# NAI GPT Pro Bundle

Product Code: CNS-NAI-GPT-B-PRO

## At-a-Glance

**Stage: Plan, Design, Deploy and Optimize**

The Nutanix Enterprise AI (NAI) generative pretrained transformer (GPT) Bundle offers a streamlined, full-stack solution for orchestrating AI/ML workloads with Nutanix Enterprise AI GPT-in-a-Box. Purpose-built for AI/ML teams, this offering simplifies the deployment and management of large language models (LLMs) in a secure, scalable environment.

NAI GPT-in-a-Box integrates the following core components:

- Nutanix Enterprise AI (NAI) – Optimized for AI/ML model orchestration and performance
- Nutanix Cloud Infrastructure (NCI) – Delivers a resilient, high-performance foundation
- Nutanix Unified Storage (NUS) – Supports high-throughput, scalable data storage
- Nutanix Kubernetes Platform (NKP) – Enables containerized deployment and lifecycle management of LLMs

Covering the key stages of the AI journey—Plan, Design, Deploy, and Optimize—the NAI GPT Bundle empowers teams to reduce complexity, accelerate time-to-value, and confidently build a future-ready AI infrastructure.

## Service Scope

The Nutanix AI (NAI) GPT Pro Bundle delivers a comprehensive, end-to-end approach for deploying AI/ML workloads on Nutanix Cloud Infrastructure (NCI). Designed to guide organizations through every stage of implementation, this structured offering ensures optimized performance, scalability, and operational efficiency for NAI GPT environments.

Led by seasoned consultants with deep expertise in both AI/ML technologies and Nutanix platforms, the NAI Bundle follows a proven methodology:

- Discovery & Planning – Assess AI/ML use cases, technical requirements, and current infrastructure to define a clear implementation path.
- Design – Develop a tailored architecture optimized for customer-specific AI/ML workloads and business goals.
- Deployment – Implement the solution in alignment with Nutanix-recommended practices and validated design specifications.
- Knowledge Transfer – Empower customer AI/ML teams with the skills and documentation needed to manage and scale the environment confidently

For customers with the NAI GPT Pro Edition software license for their on-premises NCI environment.

This bundle includes the following activities:

### AI/ML Planning Workshop

- Gather and analyze customer AI/ML use case, requirements, and expectations

- Assess and summarize the current state of AI/ML
- Identify data management considerations, including the data source, data preparation for AI/ML use, data protection, and security
- Look at the client's current Data Governance
- Learn the importance of choosing the right LLM
- Identify GPU selection and configuration options
- Explore options for training the model, such as using virtual machines (VMs) or container services
- Review capacity planning and scalability considerations for the number of end users who will leverage the GPT-based application
- Develop a Risk Management Plan

## AI/ML Design Workshop

- Gather and document solution requirements, constraints, assumptions, dependencies, and decisions in a series of initial high-level workshops
- Develop NCI cluster design for AI workloads
  - Develop NCI/Nutanix Kubernetes Platform (NKP)/NAI architecture, including interoperability, security, and scalability for future growth
  - Define integration with Active Directory (AD)/lightweight directory access protocol (LDAP) and domain name service (DNS) environments
  - Review the customer’s current data governance
  - Develop the NCI cluster design
  - Gather graphics processing unit (GPU) requirements based on use cases
  - Design virtual networking, including integration with the physical network
  - Design virtual storage, including container layout, compression, and de-duplication
  - Design NAI, including documenting the number of instances and identifying the LLM to use based on the use cases

## NUS Design Workshop

- Gather and document solution requirements, constraints, assumptions, dependencies, risks, mitigations, and decisions in the workshop
- Deliver an overview of high-level architecture and concepts of either NUS Files and Objects
- Review the customer’s current landscape, use cases, and operations, and identify how NUS data services fit into the existing environment
- Assess resources required for the on-premises NCI environment
- Evaluate and define the integration of other infrastructure services required for the deployment with a focus on NUS (AD, DNS, network time protocol (NTP), directory services, identity services, etc.)
- Define NUS File shares and share types based on the use case
- Plan security hardening and compliance as per the *Nutanix Security Operations Guide*
- Develop a Validation Plan that addresses the access and management of NUS Files

- Define Nutanix Objects store and bucket/s
- Plan security hardening and compliance as per the *Nutanix Security Operations Guide*

