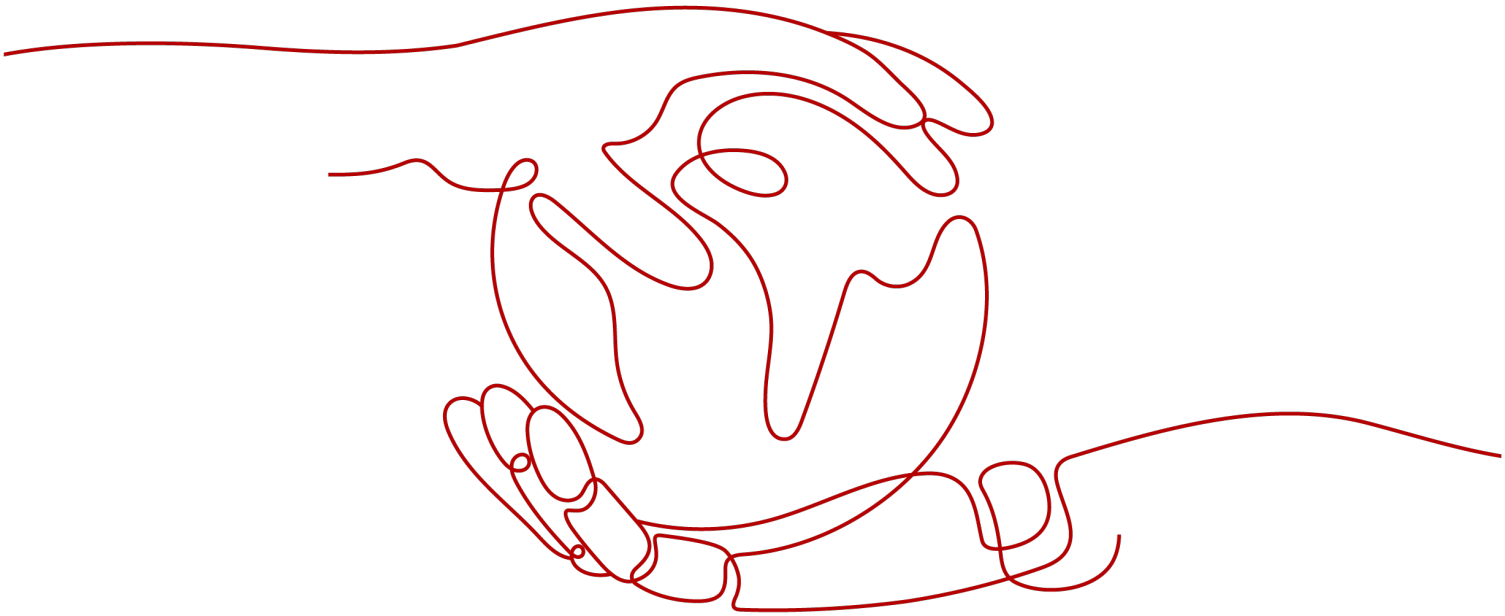
PanguLM Service

Service Overview

Issue	01
Date	2025-07-28



HUAWEI

HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.

Trademarks and Permissions


and other Huawei trademarks are the property of Huawei Technologies Co., Ltd.
All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei Cloud and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Cloud Computing Technologies Co., Ltd.

Address: Huawei Cloud Data Center Jiaoxinggong Road
Qianzhong Avenue
Gui'an New District
Gui Zhou 550029
People's Republic of China

Website: https://www.huaweicloud.com/intl/en-us/

Contents

1 What Is PanguLM?..... 1
2 Product Advantages..... 3
3 Use Cases..... 4
4 Functions..... 5
4.1 Workspace Management..... 5
4.2 Data Engineering..... 6
4.3 Model Development..... 7
4.4 Agent Development..... 8
5 Model Capabilities and Specifications..... 10
5.1 Third-Party Large Models..... 10
6 Basic Knowledge..... 16
6.1 Basic Process of Large Model Development..... 16
6.2 Basic Concepts..... 17
7 Security..... 20
7.1 Shared Responsibilities..... 20
7.2 User Authentication and Access Control..... 22
7.3 Data Protection..... 22
7.4 Auditing..... 22
8 Permissions Management..... 23
9 Notes and Constraints..... 27
10 Related Services..... 29

1 What Is PanguLM?

Introduction

ModelArts Studio is a comprehensive model development platform that supports the creation of various models and applications, offering engines for data, model, and application development.

- **Data engineering toolchain**

  Data serves as the foundation for large model training, providing the essential knowledge and information necessary for these models. As an important part of the PanguLM service, the data engineering toolchain provides data acquisition, cleaning, synthesis, labeling, evaluation, combination, publishing, and management.

  This toolchain efficiently collects and processes data in various formats to meet the requirements of different training and evaluation jobs. It optimizes raw data by providing automatic quality check and data cleaning capabilities to ensure data quality and consistency. Additionally, the data engineering toolchain provides robust data storage and management capabilities, providing high-quality data support for large model training.

- **Model development toolchain**

  The model development toolchain is the core module of the PanguLM service, offering a one-stop solution from model creation to deployment.

  This toolchain provides functions such as model training, compression, deployment, evaluation, and inference. With efficient inference performance and cross-platform migration tools, the model development toolchain can ensure efficient application of models in different environments.

