



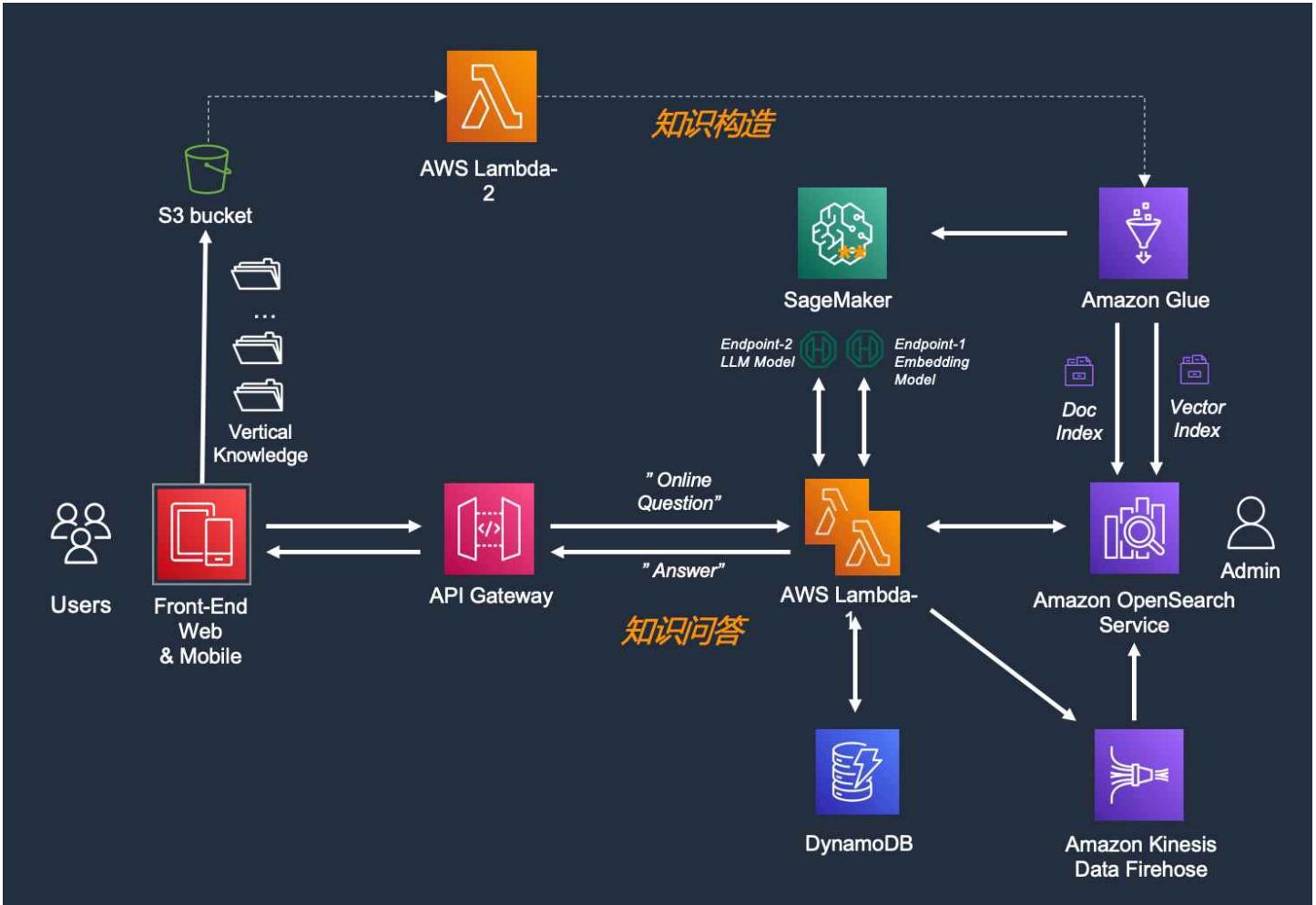
# RAG+LLM Chatbot 部署指南

修订记录：

版本	修订时间	说明
v1.1.3	2023/11/7	跟release 1.1.3 对应

## 1. 整体架构说明

- 采用serverless架构，主要使用的AWS 服务如下图所示
- 前端和后端独立部署，通过api gateway或者lambda function调用。如果前后端都在同账号/Region下，通过lambda function call直接调用，如果跨账号或者region，则通过api gateway endpoint调用。



