solution

Building a DeepSeek Inference System

lssue	1.0.0
Date	2025-02-11





HUAWEI TECHNOLOGIES CO., LTD.

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

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

Contents

1 Solution Overview	1
2 Resource Planning and Costs	3
3 Procedure	7
3.1 Preparations	
3.2 Quick Deployment	
3.3 Getting Started	
3.4 Quick Uninstallation	
4 Appendix	27
5 Change History	

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 highperformance 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**.

Huawei Cloud Service	Resource Name	Configuratio n Example	Quantity	Estimated Monthly Cost
Virtual Private Cloud (VPC)	building-a- deepseek- Inference- system-demo	 Region: CN-Hong Kong CIDR Block: 172.16.0.0/ 16 	1	USD 0.00
Subnet	building-a- deepseek- Inference- system-demo- subnet	 Region: CN-Hong Kong IPv4 CIDR Block: 172.16.1.0/ 24 	1	USD 0.00
SecurityGroup	building-a- deepseek- Inference- system-demo	 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

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

Huawei Cloud Service	Resource Name	Configuratio n Example	Quantity	Estimated Monthly Cost
Flexus X Instance	building-a- deepseek- Inference- system-demo	 Region: CN-Hong Kong Pay-per- use: USD 0.14/hour Specificatio ns: Flexus X Instance Performan ce 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	 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

Huawei Cloud Service	Resource Name	Configuratio n Example	Quantity	Estimated Monthly Cost
Virtual Private Cloud (VPC)	building-a- deepseek- Inference- system-demo	 Region: CN-Hong Kong CIDR Block: 172.16.0.0/ 16 	1	USD 0.00
Subnet	building-a- deepseek- Inference- system-demo- subnet	 Region: CN-Hong Kong IPv4 CIDR Block: 172.16.1.0/ 24 	1	USD 0.00
SecurityGroup	building-a- deepseek- Inference- system-demo	 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

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

Huawei Cloud Service	Resource Name	Configuratio n Example	Quantity	Estimated Monthly Cost
Flexus X Instance	building-a- deepseek- Inference- system-demo	 Region: CN-Hong Kong Specificatio ns: Flexus X Instance Performan ce 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	 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

IAM	Agencies ③	Agencies ①						
Users	Delete Agencies available for cr	ation: 36		All	▼ rf_admin_trust	X Q		
User Groups	Agency Name/ID ↓≣	Delegated Party J≣	Validity Period ↓Ξ	Created 4F	Description ↓Ξ	Operation		
Permissions • Projects	rt_edmin_trust	Cloud service RFS	Unimited	Mar 13, 2023 14:49:16 GMT+08:00	-	Authorize Modify Delete		
Agencies								
Identity Providers								
Security Settings								

- 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

encies / Create Agen	су
★ Agency Name	rf_admin_trust
★ 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	RFS
* Validity Period	Unlimited -
Description	Enter a brief description.
	0/255
	Next Cancel

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

Figure 3-5 Selecting a policy/role

< Authorize Agency								
Select PalayRide (2) Select Scope (3) Fetals								
Assign selected permissions to rf_edmin_frust1.								
View Selected (1) Copy Permissions from Another Project	All policies/toles All services Tenant Administrator X Q							
Policy/Role Name	Туре							
DME AdministratorAccess Excortantedda Data Model Exgline transit administrator with full permissions.	System-defined policy							
Tenart Administrator (Exclude IAM)	System-defined note							
CS Tenart Admin Cloud Stream Service Tenart 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					
	Select Policy/Role Select Scope ③ Finish					
	1 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.

I Iguic 3 / / geneicspe	Figure	3-7	Agenciespe
-------------------------	--------	-----	------------

IAM	Agen	Agencies ①							
Users		Delete Agencies available for creat	ion: 32			Al	▼ rf_admin_trust	X Q	
User Groups		Agency NameID ↓≣	Delegated Party ↓≣	Validity Period ↓≡	Created 4F	Description J≣	Operation		
Permissions • Projects		rt_admin_trust	Account op_svc_IAC	Unlimited	Jan 16, 2023 17:57:41 GMT+08:00	Created by RF, Not delete.	Authorize Modify Delete		
Agencies									
Identity Providers									
Security Settings									