- **Agent development toolchain**

  The agent development toolchain is another important module of the PanguLM service. It supports prompt engineering and intelligent agent creation. This toolchain offers tools for designing and managing input prompts to optimize model performance, enhancing the accuracy and relevance of outputs. Additionally, visual orchestration tools and an application development toolchain accelerate the development of large model applications to meet complex service requirements.

Relationship Between Third-Party Large Models and ModelArts Studio

ModelArts Studio supports third-party models, allowing you to use your own data for incremental training and fine-tuning. You can also compress, evaluate, and deploy trained models, as well as create your own agents.

ModelArts Studio is a one-stop platform for developing large models, launched by the PanguLM service. It integrates data management, model development, and agent development, offering comprehensive toolchains throughout the entire lifecycle of large model development. The data engineering toolchain provides more than 80 AI operators to meet multi-scenario requirements, greatly improving data processing efficiency and providing high-quality data for large model training. The model development toolchain supports fine-tuning, compression, evaluation, and deployment of third-party models, greatly lowering the threshold for large model development. The agent development toolchain provides visualized process orchestration based on large language models (LLMs) and abundant function plug-ins, greatly improving the development efficiency of large model applications. Additionally, ModelArts Studio manages data, models, and agents on a unified portal. You can quickly learn about the asset usage, version information, and source tracing information, facilitating unified asset management.

2 Product Advantages

Complete Product Series

The PanguLM service supports the inference and deployment of third-party large models. Currently, it is preconfigured with DeepSeek R1/V3 models and will gradually integrate multimodal capabilities.

Most Comprehensive Toolchain Platform

ModelArts Studio is the most comprehensive large model development platform in the industry, integrating data engineering, evaluation centers, model development, and agent development. It leverages Huawei's advanced experience in large model development. As such, it is the preferred toolchain platform in China.

Zero-Code Development Platform

ModelArts Studio supports zero-code and low-code development. Applications can be loaded with knowledge bases and plug-ins, and workflows can be visually edited through drag-and-drop functionality. Even beginners can easily create applications.

High Performance and Low Cost

ModelArts Studio is built on the high-performance Ascend inference framework, supporting data acceleration, training acceleration, and inference acceleration, as well as distributed efficient training and inference. It provides cost-effective computing power.

Robust Security Engineering

ModelArts Studio intercepts multi-language content (such as English), malicious operator scripts, and malicious prompts during inference service calls, ensuring secure access to AIGC content on the large model platform.

3 Use Cases

Customer Service Assistant

A third-party model is used to intelligently upgrade traditional customer service systems, enhancing the effectiveness of intelligent customer service.

Original intelligent customer service system:

- Only basic FAQs can be answered, with no capability for semantic generalization.
- The intent understanding capability is weak, resulting in a high transfer rate to manual service.
- The system cannot provide up-to-date responses in timeliness-sensitive scenarios, such as activities.

Customer service assistant:

- Improved service efficiency: The LLM-powered intelligent customer service system offers 24/7 services. This new system can manage a larger volume of customer inquiries and deliver quicker responses compared to the traditional manual customer service.
- Reduced OPEX: The LLM-powered intelligent customer service system can handle most common issues, freeing up manual customer service agents to tackle more complex and personalized customer needs.
- Personalized services: The LLM-powered customer service system can learn and adapt to users' behavior patterns and preferences to provide more personalized services.

4 Functions

4.1 Workspace Management

ModelArts Studio provides flexible and efficient workspace management. You can create multiple workspaces based on different application scenarios, project types, or team requirements. Each workspace is independent, ensuring that assets in the workspace are not affected by other workspaces. This ensures data and resource isolation and security. You can flexibly divide workspaces as required to implement orderly resource management and optimal configuration, maximizing resource utilization in different scenarios. To further optimize resource management, the platform provides multiple role permission systems. You can configure permissions for different roles, from managers to module personnel, to ensure that each user has appropriate access and operation permissions in the specified workspace. This fine-grained permission management mode ensures data security and improves resource utilization.

On the platform, workspace assets refer to all stored resources within the workspace, including data assets and model assets. These assets are the basis for users to develop and manage data on the platform. Centralized storage and unified management enhance operation efficiency while ensuring standardized and secure resource handling.

- **Data assets**: refer to all datasets published by users on the platform. These datasets are stored in the data asset repository, where you can view detailed information such as format, size, and ratio at any time. The platform also automatically logs the operational history of each dataset, including creation, publishing, and rollout. To simplify management, the platform supports dataset deletion, allowing you to flexibly manage and adjust your datasets. During model training and data analysis, these datasets can be called as needed to ensure data accuracy and security and enhance data asset utilization. Additionally, datasets can be published to and subscribed from the Gallery.
- **Model assets**: include models used for trials, subscriptions, or trained and published on the platform. These models are stored in the model asset repository for centralized management. You can review all historical versions and operation records of a model to track its evolution. The platform also offers a range of convenient operations, such as model training, compression,

and deployment, simplifying the model development and application process. Additionally, the platform provides import and export functionalities, enabling you to migrate Pangu models between different sites, enhancing flexibility and efficiency in model asset sharing and management. Models can also be published to and subscribed from the Gallery.

By centrally managing workspace assets, the platform enables you to efficiently organize and utilize resources while ensuring their security, consistency, and flexibility. This integrated approach guarantees effective resource utilization and intelligent configuration, offering you a more convenient development and management experience.