## 2. 环境准备

### 2.1 AWS账号准备

**区域说明：**

本手册涉及创建的资源可放到中国区或者global区，请确保您的账号满足如下要求。

**可用资源数量：**

- VPC: 1 个 VPC，因此，您需要确认当前区域 VPC 的额度足够（默认上限为 5 / Region）。
- EIP: 2 个 EIP，因此，您需要确认当前区域 EIP 的额度足够（默认上限为 5 / Region）。

### 2.2 创建EC2实例用于运行部署CDK代码,并且host前端代码

(如果本地机器有开发环境，且有配置AWS管理员账号的ak/sk，也可以直接在本机上操作部署)

**操作系统：Amazon Linux 2023**

**实例类型：t3.large 或者c5.large**

**EBS卷：gp3 40G**

**是否创建密钥对：否，如果需要使用本地terminal登录则选是**

**网络安全组要求：开启22端口，能SSH到该实例**

## EC2 附加一个 IAM role,

对于开发人员，赋予**AdministratorAccess**，会避免各种权限引起的错误。

### 1. 选中ec2实例，修改IAM角色

The screenshot shows the AWS EC2 Instances page. A single instance named 'botdeploy' (ID: i-03be99b4c85898dd4) is selected. In the bottom right corner of the instance card, there is a blue-bordered button labeled '修改 IAM 角色' (Modify IAM Role), which is the target of a red box.

### 2. 创建新角色

This screenshot shows the 'Modify IAM Role' dialog for the selected EC2 instance. On the left, there's a dropdown menu for selecting an IAM role, currently set to '无 IAM 角色' (No IAM Role). To its right is a blue-bordered button labeled '创建新的 IAM 角色' (Create New IAM Role). A yellow warning box at the bottom asks if the user wants to remove all existing IAM roles from the instance if they choose to do so. On the right side of the dialog, there's a sidebar titled '选择可信实体' (Select Trusted Entity) with several options like '亚马逊云科技服务' (Amazon Web Services) and 'Web 服务' (Web Services), and a section for '使用案例' (Use Cases) with a dropdown set to 'EC2'.

### 3. 追加admin权限

This screenshot shows the 'Permissions' tab of the IAM Roles page. It lists two policies: 'AdministratorAccess' and 'AmazonSSMManagedInstanceCore'. The 'AdministratorAccess' policy is highlighted with a red box. At the top of the page, there are tabs for '权限' (Permissions), '信任关系' (Trust Relationships), '标签' (Tags), and '撤消会话' (Logout).

### 4. 重新回到EC2 console，刷新角色列表，选中新增的角色

## 修改 IAM 角色 信息

将 IAM 角色附加到您的实例。

实例 ID  
i-03be99b4c85898dd4 (botdeploy)

IAM 角色  
选择要附加到您的实例的 IAM 角色，如果您尚未创建任何 IAM 角色，也可以新建一个。您选择的角色会取代当前附加到您的实例的任何角色。

adminrole-for-ec2  [创建新的 IAM 角色](#)

## 2.3 登陆EC2， 安装环境

### 1. 在EC2 控制台使用EC2 connect to instance登录

EC2 > Instances > i-03be99b4c85898dd4 > Connect to instance

### Connect to instance Info

Connect to your instance i-03be99b4c85898dd4 (botdeploy) using any of these options

[EC2 Instance Connect](#) [Session Manager](#) [SSH client](#) [EC2 serial console](#)

### 2. 安装nodejs 18

```
1 sudo yum install https://rpm.nodesource.com/pub_18.x/nodistro/repo/nodesource-re
2 sudo yum install nodejs -y --setopt=nodejs.module_hotfixes=1
```

### 3. 安装 &启动 docker

```
1 sudo yum install docker -y
2 sudo service docker start
3 sudo chmod 666 /var/run/docker.sock
```