----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

Solution Architecture

Huawei Cloud	Subnet	Building a DeepSee Version: 1.0.0 Last Updated: February 22 Built By: Huawei Cloud Time Required for Deploy Time Required for Uninsta Estimated Cost Supported Regions:	K Inference System
User Elastic IP(EIP) DeepS	Security group eek-R1-Distill Server	View D	eployment Guide
		Deple	y (GPU-based)

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

Figure 3-9 Selecting a solution

< Create Stack	
1 Select Template	2 Configure Parameters 3 Configure Stack 4 Confirm Configurations
* Creation Mode	Existing Templates Visual Designer
* Template Source	Private Templates URL Upload Template A stack is created using a template. The template must contain the deployment code file which file name extension is if or tf.json. If a stack is created using a template.
* Template URL	https://documentation-samples-5.obs.ap-southeast- The URL must contain at least the deployment code file, and the file size cannot exceed 1 MB. Image: The URL must contain at least the deployment code file, and the file size cannot exceed 1 MB. Image: The URL must contain at least the deployment code file, and the file size cannot exceed 1 MB. Image: The URL must contain at least the deployment code file, and the file size cannot exceed 1 MB. Image: The URL must contain at least the deployment code file, and the file size cannot exceed 1 MB. Image: The URL must contain at least the deployment code file, and the file size cannot exceed 1 MB. Image: The URL must contain at least the deployment code file, and the file size cannot exceed 1 MB. Image: The URL must contain at least the deployment code file, and the file size cannot exceed 1 MB. Image: The URL must contain at least the deployment code file, and the file size cannot exceed 1 MB. Image: The URL must contain at least the deployment code file, and the file size cannot exceed 1 MB. Image: The URL must contain at least the deployment code file, and the file size cannot exceed 1 MB. Image: The URL must contain at least the deployment code file, and the file size cannot exceed 1 MB. Image: The URL must contain at least the deployment code file, and the file size cannot exceed 1 MB. Image: The URL must code file, and the file size cannot exceed 1 MB. Image: The URL must code file, and the file size cannot exceed 1 MB. </th
	Next

Step 3 On the **Configure Parameters** page, enter a stack name, configure parameters based on **Table1 Parameter description**, and click **Next**.

Figure 3-10 Configuring parameters

Create Stack			
Select Template	2 Configure Parameters 3 Configure Stack 4	Confirm Configurations	
* Stack Name building-a-de	epseek-inference-system	s, digits, underscores (), ar	nd hyphens (-) The stack name must be unique.
Description Building a De	eepSeek Interence System 36/255 //		
Configure Paramete	Q V Encrypt some resources based on the template	requirements. (?)	
Parameter	Value	Туре	Description
* vpc_name	building-a-deepseek-inference-system-demo	string	Virtual Private Cloud (VPC) name. This template uses a newly created VPC and the VPC name must be unique. It can contain 1 to 5
* security_group_name	building-a-deepseek-inference-system-demo	string	Security group name. This template uses a newly created security group. For details about how to configure security group rules, ref
* ecs_name	building-a-deepseek-Inference-system-demo	string	Cloud server name, which must be unique. It consists of 1 to 64 characters and can include letters, digits, underscores (_), hyphens (
			Previous Next

Table 3-1 Parameter description

Parameter	Туре	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	Туре	Mandatory	Description	Default Value
security_grou p_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_mod el	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	Туре	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	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. NOTE For GPU- based solution, GPU- based solution, GPU- accelerated type is required. Value can be found from the specification list page of the documentatio n. (Before executing the plan, please ensure the resource is available in the corresponding region).	DeepSeek-R1- Distill- Qwen-1.5B

Parameter	Туре	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_si ze	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	Туре	Mandatory	Description	Default Value
charging_mod e	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_mod e 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_mod e is set to prePaid (yearly/ monthly). The default value is 1. Value range: 1-9 (charging_uni t set to month); 1-3 (charging_uni t 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

< Create Stack						
Select Template	Select Template (Configure Parameters (Configure Stack) Configure Stack (Configure Stack)					
RFS is free of charge, but the resource	urces in the stack are not. Currently, you need to create an ex	ecution plan (free of charge) to	o obtain the estimated price.			
Template Info						
Stack Name	building-a-deepseek-inference-system		Description Building a DeepBeek Inference System			
Parameters 🖉						
Parameter Name	Value	Туре	Description			
vpc_name	building-a-deepseek-inference-system-demo	string	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, inclu			
security_group_name	building-a-deepseek-inference-system-demo	string	Security group name. This template uses a newly created security group. For details about how to configure security group rules, refer to the deployme			
ecs_name	building-a-deepseek-Inference-system-demo	string	Cloud server name, which must be unique. It consists of 1 to 64 characters and can include letters, digits, underscores (_), hyphens (-), and periods (.)			
parameter_quantity	1.5b	string	DeepSeek-R1-Distil Model Parameter Size. Supports 1.5b, 7b, 8b. Default is 1.5b.			
ecs_flavor	x1.4u.4g	string	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			
Estimated fee: You can obtain the estimate	d fee after creating an execution plan (free of charge).		Previous Create Execution Plan Directly Deploy Stack			