## NKP Design Workshop

- Gather and document solution requirements, constraints, assumptions, dependencies, risks, mitigations, and decisions in the workshop
- Review the containerized workload use case system resource requirements with customer application owners
- Assess NKP Control Plane, workers, and infrastructure nodes quantity and capacity requirements based on solution sizing
- Review GPU-targeted containerized workload use cases and assess NKP worker node pools, quantity, and capacity requirements based on solution sizing
- Develop NKP architecture, including cluster API (CAPI) provisioning method, interoperability, security, and scalability for future growth
- Identify integration required for the customer-supplied identity provider (IdP) used for cluster-based authentication
- Identify virtual networking for NKP nodes (east-west/north-south)
- Identify virtual storage for NKP nodes and containerized workload design
- Plan default container storage interface (CSI) integration-based solution for NKP
- Identify an NKP image registry solution as needed
- Plan SSL certificate strategy
- Identify an NKP compute base machine template, as needed
- Design an NKP backup strategy using the default provided NKP backup solution
- Develop a Validation Plan for NKP
- Integrate and finalize all design documents (NCI, NUS, NKP, NAI, networking, storage, security, etc.)
- Develop a comprehensive Validation Plan (leveraging plans from T3/T4 and adding overall system tests)
- Review and update the Risk Management Plan

## Infrastructure Deployment for AI Workloads

- Review customer-provided Design and Configuration documentation
- Review and validate deployment prerequisites (questionnaires, binaries, virtual networks, IP addresses, existing environment, availability of other infrastructure services such as AD, DNS, NTP, etc.)
- Deploy and configure NCI cluster via LCM, including recommended firmware and Acropolis operating system (AOS)
- Deploy and configure the hypervisor cluster on the deployed NCI cluster, including integration into existing vCenter/virtual machine manager (VMM) or vCenter Service Appliance (VCSA) deployment
- Configure LCM for automatic updates

- Configure layer 2 virtual networking and virtual switches on hypervisor hosts
- Deploy and integrate with Prism Central
- Install and configure hardware and drivers for GPU, including host drivers
- Test and validate the deployed NCI clusters

### Optional Activities for Infrastructure Deployment for AI Workloads

- Enable local key management service (KMS) for encryption
- Install and configure non-factory installed supported hardware (random-access memory (RAM), local area network (LAN), sold-state drive (SSD), hard-disk drive (HDD), etc.)
- Harden Nutanix Controller VM and AHV according to the *Nutanix Security Guide*
- Optional activities for vGPU
  - Deploy GPU license server
  - Configure a single test VM for vGPU

## NUS Deployment

- Deploy NUS Files and Objects
- Deploy and configure the NUS data service per the customer-provided Design document
- Configure NUS File servers and shares
- Configure NUS Objects Store and buckets
- Assign IP addresses for NUS data services
- Configure security for NUS data services
- Configure certificates for NUS Objects
- Configure Internet content adaption protocol (ICAP)
- Configure bucket options (policies, versioning, lifecycle, WORM)
- Configure NUS Files to support workload-specific needs
- Verify NUS Files and Objects data service is accessible

### Optional Activities for NUS Deployment

- Configure standard smart tiering
- Deploy and configure File Analytics
- Deploy and configure Data Lens

## NKP Deployment Advanced Edition

- Set up/configure NKP deployment host for NKP cluster deployments
- Deploy a single NKP-supported CAPI-provisioned NKP cluster
- Configure an NKP Compute node machine set
- Set up NKP load balancer with fully qualified domain name (FQDN)
- Configure SSL certificates for NKP services

- Configure the default supported CSI based on the CAPI provisioning method used
- Install and configure GPU operator on NKP
- Configure NKP-based IdP authentication
- Deploy NKP platform applications
- Configure NKP licensing
- Configure NKP banner
- Seed a single NKP image registry, if required
- Review the NCI cluster configuration that runs a supported Kubernetes platform, including:
  - Verification of GPU support
  - Installation of GPU operator
  - NKP version and configuration
  - Set up/configuration of NUS files with CSI in NKP
  - Set up/configuration of NKP load balancer with FQDN and SSL certification
- Provide knowledge transfer (KT) session on the following topics:
  - NKP user interface (UI), including cluster creation, identity provider authentication, user token generation, and resource alerts
  - NKP disaster recovery (DR) backups, including example namespace backup and restore
  - NKP observability, including monitoring and logging

### NAI Deployment

- Install NAI on the NKP cluster
- Add and update Nutanix helm repository
- Set up Hugging Face
- Import LLM
- Configure endpoint
- Demonstrate the LLM with a sample application

## Limitations

- For each quantity purchased, deployment is limited to 1 node. A maximum of 31 nodes on a single on-premises NCI cluster.

