

solution

Building a DeepSeek Inference System

Issue 1.0.0
Date 2025-02-11



Copyright © Huawei Technologies Co., Ltd. 2025. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Technologies Co., Ltd.

Trademarks and Permissions



HUAWEI and other Huawei trademarks are trademarks 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 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.

Security Declaration

Vulnerability

Huawei's regulations on product vulnerability management are subject to the *Vul. Response Process*. For details about this process, visit the following web page:

<https://www.huawei.com/en/psirt/vul-response-process>

For vulnerability information, enterprise customers can visit the following web page:

<https://securitybulletin.huawei.com/enterprise/en/security-advisory>

Contents

| | |
|---|-----------|
| 1 Solution Overview..... | 1 |
| 2 Resource Planning and Costs..... | 3 |
| 3 Procedure..... | 7 |
| 3.1 Preparations..... | 7 |
| 3.2 Quick Deployment..... | 10 |
| 3.3 Getting Started..... | 18 |
| 3.4 Quick Uninstallation..... | 25 |
| 4 Appendix..... | 27 |
| 5 Change History..... | 28 |

1 Solution Overview

Application Scenarios

The explosive growth of internet information presents enterprises and individuals with the challenge of managing and efficiently retrieving massive datasets. While traditional search engines suffice for basic needs, they often fall short when confronted with diverse data types and personalized requirements. This is where DeepSeek emerges. As a Chinese developed AI large language model, DeepSeek has risen to prominence in the AI field, leveraging its high performance, low cost, and multi-modal capabilities to demonstrate significant application potential across various sectors.

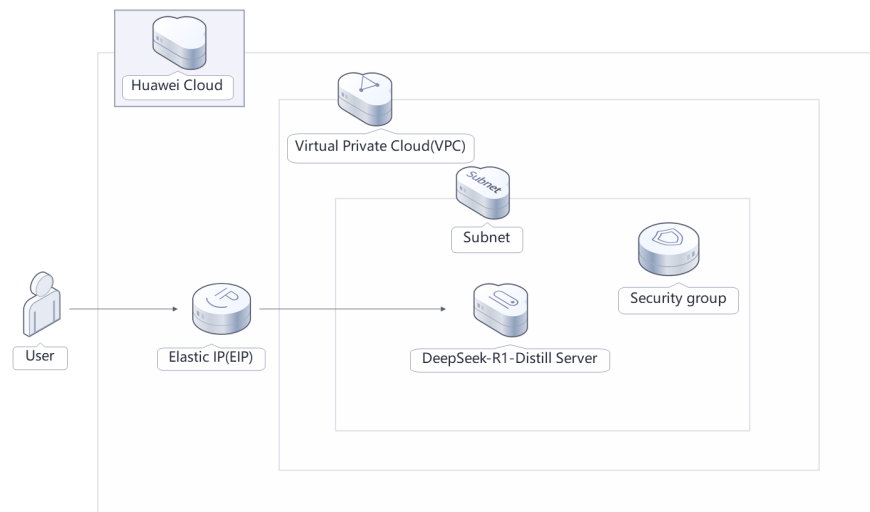
This solution enables the rapid deployment of a DeepSeek inference system on Huawei Cloud Flexus X instances (Elastic Cloud Server, ECS). DeepSeek-R1, a high-performance AI inference model specializing in mathematical, code, and natural language reasoning tasks, is deployed via Ollama on the cloud server using its distilled version. This quickly creates your private AI assistant, ideal for the following applications:

- Natural Language Processing (NLP): Understands and generates natural language text, suitable for tasks such as dialogue, translation, and summarization.
- Text Generation: Produces coherent and logically sound text, applicable to content creation and story writing.
- Question Answering System: Answers user queries, ideal for customer service and knowledge base searches.
- Sentiment Analysis: Analyzes the emotional tone of text, useful for market research and public opinion monitoring.
- Text Classification: Categorizes text, applicable to spam filtering and news categorization.
- Information Extraction: Extracts key information from text, suitable for data mining and knowledge graph construction.

Architecture

This solution helps you quickly set up the DeepSeek-R1 distilled models on the Huawei Cloud Flexus X instance (Elastic Cloud Server (ECS)).

Figure 1-1 Architecture



This solution will:

- Create a **FlexusX instance(Elastic Cloud Server (ECS))** to set up the DeepSeek-R1 distilled models.
- Create one **Elastic IP (EIP)** for internal and external communication.
- Create a security group and configure security group rules to protect Huawei Cloud cloud servers.

Advantages

- High performance
DeepSeek significantly enhances inference capabilities through reinforcement learning technology, supports multi-step logical inference, and can gradually decompose complex problems and solve them.
- Cost-effectiveness
Provide high-cost-performance cloud servers, users can customize different specifications of cloud servers according to actual needs.
- Easy deployment
In just a few clicks, you can easily deploy and complete the quick provisioning of cloud servers, public IP, and other resources, as well as the setup of the DeepSeek-R1 distilled model.

Constraints

- Before deploying this solution, ensure that you have created a Huawei ID with access to the target region and enabled Huawei Cloud services.
- If you select the yearly/monthly billing mode, ensure that your account has sufficient balance. If you do not have sufficient balance, you can go to the **Billing Center** to manually pay for the order.

2 Resource Planning and Costs

This solution will deploy the resources listed in the following table. The costs are only estimates and may differ from the final prices. For details, see [Price Calculator](#).

Table 2-1 Resource planning and costs (pay-per-use)

| Huawei Cloud Service | Resource Name | Configuration Example | Quantity | Estimated Monthly Cost |
|-----------------------------|--|--|----------|------------------------|
| Virtual Private Cloud (VPC) | building-a-deepseek-Inference-system-demo | <ul style="list-style-type: none">Region: CN-Hong KongCIDR Block: 172.16.0.0/16 | 1 | USD 0.00 |
| Subnet | building-a-deepseek-Inference-system-demo-subnet | <ul style="list-style-type: none">Region: CN-Hong KongIPv4 CIDR Block: 172.16.1.0/24 | 1 | USD 0.00 |
| SecurityGroup | building-a-deepseek-Inference-system-demo | <ul style="list-style-type: none">Region: CN-Hong KongAllow ping: 0.0.0.0/0Open port 22 to allow Cloud Shell login: 119.8.43.48/32 | 1 | USD 0.00 |

| Huawei Cloud Service | Resource Name | Configuration Example | Quantity | Estimated Monthly Cost |
|----------------------|---|--|----------|--|
| Flexus X Instance | building-a-deepseek-Inference-system-demo | <ul style="list-style-type: none"> ● Region: CN-Hong Kong ● Pay-per-use: USD 0.14/hour ● Specifications: Flexus X Instance Performance Mode (Disabled) x1.4u.4g 4vCPUs 4GB ● Image: Ubuntu 22.04 server 64bit ● System Disk: General Purpose SSD 40GB | 1 | USD 98.41 |
| Elastic IP (EIP) | building-a-deepseek-Inference-system-demo-eip | <ul style="list-style-type: none"> ● Region: CN-Hong Kong ● Pay-per-use: USD 0.16/GB/hour ● Routing Type: Dynamic BGP ● Billed By: traffic ● Bandwidth: 300Mbit/s | 1 | USD 0.16/GB/hour |
| Total | - | - | | USD 98.41+ Public network traffic price |

