

# LLM+RAG Local Platform

MVP Test Results

*Arx Industries*



192.168.4.34:8000



## Local RAG AI System i



what is inference and why can it be such a large expense?



# Local RAG AI System i



You

what is inference and why can it be such a large expense?

Assistant

Thinking...

Ask me anything...





192.168.4.34:8000



## Local RAG AI System i



what is inference and why can it be such a large expense?

Assistant

Direct Answer: Inference refers to the process of making predictions with a trained machine learning model. It can be a significant expense because it requires computational resources, such as GPU time, to run the model on new data. Supporting Details from the Context: The context mentions that inference is an important aspect of machine learning, and it can be a large expense. For example, in the section on "Compute" under the "AI Glossary", it states that compute refers to the computational resources used in training or running AI models, including GPU time. Additionally, in the section on "Neural Network" under the same glossary, it explains that neural networks are trained using large amounts of data and computational power. This suggests that inference, which is a critical step in the machine learning process, requires significant computational resources. Implications or Considerations: The expense of inference can have implications for organizations that rely heavily on AI models to make predictions or decisions. For example, if an organization has a large-scale AI model that needs to be run repeatedly to generate predictions, it may require significant computational resources and infrastructure investments. This could lead to increased costs and potentially impact the organization's bottom line. Furthermore, the expense of inference can also have implications for data privacy and security. As AI models become more complex and rely on larger amounts of data, they may require more sensitive information to be processed, which can raise concerns about data protection and security.

SOURCES

Ask me anything...



# Local RAG AI System i



or decisions. For example, if an organization has a large-scale AI model that needs to be run repeatedly to generate predictions, it may require significant computational resources and infrastructure investments. This could lead to increased costs and potentially impact the organization's bottom line. Furthermore, the expense of inference can also have implications for data privacy and security. As AI models become more complex and rely on larger amounts of data, they may require more sensitive information to be processed, which can raise concerns about data protection and security.

## SOURCES

- Chunk 1 · Score 0.424

**AI Glossary \_ Andreessen Horowitz.pdf**

Explainable AI (XAI) A subfield of AI focused on creating transparent models that provide clear and understandable explanations of their decisions. Fine-tuning The process of taking a pre-trained machine learning model that has already been trained on a large dataset and adapting it for a slightly d...
- Chunk 2 · Score 0.417

**AI Glossary \_ Andreessen Horowitz.pdf**

AI Glossary by a16z ABCDEFGHILMNOPRSTUVXZ Accelerator A class of microprocessor designed to accelerate AI applications. Agents Software that can perform certain tasks independently and proactively without the need for human intervention, often utilizing a suite of tools like calculators or web brows...
- Chunk 3 · Score 0.403

**AI Glossary \_ Andreessen Horowitz.pdf**

NeRF (Neural Radiance Fields) A method for creating a 3D scene from 2D images using a neural network. It can be used for photorealistic rendering, view synthesis, and more. Neural Network A type of AI model inspired by the human brain. It consists of connected units or nodes—called neurons—that are ...
- Chunk 4 · Score 0.386

**nist.ai.100-1.pdf**

These actors are parties that are concerned with the impact and sustainability of the organization as a whole. Additional AI Actors Third-party entities include

Ask me anything...



## Local RAG AI System



Chunk 1 · Score 0.424

### AI Glossary \_ Andreessen Horowitz.pdf

Explainable AI (XAI) A subfield of AI focused on creating transparent models that provide clear and understandable explanations of their decisions. Fine-tuning The process of taking a pre-trained machine learning model that has already been trained on a large dataset and adapting it for a slightly d...

Chunk 2 · Score 0.417

### AI Glossary \_ Andreessen Horowitz.pdf

AI Glossary by a16z ABCDEFGHILMNOPRSTUVXZ Accelerator A class of microprocessor designed to accelerate AI applications. Agents Software that can perform certain tasks independently and proactively without the need for human intervention, often utilizing a suite of tools like calculators or web brows...

Chunk 3 · Score 0.403

### AI Glossary \_ Andreessen Horowitz.pdf

NeRF (Neural Radiance Fields) A method for creating a 3D scene from 2D images using a neural network. It can be used for photorealistic rendering, view synthesis, and more. Neural Network A type of AI model inspired by the human brain. It consists of connected units or nodes—called neurons—that are ...

Chunk 4 · Score 0.386

### nist.ai.100-1.pdf

These actors are parties that are concerned with the impact and sustainability of the organization as a whole. Additional AI Actors Third-party entities include providers, developers, vendors, and evaluators of data, al- gorithms, models, and/or systems and related services for another organization ...

Chunk 5 · Score 0.383

### AI\_RMF\_Playbook-1.pdf

2022. Inioluwa Deborah Raji and Jingying Yang. "ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles." arXiv preprint, submitted January 8, 2020. Andrew Smith. "Using Artificial Intelligence and Algorithms." Federal Trade Commission Business B...

Copy Regenerate

Ask me anything...