### 4. 安装git

```
1 sudo yum install git -y
```

### 5. 安装aws-cdk

```
1 sudo npm install -g aws-cdk  
2 sudo npm install --global yarn
```

### 3. 部署说明-后端部分

#### 3.1 下载代码

##### 3.1.1 通用方式

登陆到EC2以后，通过如下命令下载代码

```
1 git clone https://github.com/aws-samples/private-llm-qa-bot.git
```

##### 3.1.2 中国区Wordaround(Global Region直接跳过)

如果因为在中国区，网络原因导致下载失败，可以执行下面代码先下载到本地再通过S3 Bucket进行中转。

###### 1. 下载repo到本地，**并配置好aws的credentials**

```
1 cat ~/.aws/credentials  
2  
3 [default]  
4 aws_access_key_id = *****  
5 aws_secret_access_key = *****  
6  
7 [BJS_ACC]  
8 aws_access_key_id = *****  
9 aws_secret_access_key = *****
```

###### 2. 本地命令行执行如下脚本，同步代码到S3中，**脚本中的region需要进行相应修改**

```
1 region="cn-northwest-1"  
2 profile_name="BJS_ACC"  
3 timestamp=$(date +%s)  
4 bucket_name="code-temp-dir-$timestamp"  
5 aws s3 mb "s3://$bucket_name" --region $region --profile $profile_name  
6  
7 aws s3 sync ./private-llm-qa-bot "s3://$bucket_name/private-llm-qa-bot" --  
profile "$profile_name" --region "$region"
```

### 3. 切换到Ec2，执行如下脚本，从S3上同步下来, 注意bucket\_name和region需要手动设定

```
1 aws s3 sync "s3://$bucket_name/private-llm-qa-bot" private-llm-qa-bot --region $region
```

## 3.2 自动化部署

### 3.2.1 通用方式

在命令行执行如下代码

```
1 # 进入到deploy目录
2 cd private-llm-qa-bot/deploy/
3
4 #设置当前region, 需手动修改这个变量
5 export region=${region}
6
7 #生成环境变量 (报错可以用bash执行, Ubuntu系统可能会被替换成dash造成语法问题)
8 sh gen_env.sh ${region}
9
10 #创建Amazon OpenSearch可能需要的role, 账号内运行一次即可, 第二次运行会报错 (忽略即可)
11 aws iam create-service-linked-role --aws-service-name es.amazonaws.com
12
13 npm install
14
15 #如果这个region有其他项目执行过这个bootstrap, 会提示报错, 忽略即可
16 cdk bootstrap
17 cdk synth
18 #执行cdk部署程序
19 cdk deploy --require-approval never --all
```

注意：整个部署时间约25-30分钟，部署完毕以后，可以在CloudFormation中看到如下Stack信息

The screenshot shows the AWS CloudFormation console. On the left, there's a sidebar with sections like 'CloudFormation', 'Stacks (12)', 'Designer', 'Registry', 'Feedback', and 'Spotlight'. The main area shows a list of stacks, with one stack named 'QAChatDeployStack' expanded. This stack has three nested stacks: 'stackResource020D7F3E-2GOUWSHA151N', 'Ec2StackNestedStackEc2StackNestedSt', and 'QAChatDeployStack-vpcstackNestedStackvpcstackNestedSt'. The 'QAChatDeployStack' stack itself was created on 2023-05-18 15:34:45 UTC+0800 and updated on 2023-05-18 15:14:40 UTC+0800. The 'Outputs' tab is selected, displaying 13 outputs with their keys and values.

### 3.2.2 中国区Wordaround(Global Region直接跳过)

如果是在中国区部署，包安装很可能会失败，请参考下面建议：

在多个private-llm-qa-bot/code/{main, query\_rewriter, intention\_detect, chat\_agent}目录中 Dockfile 文件，需要进行如下改动指定包安装的源

```

1 ...
2 COPY requirements.txt .
3 +RUN pip3 config set global.index-url https://mirrors.ustc.edu.cn/pypi/web/simple
4 RUN pip3 install -r requirements.txt --target "${LAMBDA_TASK_ROOT}"
5 ...