4.2 Data Engineering

ModelArts Studio provides comprehensive data engineering functions. These functions cover key stages such as data acquisition, processing, labeling, evaluation, and publishing, enabling you to efficiently build high-quality training datasets and facilitate the successful implementation of AI applications. The functions are as follows:

- **Data acquisition:** You can easily import various types of data to ModelArts Studio, including text, image, video, and weather data. ModelArts Studio supports flexible data ingestion and multiple file formats to cater to the needs of different service scenarios..
- **Data processing:** ModelArts Studio provides powerful data processing functions, including data extraction, filtering, conversion, tagging, and scoring for text, video, image, and weather data. The platform provides dedicated cleaning operators for different types of datasets and allows you to create custom operators to meet personalized data cleaning requirements. These functions ensure the generation of high-quality training data to meet both business and model training needs. You can flexibly adjust the operator sequence and customize cleaning templates to enhance data cleaning efficiency, support large-scale data processing, and ensure that generated datasets meet training standards.
- **Data synthesis:** The platform allows you to use preset or custom data instructions to process pre-trained text, single-turn Q&A, and single-turn Q&A (with a system persona) datasets, and generate new data based on a specified number of epochs. Data synthesis generates a large volume of high-quality training data, which can be used for pre-training large models to enhance their generalization and performance.
- **Data labeling:** The platform allows you to label or re-label data to improve the quality of dataset annotations. You can flexibly choose from various labeling options tailored to different datasets, and facilitate labeling, review, and labeling task transfer. Additionally, the platform offers AI pre-labeling capabilities for both text and image datasets. Leveraging the intelligent capabilities of the Pangu models, this feature significantly reduces the workload and costs associated with manual labeling, thereby greatly improving overall labeling efficiency.
- **Data evaluation:** The platform evaluates the quality of processed data in multiple formats, such as text, images, and videos. It provides preset basic evaluation criteria that you can either adopt directly or customize to meet personalized data quality requirements. Detailed quality evaluation reports

are then generated, enabling you to verify the accuracy, integrity, and consistency of your data. This ensures high-quality data prior to model training and guarantees the reliability and stability of models in real-world applications.

- **Data proportioning:** The platform allows you to flexibly adjust the data proportions in text or image datasets. You can select multiple datasets and adjust the proportions of data from different sources or types to optimize the model training process. The purpose of data splitting is to ensure that the model can more thoroughly learn and understand diverse data, thereby enhancing its generalization capability and performance.
- **Data publishing:** The platform allows you to publish datasets. You can publish a processed dataset in a variety of formats, including standard and Pangu formats. Especially for text and image datasets, the platform can convert them into Pangu format for training Pangu models, providing efficient data support for subsequent model training.
- **Data management:** The platform supports full-link lineage tracing. You can click a dataset name to view the operations performed on the dataset on the **Data Lineage** tab page. Full-link lineage tracing helps you analyze the impact of data sets in both forward and backward directions, quickly identify issues, and improve data O&M and governance efficiency. This also helps you better trace data sources. Moreover, the platform offers a comprehensive labeling system that supports data classification based on industry standards for both industry sectors and security levels, as well as built-in scenario classification labels. This facilitates data classification, data quality control, and data asset management, thereby enhancing the efficiency and effectiveness of data governance.

By integrating these features, data engineering not only enables you to efficiently create high-quality training datasets for AI research and development but also explores the intrinsic relationships between data and model performance through end-to-end data processing and management. This provides a robust data foundation for model training and application, promoting precise model training and continuous optimization, and ultimately improving the efficiency of AI application development and the reliability of outcomes.

4.3 Model Development

ModelArts Studio provides the model development function, covering all phases from model training to model calling. The platform supports full-process model lifecycle management, ensuring efficient and accurate execution of every phase from data preparation to model deployment, providing powerful intelligent support for real-world applications.

- **Model training:** ModelArts Studio provides abundant training tools and flexible configuration options for model training, which is the first step of model development. You can select an appropriate model architecture based on actual requirements and perform refined training using different training data. The platform supports distributed training and can process large-scale datasets, helping you quickly improve model performance. It supports various training types including pre-training, full-tuning, LoRA fine-tuning.
- **Model evaluation:** The platform provides comprehensive model evaluation functions to ensure the effectiveness of models in real-world applications. The

automatic evaluation function allows you to continuously monitor key metrics such as model precision and recall rate during training to promptly identify and address potential issues. Model evaluation functions help you verify the accuracy and reliability of models in diverse application scenarios. It supports rule-based automatic evaluation, manual evaluation, and customization of evaluation metrics. Additionally, it allows you to score model performance from various evaluation metrics on the manual evaluation page.

- **Model compression:** Before model deployment, compression is a key step to improve inference performance. Through model compression, the memory occupied in the inference process can be effectively reduced, thereby saving resources and improving the computing speed.
- **Model deployment:** The platform provides a one-click deployment feature, allowing you to easily deploy trained models to either a cloud or on-premises environment. It supports multiple deployment options to meet the requirements of different scenarios. With flexible APIs, models can be seamlessly integrated into various applications.
- **Model calling:** After a model is deployed, you can use the model calling function to quickly access the model's services. The platform provides efficient APIs to help you easily integrate models into your applications and implement functions such as intelligent dialogue and text generation.