Table 2-2 Resource planning and costs (yearly/monthly)

| Huawei Cloud Service | Resource Name | Configuration Example | Quantity | Estimated Monthly Cost |
|-----------------------------|--|--|----------|------------------------|
| Virtual Private Cloud (VPC) | building-a-deepseek-Inference-system-demo | <ul style="list-style-type: none"> Region: CN-Hong Kong CIDR Block: 172.16.0.0/16 | 1 | USD 0.00 |
| Subnet | building-a-deepseek-Inference-system-demo-subnet | <ul style="list-style-type: none"> Region: CN-Hong Kong IPv4 CIDR Block: 172.16.1.0/24 | 1 | USD 0.00 |
| SecurityGroup | building-a-deepseek-Inference-system-demo | <ul style="list-style-type: none"> Region: CN-Hong Kong Allow ping: 0.0.0.0/0 Open port 22 to allow Cloud Shell login: 119.8.43.48/32 | 1 | USD 0.00 |

| Huawei Cloud Service | Resource Name | Configuration Example | Quantity | Estimated Monthly Cost |
|----------------------|---|--|----------|--|
| Flexus X Instance | building-a-deepseek-Inference-system-demo | <ul style="list-style-type: none"> ● Region: CN-Hong Kong ● Specifications: Flexus X Instance Performance Mode (Disabled) x1.4u.4g 4vCPUs 4GB ● Image: Ubuntu 22.04 server 64bit ● System Disk: General Purpose SSD 40GB | 1 | USD 73.22 |
| Elastic IP (EIP) | building-a-deepseek-Inference-system-demo-eip | <ul style="list-style-type: none"> ● Region: CN-Hong Kong ● Billing Mode: Pay-per-use ● Pay-per-use: USD 0.16/GB/hour ● Routing Type: Dynamic BGP ● Billed By: traffic ● Bandwidth: 300Mbit/s | 1 | USD 0.16/GB/hour |
| Total | - | - | | USD 73.22+ Public network traffic price |

3 Procedure

- [3.1 Preparations](#)
- [3.2 Quick Deployment](#)
- [3.3 Getting Started](#)
- [3.4 Quick Uninstallation](#)

3.1 Preparations

When you log in with your Huawei Cloud account, you do not need to perform this preparation step. If you are using an IAM user account, please confirm whether you are in the admin group. If you are not in the admin group, you will need to [\(Optional\) Creating the rf_admin_trust Agency](#) to your IAM account and complete the following preparation steps.

(Optional) Creating the rf_admin_trust Agency

- Step 1** Log in to the Huawei Cloud official website, open the [console](#), hover over the account name, and choose **Identity and Access Management**.

Figure 3-1 Console page

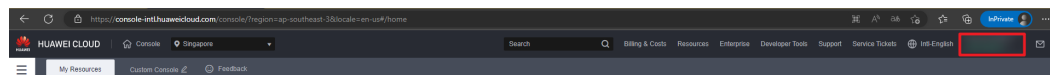
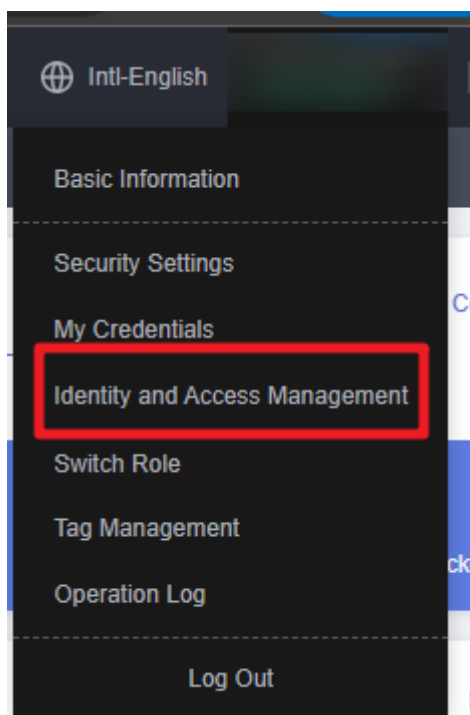
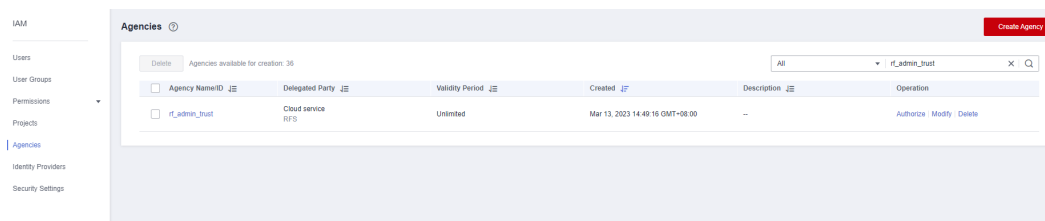


Figure 3-2 Identity and access management page



Step 2 Choose **Agencies** in the left navigation pane and search for the **rf_admin_trust** agency.

Figure 3-3 Agencies



- If the agency is found, skip the following steps.
- If the agency is not found, perform the following steps to create it.

Step 3 Click **Create Agency** in the upper right corner of the page. On the displayed page, enter **rf_admin_trust** for **Agency Name**, select **Cloud service** for **Agency Type** and **RFS** for **Cloud Service**, and click **Next**.

Figure 3-4 Creating the rf_admin_trust agency

Agencies / Create Agency

* Agency Name

* Agency Type Account
Delegate another HUAWEI CLOUD account to perform operations on your resources.
 Cloud service
Delegate a cloud service to access your resources in other cloud services.

* Cloud Service

* Validity Period

Description
0/255

Step 4 Search for **Tenant Administrator**, select it in the search results, and click **Next**.

Figure 3-5 Selecting a policy/role

Authorize Agency

1 Select Policy/Role 2 Select Scope 3 Finish

Assign selected permissions to rf_admin_trust1. Create Policy

| Policy/Role Name | Type |
|---|-----------------------|
| <input type="checkbox"/> DME AdministratorAccess Data Model Engine tenant administrator with full permissions. | System-defined policy |
| <input checked="" type="checkbox"/> Tenant Administrator Tenant Administrator (Exclude IAM) | System-defined role |
| <input type="checkbox"/> CS Tenant Admin Cloud Stream Service Tenant Administrator, can manage multiple CS users | System-defined role |

Step 5 Select **All resources** and click **OK**.

Figure 3-6 Setting the authorization scope

Authorize Agency

1 Select Policy/Role 2 Select Scope 3 Finish

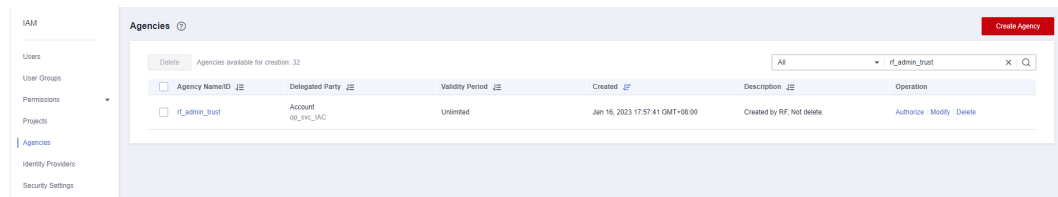
i The following are recommended scopes for the permissions you selected. Select the desired scope requiring minimum authorization.

Scope

All resources
IAM users will be able to use all resources, including those in enterprise projects, region-specific projects, and global services under your account based on assigned permissions.
[Show More](#)

Step 6 Check that the **rf_admin_trust** agency is created and displayed in the agency list.