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	×
1 To preview your reso	urce billing information, you can create an execution plan.	
* Execution Plan Name	executionPlan_20250208_1643_83kb	J
Description	Enter a description of the execution plan.	
	0/255 /	J
	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	te Template/Parameter C
Basic Information Res	ources Outputs	Events	Template	Execution Plans			
Deploy						Enter a keyword.	Q
Execution Plan Na	me/ID	Status		Estimated Price ③	Created	Description	Operation
executionPlan_202 €	50208_1643_83kb I	Available	:	View Details	Feb 08, 2025 16:43:33 GMT+08:00	-	Deploy Delete

Figure 3-15 Confirming the execution plan

Execution Plan	\times
Are you sure you want to execute the plan?	
Execution Plan Name Status Created	
executionPlan_20250208_16 Available Feb 08, 2025 16:43:33 (GMT+
After the plan is executed, the stack is updated accordingly, and resource template are enabled, which may incur fees based on resource payment requirements.	s in the
Execute	No

- 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 Table2 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

< building-a-deepseek-infe	rence-system			Delete Update Template/Parameter C
Basic Information Resources	Outputs Events Templa	te Execution Plans		
			Resource Name	 Enter a keyword.
Time 🖨	Туре 🖓	Description	Resource Name/Type	Associated Resource ID
Feb 08, 2025 16:46:01 GMT+08:00	Log	Apply required resource success.	-	-

----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 Configurate the connection

CloudShell View Terminal	Help	>_		RemoteShell, not only a resource access tool	← Download
Connecting to a Remote Server ×					
		• Region :	ap-southeast-1		 ✓ Refresh
		• EUS .	O T2.16.1.149 (Private IP)		Ť
		Port : • User :			
		* Auth-Type:			
		* Password :			
		Session Name :	root@		
		🗹 Open Remot	e Host Filesystem		
		Note: - To ensure the minutes. - Please make s ECSs in the sect - When operation notifications wh - Huawei Cloud	security of the connection, the system will automatically disconnect sess ure to add inbound rules to allow external network traffic from CloudSh urity group. ons get stuck after remote login, please check the CPU and memory of t en abnormal ECS events occur. Shell will not save your password, please keep it properly.	iions that have not been active for more th ell Proxy Server (SSH default port 22) to be the machine. Please configure <u>Cloud Eye</u> to	an 20 e sent to the send alarm
			Connect Cancel		

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 commad start the dialog test.

Figure 3-20 Dialog test

root@building-a-deepseek-inference-system-demo:~# ollama run deepseek-r1:1.5b >>> why the sky is blue? <think></think>
The color of the sky, known as its blue, is primarily due to a combination of factors that create a soft, uniform lightness. Here are the key reasons why the sky appears blue:
1. **Scattering of Light**: When sunlight enters Earth's atmosphere, it passes through countless tiny particles called atmospheric molecules (such as nitrogen and oxygen). These molecules scatter visible light in all directions. Since shorter (violet) light waves have a higher wavelength than longer (blue) light waves, they scatter less efficiently. This process is known as Rayleigh scattering.
2. **Air Mass Effect**: As sunlight passes through the atmosphere from the Earth's surface up to the upper layer, it has to pass through more air mass (the mass of the air column above a given point). The longer the path of light through this medium, the more scattering occurs, which results in more red and orange hues dominating the sky. However, at higher altitudes, these effects are minimized due to less atmospheric oxygen content.
3. **Turbulence**: In the upper atmosphere, sunlight interacts with turbulence, causing it to scatter and lose energy as it travels through regions of high wind speed and turbulence. This further reduces the intensity of scattered light and results in a redder or bluish appearance.
4. **Earth's Surface Conditions**: The color can also be influenced by factors such as temperature, humidity, and cloud cover. On sunny days, when the sky is clear, it appears blue to an observer on Earth. However, on cloudy days with lots of rain or smoke, the sky may appear darker or redder.
5. **Local Weather Conditions**: The atmosphere's composition and density vary slightly depending on local weather conditions, which can alter the scattering effect and hence the observed color of the sky.
In summary, while the sky appears blue primarily due to Rayleigh scattering and the air mass effect, it is influenced by a complex interplay of atmospheric conditions, including temperature, humidity, and cloud cover.
>>>> Bend a message (/? for help)