4.4 Agent Development

The Agent development platform provides a comprehensive tool set to help you efficiently develop, optimize, and deploy agents. Whether you are a beginner or an experienced developer, you can quickly develop and deploy agents using the prompt engineering, plug-in extension, flexible workflow design, and full-link debugging functions provided by the platform, accelerating the innovation and deployment of industry AI applications.

- **For zero-code developers who have no coding experience**:
  - The platform provides functions such as prompt engineering and plug-in customization, enabling you to quickly build, fine-tune, and run your own large model applications without writing code. Through simple configuration, you can effortlessly create an agent application and conveniently explore AI applications.
  - The platform provides a knowledge import function that enables you to store and manage data and interact with AI applications. Local documents in various formats, such as .docx, .pptx, and .pdf, can be easily imported into knowledge bases, providing personalized data support for agent applications.
  - The platform also offers comprehensive observation and debugging tools to enable developers to thoroughly analyze each phase of the agent's execution process. Hierarchical information display helps developers optimize the performance and stability of AI applications and ensures smooth operation of the applications across different environments.
- **For low-code developers (who have certain coding experience)**:
  - Based on the preceding functions, the platform also provides flexible workflow design capabilities, allowing you to compile minimal code to develop sophisticated and highly stable agent applications. Developers

can swiftly set up workflows and achieve more efficient application development by combining various components, such as foundational models, code, and intent recognition, through a drag-and-drop interface.

- The platform also offers comprehensive observation and debugging tools to enable developers to thoroughly analyze each phase of the workflow's execution process. Hierarchical information display helps developers optimize the performance and stability of AI applications and ensures smooth operation of the applications across different environments.

5 Model Capabilities and Specifications

5.1 Third-Party Large Models

Specifications of Third-Party Large Models

In addition to Pangu models, ModelArts Studio integrates popular open-source third-party NLP models.

For example, DeepSeek V3 was released on December 26, 2024. It is a Mixture-of-Experts (MoE) language model with 671B parameters. DeepSeek V3 outperforms GPT-4.5 on mathematical and coding evaluation benchmarks. DeepSeek R1 has a similar structure to DeepSeek V3. It was officially open-sourced on January 20, 2025. As an outstanding representative of models with strong reasoning capabilities, DeepSeek R1 has attracted great attention. DeepSeek R1 has achieved the same or even better performance than top closed-source models such as GPT-4o and GPT-4o1 in core tasks such as mathematical reasoning and code generation, and is recognized as a leading LLM in the industry.

ModelArts Studio provides you with third-party NLP models of different specifications to meet different scenarios and requirements. The following table lists the supported models. You can choose the most suitable model based on your requirements for development and application.

Supported Region	Model Name	Maximum Context Length	Maximum Output Length	Description
CN-Hong Kong	DeepSeek-R1-32K-0.0.2	32K	8K	It was released in June 2025. It supports inference for a context length of 32K tokens. 16 inference units are required to deploy the model. Inference with a context length of 32K tokens supports up to 256 concurrent calls. The base model of this version is the open-source model DeepSeek R1-0528.
	DeepSeek-V3-32K-0.0.2	32K	8K	It was released in June 2025. It supports inference for a context length of 32K tokens. 16 inference units are required to deploy the model. Inference with a context length of 32K tokens supports up to 256 concurrent calls. The base model of this version is the open-source model DeepSeek V3-0324.
	DeepSeek-R1-Distil-Qwen-32B-0.0.1	32K	8K	DeepSeek-R1-Distill-Qwen-32B is a model fine-tuned based on the open-source model Qwen2.5-32B using data generated by DeepSeek-R1.
	DeepSeek-R1-Distill-LLama-70B-0.0.1	32K	8K	DeepSeek-R1-Distill-Llama-70B is a model fine-tuned based on the open-source model Llama-3.1-70B using data generated by DeepSeek-R1.
	DeepSeek-R1-Distill-LLama-8B-0.0.1	32K	8K	DeepSeek-R1-Distill-Llama-8B is a model fine-tuned based on the open-source model Llama-3.1-8B using data generated by DeepSeek-R2.
	Qwen3-235B-A22B-0.0.1	32K	8K	Qwen3-235B-A22B uniquely supports seamless switching between thinking mode and non-thinking mode, allowing users to switch between the two in a dialogue. The model's inference capability significantly outperforms that of QwQ, and its general capability far exceeds that of Qwen2.5-72B-Instruct, achieving SOTA performance among models of the same scale in the industry.

Supported Region	Model Name	Maximum Context Length	Maximum Output Length	Description
	Qwen3-32B-0.0.1	32K	8K	Qwen3-32B uniquely supports seamless switching between thinking mode and non-thinking mode, allowing users to switch between the two in a dialogue. The model's inference capability significantly outperforms that of QwQ, and its general capability far exceeds that of Qwen2.5-32B-Instruct, achieving SOTA performance among models of the same scale in the industry.
	Qwen3-30B-A3B-0.0.1	32K	8K	Qwen3-30B-A3B uniquely supports seamless switching between thinking mode and non-thinking mode, allowing users to switch between the two in a dialogue. The model's inference capability significantly outperforms that of QwQ, and its general capability far exceeds that of Qwen2.5-32B-Instruct, achieving SOTA performance among models of the same scale in the industry.
	Qwen3-14B-0.0.1	32K	8K	Qwen3-14B uniquely supports seamless switching between thinking mode and and non-thinking mode, allowing users to switch between the two in a dialogue. The model's inference capability reaches the SOTA level among models of the same scale in the industry, and its general capability significantly surpasses that of Qwen2.5-14B.
	Qwen3-8B-0.0.1	32K	8K	Qwen3-8B uniquely supports seamless switching between thinking mode and non-thinking mode, allowing users to switch between the two in a dialogue. The model's inference capability reaches the SOTA level among models of the same scale in the industry, and its general capability significantly surpasses that of Qwen2.5-7B.