Figure 3-7 Agenciesespe



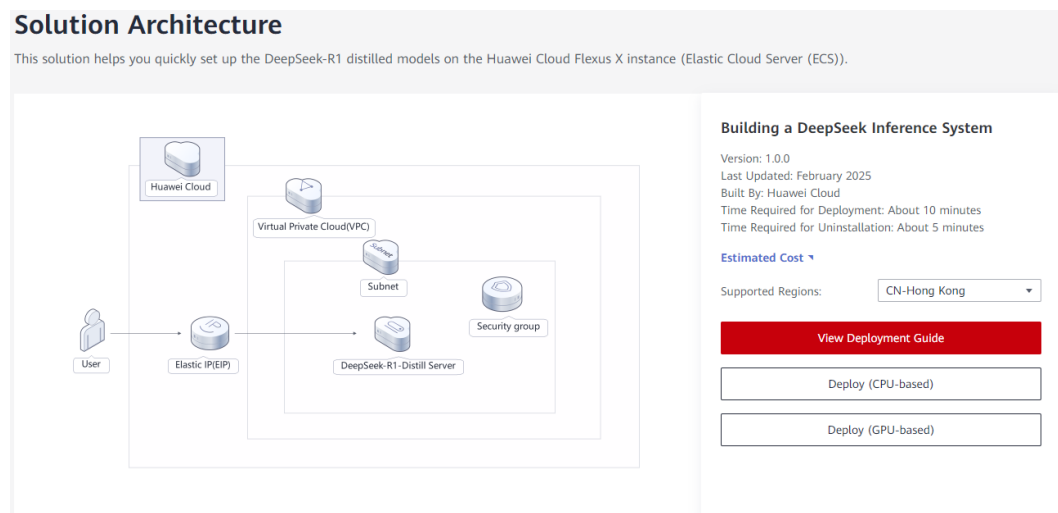
----End

3.2 Quick Deployment

This section helps you quickly **building-a-deepseek-inference-system** on Huawei Cloud.

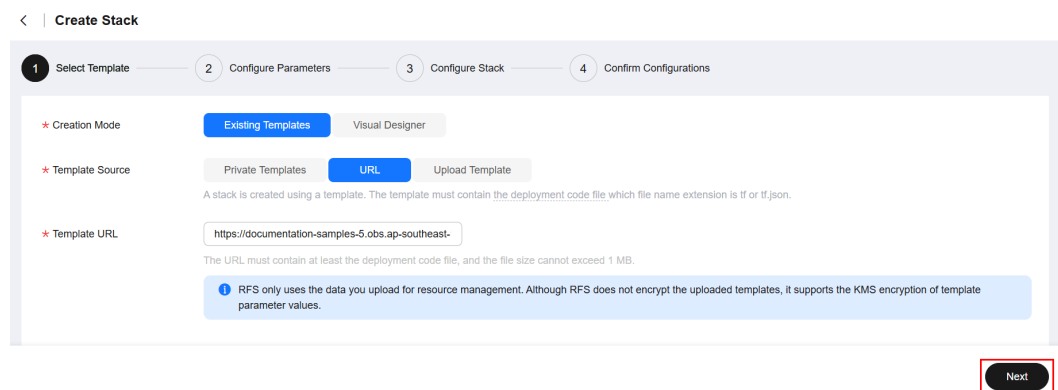
- Step 1** Log in to [Huawei Cloud Quick-Start Guides](#) and choose **Building a DeepSeek Inference System**. Select a region from the Data Center drop-down list and click Deploy.

Figure 3-8 Selecting a solution



- Step 2** On the **Select Template** page, click **Next**.

Figure 3-9 Selecting a solution



Step 3 On the **Configure Parameters** page, enter a stack name, configure parameters based on **Table 1 Parameter description**, and click **Next**.

Figure 3-10 Configuring parameters

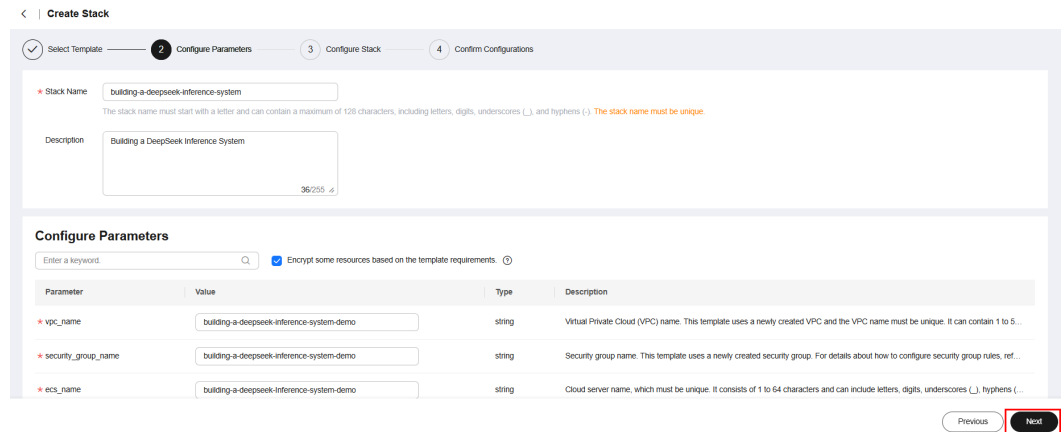


Table 3-1 Parameter description

| Parameter | Type | Mandatory | Description | Default Value |
|-----------|--------|-----------|---|---|
| vpc_name | string | Yes | Virtual Private Cloud (VPC) name. This template uses a newly created VPC and the VPC name must be unique. It can contain 1 to 54 characters, including only letters, digits, underscores (_), hyphens (-), and periods (.). | building-a-deepseek-inference-system-demo |

| Parameter | Type | Mandatory | Description | Default Value |
|---------------------|--------|-----------|--|---|
| security_group_name | string | Yes | Security group name. This template uses a newly created security group. For details about how to configure security group rules, see (Optional) Modifying Security Group Rules . It can contain 1 to 64 characters, including only letters, digits, underscores (_), hyphens (-), and periods (.). | building-a-deepseek-inference-system-demo |
| ecs_name | string | Yes | Cloud server name, which must be unique. It consists of 1 to 64 characters and can include letters, digits, underscores (_), hyphens (-), and periods (.). | building-a-deepseek-inference-system-demo |
| distilled_model | string | Yes | DeepSeek-R1-Distill model. Supports Qwen-1.5B, Qwen-7B, Llama-8B. Default is Qwen-1.5B. | DeepSeek-R1-Distill-Qwen-1.5B |

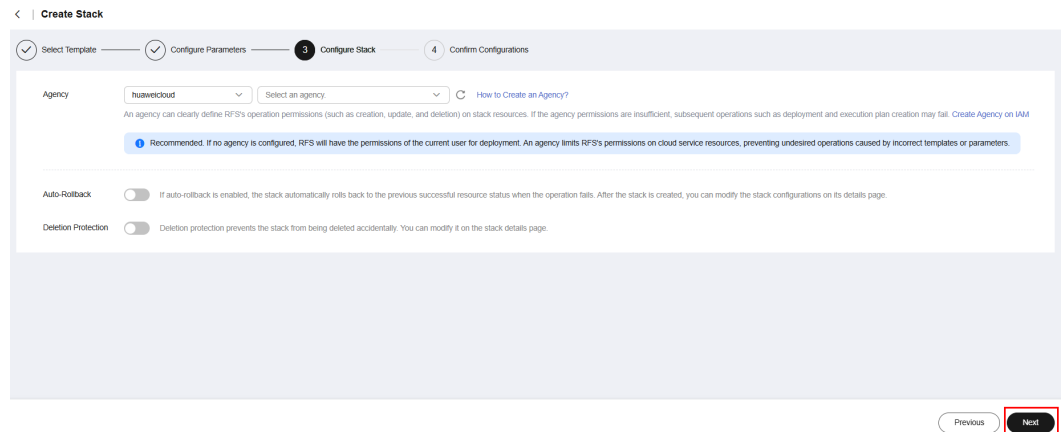
| Parameter | Type | Mandatory | Description | Default Value |
|-------------|--------|-----------------|---|-------------------------------|
| dify_enable | string | Yes (GPU-based) | Whether installing Dify along with Ollama. Dify provides the out-of-box web application to interact with the model. | enable |
| ecs_flavor | string | Yes | <p>Cloud Server Instance Specifications: For 1.5B model, it is recommended to use x1.4u.4g or higher; for 7B and 8B models, it is recommended to use x1.16u.16g or higher.</p> <p>NOTE For GPU-based solution, GPU-accelerated type is required. Value can be found from the specification list page of the documentation. (Before executing the plan, please ensure the resource is available in the corresponding region).</p> | DeepSeek-R1-Distill-Qwen-1.5B |

