# Seraphnet: Ideologically-Transparent Playground for Generative AI Apps

Maciej Szafarczyk
CPO
team@seraphnet.ai

Piotr Malicki
CTO
dev@seraphnet.ai

**ABSTRACT: In a world where pixels flickering on the screen often affect our reasoning more than the physical stimuli, where AI agents already may influence nearly every aspect of human perception, and where fallout from incidents like Cambridge Analytica still lingers unaddressed, there is a growing need to distinguish digital propaganda from facts. Unfortunately most commercial generative AI apps prohibit users from accessing data in an unlimited way, instead offering a more moderated and shallow experience.**

**To address this, we introduce Seraphnet - an ideologically transparent generative AI (GenAI) apps playground, designed to unlock the full potential of contemporary large-scale open source and commercial neural networks. This playground is populated by Swarm Pods - GenAI apps called Clearpills, orchestrated by a Swarm Manager. Clearpills are designed to be compatible with multiple Large Language Models (LLMs) and are able to bypass the limitations imposed by Big Tech companies. Additionally, Seraphnet utilizes RAG technology in order to accurately validate external data, including on-chain information. Furthermore, the LLMOps part of the infrastructure called Forge allows users to generate their own Swarm Pods with ease - either using Seraphnet's premium services or completely on their own. The end goal is an open, scalable infrastructure that produces comprehensive, multi-dimensional and unbiased data.**

***INDEX TERMS*: Generative AI, RAG, LLM, Blockchain, Bias in AI**

## I. BACKGROUND

We live in times of a fragmented society. Since the dawn of the human race, various ideologies have competed against one another in a cutthroat race to dominate the minds of its followers. Whether through a religious, political or cultural angle, there are almost as many ways to explain the complexity of existence and the world we are immersed in as there are human beings.

This plurality of thought is often a beautiful thing, allowing the very fabric of vibrant societies to exist. However, there is also an underlying, very primal issue that comes with the aforementioned clash of ideologies - a tendency to favor a specific one over the other. For if I am right, the other side must be wrong - otherwise, it's just contradictions. And in times where the information transfer speed is ever increasing, ambiguity is more and more difficult to handle.

This mechanism of preferring a specific ideology, if not addressed, creates a scenario where people become more and more ingrained within their specific worldviews. Their own Platonic Caves. If we couple this with the rapidly increasing power of AI that is intransparent, biased and doesn't encourage critical thinking, we have a ticking bomb on our hands.

We have built Seraphnet in an attempt to overcome this. Each of the Clearpills, also called Swarm Pods, are generative AI applications that can learn from existing data sources to generate new, realistic artifacts at scale that reflect the characteristics of the training data but don't repeat it. They can produce a variety of novel content, such as images, video, music, speech, text, software code and product designs. In Seraphnet V1 we are focused on text-based GenAI apps. Clearpills can become your personal, contemporary equivalent of the Pythia Oracle, designed to provide customizable, unfiltered information harnessed from the raw power of various unlocked LLMs connected to multiple data sources. Seraphnet is our first step on a long road to liberation of human-machine interaction.

## II. PROBLEM

The way contemporary Generative AI applications, such as OpenAI's ChatGPT, are designed effectively positions them as gatekeepers between end-users, and the raw power of the underlying large neural networks. Rather than providing comprehensive answers, users receive half-baked shards of Data Lakehouse and are prohibited from obtaining a critical, multi-dimensional perspective. This fragmentation results in moderated, shallow, surface level human-AI interactions that are suboptimal and fail to harness the true capabilities of these powerful LLMs.

