FAQ

What are the minimum hardware requirements to run it locally?

To ensure a smooth operation of Cardano Smart locally, the following minimum system requirements are recommended:

Processor: A minimum of a 4-core processor is recommended. An 8-core processor or better is ideal if you plan to run the system without a dedicated GPU.
RAM: At least 16 GB of RAM is necessary to handle larger models and maintain optimal performance.
Disk Space: Ensure you have at least 10 GB of free space available for storing models and necessary updates.
GPU (Optional): For enhanced performance, an Nvidia GPU with a compute capability of 6.0 or higher is advisable. You can verify the compute capability of your Nvidia GPU here (https://developer.nvidia.com/cuda-gpus). If a suitable GPU is not available, the system can still operate in CPU mode, although this will be significantly slower.

Using a GPU accelerates processing and is highly recommended for efficiency, but is not strictly required if the CPU is sufficiently powerful.

For a cloud set-up, what specs for the server are recommended?

For optimal cloud deployment of the Cardano Smart project on a Kubernetes cluster, the following specifications are recommended for the servers:

GPU Nodes: At least one node equipped with an Nvidia Tesla T4 GPU or equivalent, complemented by at least 16 GB of regular RAM and 2 CPU cores. This setup is crucial for handling compute-intensive tasks efficiently.
Standard Nodes: Additional nodes should have at least 16 GB of RAM and 2 CPU cores to support other essential workloads seamlessly.
Lightweight Workload Node: For less demanding tasks, a node with a minimum of 2 GB RAM may suffice.

The project leverages Kubernetes' capabilities to autoscale both vertically and horizontally, allowing for dynamic resource allocation based on workload demands. For environments expecting a high number of concurrent users, it is advisable to deploy additional GPU nodes. Consider configuring three or more GPU nodes to ensure robust performance and availability.

How expensive is it to run on a cloud server?

Running the Cardano Smart project on a cloud server such as Google Cloud Platform (GCP) can be quite costly, particularly due to the resources required for optimal performance. For a typical setup on a private dedicated Google Kubernetes Engine (GKE) cluster with three GPU nodes, the costs can exceed $1000 per month. This estimate is based on continuous usage and includes the expenses associated with high-performance GPU nodes.

The actual cost may vary depending on usage patterns and the number of active users, as Kubernetes' auto-scaling features can increase or decrease resource allocation dynamically, thus affecting the overall cost. Given these factors, deploying on a local Kubernetes cluster or a powerful local PC might be a more cost-effective solution.

In some places, DeepSeek is blocked or banned for use. Can you provide a link to the previous version with Llama?

If certain models like DeepSeek are restricted in your region, you can seamlessly switch to alternative models such as Llama, ensuring compliance with local regulations. It's important to note that our project uses open source versions of these models that are hosted directly on your own hardware or private cloud infrastructure, not through any third-party APIs. This setup provides full control over the models and data, enhancing security and privacy.

To switch models in your deployment, you can specify your preferred model when running the build script for the privateGPT component. For example, to use the Llama model, execute the following command:

```
./build_scripts/build_private_gpt.sh "llama3.2"
```

This command configures the privateGPT component to deploy the "llama3.2" model. You can update or change the model at any time by rebuilding and redeploying using the build script. Detailed instructions for building with different models are available in the `README.md` file of the project repository. By hosting these models yourself, you maintain autonomy over your deployment and avoid any dependencies on external model providers.

Is an Nvidia GPU needed to run locally?

No, an Nvidia GPU is not strictly required to run the project locally, but it is highly recommended for optimal performance, especially when processing larger models. If you do not have an Nvidia GPU, you can still run the project using a more powerful CPU. Here are the recommendations:

With GPU: An Nvidia GPU with Compute Capability 6.0 or higher significantly enhances performance, allowing for faster processing and better handling of complex computations.

Without GPU: If you do not have access to an Nvidia GPU, you can run the project on a CPU. For a smooth experience, it is advisable to use a CPU with at least 8 cores. However, keep in mind that running on CPU mode will be much slower compared to using a GPU.

Can you add Helios? (this is what they mean https://www.hyperion-bt.org/)

Yes, Helios integration is on our roadmap. We are soon developing a new scraping spider specifically for Helios, making use of their comprehensive documentation. This update will be available shortly.

Is it better to use it for a complete dApp? Or is it better to ask particular questions?

This tool is best utilized as an assistant for developers rather than a replacement. It excels at providing real-time, up-to-date Cardano documentation and generating code snippets, which can significantly streamline the development process. While it supports various aspects of developing a full dApp, it's particularly effective for answering specific questions and solving challenges as they arise during development. This ensures that developers have access to the latest information and support right when they need it.