| Parameter | Type | Mandatory | Description | Default Value |
|------------------|--------|-----------|---|---------------|
| ecs_password | string | Yes | Initial password of the cloud server. The password can include 8 to 26 characters and must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters (!@#\$%^&*_=-+[]{};,:/?.). The password cannot contain any username or the username spelled backwards. The administrator username is root. | false |
| system_disk_size | number | Yes | System disk size of the cloud server. The default disk type is General Purpose SSD, and the unit is GB. The system disk can only be increased. The default value is 100. Value range: 40-1,024. | 40 |

| Parameter | Type | Mandatory | Description | Default Value |
|---------------|--------|-----------|--|---------------|
| charging_mode | string | Yes | Billing mode. By default, expenses are automatically deducted. The value can be postPaid (pay-per-use) or prePaid (yearly/monthly). | postPaid |
| charging_unit | string | Yes | Subscription period type. This parameter is valid only when the charging_mode is set to prePaid (yearly/monthly). The value can be month or year. | month |
| charge_period | number | Yes | Subscription period. This parameter is valid only when charging_mode is set to prePaid (yearly/monthly). The default value is 1. Value range: 1-9 (charging_unit set to month); 1-3 (charging_unit set to year). | 1 |

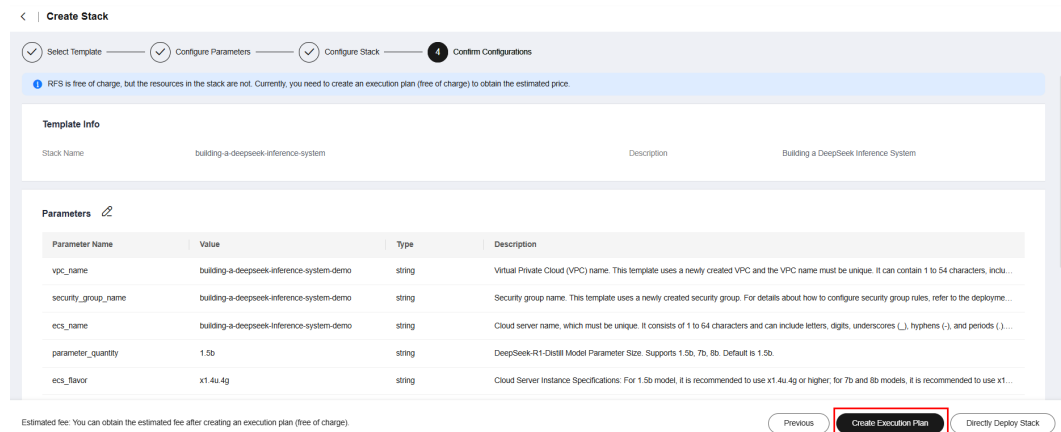
Step 4 On the **Configure Stack** page, select **rf_admin_trust** from the **Agency** drop-down list and click **Next**. This step is optional if you use an account (HUAWEI ID) or use an IAM user in the admin user group.

Figure 3-11 Configuring a stack



Step 5 On the **Confirm Configurations** page, confirm the configurations and click **Create Execution Plan**.

Figure 3-12 Confirming the configurations



Step 6 In the displayed **Create Execution Plan** dialog box, enter an execution plan name and click **OK**.

Figure 3-13 Creating an execution plan

Create Execution Plan ✕

i To preview your resource billing information, you can create an execution plan.

✖ Execution Plan Name

Description 0/255 ↗

OK **Cancel**

Step 7 Wait until the status of the execution plan changes to **Available** and then click **Deploy** in the **Operation** column. In the displayed dialog box, click **Execute**.

Figure 3-14 Execution plan page

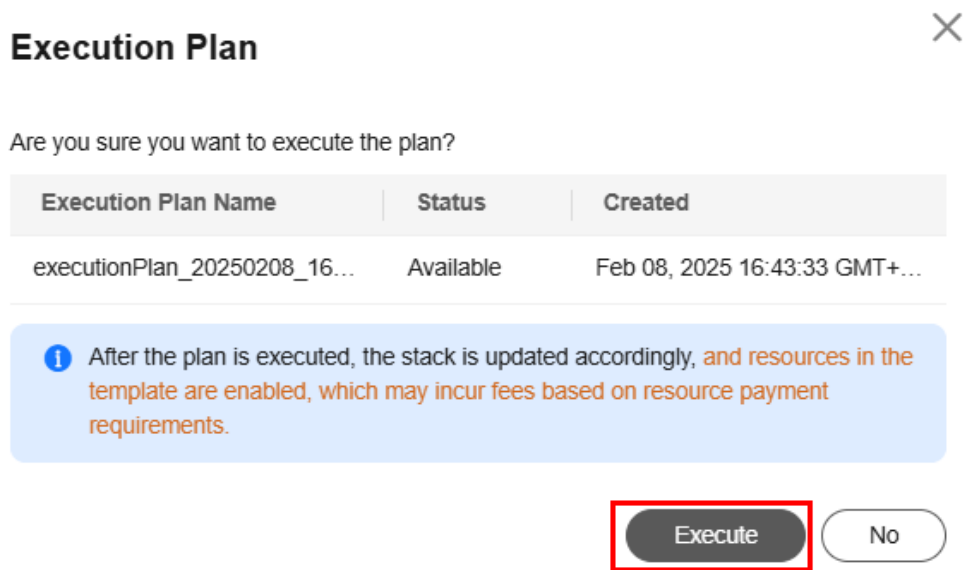
building-a-deepseek-inference-system Delete Update Template/Parameter ⌂

Basic Information Resources Outputs Events Template **Execution Plans**

Deploy Q

| Execution Plan Name/ID | Status | Estimated Price | Created | Description | Operation |
|----------------------------------|-----------|------------------------------|---------------------------------|-------------|---|
| executionPlan_20250208_1643_83kb | Available | View Details | Feb 08, 2025 16:43:33 GMT+08:00 | - | Deploy Delete |

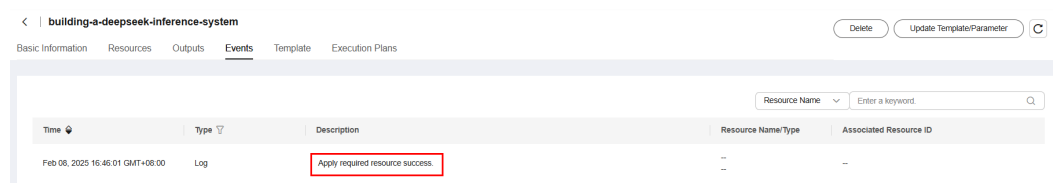
Figure 3-15 Confirming the execution plan



Step 8 (Optional) If you select the yearly/monthly billing mode and your account balance is insufficient, log in to the Billing Center to manually pay for the order. You can refer to [Table 2 Resource planning costs \(yearly/monthly\)](#) to see the total price.