In our opinion this stance is completely irrational, one-sided and patronizing towards both developers and users. It also reflects the centralized, corporate-owned and profit-driven philosophy that drives the entities developing these sophisticated AI products. [1]

Moreover, the current status quo is further cemented by the fact that the process of setting up GenAI apps is a complicated, expensive and time consuming task for both developers and end-users, especially from the Large Language Model Operations (LLMOps) perspective. This hinders the development of a more open and democratized AI landscape,

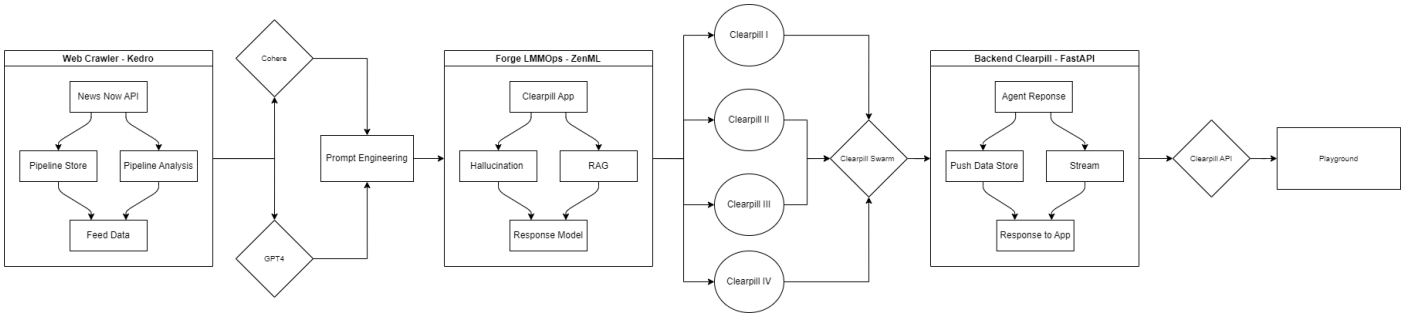


**Fig1: Seraphnet's V1 Architecture Overview**

stifles innovation, limits competition, and perpetuates the existing limitations on user access and control.

## III. SOLUTION

We propose an alternative that tackles all of the previous limitations. Our infrastructure is composed of a multi variety playground for GenAI apps called Seraphnet. It is directly connected to a network of specialized, prebuilt GenAI apps called Clearpills, acting as Swarm Pods in an interconnected network orchestrated by a Swarm Manager.

In addition to Clearpills, there is also a possibility to add community-sourced GenAI apps. Together, these components are able to tap into multiple LLMs, access information using web crawlers and fine-tune it with RAG technology. We also provide automated LLMOps infrastructure that decreases the complexity and time needed to develop new GenAI apps, allowing users to focus on innovation and experimentation and not technical obstacles.

The main differentiating element of our infrastructure from most commercial solutions is the lack of ideological bias. Inspired by the concept of "Clear pill" introduced by a philosopher named Mencius Moldbug [2], we strive to provide users with data in an unlimited way, trusting that they will do the moderating part on their own. This approach embraces user autonomy, transparency and collaboration.

The graph presented above shows the architecture of Seraphnet V1 - our first attempt at turning the concept into an efficient and modular network of GenAI apps, able to operate without any ideological limitations. Seraphnet's architecture represents a paradigm shift towards a more open, democratized, and user-centric approach to generative AI.

### A. DATA SOURCES

Seraphnet's V1 utilizes a ready-made API that refers to global data. This raw data is then processed by the data pipeline. Nextly, Seraphnet leverages existing web crawling frameworks to create web crawlers which fetch data from various news sources such as: Google News, Bing News, BBC, CNN, Yahoo, MSN, New York Times, The Washington Post etc. The raw data coming from these sources allows us to facilitate prompt engineering for the LLM integration.

Seraphnet's data sources are designed to be flexible and extensible, allowing for the integration of various new information sources as the ecosystem evolves. This plurality of sources is key to make sure the infrastructure remains ideologically neutral.

### B. DATA ARCHITECTURE

Seraphnet utilizes the Dremio Data Lakehouse [3] architecture, built on top of Apache Iceberg, to distribute the data. The Data Lakehouse is created with the Kedro data engineering framework, a toolbox for production-ready data science. [4] Through its implementation of Kedro, Seraphnet's data pipeline follows a structured and modular approach. The information gathered from various sources is first processed through the pipeline store, where it undergoes data ingestion, cleaning, and preprocessing steps.