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

Dify_		
	Hey, let's get star	rted!
	Selecome to Dify, please los	g in to continue.
	Email address	
	,	
	Password	Encoder and a second se
	Password	Porgot your password?
		W
	Sign	n in
	By signing up, you agree to our Terr	ms of Service & Privacy Policy
	If you have not initialized the account	int, please go to the initialization

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

Figure 3-24 Click Settings

Dify_ O Star 65,562	🕰 Explore 🌰 Studio 🖾 Knowledge 🚏 Tools	A admin ~
왕 All	Created by	y me 🗞 All Tags 🗸 🔥 admin wk1519147@163.com
CREATE ADD		Workspace
Create from Blank		Account
Create from Template Import DSL file		Settings
		Community
		Help
		Roadmap About 0.15.3 •
	No apps found	Log out

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

Figure 3-25 Add Model

Settings	Model Provider			×	
WORKSPACE	Models provided by Upstage, such as Solar-1-mini-chat.	AWS Bedrock's models.			
S Model Provider					
Members Members	LLM	LLM TEXT EMBEDDING RERANK			ł
Data Source	<pre>CopenRouter</pre>	ි Ollama	MISTRAL AI_		I
ය API Extension			Models provided by MistralAI, such as open-mistral-7b and mistral- large-latest.		l
GENERAL	LLM	+ Add Model	LLM		I
≭ _A Language					ł
	GroqCloud GroqCloud provides access to the Groq Cloud API, which hosts models like LLama2 and Mixtral.	lFeplicate	😑 Hugging Face		
		LLM TEXT EMBEDDING	LLM TEXT EMBEDDING		
	Korbits Inference		💥 ZHIPU-AI		
	(LLM) (TEXT EMBEDDING) (RERANK) (SPEECH2TEXT) (TTS)	LLM	LLM TEXT EMBEDDING		

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.

ь.

Add Ollama	ි Ollar
Model Type *	
	O Text Embedding
Model Name *	
deepseek-r1:14b	
Base URL *	
http://172.16.1.116:11434	
Completion mode *	
Chat	
Model context size *	
4096	
Upper bound for max tokens *	
4096	
Vision support	
	O No
How to integrate with Ollama 🖸	Cancel

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

Dify_ O Star 65,565	🔕 Explore	🖨 Studio	C Knowledge	1 Tools			(🔥 admin ~
All Chatbot R Agent S Workflow					Created by me	🛇 All Tags 🗵	Q Search	
CREATE ADP C. Create from Blank C. Create from Blank C. Create from Tempfate D. Import DSL file								
	1	No apps found						

Figure 3-28 Create application

Create from Blank			CHAT Ouickl	BOT v build an LLM-b	ased chatbot with simpl	e configuration.		
Choose App Type			You ca	n switch to Chatf	low later. Learn more			
OR BEGINNERS								
	_		•	Orchestrate		🔘 GP	T-lo curr li e - ta Publik S	,
Chatbot LLM-based chatbot with simple setup	Agent Intelligent agent with reasoning and autonomous tool use	Text Generator Al assistant for text generation tasks	5 5 6 6 9	INSTRUCTIONS ()	2002 (D) 🗋	CERUG & PREVIEW Hello, 1 am L. Lan anover your questions related Heavity natis a lowed set of the Tips for analyzing competitors in marks © Conversation Openar - Bet	Features Extranse web-sage user segeritence Image: Conservation Operater Image: The second	
B BETA Chatflow Workflow for complex multi-turn dialogues with memory	BETA Workflow Orchestration for single-turn automation tasks			- 'MARABLES (d) expert_Rame The name	+ X44 of the strategic consulting		Fe Lyked Samoti r LE 1495 Samoti r LE 149 Samoti r LE 149	
pp Name & Icon				KHOWLEDGE Marketing Basics Reand Studegy Market Research Oudde	2: Renew Cettings + Add eq. Interes (iq. Interes) (iq. Interes)		Setting up next questions suggestion can give use a better chet. Test to Speech Conversation messages can be converted to speec	ek Kh
escription (Optional)			•	Vision O	RESOLUTION () High Low	Talk to Difyillot	Citations and Ats/Bucians Drow source document and attributed sector of t generated context. Annotation Rophy	the second
Enter the description of the app								
No ideas? Check out our templates \rightarrow		Cancel Create ¥ @						