Step 9 Wait until the message "Apply required resource success" is displayed on the **Events** tab page. This means the deployment is complete. The deployment takes about 10 minutes, which will be delayed by network fluctuations.

Figure 3-16 Resources created



----End

3.3 Getting Started

This solution utilizes CloudShell for remote login to the cloud server via port 22. An IP address whitelist is pre-configured. To access the server remotely, simply use CloudShell.

Following successful deployment, environment initialization, including downloading ollama and DeepSeek-R1-Distill model, is estimated to take 5-10 minutes. Network and bandwidth conditions may affect this time; Service can only be available after deployment is complete.

(Optional) Modifying Security Group Rules

A security group is a collection of access control rules to control traffic to and from cloud resources, such as cloud servers, containers, and databases. Cloud

resources associated with the same security group have the same security requirements and are mutually trusted within a VPC.

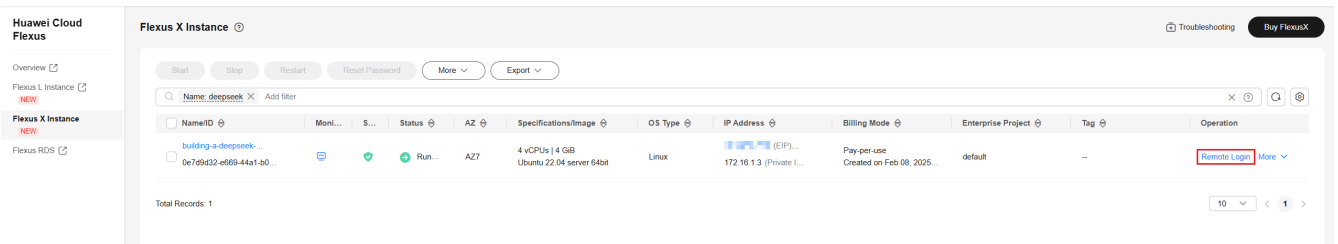
You can modify security group rules, for example, by adding, modifying, or deleting a TCP port, as follows:

- Adding a security group rule: [Add an inbound rule](#) and enable a TCP port if needed.
- Modifying a security group rule: Inappropriate security group settings may introduce serious security risks. You can [modify security group rules](#) to ensure the network security of your ECSs.
- Deleting a security group rule: If the source or destination IP address of an inbound or outbound security group rule changes, or a port needs to be disabled, you can [delete the security group rule](#).

CPU-based Solution

Step 1 Log in to the [Huawei Cloud Flexus X](#) console, select the server created using this solution, and click **Remote Login**.

Figure 3-17 Click Remote Login



Step 2 Click **Log In** button, insert the server's password on the CloudShell page and click **Connect**.

Figure 3-18 Click Log In

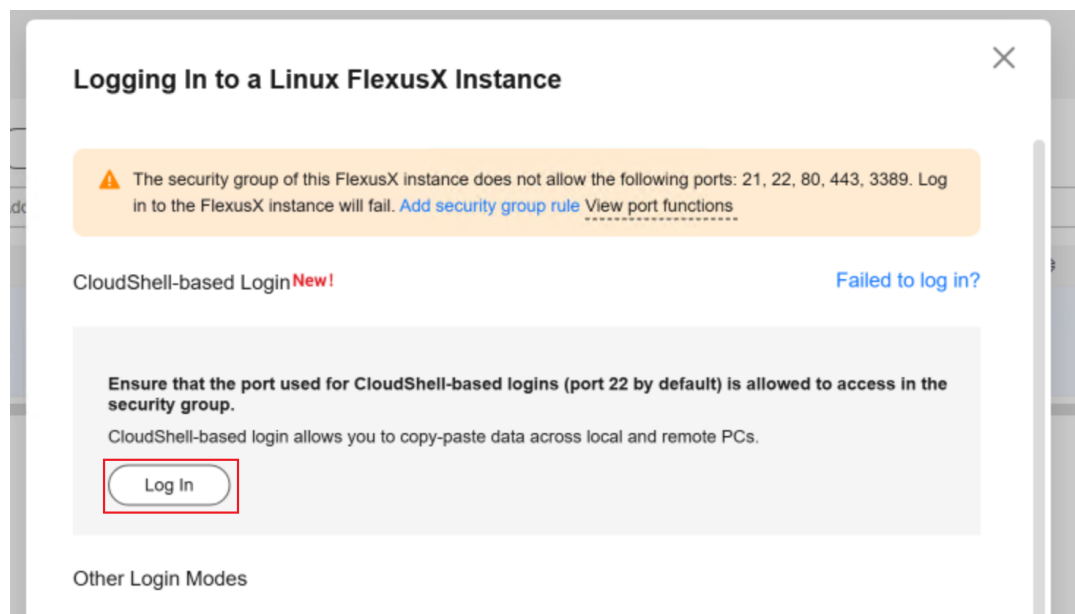
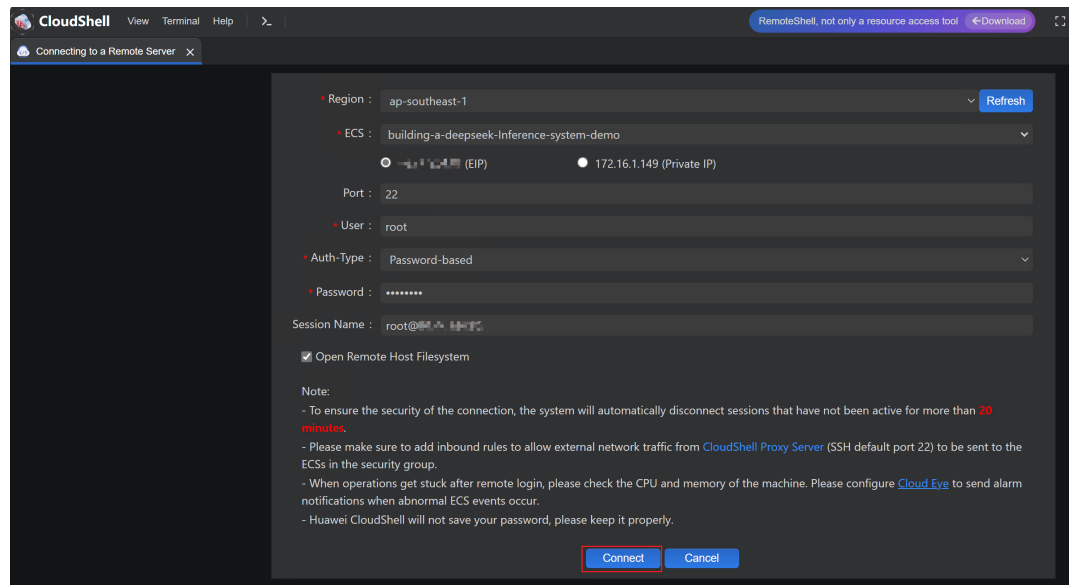
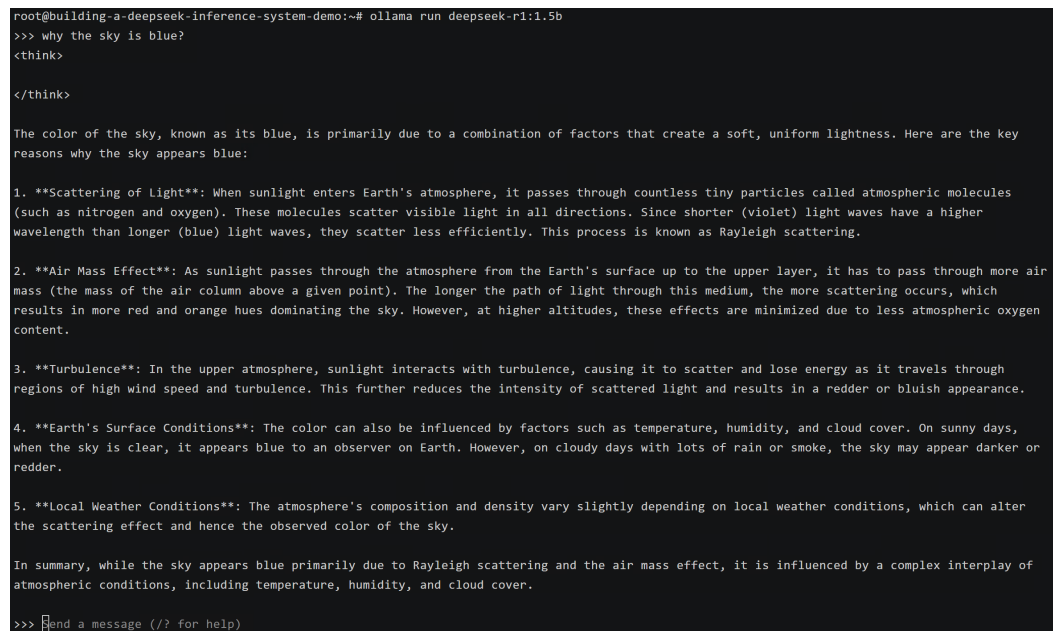