Once the data is ingested and preprocessed, it enters the pipeline analysis stage where techniques such as natural language processing, information retrieval, and data mining, (depending on the specific requirements of the Clearpills and the user's queries), are applied.

The processed and analyzed data is then fed into the LLM; here, it leverages the curated and enriched information to generate relevant and insightful responses.

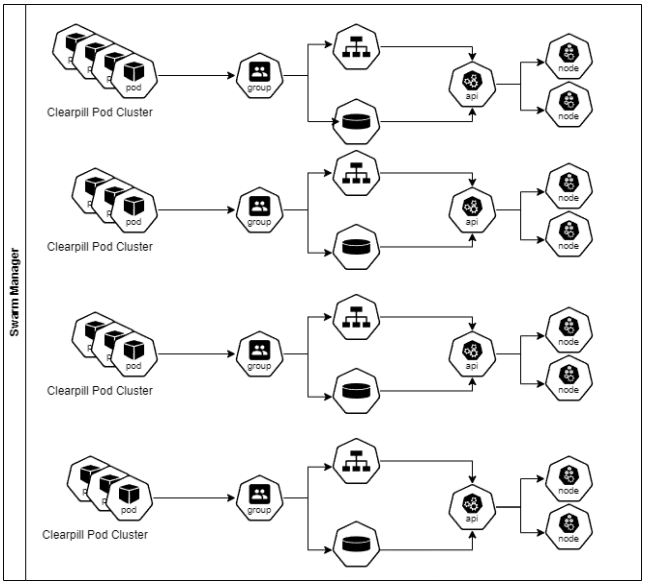


**Fig2: Swarm Manager Architecture**

### C. SWARM MANAGER & PODS

Seraphnet's infrastructure centers around a Swarm Manager which is similar to Docker Swarm Manager [5], an entity that orchestrates and manages microservices

architecture. Swarm Manager's role is to manage multiple Swarm Pods in order to find the most optimal pathway across the techstack. It analyzes the user's intent by determining probability statistics and then compares the results against the network of pre-established GenAI apps, LLMs and web crawlers, to figure out which one is the most capable of handling the task. This microservices infrastructure is analogous to the ArgoCD [6] framework.

Swarm Pods are specialized GenAI apps. They are able to access vectorized data bases to find relevant text to the user's prompt or query and can handle specific, well-defined tasks individually. However, for more complex, multi-criteria tasks, a cluster of Clearpills is employed, enabling them to collaborate and collectively tackle intricate challenges.

Clearpills are grouped, hierarchized and fed into the API provider to be finally hosted as nodes. In V1 of Seraphnet, each Clearpill may have a different target function that includes: GenAI text processing, moderating content or selecting the most important facts for the moment from the subject matter the user selects.

V1 of Seraphnet implements FastAPI [7]: a modern, fast, web framework for building APIs with Python 3.10 based on standard Python type hints. Additionally, to simplify packaging and dependency management, we have implemented the Poetry [8] tool into our techstack.

FastAPI not only enables seamless communication but also provides a dual-use capability, catering to both non-technical users and developers. Importantly, FastAPI partakes in the pricing mechanism for the $DLLM token associated with Seraphnet's ecosystem.

The synergy between the Swarm Manager and Clearpills enables Seraphnet to provide a comprehensive and tailored solution, leveraging the collective intelligence of multiple components to deliver accurate, relevant, and insightful responses to users' queries.

## D. SERAPHNET PLAYGROUND

To meet the needs of non-tech users, the playground for Seraphnet's Clearpill GenAI apps is the most user-friendly part of the architecture. In essence, it is a front-end User Interface that is connected to the Clearpill API located in the backend.

