Architecture Secret AI is implemented as an off-chain Confidential Computing layer on top of Secret Network.

The Confidential Computing Layer consists of off-chain workers that utilize the NVIDIA Confidential Computing technology paired with Intel TDX and AMD SEV confidential VMs to perform AI tasks, such as inference, fine-tuning and training.

Secret AI supports multiple Services, with each Service offering a particular set of LLM models and functionality, pricing, access policy and more. Each Service can have multiple Confidential GPU workers that share the load of serving the users.

To join the network, Workers register with the WorkerManager contract and perform attestation that is stored on-chain and available for any interested third party to validate the authenticity of the workers' hardware and software.

To access a Service, the users interact with the WorkerManager contract in order to get the address of the next available GPU worker, and establish direct connection with it. The Worker validates that the user has a valid subscription, and provides the required service.

Workers report the amount of work (e.g. number of input and output tokens) to the RewardsManager contract, and the user payments are distributed between the workers' operators according to their share of work performed.

The architecture also includes a decentralized and encrypted PRAG (Private RAG) database, that will allow users to share confidential data that can be later used for Retrieval-Augmented Generation on the participant LLMs. Use cases for such PRAG data may include medical data, private email archives, financial transaction data and more.

Next, dive into the Secret AI SDK to learn how to get started!

Previous Introduction Next Secret AI SDK

Last updated11 days ago

Was this helpful?