Figure 3-19 Configure the connection



Step 3 In the shell, insert "**ollama run deepseek-r1:\$parameter_quantity**". **\$parameter_quantity** supports 1.5b, 7b, 8b. Please replace with the actual value of parameter "parameter_quantity" in [3.2 Quick Deployment](#). Execute the command to start the dialog test.

Figure 3-20 Dialog test



In the interactive mode, you can test the model under various scenarios, for example:

- Intelligent Customer Service: Input common customer questions, such as "How do I install nginx?"
- Content Creation: Input prompts like "Write an advertisement for a smart watch."

- Programming Assistance: Input requests such as "Implement quicksort in Python."
- Educational Assistance: Input requests for explanations, such as "Explain Newton's Second Law."

Instead of CLI, you can also use Ollama API to interact with the model.

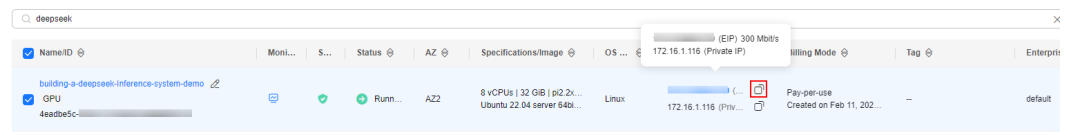
----End

GPU-based Solution

If you enable the dify installation during the **3.2 Step3**, then you can walk through the following instructions:

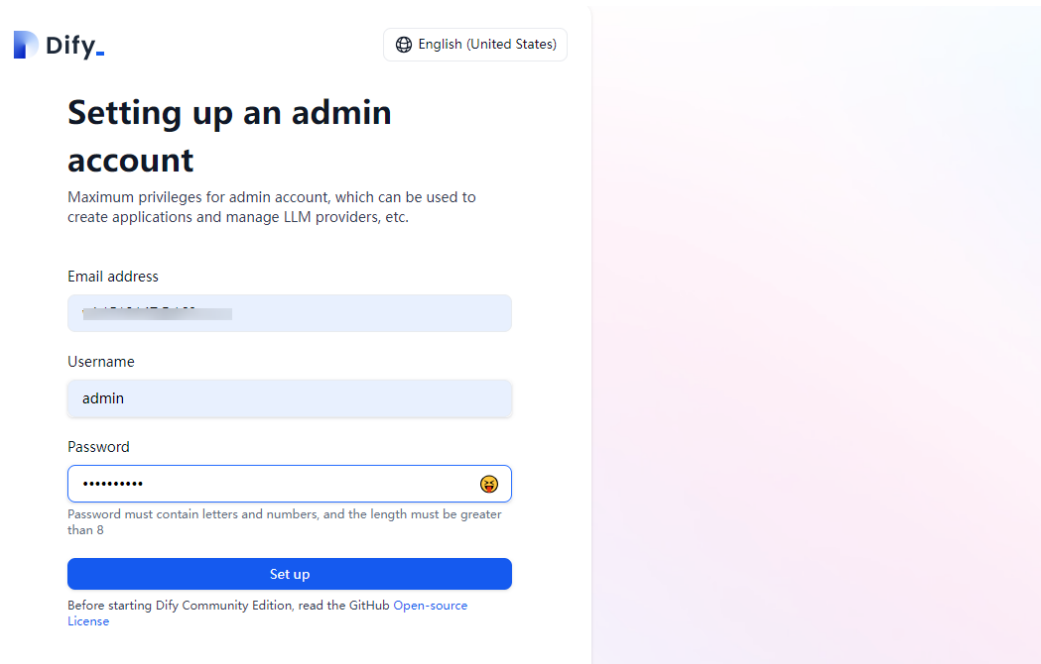
- Step 1** Log in to the **ECS console**, and get the EIP and private IP addresses of the instance deployed in the **3.2 Step3**.

Figure 3-21 Get EIP and private IP addresses



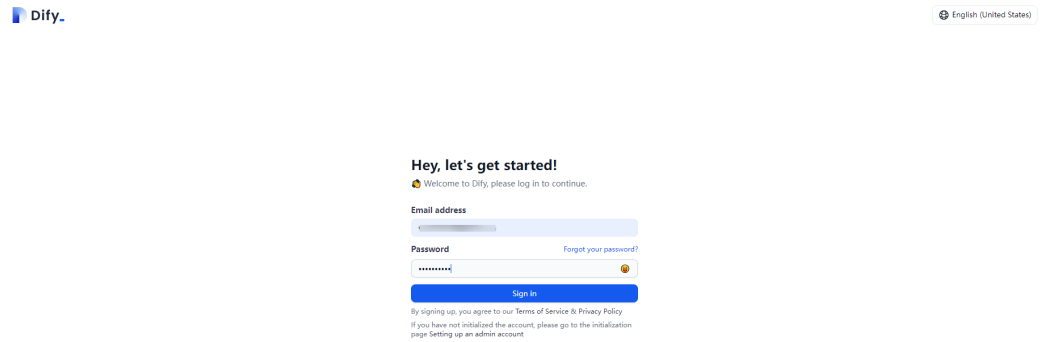
- Step 2** Access the Dify application by typing "http://[your-instance-EIP]" in the browser. For the first login, you need to register an administrator account by sequentially filling in your email, username, and password.