Supported Region	Model Name	Maximum Context Length	Maximum Output Length	Description
	Qwen2.5-72B-0.0.1	32K	8K	Compared to Qwen2, Qwen2.5 has acquired significantly more knowledge and has greatly improved capabilities in coding and mathematics. Additionally, the new models achieve significant improvements in instruction following, generating long text, understanding structured data (e.g, tables), and generating structured outputs especially JSON.
	Qwen-QWQ-32B-0.0.1	32K	8K	This model is the QwQ reasoning model trained based on Qwen2.5-32B. Reinforcement learning greatly improves the model's inference capability. The core metrics of the model, including mathematical code (AIME 24/25, LiveCodeBench) and some general metrics (IFEval, LiveBench, etc.), reach the level of the full version of DeepSeek-R1, with all metrics significantly surpassing those of DeepSeek-R1-Distill-Qwen-32B, which is also based on Qwen2.5-32B.

Supported Platform Operations

Table 5-1 Platform operations supported by third-party large models

Model Name	Model Evaluation	Real-Time Inference	Model Commissioning in Experience Center
DeepSeek-V3-32K-0.0.2	√	√	√
DeepSeek-R1-32K-0.0.2	√	√	√
DeepSeek-R1-Distil-Qwen-32B-0.0.1	√	√	√
DeepSeek-R1-Distill-LLama-70B-0.0.1	√	√	√
DeepSeek-R1-Distill-LLama-8B-0.0.1	√	√	√

Model Name	Model Evaluation	Real-Time Inference	Model Commissioning in Experience Center
Qwen3-235B-A22B-0.0.1	√	√	√
Qwen3-32B-0.0.1	√	√	√
Qwen3-30B-A3B-0.0.1	√	√	√
Qwen3-14B-0.0.1	√	√	√
Qwen3-8B-0.0.1	√	√	√
Qwen2.5-72B-0.0.1	√	√	√
Qwen-QWQ-32B-0.0.1	√	√	√

Dependency of Third-Party Large Models on Resource Pools

Table 5-2 Dependency of third-party large models on resource pools

Model Name	Cloud-based Deployment
	Arm+Snt9B3
DeepSeek-V3-32K-0.0.2	Supported
DeepSeek-R1-32K-0.0.2	Supported
DeepSeek-R1-Distil-Qwen-32B-0.0.1	Supported
DeepSeek-R1-Distill-LLama-70B-0.0.1	Supported
DeepSeek-R1-Distill-LLama-8B-0.0.1	Supported
Qwen3-235B-A22B-0.0.1	Supported
Qwen3-32B-0.0.1	Supported
Qwen3-30B-A3B-0.0.1	Supported
Qwen3-14B-0.0.1	Supported
Qwen3-8B-0.0.1	Supported
Qwen2.5-72B-0.0.1	Supported

Model Name	Cloud-based Deployment
	Arm+Snt9B3
Qwen-QWQ-32B-0.0.1	Supported

6 Basic Knowledge

6.1 Basic Process of Large Model Development

Large models are characterized by their numerous parameters and complex architectures. The process of developing a large model consists of the following steps:

- **Dataset preparation**: The performance of a large model depends on a large amount of training data. Therefore, dataset preparation is the first step of model development. First, collect raw data based on service requirements to ensure data coverage and diversity. For example, an NLP task may require a large amount of text data, and a CV task requires image or video data.
- **Data preprocessing**: Data preprocessing is an important part of data preparation to improve data quality and adapt to model requirements. Common data preprocessing operations are as follows:
  - Deduplication: Ensure that each data record in the dataset is unique.
  - Filling missing values: Fill missing parts in data. Common methods include mean filling, median filling, and missing data deletion.
  - Data standardization: Convert data into a unified format or range, especially when processing numeric data (such as normalization or standardization).
  - Denoising: Remove irrelevant or abnormal values to reduce interference to model training.

  The purpose of data preprocessing is to ensure the quality of datasets, enabling effective model training and reducing negative impacts on model performance.
- **Model development**: Model development is the core phase of a large model project and usually includes the following steps:
  - Model selection: Select a proper model based on the task objective.
  - Model training: Use the processed dataset to train a model.
  - Hyperparameter tuning: Select proper hyperparameters such as the learning rate and batch size to ensure that the model can quickly converge and achieve good performance during training.

The key to the development phase is to balance the model complexity and compute resources, avoid overfitting, and ensure that the model can provide accurate predictions in actual applications.

- **Application and deployment**: After a large model is trained and verified, it enters the application phase, which includes the following:
  - Model optimization and deployment: Trained large models are deployed in production environments and inference services can be provided via cloud services or local servers. In this case, the response time and concurrency capabilities of the model must be considered.
  - Model monitoring and iteration: Continuously monitor the performance of a deployed model, and periodically update or retrain the model based on feedback. As new data is introduced, the model may require adjustments to maintain stable performance in real-world applications.