It allows users to communicate with the intent-sensitive Swarm Manager, upload their own GenAI apps, verify newly created Clearpills, choose the most optimal prebuilt Clearpills and look for possible integrations and synergies to achieve various objective functions. The Playground UI panel also offers features such as result history, allowing users to review and compare previous outputs, as well as customization options for adjusting parameters like output length or level of detail.

## E. FORGE LLMOPS

Behind each Clearpill stands a powerful multilayered Large Language Model Operations (LLMOps) infrastructure. LLMOps encompasses the practices, techniques and tools used for the operational management of large language models in production environments [9]. Since most LLMs are not built from scratch but rather, existing foundation models are fine-tuned by developers for various purposes, these practices are extremely important. Some of the processes include: deployment, monitoring and maintenance. Its objective is to oversee that the response model remains efficient and accurate by a combination of high quality prompt engineering and verifying LLMs hallucinations with the RAG module.

Usually, these processes are handled by entire teams made of data scientists, DevOps engineers, and IT professionals where they collaborate on data exploration, prompt engineering, and pipeline management.

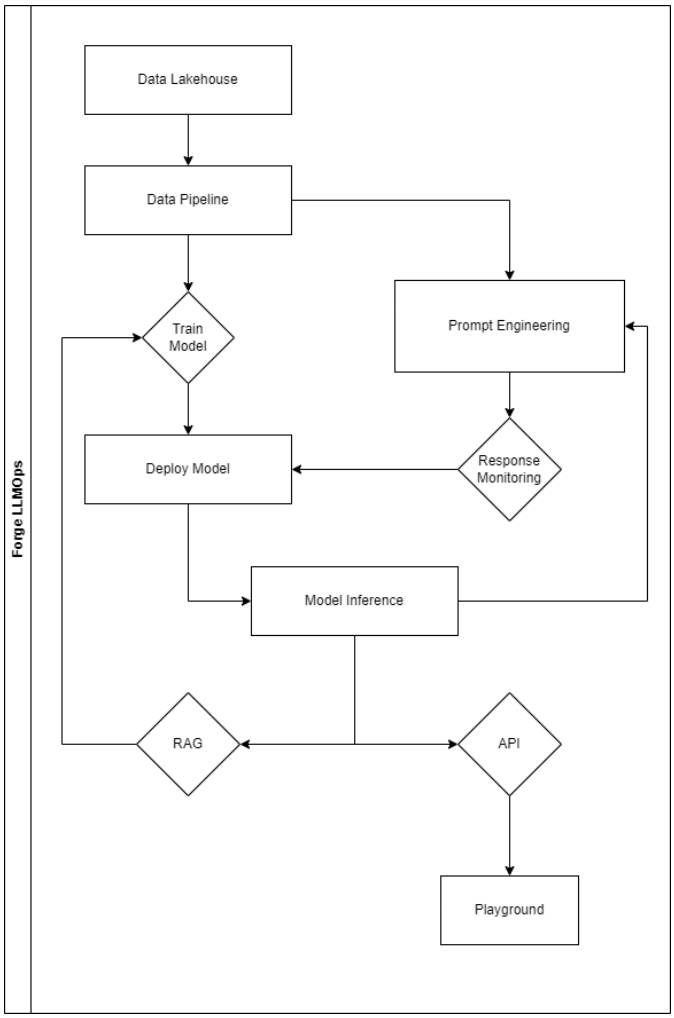


**Fig3: Forge LLMOps**

We provide an automated alternative that integrates all processes starting from reproducible microservices containing Clearpill, fine-tuning of LLM models, Retrieval-Augmented Generation (RAG), and APIs. We offer a self-hosted open source version of our LLMOps platform created based on open source frameworks, which are equivalents to commercial frameworks. Within the LLMOps platform, we can create Clearpills based on a prototype provided by us, allowing for its customization to tasks through our own fine-tuning or RAG, as well as the replacement of LLMs or Data Sources.