Figure 3-22 Setting up an admin account



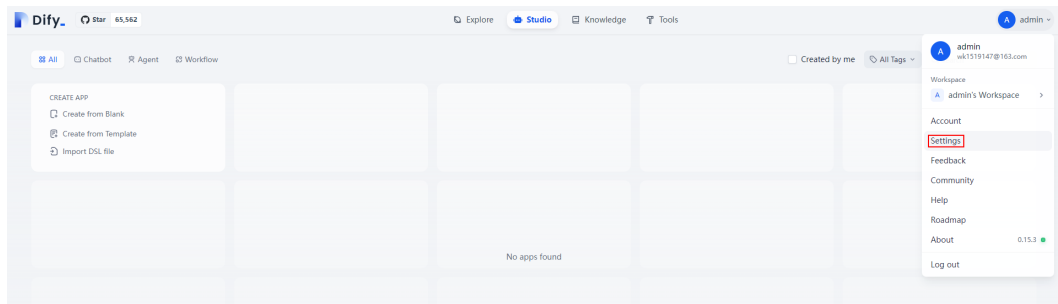
- Step 3** Log in to the Dify platform using the email and password from the previous step.

Figure 3-23 Log in to the Dify platform



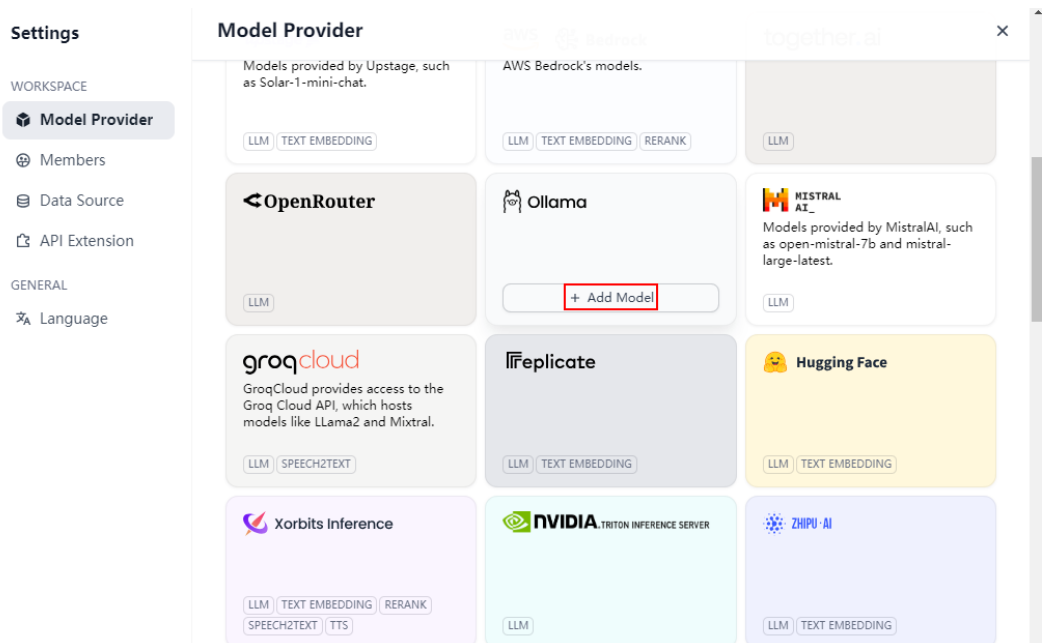
Step 4 Click the username in the top right corner and click **Settings** in the dropdown menu.

Figure 3-24 Click Settings




Step 5 Click **Model Provider** in the left panel. Find the Ollama box and click **Add Model**.

Figure 3-25 Add Model



Step 6 Type "deepseek-r1:\${quantity}b" for Model Name, and "http://\${your-instance-private-ip}:11434" for Base URL. Click **Save** and close the settings window.

Figure 3-26 Add Ollama

Add Ollama 

Model Type *
 LLM Text Embedding

Model Name *

Base URL *

Completion mode *

Model context size *

Upper bound for max tokens *

Vision support
 Yes No

[How to integrate with Ollama](#)

Step 7 Click **Create from Blank**, choose "Chatbot" and fill the application name and icon, and then click **Create**.

Figure 3-27 Create from Blank

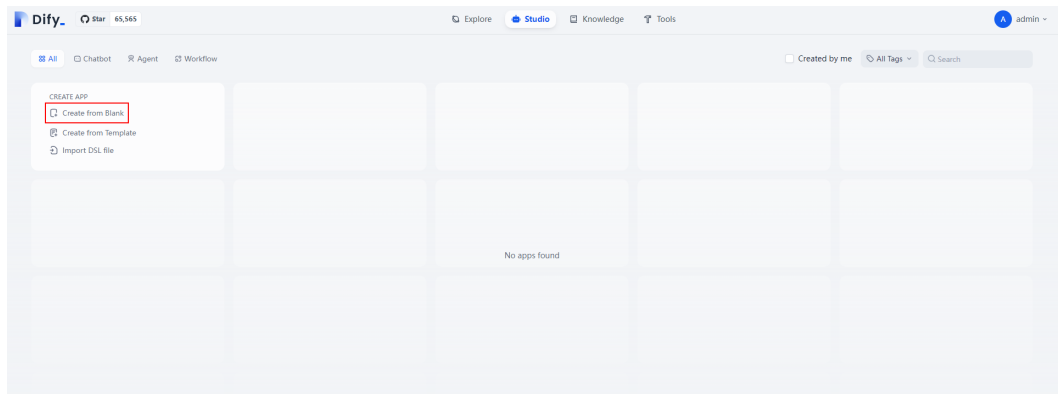
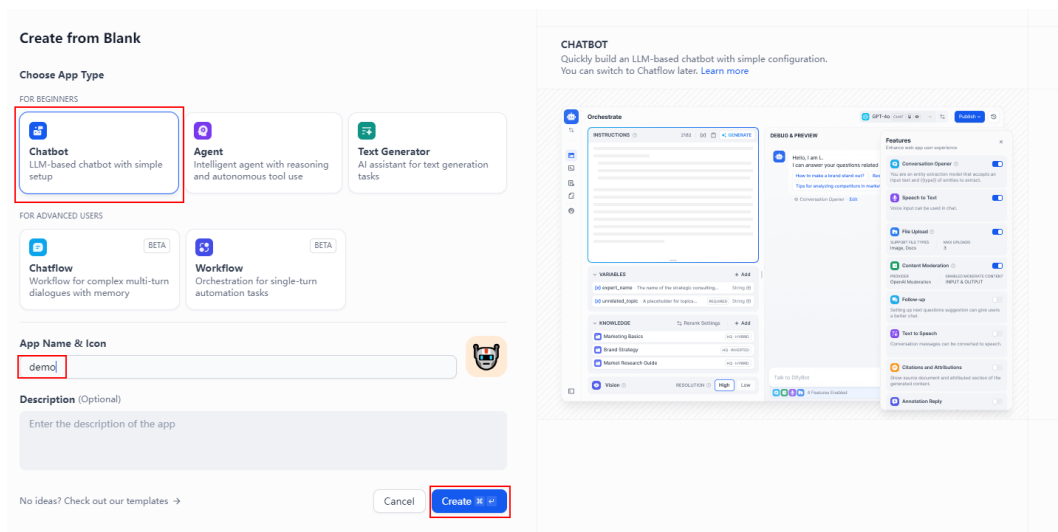
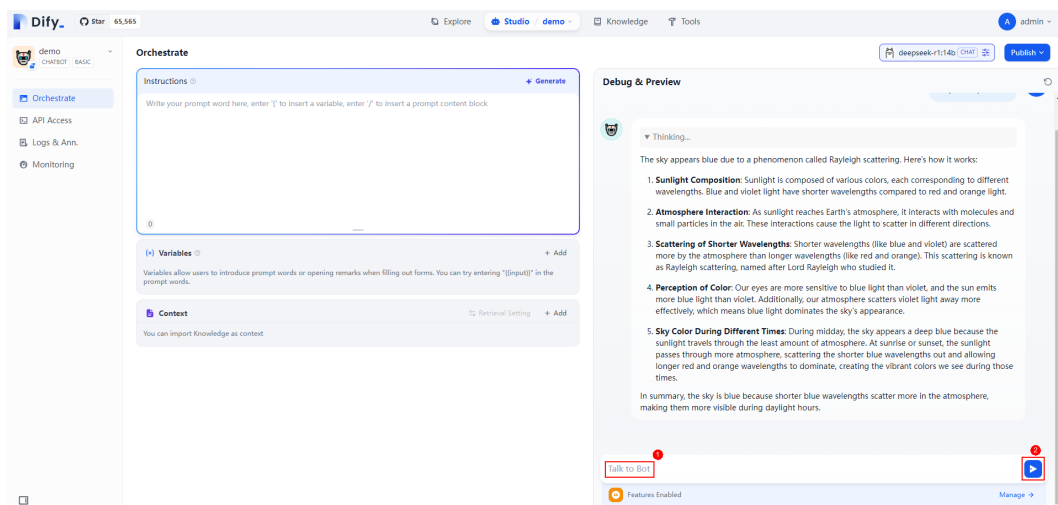


