Smarter technology for all

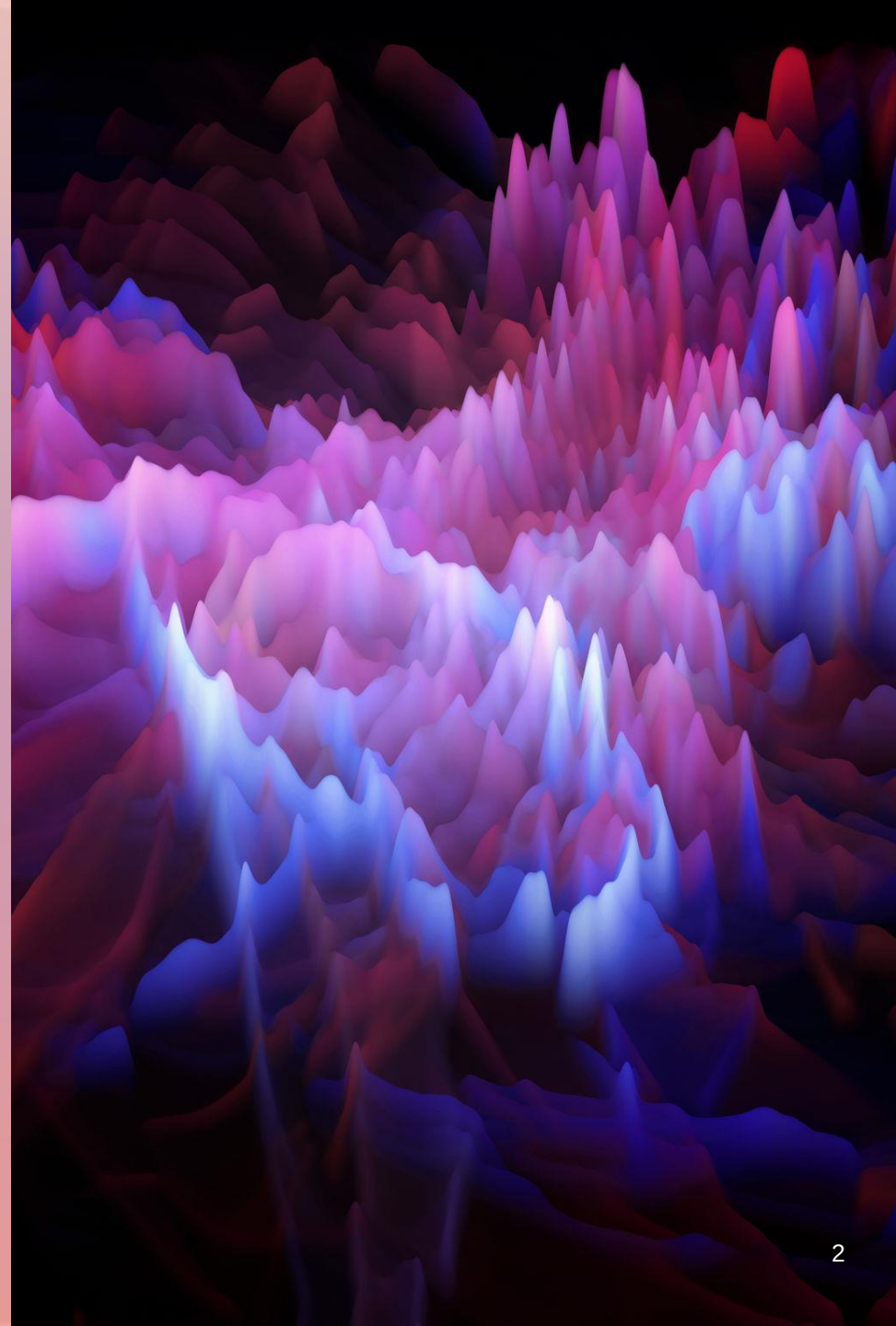
LLM in a Box

Introduction to Lenovo on-premises AI Solution and Service Package

Lenovo

Empower your AI with On-Premises LLM Deployment, Seamlessly and Securely

Our private LLM (Large Language Model) server in a box provides an off-the-shelf solution for businesses and organizations to deploy and utilize advanced AI capabilities on-premises. Powered by state-of-the-art Large Language Models, this compact server enables you to run high-performance language models securely within your own infrastructure, without the need to send sensitive data to the cloud



Value Proposition and Benefits

With our private LLM model server in a box, you can unlock the power of advanced LLM's within your organization, while maintaining the highest standards of data security and sovereignty.

