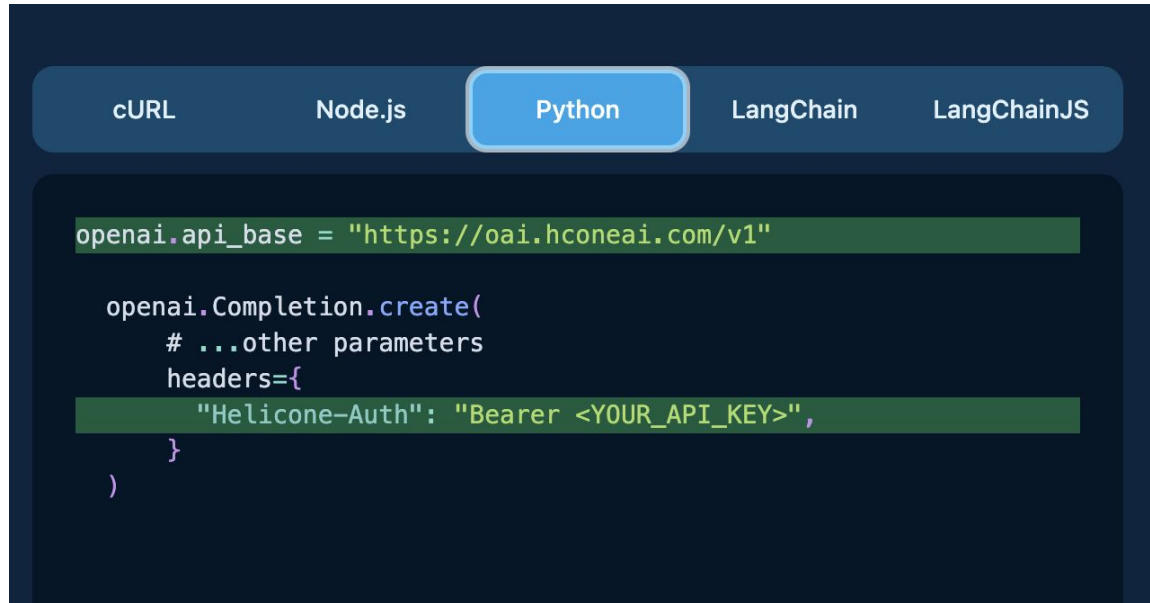
Suggested Workflows

Based on your tech stack

Actions Importer Set up

Automatically convert CI/CD files to

Helicone let's you easy capture your LLM requests

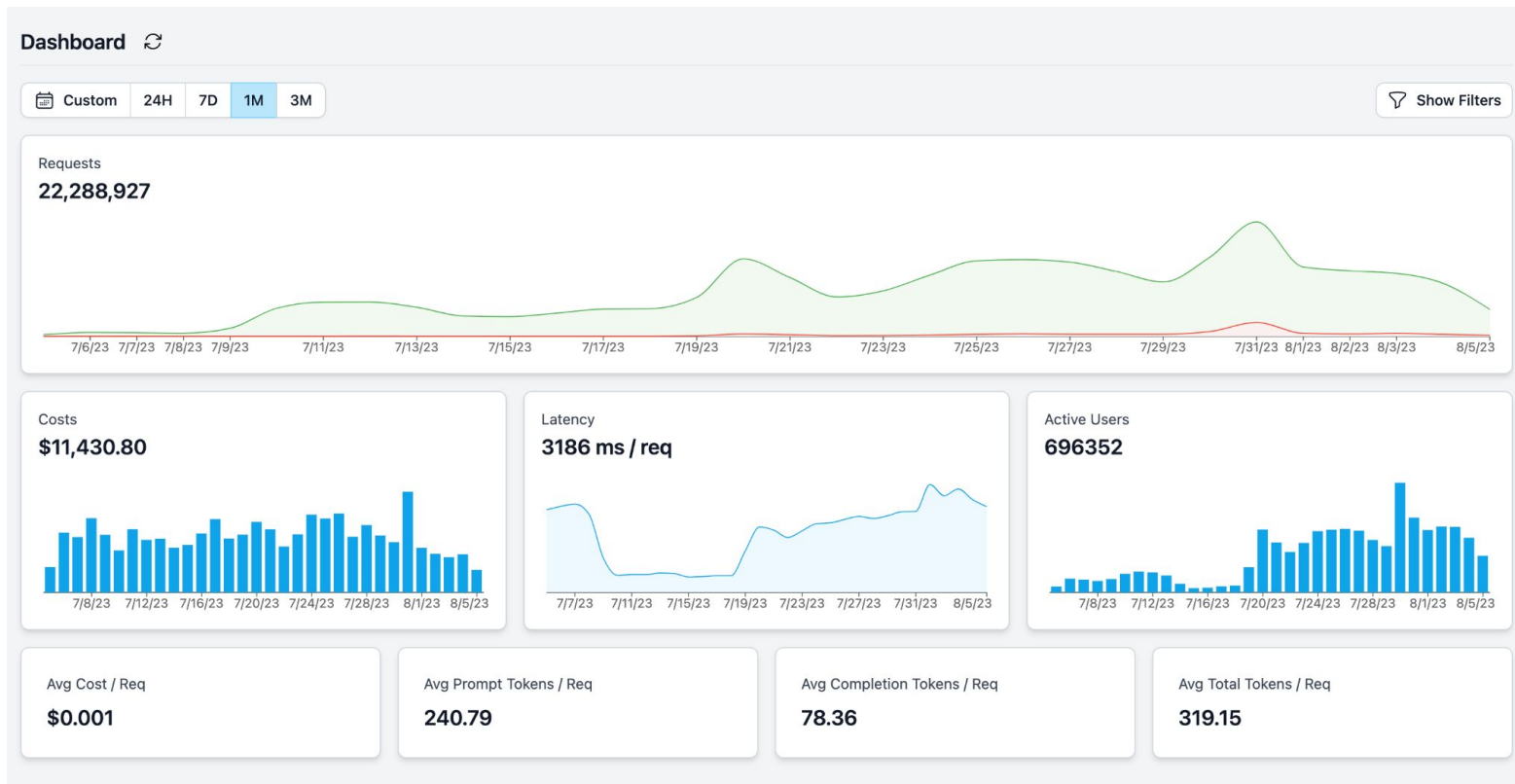


The image shows a code editor interface with a dark blue background. At the top, there is a horizontal bar with five tabs: 'cURL', 'Node.js', 'Python', 'LangChain', and 'LangChainJS'. The 'Python' tab is selected and highlighted with a light blue border. Below the tabs, there is a code editor area with a dark blue background. The code is written in Python and shows how to configure the OpenAI API base and headers for Helicone. The code is as follows:


```
openai.api_base = "https://oai.hconeai.com/v1"


openai.Completion.create(
    # ...other parameters
    headers={
        "Helicone-Auth": "Bearer <YOUR_API_KEY>",
    }
)
```

Build rich dashboards automatically



And get requests

Requests 

Live 

Custom


24H


7D


1M

3M

All

 Hide Filters

 View (16 / 16)

 Export

Filters

+ Add Filter

Created At	Status	Request	Response	Model	Total Tokens	Prompt Tokens	Completion Tokens	Latency	User