  In the application phase, the model can be integrated into a specific service process and also needs to be continuously optimized based on service requirements to make the model more accurate and efficient.

6.2 Basic Concepts

Concepts Related to Large Models

Concept	Description
LLM	Large language models (LLMs) are a category of foundation models pretrained on immense amounts of data. A pre-trained model means that the model is trained on an original task and continuously fine-tuned on downstream tasks. This improves the model accuracy on downstream tasks. A large-scale pre-trained model is a pre-trained model whose model parameters reach a level of 100 billion or 1 trillion. These models have stronger generalization capabilities and can accumulate industry experience and obtain information more efficiently and accurately.
Token	A token is the smallest unit of text a model can work with. A token can be a word or part of characters. An LLM converts input and output text into tokens, generates a probability distribution for each possible word, and then samples tokens according to the distribution. Some compound words are split based on semantics. For example, overweight is made up of two tokens: "over" and "weight". Take Pangu N1 models as an example. One token represents approximately 0.75 English words and 1.5 Chinese characters. For details about the word-to-token ratios, see **Table 6-1**.

Table 6-1 Word-to-token ratios

Model Specifications	English Word-to-Token Ratio	Chinese Character-to-Token Ratio
N1 series models	0.75	1.5
N2 series models	0.88	1.24
N4 series models	0.75	1.5

Training

Table 6-2 Training-related concepts

Concept	Description
Self-supervised learning	Self-supervised learning (SSL) is a subset of unsupervised learning that utilizes pretext tasks to derive supervision signals from unlabeled data. These pretext tasks are self-generated challenges that the model solves to learn from the data, thereby creating valuable representations for downstream tasks. SSL does not require additional manually labeled data because the supervisory signal is derived from the data itself.
Supervised learning	Supervised learning is a machine learning task that infers a function from labeled training data to make predictions. Each sample in labeled training data includes an input and an expected output.
LoRA	Low-rank adaptation (LoRA) fine-tuning is an optimization technology used to update only some parameters of a deep learning model during fine-tuning. This approach can significantly reduce the computational resources and time required for fine-tuning while maintaining or approaching the optimal performance of the model.
Overfitting	Overfitting occurs when a model tries to fit the training data so closely that it does not generalize well to new data.
Underfitting	Underfitting occurs when a model performs poorly on the training data or the model is too simplistic to capture the underlying patterns of the data.
Loss function	A loss function, often defined as L(Y, f(x)), is a mathematical function that measures the error between the predicted value f(x) and the actual value Y of the sample x. It is a non-negative real value function. A smaller loss indicates better robustness of the model.

Inference

Table 6-3 Inference-related concepts

Concept	Description
Temperature	The **temperature** parameter controls the randomness and creativity of the generated text in a generative language model. It is used to adjust the probabilities of the predicted words in the softmax output layer of the model. Higher temperature indicates a smaller variance of probabilities of the predicted words. That is, there is a higher probability that many words are more likely to be selected, facilitating the diversity of the generated text.
Diversity and consistency	Diversity and consistency are two important dimensions of evaluating text generated by LLMs. Diversity refers to the difference between different outputs generated by a model. Consistency refers to the consistency between different outputs corresponding to the same input.
Repetition penalty	Repetition penalty is a technique used in model training or text generation. It discourages the repetition of tokens that have appeared recently in the generated text. This is done by adding a penalty for repetitive output during loss calculation. (The loss function is essential for model optimization.) If the model generates repetitive tokens, its loss will increase, which encourages the model to produce more diverse tokens.

Prompt Engineering

Table 6-4 Concepts related to prompt engineering

Concept	Description
Prompt	A prompt is a language used to interact with an AI model, indicating the content needed for model generation.
CoT	Chain-of-thought (CoT) is a method that simulates human problem-solving. It uses a series of natural language inference processes to gradually deduce a problem from the input to the final conclusion.
Self-Instruct	Self-Instruct is a method for aligning pre-trained language models with instructions. With Self-Instruct, language models are able to generate instruction data themselves without relying on extensive manual annotation.