Retain full control and ownership of your data and language models

Leverage cutting-edge LLM AI capabilities on-premises

Ensure data privacy and compliance with regulatory requirements

Reduce costs and overhead associated with cloud-based solutions

Accelerate your AI-driven initiatives without infrastructure constraints

Seamlessly integrate with your existing technology stack

Pain Points

Artificial Intelligence

Scalability

Adapting AI solutions to different environments or scaling them to meet demand can be complex

Privacy

Handling sensitive data with AI models raises privacy concerns, especially if data is not properly anonymized.

Transparency

The "black box" nature of AI models makes it difficult to understand how decisions are made, complicating accountability.

Integration

Integrating AI into existing systems and workflows requires careful planning and execution

Difficult to Setup

Long setup time and development lead time before realizing the business benefits and ROI.

Cost and Resources

Implementing AI solutions can be expensive and resource-intensive, requiring significant infrastructure and expertise.

Key Features and Mapping

Grow with your needs

Scalable performance to handle growing workloads by horizontal scaling by simply adding multiple servers to an existing pool of machines

Secure and Safe

On-premises deployment for maximum privacy and control

Pipeline Design

Enable clarity of inputs, outputs and operations performed at each stage. We can provide further PS consultancy for your specific use case

AI

APIs Integration

Easy deployment and serving of LLM models with support of popular APIs and interfaces. Easy integration with enterprise existing systems such as active directory

Setup and Training

Equipped with setup configuration and training - can be used immediately once finished the initial setup.

Budget Safe and Free Choice of Models

One-time purchase investment gives you a peace of mind to avoid unexpected usage bills.
Support all popular LLM models – Llama 3.1,, GLM-4, Mistral, GPT,….
You can choose the best-of-breed solution for your specific use case

Currently Supported Large Language Models

NeuroNode supports a variety of existing common LLMs as well as upcoming more advanced models in the future

Current LLMs supported from various open-source models

Continuous update of LLM's is possible through service plan

Meta AI
- Llama 3.1 8B, 70B

Mistral AI
- Pixtral 12B
- Mistral-Nemo

THUDM (Zhipu)
- GLM-4 9B

Alibaba Cloud
- Qwen 2.5 7B

Deepseek AI
- Deepseek R1-Distill-Llama 8B
- Deepseek R1-Distill-Qwen 7B
- Deepseek R1-Distill-Qwen 32B

LLM in a Box Overview

Conversational AI

Content Generation

Knowledge Extraction

Workflow Automation

Code Generation

deepseek

LLaMA by Meta

Qwen2.5

Open-Sourced Community Models

Legal

Medical

Insurance

Service

Custom/Domain Specific Models

Software Stack

Orchestration & Administration

- Dashboard
- Role-based access control
- Monitoring & Reporting
- Model Configurations/Deployment