August 5 1:29 AM	502 Error	Write a concise su...	Bad gateway.	gpt-3.5-turbo				600.562s	
August 5 1:27 AM	Success	Determine exactly ...	{ "thoughts": { "te...	gpt-4-0314	1924	1636	288	22.627s	
August 5 1:27 AM	Success	The text is a code ...		text-embedding-ada-002		78		0.273s	
August 5 1:27 AM	Success	# mypy: ignore-err...		text-embedding-ada-002		77		0.275s	
August 5 1:27 AM	Success	Write a concise su...	The text is a code ...	gpt-3.5-turbo	188	110	78	3.307s	
August 5 1:27 AM	Success	Determine exactly ...	{ "thoughts": { "te...	gpt-4-0314	1509	1334	175	15.414s	
August 5 1:26 AM	Success	Determine exactly ...	{ "thoughts": { "te...	gpt-4-0314	1257	1068	189	15.475s	
August 5 1:26 AM	Success	The text includes ...		text-embedding-ada-002		53		0.31s	
August 5 1:26 AM	Success	# mypy: ignore-err...		text-embedding-ada-002		250		0.604s	
August 5 1:26 AM	Success	Write a concise su...	The text includes ...	gpt-3.5-turbo	336	283	53	2.509s	

1 month after launch

- 3 Million requests a day



Our dashboard

- Timeouts
- Aggregations were slow (30s+)



