While tech companies play with OpenAI's API, this startup believes small, in-house AI models will win

Description

ZenML wants to be the glue that makes all the open-source AI tools stick together. This open-source framework lets you build pipelines that will be used by data scientists, machine-learning engineers and platform engineers to collaborate and build new AI models.

The reason ZenML is interesting is that it empowers companies so they can build their own private models. Of course, companies likely won't build a GPT-4 competitor. But they could build smaller models that work particularly well for their needs. And it would reduce their dependence on API providers, such as OpenAI and Anthropic.

"The idea is that, once the first wave of hype with everyone using OpenAI or closed-source APIs is over, [ZenML] will enable people to build their own stack," Louis Coppey, a partner at VC firm Point Nine, told me.

Earlier this year, ZenML raised an extension of its seed round from Point Nine with existing investor Crane also participating. Overall, the startup based in Munich, Germany has secured $6.4 million since its inception.

Adam Probst and Hamza Tahir, the founders of ZenML, previously worked together on a company that was building ML pipelines for other companies in a specific industry. "Day in, day out, we needed to build machine learning models and bring machine learning into production," ZenML CEO Adam Probst told me.

From this work, the duo started designing a modular system that would adapt to different circumstances, environments and customers so that they wouldn't have to repeat the same work over and over again — this led to ZenML.

At the same time, engineers who are getting started with machine learning could get a head start by using this modular system. The ZenML team calls this space MLOps — it's a bit like DevOps, but applied to ML in particular.

"We are connecting the open source tools that are focusing on specific steps of the value chain to build a machine learning pipeline — everything on the back of the hyperscalers, so everything on the back of AWS and Google — and also on-prem solutions," Probst said.

The main concept of ZenML is pipelines. When you write a pipeline, you can then run it locally or deploy it using open source tools like Airflow or Kubeflow. You can also take advantage of managed cloud services, such as EC2, Vertex Pipelines and SageMaker. ZenML also integrates with open source ML tools from Hugging Face, MLflow, TensorFlow, PyTorch, etc.

"ZenML is sort of the thing that brings everything together into one single unified experience — it's multi-vendor, multi-cloud," ZenML CTO Hamza Tahir said. It brings connectors, observability and

auditability to ML workflows.

The company first released its framework on GitHub as an open source tool. The team has amassed more than 3,000 stars on the coding platform. ZenML also recently started offering a cloud version with managed servers — triggers for continuous integrations and deployment (CI/CD) are coming soon.

Some companies have been using ZenML for industrial use cases, e-commerce recommendation systems, image recognition in a medical environment, etc. Clients include Rivian, Playtika and Leroy Merlin.

Private, industry-specific models

The success of ZenML will depend on how the AI ecosystem is evolving. Right now, many companies are adding AI features here and there by querying OpenAI's API. In this product, you now have a new magic button that can summarize large chunks of text. In that product, you now have pre-written answers for customer support interactions.

> "OpenAI will have a future, but we think the majority of the market will have to have its own solution." *Adam Probst*

But there are a couple of issues with these APIs — they are too sophisticated and too expensive. "OpenAI, or these large language models built behind closed doors are built for general use cases — not for specific use cases. So currently it's way too trained and way too expensive for specific use cases," Probst said.

"OpenAI will have a future, but we think the majority of the market will have to have its own solution. And this is why open source is very appealing to them," he added.

OpenAI's CEO Sam Altman also believes that AI models won't be a one-size-fits-all situation. "I think both have an important role. We're interested in both and the future will be a hybrid of both," Altman said when answering a question about small, specialized models versus broad models during a Q&A session at Station F earlier this year.

There are also ethical and legal implications with AI usage. Regulation is still very much evolving in real time, but European legislation in particular could encourage companies to use AI models trained on very specific data sets and in very specific ways.

"Gartner says that 75% of enterprises are shifting from [proofs of concept] to production in 2024. So the next year or two are probably some of the most seminal moments in the history of AI, where we are finally getting into production using probably a mixture of open-source foundational models fine tuned on proprietary data," Tahir told me.

"The value of MLOps is that we believe that 99% of AI use cases will be driven by more specialized, cheaper, smaller models that will be trained in house," he added later in the conversation.

Image Credits: ZenML

Date
02/12/2024
Date Created
24/10/2023
Author

susantswain1