```

## 3.3 构建知识库

### 1. 创建Amazon OpenSearch索引

#### a. 进入EC2服务，并连接EC2 proxy实例

The screenshot shows the AWS EC2 Instances page. It displays two instances: 'aws-cloud9-chatbot-cloud9-01-ff5d0f9d03ac4bae90...' (Stopped) and 'QAChatDeployStack/Ec2Stack/ProxyInstance' (Running). The 'Running' instance is highlighted with a red box. The page includes filters, sorting options, and actions for managing instances.

#### b. 获取知识库创建Script

```
1 curl -LJO https://raw.githubusercontent.com/aws-samples/private-llm-qa-
```

```
bot/main/deploy/setup_knowledgebase.sh
```

【中国区Wordaround】如果因为网络原因，上面的脚本下载失败，可以从S3 download

```
1 aws s3 cp s3://$bucket_name/private-llm-qa-
bot/deploy/setup_knowledgebase.sh . --region $region
```

### c. 执行知识库创建Script

```
1 # OpenSearch的endpoint，可以从上面的Cloudformation的output中获取
2 # Dimension 是向量模型的输出维度，由选择的模型确定
3 sh setup_knowledgebase.sh ${OpensearchEndpoint} ${Dimension}
```

## 2. 配置OpenSearch Dashboard (可选)

OpenSearch Dashboard主要用于获取OpenSearch索引的内部细节，是进行效果深入调优的方式。

### a. 安装nginx

```
1 sudo yum update
2 sudo amazon-linux-extras install nginx1 -y
```

### b. 创建nginx 配置文件, 进行反向代理

```
1 cd /etc/nginx/conf.d
2 sudo vim default.conf
```

### c. 修改如下, domain\_endpoint需要设置为创建的Amazon OpenSearch Endpoint(在Cloudformation获取)

```
1 # ${OpenSearch Endpoint} 需要从cloudformation的输出中获取;
2 server {
3     listen 80;
4     server_name localhost;
5
6     location / {
7         proxy_pass https://vpc-domain66ac69e0-pl89l5ymkgar-v64op6yrjkmn3phter;
```

```

8     proxy_pass https://${OpenSearch Endpoint};
9 }
10 }

```

#### d. 重启nginx

```

1 sudo nginx -t
2 sudo systemctl restart nginx

```

#### e. 访问OpenSearch Dashboard

i. 打开Dashboard网页，可以参见下图(Cloudformation Output)找到对应的网址

Key	Value	Description
APIGatewayEndpointUrl	<a href="https://acwixwmd7.execute-api.us-west-2.amazonaws.com/prod/">https://acwixwmd7.execute-api.us-west-2.amazonaws.com/prod/</a>	-
DownloadKeyCommand	aws secretsmanager get-secret-value --secret-id ec2-ssh-key/cdk-keypair/private --query SecretString --output text > cdk-key.pem && chmod 400 cdk-key.pem	-
embeddingendpoint	9.....3-05-18-07-05-20-embedding-endpoint	-
embeddingmodelname	9.....3-05-18-07-05-20-embedding	-
GlueJobName	chatbotfromstodoaf988AG33-9qjeySWm8qvz	-
KinesisFirehoseRole	QAChatDeployStack-chatbotkinesisfirehoseAADBC891-7Q5Q7TE715K	-
Ilmchatglmendpoint	.....7-23-05-18-07-05-20-ilm-chatglm-endpoint	-
modelname	.....7-23-05-18-07-05-20-ilm-chatglm	-
OpenSearchEC2ProxyAddress	<a href="http://34.2.14.15.111.8081/_dashboards/">http://34.2.14.15.111.8081/_dashboards/</a>	-
opensearchendpoint	yoc-domain66ac69e0-p18915ymkqr-v5406vrkmn9ohemmz1hwkne.us-west-2.es.amazonaws.com	-