7 Security

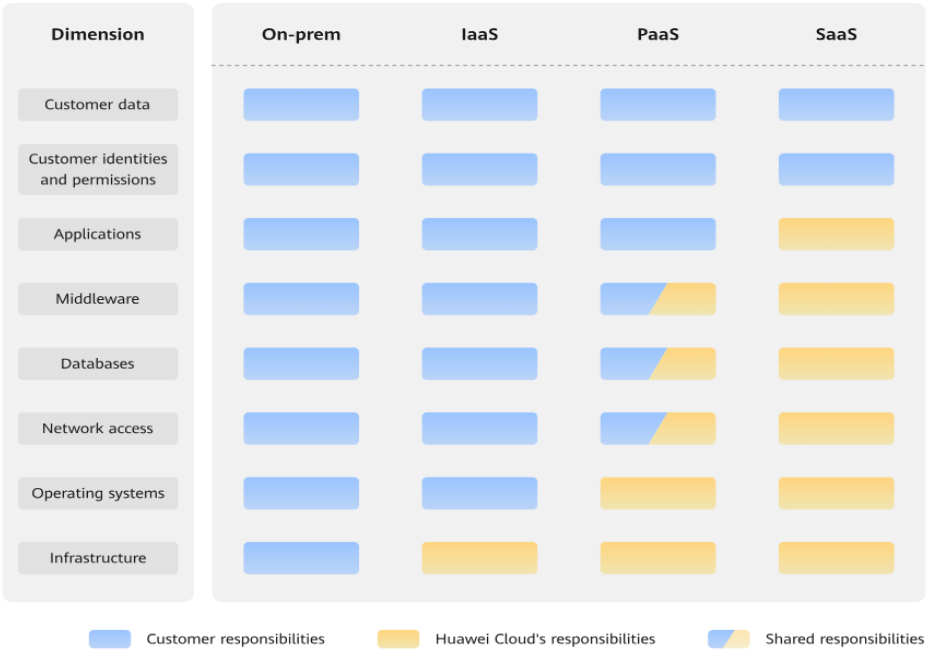
7.1 Shared Responsibilities

Huawei guarantees that its commitment to cyber security will never be outweighed by the consideration of commercial interests. To address pressing cloud security challenges, threats, and attacks, Huawei Cloud has developed a comprehensive cloud service security system that can be quickly tailored to the needs of different regions and industries. This system leverages Huawei's unique strengths in both software and hardware, as well as those of its security services partners, while ensuring compliance with applicable laws, regulations, and industry standards.

Unlike traditional on-premises data centers, cloud computing separates operators from users. This approach not only enhances flexibility and control for users but also greatly reduces their operational workload. For this reason, cloud security cannot be fully ensured by one party. Cloud security requires joint efforts of Huawei Cloud and you, as shown in **Figure 7-1**.

- **Huawei Cloud**: Huawei Cloud is responsible for infrastructure security, including security and compliance, regardless of cloud service categories. The infrastructure consists of physical data centers, which house compute, storage, and network resources, virtualization platforms, and cloud services Huawei Cloud provides for you. In PaaS and SaaS scenarios, Huawei Cloud is responsible for security settings, vulnerability remediation, security controls, and detecting any intrusions into the network where your services or Huawei Cloud components are deployed.
- **Customer**: As our customer, your ownership of and control over your data assets will not be transferred under any cloud service category. Without your explicit authorization, Huawei Cloud will not use or monetize your data, but you are responsible for protecting your data and managing identities and access. This includes ensuring the legal compliance of your data on the cloud, using secure credentials (such as strong passwords and multi-factor authentication), and properly managing those credentials, as well as monitoring and managing content security, looking out for abnormal account behavior, and responding to it, when discovered, in a timely manner.

Figure 7-1 Huawei Cloud shared security responsibility model



Cloud security responsibilities are determined by control, visibility, and availability. When you migrate services to the cloud, assets, such as devices, hardware, software, media, VMs, OSs, and data, are controlled by both you and Huawei Cloud. This means that your responsibilities depend on the cloud services you select. As shown in **Figure 7-1**, customers can select different cloud service types (such as IaaS, PaaS, and SaaS services) based on their service requirements. As control over components varies across different cloud service categories, the responsibilities are shared differently.

- In on-premises scenarios, customers have full control over assets such as hardware, software, and data, so tenants are responsible for the security of all components.
- In IaaS scenarios, customers have control over all components except the underlying infrastructure. So, customers are responsible for securing these components. This includes ensuring the legal compliance of the applications, maintaining development and design security, and managing vulnerability remediation, configuration security, and security controls for related components such as middleware, databases, and operating systems.
- In PaaS scenarios, customers are responsible for the applications they deploy, as well as the security settings and policies of the middleware, database, and network access under their control.
- In SaaS scenarios, customers have control over their content, accounts, and permissions. They need to protect their content, and properly configure and protect their accounts and permissions in compliance with laws and regulations.

7.2 User Authentication and Access Control

Requests for calling a RESTful API of PanguLM can be authenticated using either of the following methods:

- Token-based authentication: Requests are authenticated using a token.
- Access Key ID/Secret Access Key (AK/SK)-based authentication: Requests are encrypted using an AK/SK. An authenticated request must contain a signature value. The signature value is calculated based on the requestor's access key (AK/SK) as the encryption factor and the specific information carried in the request body. The platform supports authentication using an access key (AK/SK pair). It uses AK/SK-based encryption to authenticate requests.

7.3 Data Protection

PanguLM takes different measures to keep data secure and reliable.

Table 7-1 PanguLM data protection methods and features

Method	Description
Transmission encryption (HTTPS)	PanguLM uses HTTPS to secure data transmission.
OBS-based data protection	PanguLM works with OBS to store and protect user data. For details, see **OBS Data Protection**.

7.4 Auditing

Cloud Trace Service (CTS) records operations on the cloud resources in your account. You can use the logs generated by CTS to perform security analysis, audit compliance, track resource changes, and locate faults.

After you enable CTS and configure a tracker, CTS can record management and data traces of PanguLM for auditing.

For details about how to enable and configure CTS, see **CTS Getting Started**.

8 Permissions Management

If you need to assign different permissions to different personnel in your enterprise to access your PanguLM resources, Identity and Access Management (IAM) and PanguLM's role management function can be used for fine-grained permissions management.

If your Huawei Cloud account does not require individual IAM users for permissions management, skip this section.

With IAM, you can use your Huawei Cloud account to create IAM users, and grant permissions to the users to control their access to specific resources. For example, you can create IAM users and assign permissions to software developers, allowing them to call PanguLM service APIs but prohibiting model training or access to training data.