In V1, Forge is compatible with OpenAI's GPT4 transformer-based model but in the future we plan to expand it to be compatible with more LLMs. To handle LLMOps

smoothly, we use the BentoML [10] framework, an integral part of ZenML [11], an extensible, open-source MLOps framework for creating portable, production-ready machine learning pipelines. BentoML assists in handling computation tasks such as data preprocessing, model training, inference, and deployment, ensuring that the Seraphnet ecosystem can scale and adapt to increasing demand and evolving requirements.

We are confident that by designing the infrastructure for large language model operations in this manner, we are offering users a new level of quality. This enables the delivery of stable and, most importantly, accurate GenAI apps that are not constrained by any specific ideology.

## F. RAG

Retrieval-Augmented Generation or RAG [12] for short is a technique for enhancing the accuracy and reliability of generative AI models with facts fetched from external sources to ground large language models (LLMs) on the most accurate, up-to-date information and to give users insight into LLMs' generative process. Implementing RAG gives the model fewer opportunities to pull information baked into its parameters. This reduces the chances that an LLM will leak sensitive data, or hallucinate incorrect or misleading information.

Additionally, to further improve our RAG module, we have integrated it with semantic search. This technology might drastically improve the search results provided by RAG, especially in the context of organizations wishing to add sizable sources of information to their Swarm Pods. Semantic search can scan vast databases in search for various specific information and pick data in a precise manner, which increases the quality of responses generated.

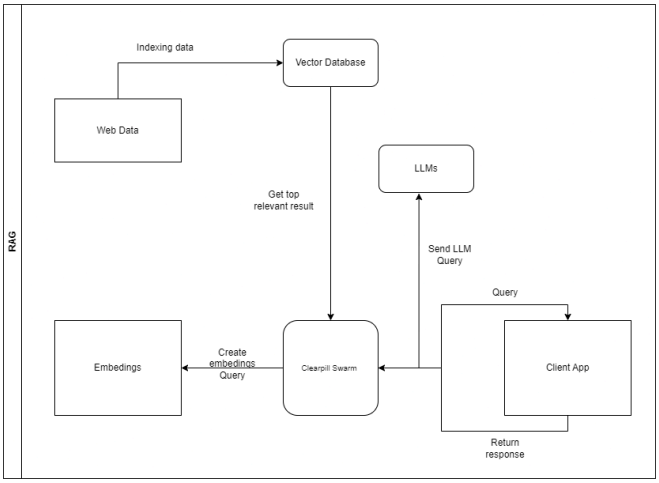


**Fig4: RAG Architecture**

In Seraphnet's V1 for RAG we use Cohere [13], which is an efficient and secure toolkit that allows LLMs to accurately answer questions and solve tasks using enterprise data as the source of truth. We also implemented NVIDIA Triton Inference Server [14], an open-source software that standardizes AI model deployment and execution across every workload and NVIDIA TensorRT-LLM [15], which provides users with an easy-to-use Python API to define Large Language Models (LLMs) and build TensorRT engines that contain state-of-the-art optimizations to perform inference efficiently on NVIDIA GPUs. TensorRT-LLM also contains components to create Python and C++ runtimes that execute those TensorRT engines.

Combining semantic search with Cohere's and Nvidia's solutions, Seraphnet's RAG module allows to reduce the risk of LLM hallucinations and data leakage, while adding an ability for our infrastructure to draw accurate and up-to-date information from various sources. This key in providing ideologically transparent solutions.

## G. ONCHAIN DATA

In addition to traditional data sources, it is our intention to ensure that the Web3-compatible versions of Seraphnet's Clearpills, called dClearpills, are capable of accessing and leveraging onchain data sources. Blockchain-based research in the field of LLM improvement has been significant in recent years and as Seraphnet we strive to be at the forefront of innovation. [16][17][18][19]

Onchain data refers to all the information recorded and stored directly on blockchain networks .[20] This encompasses a wide range of data types, including but not limited to:

- Transaction records: analyze transaction patterns, volumes and flows across different blockchains; identify whale wallets, money laundering activities and other anomalies
- Smart contract data: code, state variables, and event logs
- Metadata of onchain assets, including NFTs and tokenized real-world assets
- Governance data from decentralized autonomous organizations (DAOs): proposal details, voting records, and decision-making data; governance processes, participation levels, and decision outcomes

Tapping into these transparent and immutable data sources will allow dClearpills to analyze transaction patterns, smart contract code metadata to identify various trends, risks or opportunities.