ii. 访问页面后，可以对Opensearch进行操作，比如查询索引的schema，如图

```

1+ {
2+   "clustering": {
3+     "index": {
4+       "aliases": {},
5+       "mappings": {
6+         "properties": {
7+           "content": {
8+             "type": "text",
9+             "analyzer": "ik_max_word",
10+            "search_analyzer": "ik_smart"
11+          },
12+          "doc": {
13+            "type": "text",
14+            "analyzer": "ik_max_word",
15+            "search_analyzer": "ik_smart"
16+          },
17+          "doc_author": {
18+            "type": "keyword"
19+          },
20+          "doc_category": {
21+            "type": "keyword"
22+          },
23+          "doc_title": {
24+            "type": "keyword"
25+          },
26+          "doc_type": {
27+            "type": "keyword"
28+          },
29+          "embedding": {
30+            "type": "knn_vector",
31+            "dimension": 1024,
32+            "method": "l2",
33+            "engine": "mlslib",
34+            "space_type": "cosinesimil",
35+            "max_top_k": 100,
36+            "parameters": {
37+              "ef_construction": 128,
38+              "m": 16
39+            }
40+          }
41+        }
42+      }
43+    }
44+  }
45+
46+ }

```

## 3.4 串联整个流程

### 3.4.1 模型部署

i. 模型部署环境

## A. [Global Region] 图示步骤

## B. [China Region] 图示步骤

### ii. 模型部署步骤

#### A. LLM模型部署

如果使用Bedrock或者其他SAAS化的LLM API，这步可跳过，否则使用这个[目录](#)内notebook进行部署，默认选择[chatglm3\\_deploy.ipynb](#)进行部署。

#### B. Embedding模型部署

如果使用Bedrock内的Embedding API，则跳过这个步骤。bedrock内可使用下面三个embedding模型：

- [cohere.embed-multilingual-v3](#)
- [cohere.embed-english-v3](#)
- [amazon.titan-embed-text-v1](#)

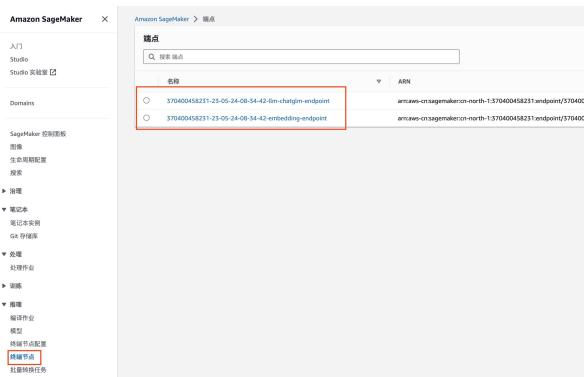
使用这个[目录](#)内notebook进行部署，默认选择[bge\\_zh\\_deploy.ipynb](#)进行部署

### 3.4.2 设置Lambda函数的环境变量

共有五个Lambda的环境变量需要设置为部署的模型Endpoint, 分别是:

- [Ask\\_Assistant](#)
- [Trigger\\_Ingestion](#)
- [Detect\\_Intention](#)
- [Query\\_Rewrite](#)
- [Chat\\_Agent](#)

具体请参考下面两个图, 环境变量名称为embedding\_endpoint和llm\_model\_endpoint。



SageMaker中部署的模型Endpoint

Environment variables (19)	
The environment variables below are encrypted at rest with the default Lambda service key.	
Key	Value
llm_endpoint	vpc://domain66a69e0-avaliabilitygroup-45d0ba3un2tryphqf5yuu-east-1.es.amazonaws.com
llm_index	chatbot-index
llm_knn_field	embedding
llm_results	2
lm256_llm_threshold_hard	2
lm256_llm_threshold_soft	10
chat_session_table	QChatDeployStock-chatbotSessionTable7C56-C4949A9C10C
embedding_endpoint	bge-dh-15-2023-09-17-01-00-27-08-endpoint
Kendra_index_id	f05f962-4ca8-4a65-9e60-08f13ef460c
Kendra_result_num	3
kmr_llm_threshold_hard	0.6
kmr_llm_threshold_soft	0.8
kmr_llm_threshold_hard	0.6
kmr_llm_threshold_soft	0.8
lambda_feedback	lambda_feedback
lm_model_endpoint	106459800180-23-10-09-13-08-15-lm-default-endpoint
neighbors	1
prompt_template_table	QChatDeployStock-promptTemplateTable32427-YOPM4aDEV1
TOP_K	4