  **Note:** For AI/ML workloads running on Bare Metal or Public Cloud, a custom statement of work (SOW) is required

### AI/ML Planning Workshop

- Planning is limited to a single AI/ML use case

### AI/ML Design Workshop

- Infrastructure design is limited to a single AI/ML inference use case in a single physical site

- Management and other cluster designs require a separate *Infrastructure Design Workshop* for each additional cluster

### NUS Design Workshop

- NUS design includes a single NUS Files and a single Objects data service

### Infrastructure Deployment

- Excludes deployment of NCI Flow Network Security, or NCI Advanced Replication

### NUS Deployment

- Excludes migration of existing data to NUS Files or NUS Objects

### NKP Deployment

- Selected CAPI provisioner must support the hardware platform
- Excludes continuous integration (CI) design of containerized workloads
- Excludes configurations requiring customization or enhancements of the existing product's capabilities
- KT session is limited to NKP out-of-box functionality
- Cluster API provisioning method is limited to the Nutanix infrastructure
- Configuration of the NKP product is limited to the features available in the NKP Pro license
- Configure up to 2 storage classes for NUS Volumes
- Configure up to 2 volume snapshot storage classes for NUS Volumes
- Configure up to 1 NKP-supported IdP source for cluster authentication

#### Optional Activities for NKP Deployment

- Configure up to 2 storage classes for NUS Files

### NAI Deployment

- Excludes training a new LLM

### Supported Hypervisors

- Nutanix AHV

## Prerequisites

- Hardware that meets all product requirements that meets all product requirements for NCI, NKP, NUS, and a supported GPU

> **Note:** For information on the requirements for NCI Clusters, see Field Installation Overview in the *Field Installation Guide* on the Nutanix Support Portal.
>
> For information on NUS Files Prerequisites, see Prerequisites in *Nutanix Files User's Guide* on the Nutanix Support Portal.
>
> For information on NUS Objects Prerequisites, see Objects Prerequisites and Limitations in *Nutanix Objects User's Guide* on the Nutanix Support Portal.

**Note**: For information on the requirements for deploying NKP, see Basic Installations by Infrastructure in the *Nutanix Kubernetes Platform Guide* on the Nutanix Support Portal

For information on the requirements for NAI, see Nutanix Enterprise AI Requirements in the *Nutanix Enterprise AI Guide on the Nutanix Support Portal*

Fully supported and functional on-premises Prism Central instance

**Note:** For information on the requirements for configuring NCM Intelligent Operations, see Prism Central Installation *or* Upgrade in *Prism Central Infrastructure Guide* on the Nutanix Support Portal.

- Required NUS certificates must be generated and made available by the customer

### Required Product Licenses

- NAI GPT Pro Only

## Delivered Artifacts

| Service | Delivered Artifacts |
|---|---|
| Planning Workshop | • Optimized Customer Strategy<br>• High-level Summary Presentation |
| Design Workshops | • Configuration Workbook<br>• Design Document |
| Deployments | • Configuration Workbook<br>• As-built Guide |

## Level of Effort

Typically up to 25 days (varies based upon the number of nodes purchased)

### Delivery Type

| Delivery Type | Delivery Activities |
|---|---|
| Virtual | • Virtual workshop<br>• Virtual documentation<br>• Virtual deployment |
| In-person | • In-person workshop<br>• Virtual documentation<br>• In-person NCI infrastructure deployment<br>• Virtual NUS, NKP, and NAI deployment |

# Related Products

- Nutanix Enterprise AI (NAI)
- Nutanix Cloud Infrastructure (NCI)
- Nutanix Unified Storage (NUS)
- Nutanix Kubernetes Platform (NKP)

## Terms and Conditions

This document contains the entire scope of the service offer. Anything not explicitly included above is out of scope. This service offer is subject to the Nutanix Services General Terms and Conditions, which can be viewed at https://www.nutanix.com/support-services/consulting-services/terms-and-conditions