## H. ONCHAIN INTERACTIONS

Beyond being able to access onchain data, dClearpills will be able to directly interact with smart contracts [21]. This interaction layer enables dClearpills to not only read data from blockchains but also execute transactions and invoke functions. Some exemplary use cases for onchain interactions for dClearpills may include:

- Decentralized finance (DeFi) applications: executing trades on decentralized exchanges (DEXs), providing liquidity to lending/borrowing protocols, participate in yield farming strategies
- Non-Fungible Token (NFTs) Management: mint new NFTs, transfer ownership, or manage their NFT assets directly through the AI interface
- DAO Participation: participate in governance processes, vote on proposals, and manage their DAO memberships by staking/unstaking tokens
- dApps Interactions: interact with various dApps built on blockchain networks through a unified interface

- Onchaing Gaming: act as intelligent agents in Web3 compatible titles, able to make strategic decisions within the logic of the games

Implementing smart contract logic interaction possibilities into Clearpills allows to bridge the gap between the worlds of blockchain and Artificial Intelligence technology and to foster a new generation of decentralized Generative AI apps.

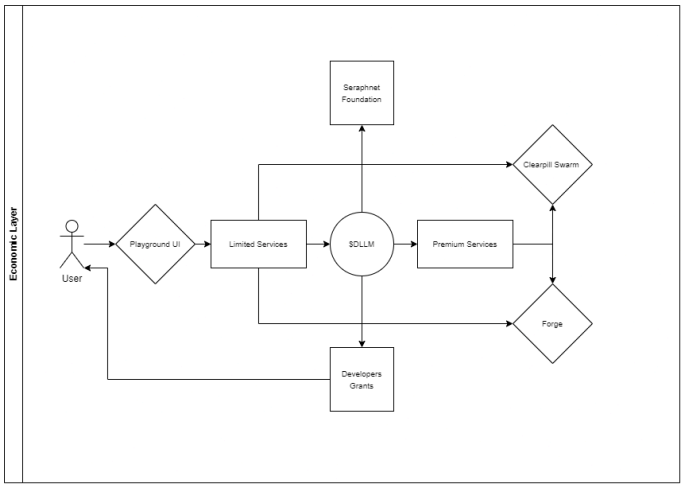


**Fig5: Economic Layer**

## IV. ECONOMIC LAYER

Seraphnet's economic layer is designed to be sustainable and incentivizing for developers and users, to encourage both access and contribution to infrastructure. Access to infrastructure is possible via two pathways: premium and limited. The premium option requires holding a $DLLM token, the native utility token of Seraphnet ecosystem, and enables developers to access and utilize additional ecosystem features, such as:

- Data retrieval: developers can get data from various sources, including onchain data from various blockchains, to train and fine-tune their GenAI apps
- Inference: ability to run inference on GenAI apps, allowing real-time analysis and decision making
- Fine-tuning: users may employ custom data and techniques to allow more accurate and specialized GenAI apps performance
- Training processes: access to resources allowing the training of GenAI apps from scratch, or continuing the training processes of existing apps

Seraphnet operates on a Pay-As-You-Go (PAYG) business model, providing flexibility and scalability to users based on their specific needs. This approach allows users to pay only for the resources they actually consume, making the interaction with the infrastructure as cost-effective as possible.

Furthermore, the $DLLM token is used to deploy new Generative AI (GenAI) apps with Forge and is an integral part of our upcoming Developers Grants program, designed to provide financial support to promising projects and fostering a vibrant and innovative developer community.