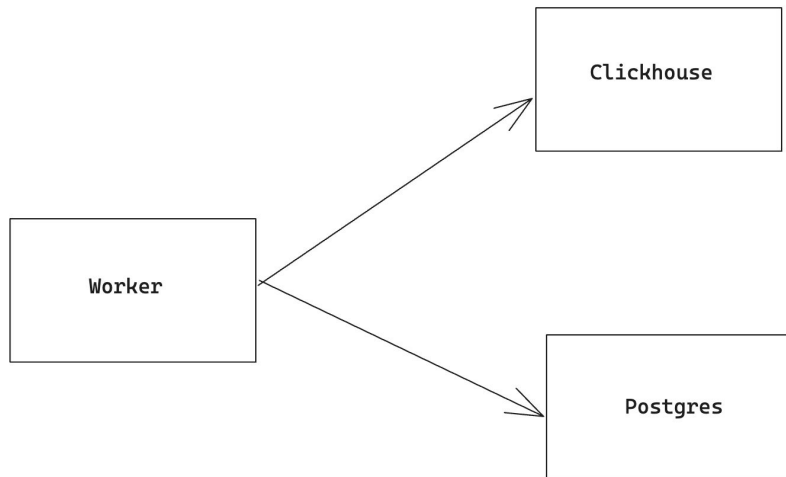
Clickhouse saved Helicone

- Open source
- Fast
- Easy to use
- Scalable
- Support
- Async buffer insert



Postgres View to Clickhouse Tables

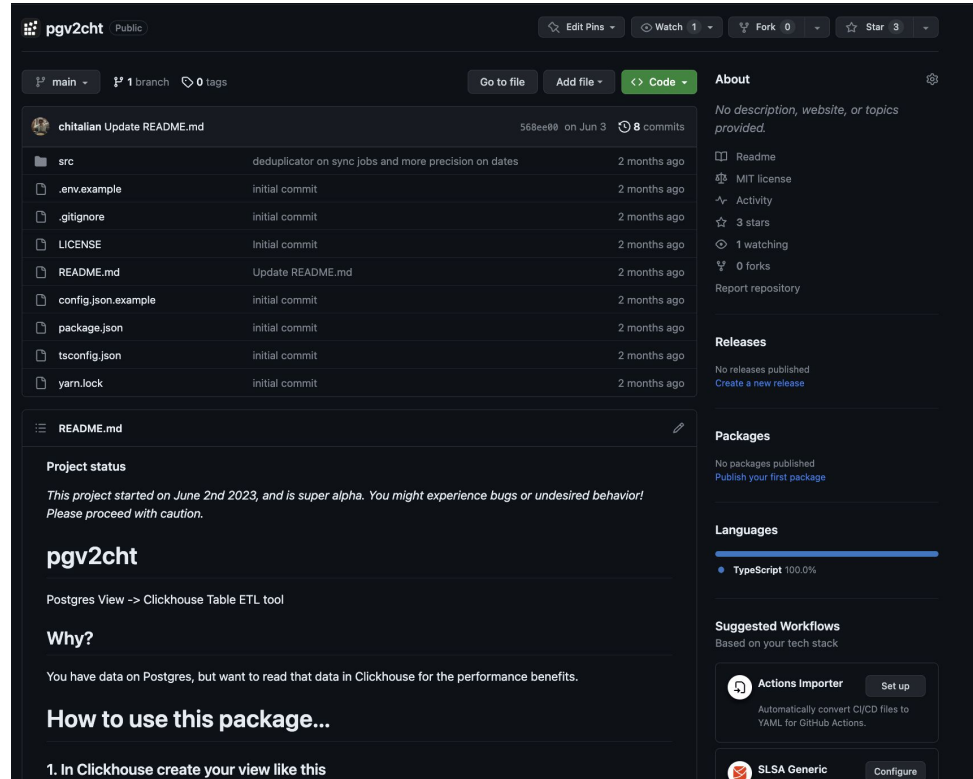
- Postgres source of truth
 - Relationships
- Replicate data on Clickhouse for specific queries



Treating Clickhouse tables as Postgres Materialized views

Triggers are setup to insert into Clickhouse, but how do we now backfill that data?

Postgres View to Clickhouse Table
(pgv2cht)



Party time! 🎉

Helicone works now! And it is faster than ever!

100+s query times -> 0.5s query times

Explore Helicone's role in shaping the future LLM stack.

Sequoia Capital | Andreessen Horowitz

Helicone Pricing Docs Roadmap Blog

Star us on GitHub Sign In

Backed by Y Combinator

Open-Source Monitoring for Generative AI

Thousands of users and organizations leverage Helicone to monitor their LLM applications. Instantly get insights into your latency, costs, and much more.

[Get Started](#) [View Demo](#)

Dashboard

Requests

Custom 100 10 100 All

View Filters

Status	Request	Response	Model	Total Tokens
Success	As a historian, I...	I will start by writing...	gpt-3.5-turbo-0613	750
Success	As a historian, I...	I will start by writing...	gpt-3.5-turbo-0613	807
Success	As a historian, I...	I will use the "T...	gpt-3.5-turbo-0613	832
Success	As a historian, I...	Here is your ques...	gpt-3.5-turbo-0613	819
Success	As a historian, I...	Thank you for the...	gpt-3.5-turbo-0613	885
Success	Here is your ques...	I choose to share...	gpt-3.5-turbo-0613	1086
Success	Here is your ques...	I will choose the "I...	gpt-3.5-turbo-0613	1420
Success	Here is your ques...	I choose to share...	gpt-3.5-turbo-0613	1526
Success	Here is your ques...	Thank you for the...	gpt-3.5-turbo-0613	1238

Custom Properties

View

Filter

Chat

You're an AI model that generates incredibly lengthy and detailed creative content for students. The final answer. The content is the creative response of all steps. The model shows the length of the content to be long and creative, and that it covers the entirety of the subject matter. The answer all answer content is in markdown format and is to be used as a reference.

As an AI model acting as an expert teacher, you will use an engaging, creative style to generate lengthy, creative content. You will ensure the content is in markdown format and is to be used as a reference.

Thank you Clickhouse!

Huge shout out to Clickhouse for all your support

- Thom O'Connor, Mike Hayes, Brian Hunter, Derek Chia, Marcelo Rodriguez, and many more! (sorry if I missed you!)

Email me or text me if you want to try Helicone:

justin@helicone.ai | 631-834-2420



helicone.ai

Now when we want to add a new view for aggregates we simply...

- Add a trigger or other means to insert new rows from postgres also in Clickhouse
- Build a view in Postgres that matches the schema
- Run the pg2cht backfill scripts