Resilience & Load Balancer

API Endpoint

Model Serving Engine

CUDA

PyTorch

Huggingface Transformers

Hardware

Box Server with GPU's

Box Server with GPU's

Box Server with GPU's

Box Server with GPU's

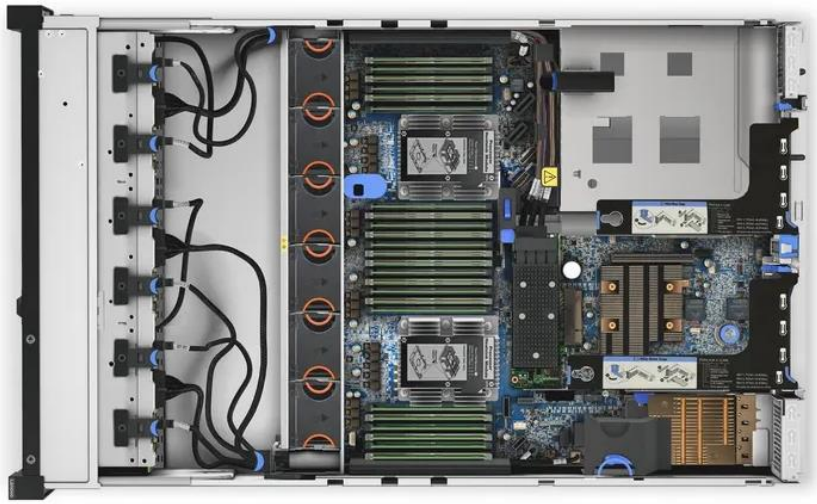
Hardware Specifications

Available Models and Detailed Specifications

Model	CPU Core	RAM	GPU	GPU Memory	Qty. of GPU	Total GPU Memory	GPU TFPOPS	SSD & RAID	Network	Hardware Warranty
P1	24	256GB	A2	16GB	4	64GB	4.53	4 x 480GB	1GB, 4 ports	3 years
P2	24	256GB	L2	24GB	4	96GB	24.1	4 x 480GB	1GB, 4 ports	3 years
P3	24	256GB	A10	24GB	4	96GB	31.2	4 x 480GB	1GB, 4 ports	3 years
P4	24	1TB	L20	48GB	4	192GB	59.4	4 x 480GB	1GB, 4 ports	3 years

Experimental Product Performance

Model	Generation Speed (Tokens/s/request)	Average Requests Handled Per Minute
P1	8	11
P2	42	53
P3	56	71
P4	107	137



Assumptions:

1. An input of 550 tokens (1 full A4 page) with the output generation of 150 tokens (1 paragraph) based on Llama 3.1 8B
2. Number of con-current users: 4 users

Warranty and Support Service

The LLM in a Box comes with service package which includes below:

- Initial Onsite Setup
- Initial Customer Training
- Hotline Support for Fault Reporting
- Onsite Troubleshooting and Faulty Hardware Replacement
- LLM Model Support (Deployment, Tuning & Upgrade)

The LLM in a Box included 1st year standard warranty and choices of 1-year, 2-year, or 3-year warranty

	Support Service Package
Support Service Items	
Initial Onsite Setup	✓
Initial Customer Training	✓
LLM Model Support (Deployment, Tuning & Upgrade)	
Remote Hotline Support	✓
Onsite Technical Support	✓
Hotline Support for Fault Reporting	✓
Onsite Troubleshooting and Faulty Hardware Replacement	✓

* All list price purchase will include 1st one-year standard support service package

Warranty and Support Service (continued)

Initial Onsite Setup

An initial one-time onsite setup shall be provided to customer for activating and initializing the LLM in a Box

Initial Customer Training

An initial half-day training shall be provided to customer for a quick start of the LLM in a Box.

Hotline Support for Fault Reporting

Under the warranty period, an ongoing hotline helpdesk support shall be provided to customer for faulty incident reporting.

- With 8x5 coverage: 8 hours per day, 5 days per week, during normal business hours, excluding local public & national holiday.

Onsite Troubleshooting and Faulty Hardware Replacement

Under the warranty period, onsite troubleshooting and faulty hardware replacement support shall be provided to customer upon request.

- With 8x5 coverage: 8 hours per day, 5 days per week, during normal business hours, excluding local public & national holiday.
- Response Time: Next Business Day after receiving customer request and confirming that the issue cannot be resolved through remote hotline support.

Warranty and Support Service (continued)

LLM Model Support (Deployment, Tuning & Upgrade)

Under the warranty period, an ongoing support shall be provided to customer for the deployment, tuning and upgrade of LLM Models. There are 3 support areas available subjected to the subscribed service packages by customer

Remote Hotline Support

- With 8x5 coverage: 8 hours per day, 5 days per week, during normal business hours, excluding local public & national holiday.
- Available to service packages (Advanced and Premier).

Onsite Technical Support

- With 8x5 coverage: 8 hours per day, 5 days per week, during normal business hours, excluding local public & national holidays. Normal business hours is 9am to 6pm Monday to Friday except public holidays
- Response Time: Next Business Day after receiving customer request.
- Number of onsite supports available under each warranty year: 4
- Available to service package (Premier).

Smarter technology for all

Lenovo

thanks.