Figure 3-28 Create application



Step 8 Click **Orchestrate** in the left panel and start testing in the Debug & Preview window.

Figure 3-29 Debug & Preview

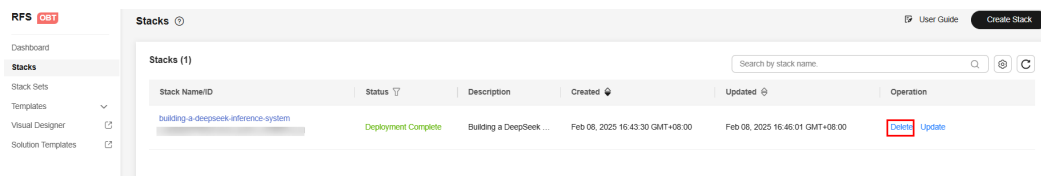


----End

3.4 Quick Uninstallation

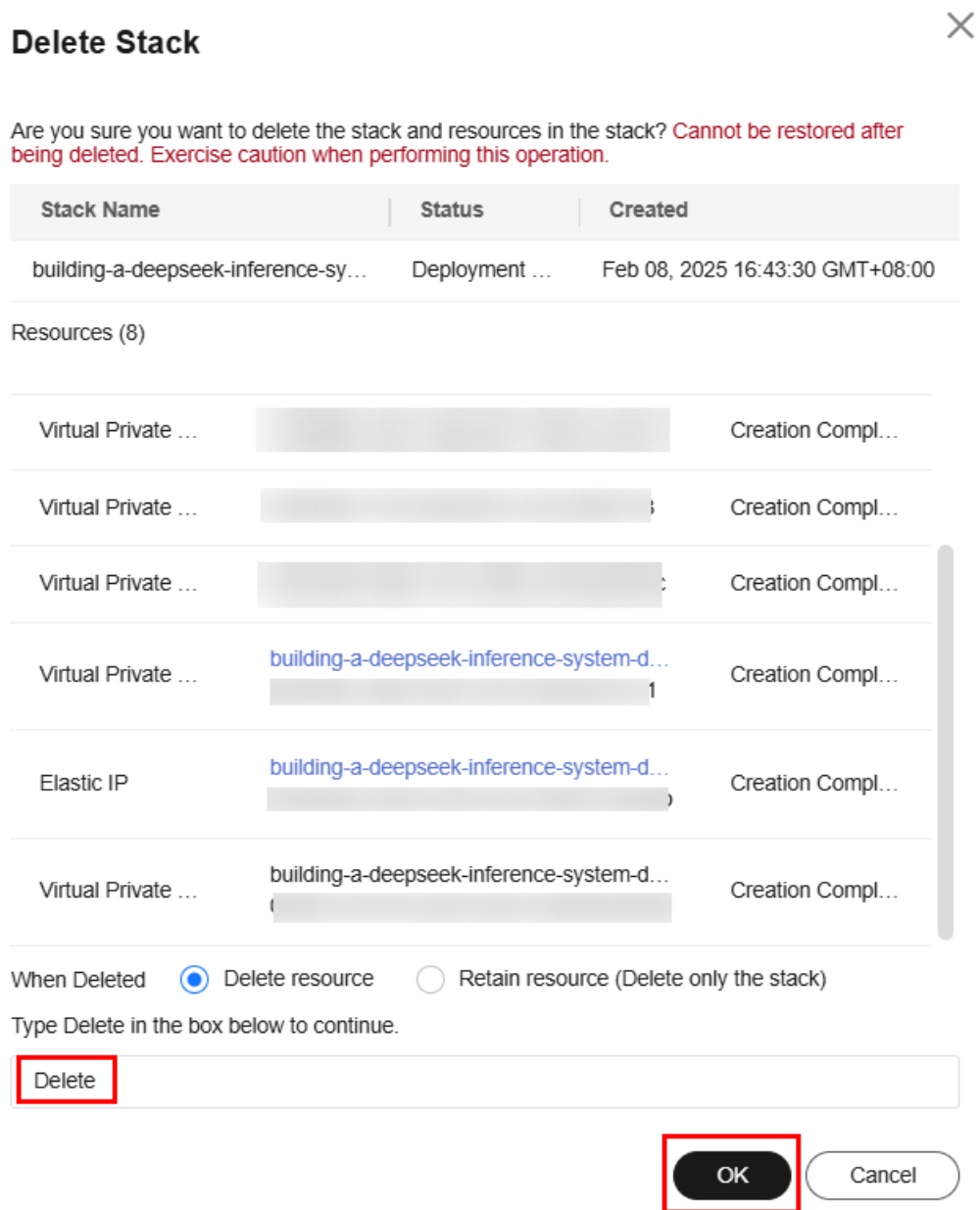
Step 1 Log in to the [RFS console](#). On the Stacks page, locate the resource stack you created and click Delete in the Operation column.

Figure 3-30 Deleting a stack



Step 2 In the displayed Delete Stack dialog box, set When Deleted to Delete resource, enter "Delete" and click OK.

Figure 3-31 Confirming the deletion



----End

4 Appendix

Terms

- **Flexus X Instance (FlexusX):** FlexusX is a next-generation flexible cloud server service designed for small- and medium-sized enterprises (SMEs) and developers. FlexusX provides functions similar to what ECS provides. In addition, with FlexusX, you can flexibly configure vCPU to memory ratios to match your specific needs and change server specifications without service interruptions. For details, see [Flexus X Instance \(FlexusX\)](#).
- **Elastic Cloud Server (ECS):** ECS provides secure, scalable, on-demand compute resources, enabling you to flexibly deploy applications and workloads.
- **Virtual Private Cloud (VPC):** VPC allows you to isolate online resources with virtual private networks. VPC enables your cloud resources to securely communicate with each other, the internet, and on-premises networks.
- **Elastic IP (EIP):** EIP provides static public IP addresses and scalable bandwidths that enable your cloud resources to communicate with the Internet. You can easily bind an EIP to a FlexusX instance, ECS, BMS, virtual IP address, NAT gateway, or load balancer, enabling immediate Internet access.

5 Change History

| Released On | Description |
|-------------|---|
| 2025-02-08 | This issue is the first official release. |