Lambda Ask\_Assistant 环境变量设置

如果没有部署Sagemaker的endpoint，使用的是bedrock，设置方面遵循下面三个要求



1. 使用Bedrock的前提条件

- a. 需要提前申请Bedrock权限, 进入Bedrock在左边框找到'Model access'进行设置
- b. 部署在us-east-1, us-west-2这两个region, 其他region的bedrock服务不全, 如ap-southeast-1
2. 如果使用Bedrock embedding, 则embedding\_endpoint设成模型名称, 如'cohere.embed-multilingual-v3', 目前可以支持以下embedding模型:
  - cohere.embed-multilingual-v3
  - cohere.embed-english-v3
  - amazon.titan-embed-text-v1
3. 如果使用Bedrock Claude, 则 llm\_model\_endpoint 留空。

## 3.5 日志回流检索

参考[workshop](#)部分章节。

## 3.6 安全防范措施

### 3.6.1 删掉OpenSearch Proxy Ec2的安全组 80端口

入站规则 (5)								
	Name	安全组规则 ID	IP 版本	类型	协议	端口范围	源	描述
<input type="checkbox"/>	-	sgr-028041fd4bddeb04...	IPv4	HTTPS	TCP	443	10.22.0.0/16	from 10.22.0.0/16:443
<input type="checkbox"/>	-	sgr-0679b9092260a45...	IPv4	SSH	TCP	22	0.0.0.0/0	Allow SSH Access
<input type="checkbox"/>	-	sgr-08279b8a111900d...	IPv4	自定义 TCP	TCP	8081	0.0.0.0/0	Allow HTTP 8081 port Access
<input type="checkbox"/>	-	sgr-08715bd02a24c2fcf...	-	所有流量	全部	全部	sg-0c41cef06e0322cb...	Allow self traffic
<input type="checkbox"/>	-	sgr-0940e77fe455d9496...	IPv4	HTTP	TCP	80	0.0.0.0/0	Allow HTTP Access

这个80端口主要是用于从外网登陆OpenSearch Dashboard, 用于定制化修改对应的知识库 Schema信息。一般情况下无需打开, 如果有访问需要, 可以限制源为指定的外网IP。

## 4. 部署说明-前端部分 (可选)

### 4.1 下载代码

如果是**中国区**, 因为网络原因导致下载失败, 可以先下载到本地再通过S3 Bucket进行中转, 参考前面的3.1章节

```
1 git clone https://github.com/xiehust/chatbotFE.git
2 cd chatbotFE/deploy
```

进入chatbotFE/deploy目录后，修改env.sample。只需修改以下4项，根据自己的账号修改：

其中 **MAIN\_FUN\_ARN** 为后端部署完成之后主lambda的arn，可以在lambda控制台获取，也可以按以下格式拼写。**UPLOAD\_BUCKET**是部署后端时定义的文档存放S3桶名

例如：

CDK\_DEFAULT\_ACCOUNT=6310xxxxx15

CDK\_DEFAULT\_REGION=cn-northwest-1

如果是中国区部署：

MAIN\_FUN\_ARN=arn:aws-cn:lambda:{region}:{aws account id}:function:Ask\_Assistant

非中国区域：

MAIN\_FUN\_ARN=arn:aws:lambda:{region}:{aws account id}:function:Ask\_Assistant

UPLOAD\_BUCKET=是部署后端时定义的文档存放S3桶名

```
CDK_DEFAULT_ACCOUNT=631023...  
CDK_DEFAULT_REGION=us-...  
TOKEN_KEY=0001.chat-test_20250502  
OPENAI_API_KEY=  
START_CMD=/rs  
UPLOAD_BUCKET=6310...atbot-bucket  
UPLOAD_ORI_PREFTX=ai-content/  
MAIN_FUN_ARN=arn:aws:lambda:us-...-2:6310...15:function:Ask_Assistant  
embedding_endpoint=  
sd_endpoint_name=  
all_in_one_api=
```