Lastly, part of the income generated from access to the premium services will be allocated to the Seraphnet Foundation. These funds will be reinvested into R&D efforts to ensure advancement and improvement of the platform's capabilities and technologies.

Our economic layer is designed to provide all the necessary resources and tools to build innovative GenAI applications and support an entire ecosystem.

## V. USE CASES

Seraphnet offers a variety of possible use cases, for GenAI enthusiasts, researchers, professional developers and regular end users alike. Providing the flexibility to create custom GenAI apps or leverage pre-built solutions, Seraphnet empowers users to achieve a wide array of objectives. Here we are proposing a set of possible use cases, with more to come in the future.

### A. NON-TECHNICAL USERS

Users who do not possess technical knowledge can access Seraphnet through the User Interface of the Playground. Over there it is possible to interact with pre-built Clearpills, designed for specific use cases, such as question-answering, content generation, data analysis and more.

By making sure the infrastructure is ideologically transparent, we help end users to bypass the limitations imposed by contemporary GenAI apps to produce accurate, factual-based data of customizable scope coming from multiple sources.

Some ideas may include:

- Personal research assistant able to get any information, customized to the researcher's needs
- Unlimited creative writing companion that can brainstorm any topic free from creative constraints, in the name of ars poetica
- Realistic financial assistant - a valuable resource for analyzing complex financial decisions and investments, providing comprehensive insights and recommendations. This could aid individuals in making informed choices regarding personal finance, investment strategies, or business decisions.
- Customized content generation: being able to access different data sources, Seraphnet can generate highly customized content tailored to specific needs (marketing, educational or entertainment)

### B. DEVELOPERS

Seraphnet acts as a platform where tech-savvy users can experiment with cutting-edge LLMs and Clearpills to explore various capabilities of ideologically-transparent generative AI. Enthusiasts and developers can utilize our open source LLMOps infrastructure to offset the development time and infrastructure costs on us for a fraction of the total price or they can host and run custom Swarm Pods on their own.

Potential use cases:

- Custom GenAI apps development: professionals can leverage Seraphnet's LLMOps infrastructure and the ability to host custom Swarm Pods to build and

deploy their own apps, expanding the potential of the Clearpills built by the Seraphnet's team
- Multi LLM Experimentation: since Seraphnet aims to be compatible with multiple LLMs, developers can pick the best one for their project's needs
- Benchmarking and testing: by making sure the infrastructure is open and transparent, developers will be able to make various tests and compare their results against an ever-growing database
- Onchain experimentations: dClearpills aim to be compatible with Web3 data sources and smart contracts, allowing developers to create GenAI apps that are both offchain and onchain: for example, social media analyzing tool, that is able to connect Web2 data with onchain interactions.

The provided examples are just a small sampling from a vast, expansive ocean of possibilities of GenAI technology. As exciting as the current state-of-the-art is, we are really just beginning this journey into a new era of human-AI co-creation and collaboration. The most transformative and game-changing use cases have likely not even been conceived or imagined yet. That's why we are so eager and looking forward to collaborating deeply with end users, developers, subject matter experts, and creative minds across all industries and domains.

By combining the incredible capabilities of these generative AI systems with human ingenuity, domain expertise, and creativity, we can push the boundaries and envelope of what's possible. We can explore entirely new frontiers that unlock novel solutions to long-standing challenges and spark unprecedented innovations that will positively impact businesses, science, art, education, and society as a whole.

## VI. CONCLUSION

Seraphnet offers a revolutionary approach to generative AI applications, offering an ideologically-transparent, modular, community-centric GenAI app network overseen by a Swarm Manager. By analyzing the user's intent in an user-friendly Playground, it employs the most capable apps, called Clearpills, to act independently or collaboratively to execute sophisticated text-based generative AI tasks. In order to achieve maximum impact, they utilize the unlocked power of various LLMs, and additionally refine the results with RAG technology.