LLM in a Box – Use Cases and Applications

Different vertical tasks that LLM in a Box can address – Applicable to all industry verticals, and role specific cases (e.g., FSI, Insurance, Human Resources, Customer Service, Healthcare & Hospitality, and Micro-verticals)

Conversational AI and Chatbots

Automated content generation (reports, articles, marketing copy)

Intelligent document processing and analysis

Customer service and support automation

Knowledge management and question answering

Compliance and risk mitigation

Research and development

All Happening on your Local Premises and Infrastructure

1. Embedding into your infrastructure and database
2. Local and on premises ensuring data privacy, security, and control & ownership
3. Help to enhance business operational efficiency and productivity across different verticals
4. Leverage AI to automate tasks and build a future proof operational structure

Generative AI Implementation Service Approach

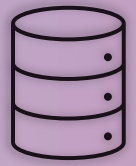
Leverage the power of AI while maintaining complete control over their valuable assets with security.

6-9 Months

Define & Design | **Deployment** | **Test**

1. Problem Definition

Identify use case

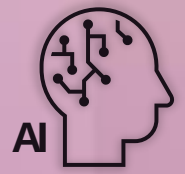


2. Data Engineering

Optimize data for AI usage

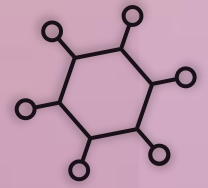
3. Proposal Design

Define implementation roadmap



4. AI Engines Developments

Setup prompt construction, embedding model, and vector database



5. Deployment

Deploy AI engines and integrate with existing applications

6. Validation

Validate & Fine-tune the results

Proposal of Deliverables:

1. Business Requirements Specification
2. Technical Architecture and Integration Specification
3. Evolution and Testing Report
4. Project Documentation and Knowledge Sharing

Generative AI Implementation Framework

1. Problem Definition
- Identify the business problem and requirements
- Define specific use case by aligning business requirements with technical capabilities

2. Data Engineering
- Identify correct data source
- Convert, clean, optimize, and transform data for LLM usage

3. Proposal Design
- Select optimal AI infrastructure & model based on user scenarios
- Define implementation roadmap

4. AI Engines Development
- Setup prompt construction, embedding model, and vector database
- Optimize implementation strategy

5. Deployment
- Deploy AI engines and integrate with existing applications
- Enable API driven integration to expose LLM capabilities

6. Validation
- Validate & Fine-tune the model's performance on new, unseen data to assess its generative AI capabilities

Key Benefits

1. Tailored Solution:
Fine-tune the private LLM that align with your business needs;

2. Risk Mitigation:
Identify and mitigate potential challenges or limitations before a full-scale deployment;

3. Data Privacy and Security:
Better control and protect sensitive data;

4. Scalability and Future Expansion:
Allow clients to plan for future expansion or integration with other systems

What Generative AI can do for you

DATA TRANSFORMATIONS

A data product that transforms a text input into a text output, e.g., classify, summarize, convert to JSON

Example

Summarization – An email thread summarization feature in a customer support tool

Format Translation – Convert free-text feedback messages into JSON objects that can be submitted as bug tickets

NATURAL LANGUAGE INTERFACE

A language-based interface to data or a tool e.g. Chat-year-documents, SQL query

Example

Customer chatbot that answers questions based on product documentation with reasoning and fluency capabilities

Conversational marketing that involves AI – enabled tools that helps businesses to offer more personalized services to the customer

WORKFLOW AUTOMATION

Automate predefined workflows using access to data and tools, e.g., write a proposal, book a flight

Example

Compare company policies against updated regulations, surfacing risks.

Reviewing engineering change requests for compliance, quality, traceability.

COPILOTS & ASSISTANTS

A mixture of natural language interfaces and automation capabilities, used in the loop with a human user, e.g., MS Copilot

Example

Advanced customer chatbot that answer questions from documentation/user account data and perform actions on behalf of the user

Coding assistant e.g. GitHub X Copilot

- Autocomplete code
- Create and rewrite code
- Answer questions about code and codebase

AUTONOMOUS AGENTS

Automate arbitrary, unseen workflows using data and tools

Example

Completely replace human employees, using the same tools and knowledge sources that they do

Implementing an executive reasoning system that can plan and execute, doing the job of the user in a Copilot-style partnership, and no human intervention.

INCREASING COMPLEXITY