IAM Permissions

By default, a new IAM user created by the administrator does not have any permissions assigned. New users must be added to one or more groups, and permission policies or roles must be attached to these groups. The users then inherit permissions from the groups and can perform specified operations on cloud services based on the permissions they have been assigned.

PanguLM uses OBS to store training data and evaluation data. If fine-grained control over OBS access is required, you can add the Pangu OBSWriteOnly and Pangu OBSReadOnly policies to the agency of PanguLM to control the read and write permissions on OBS.

Table 8-1 Policy information

Policy Name	Fine-grained Permissions/Action	Description
Pangu OBSWriteOnly	obs:object:AbortMultipartUpload obs:object:DeleteObject obs:object:DeleteObjectVersion obs:object:PutObject	Write permission on OBS buckets
Pangu OBSReadOnly	obs:bucket:GetBucketLocation obs:bucket:HeadBucket obs:bucket:ListAllMyBuckets obs:bucket:ListBucket obs:object:GetObject obs:object:GetObjectAcl obs:object:GetObjectVersion obs:object:GetObjectVersionAcl obs:object:ListMultipartUploadParts	Read-only permission on the user's OBS bucket

Pangu User Roles

Pangu model users can be assigned different roles to implement refined control over platform resources.

Table 8-2 Role definition

Role Name	Role Description
Super Admin	Subscribes to the service and has all permissions on all workspaces on the current platform.
Administrator	Has full access to the workspace, including viewing, creating, editing, and deleting (when applicable) assets in the workspace, adding and removing workspace members, and editing workspace member roles.
Model development engineer	Has permissions to perform all operations on the model development toolchain module, but cannot create or delete compute resources or modify the workspace where it belongs.

Role Name	Role Description
Application development engineer	Has permissions to perform all operations on the application development toolchain module. Other roles do not have such permissions.
Annotation administrator	Has permissions on the following modules: • **Data Engineering** > **Data Processing** > **Data Labeling** > Task management • **Data Engineering** > **Data Processing** > **Data Labeling** > Labeling jobs • **Data Engineering** > **Data Processing** > **Data Labeling** > Labeling review • **Data Engineering** > **Data Management** > **Datasets**
Annotation operator	Has permissions on the following modules: • **Data Engineering** > **Data Processing** > **Data Labeling** > Labeling jobs
Annotation auditor	Has permissions on the following modules: • **Data Engineering** > **Data Processing** > **Data Labeling** > Labeling review
Evaluation administrator	Has permissions on the following modules: • **Data Engineering** > **Data Management** > **Datasets** • **Data Engineering** > **Data Management** > **Data Evaluation** > **Manual Evaluation** • **Data Engineering** > **Data Management** > **Data Evaluation** > **Manual Evaluation Standard**
Evaluation operator	Has permissions on the following modules: • **Data Engineering** > **Data Management** > **Data Evaluation** > **Manual Evaluation**
Data importer	Has permissions on the following modules: • **Data Engineering** > **Data Acquisition** > **Data Import** > **Import Task** • **Data Engineering** > **Data Management** > **Datasets**
Data processor	Has permissions on the following modules: • **Data Engineering** > **Data Processing** > **Processing Tasks** • **Data Engineering** > **Data Synthesis** > **Synthesis Task** • **Data Engineering** > **Data Processing** > **Data Combination** > **Data Combine Task** • **Data Engineering** > **Data Management** > **Data Instruction** • **Data Engineering** > **Data Management** > **Datasets**

Role Name	Role Description
Data publisher	Has permissions on the following modules: • **Data Engineering** > **Data Publishing** > **Publishing Task** • **Data Engineering** > **Data Management** > **Datasets**

9 Notes and Constraints

This section describes some limitations and constraints on using PanguLM.

Specifications

Table 9-1 lists the specifications of the PanguLM service.

Table 9-1 Specifications

Asset and Resource Type	Specifications	Description
Model assets, data resources, training resources, and inference resources	All model assets, data resources, training resources, and inference resources in pay-per-use or yearly/monthly billing mode	All types of purchased assets and resources can only be used in the **CN Southwest-Guiyang1** region.

Quota Limits

Table 9-2 lists the quota limits of the PanguLM service.

Table 9-2 Quota limits

Resource Type	Default Quota	Adjustable
Model instance	On ModelArts Studio, a single user can create and manage a maximum of 2,000 model instances.	Yes If you want to apply for a higher quota, contact customer service.

Function Constraints

Table 9-3 lists the function constraints of the PanguLM service.

Table 9-3 Function constraints

Function Type	Constraints
Data Engineering - Data Format Requirements	Data that can be access on ModelArts Studio must meet the format requirements, including the file format, size of a single file, size of all text, and number of files. For details, see *User Guide* > "Using Data Engineering to Create a Dataset" > "Dataset Format Requirements."
Model Development - Minimum Data Volume for Training and Evaluation	Data volume restrictions vary depending on models trained or evaluated on ModelArts Studio.
Model Development - Minimum Model Training Unit	The minimum training unit of different models varies. For details, see **Model Capabilities and Specifications**.

10 Related Services

OBS

PanguLM uses Object Storage Service (OBS) to securely and reliably store data and models at low costs.

ModelArts

PanguLM uses ModelArts for algorithm training and deployment, helping users quickly create and deploy models.

CSS

PanguLM leverages Cloud Search Service (CSS) to offer search capabilities, improving the accuracy and timeliness of its responses.