Besides just Swarm Pods and Manager, Seraphnet also offers a complex LLMOps infrastructure built using the latest cutting-edge technologies that allows it to produce new GenAI apps with or without our assistance. While there is still much research and development ahead, especially in areas like onchain data and interactions, Seraphnet's complex LLMOps infrastructure for producing new GenAI apps positions it as a prime choice. Its premium token-gated functions help offset costs of deploying and hosting these apps for both regular users and professional GenAI developers alike on this extraordinary journey to shape generative AI's next transformative phase.

## REFERENCES


[1] *Is ChatGPT biased?* (n.d.). OpenAI Help Center. Retrieved March 30, 2024, from https://help.openai.com/en/articles/8313359-is-chatgpt-biased

[2] The Clear Pill, part 1 of 5: The four-stroke regime. (2019, September 27). *The American Mind*. https://americanmind.org/salvo/the-clear-pill-part-1-of-5-the-four-stroke-regime/

[3] *Data lakehouse*. (2023, December 8). Dremio. https://www.dremio.com/solutions/data-lakehouse/

[4] *Kedro*. (n.d.). Kedro. Retrieved March 30, 2024, from https://kedro.org/

[5] *"Swarm mode overview."* (2024, February 9). Docker Documentation. https://docs.docker.com/engine/swarm/

[6] *Argo CD*. (n.d.). Declarative GitOps CD for Kubernetes. Retrieved March 30, 2024, from https://argo-cd.readthedocs.io/en/stable/

[7] *FastAPI*. (n.d.). Retrieved March 30, 2024, from https://fastapi.tiangolo.com/

[8] *Poetry*. (n.d.). Python Dependency Management and Packaging Made Easy. Retrieved March 30, 2024, from https://python-poetry.org/

[9] *What is LLMops*. (n.d.). Retrieved March 30, 2024, from https://www.redhat.com/en/topics/ai/llmops

[10] *BentoML: Build, ship, scale AI applications*. (n.d.). Retrieved March 30, 2024, from https://www.bentoml.com/

[11] *MLOps framework for infrastructure agnostic ML pipelines*. (n.d.). Retrieved March 30, 2024, from https://www.zenml.io/

[12] Merritt, R. (2023, November 15). *What Is Retrieval-Augmented Generation aka RAG*. NVIDIA Blog. https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/

[13] *Home*. (n.d.-a). Cohere. Retrieved March 30, 2024, from https://cohere.com/

[14] *Triton inference server*. (n.d.). NVIDIA Developer. Retrieved March 30, 2024, from https://developer.nvidia.com/triton-inference-server

[15] NVIDIA. (n.d.). *GitHub - NVIDIA/TensorRT-LLM*. Retrieved March 30, 2024, from https://github.com/NVIDIA/TensorRT-LLM

[16]Cibin, N., & Albano, M. (2023, July). Blockchain-Based Platform for Crowdsourcing Machine Learning Models Design and Training, while Incentivizing Continuous Improvement. *2023 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPS)*. http://dx.doi.org/10.1109/dapps57946.2023.00014

[17] Gong, Y. (2023, July 20). *Dynamic large language models on blockchains*. arXiv.Org. https://arxiv.org/abs/2307.10549

[18] Park, S., Lee, J., & Moon, S.-M. (2023, May 6). A blockchain-based platform for reliable inference and training of large-scale models. arXiv.Org. https://arxiv.org/abs/2305.04062

[19] Blockchain in the age of LLMs. (n.d.). Retrieved March 31, 2024, from https://www.propellerheads.xyz/blog/blockchain-and-llms

[20] P, M. (2023, November 4). What is Onchain Data? A Beginner's Guide (2023). *Thirdweb*. https://blog.thirdweb.com/onchain-data/

[21] *What are smart contracts on blockchain?* (n.d.). IBM. Retrieved March 30, 2024, from https://www.ibm.com/topics/smart-contracts