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



Dify_	O Star 65,56	s 🗅 Explore de Studio / d	emo 🕤 🖾 Kr	nowledge 🕆 Tools	🔺 admin ~
demo CHATBOT E	BASIC	Orchestrate			deepseek-r1:14b (CHAT) 🚓 Publish 🗸
		Instructions © + Ger	erate De	bug & Preview	0
Orchestrate		Write your prompt word here, enter '(' to insert a variable, enter '/' to insert a prompt content block			
API Access				Thinking	
🖪 Logs & Ann.					
Ø Monitoring				 Sunlight Composition: Sunlight is composed of various coll wavelengths. Blue and violet light have shorter wavelengths. 	Itening. Here's now it works: lors, each corresponding to different s compared to red and orange light.
		0		Atmosphere Interaction: As sunlight reaches Earth's atmos small particles in the air. These interactions cause the light f	sphere, it interacts with molecules and to scatter in different directions.
		(c) Variables :: Variables is users to introduce prompt words or opening newska when filing out forms. You can try entering "(input))" in the prompt words.	- Add	 Scattering of Shorter Wavelengths: Shorter wavelengths more by the atmosphere than longer wavelengths (like red as Rayleigh scattering, named after Lord Rayleigh who stud Perception of Color: Our eyes are more sensitive to blue lip more built built that wide! Arditionally are more sensitive atmospheres of the light than vide! Arditionally are more sensitive atmospheres of the sensitive sensitive and the sensitive atmospheres of the sensitive more built that the vide! Arditionally atmospheres of the sensitive attractive atmosphere of the sensitive sensitive atmospheres of the sensitive atmosphere atmosphere at the sensitive atmosphere of the sensitive attractive atmosphere atmosphere at the sensitive atmosphere at the sensitive atmosphere atmosphere at the sensitive atmosphere at the sensitive atmosphere atmosphere atmosphere atmosphere at the sensitive atmosphere atmosphere atmosphere	like blue and violet) are scattered and orange). This scattering is known lied it. ght than violet, and the sun emits catters violet light away more
		5 Context 55 Retrieval Setting	- Add	effectively, which means blue light dominates the sky's app	earance.
		You can import floowledge as context		5. Sky Color During Different Times: During midday, the sky sunlight travels through the least amount of atmosphere. A passes through more atmosphere, scattering the shorter bi longer red and orange wavelengths to dominate, creating t times.	appears a deep blue because the it surrise or sunset, the sunlight ue wavelengths out and allowing the vibrant colors we see during those
				In summary, the sky is blue because shorter blue wavelengths s making them more visible during daylight hours.	catter more in the atmosphere,
			T	alk to Bot	
					Manage 19
Er	nd				

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

RFS OBT		Stacks ③			🕼 User Guide	Create Stack		
Dashboard		Al-1-141						
Stacks		Stacks (1)				Search by stack name.		Q®C
Stack Sets		Stack Name/ID	Status 🖓	Description	Created 🖨	Updated 🕀	Operation	
Templates	~							
Visual Designer	C	building-a-deepseek-interence-system	Deployment Complete	Building a DeepSeek	Feb 08, 2025 16:43:30 GMT+08:00	Feb 08, 2025 16:46:01 GMT+08:00	Delete Update	
Solution Templates	ß							

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

 \times

Figure 3-31 Confirming the deletion

Stack Name		Status	Created	
building-a-deepseek-	inference-sy	Deployment	Feb 08, 20	25 16:43:30 GMT+08:0
Resources (8)				
Virtual Private				Creation Compl
Virtual Private			;	Creation Compl
Virtual Private			:	Creation Compl
Virtual Private	building-a-d	eepseek-inference	-system-d… 1	Creation Compl
Elastic IP	building-a-d	eepseek-inference	e-system-d…	Creation Compl
Virtual Private	building-a-d	eepseek-inference	-system-d	Creation Compl
When Deleted 💿 [Delete resource	Retain res	ource (Delete o	nly the stack)

----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.