再另存为.env文件

```
1 cp env.sample .env
```

## 4.2 运行CDK

1. 仍然在**chatbotFE/deploy**目录中，执行

```
1 npm install  
2 cdk synth  
3 cdk deploy --require-approval never
```

2. 执行成功之后，会在账号中部署2个api gateway，其中https开始的是rest apigateway， wss开始的是websocket apigateway.

```
Deployment time: 101.66s  
Outputs:  
BackendCdkStack.APIGatewayendpointurl = https://g1w[REDACTED].execute-api.cn-northwest-1.amazonaws.com.cn/prod/  
BackendCdkStack.ChatBotWsApiURL = wss://1df[REDACTED].cn-northwest-1.amazonaws.com.cn/Prod
```

3. 回到上一级目录，在chatbotFE目录下，修改env.sample文件

```
1 cd ..
```

📌 **REACT\_APP\_API\_http**=上一步的中输出的https开头的地址

**REACT\_APP\_API\_socket**=上一步的中输出的wss开头的地址

**REACT\_APP\_DEFAULT\_UPLOAD\_BUCKET**=是部署后端时定义的文档存放S3桶名

```
app > chatbotFE > .env  
REACT_APP_API_http=https://dv[REDACTED].execute-api.us-west-2.amazonaws.com/prod  
REACT_APP_API_socket=wss://gje7[REDACTED].execute-api.us-west-2.amazonaws.com/Prod  
REACT_APP_DEFAULT_UPLOAD_BUCKET=631023[REDACTED]-chatbot-bucket
```

4. 另存为.env 文件

```
1 cp env.sample .env
```

5. Build 前端js文件，在目录chatbotFE/下运行

```
1 yarn install  
2 yarn build
```

## 4.3 部署hosting

前端js code可以托管到S3，也可以继续在此EC2上部署。

### 如果托管S3：

需要开启S3的static website hosting, 设置 index document, permissions. 可以参考更多 [AWS doc Reference](#).

1. Create an S3 bucket named **bucket-name** on the Amazon S3 console.
2. Enable static website hosting of this bucket. In Index document, enter the file name of the index document `index.html`.
3. By default, the S3 bucket blocks public access. You need to change the setting by unchecking the option in the "Permissions" tab of the bucket detail page.
4. Add the policy below to the bucket policy to allow public access.

```
1 {
2     "Version": "2012-10-17",
3     "Statement": [
4         {
5             "Sid": "PublicReadGetObject",
6             "Effect": "Allow",
7             "Principal": "*",
8             "Action": [
9                 "s3:GetObject"
10            ],
11             "Resource": [
12                 "arn:aws:s3:::bucket-name/*"
13            ]
14        }
15    ]
16 }
```

5. 把chatbotFE/build 目录的内容上传到s3

```
1 cd chatbotFE
2 aws s3 sync ./build/ s3://bucket-name/
```

最后可以通过以下两种url格式访问：

- <http://bucket-name.s3-website-Region.amazonaws.com>
- <http://bucket-name.s3-website.Region.amazonaws.com>

**如果在EC2上部署：**

### 1. 安装pm2工具

在chatbotFE目录中，启动pm2

```
1 sudo yarn global add pm2
```

```
2 cd chatbotFE  
3 pm2 start yarn --name "chatbotFE" -- start
```

查看后台进程是否加入了运行

```
1 pm2 list
```

id	name	namespace	version	mode	pid	uptime	σ	status	cpu	mem	user	watching
0	chatbotFE	default	N/A	fork	32277	82s	0	online	0%	77.6mb	ec2-user	disabled

## 2. 设置默认启动

```
1 pm2 startup systemd
```

## 3. 设置ALB

- 在EC2控制台的负载均衡中，创建一个目标组，并选择之前的ec2实例

# Specify group details

Your load balancer routes requests to the targets in a target group and performs health checks on the targets.

## Basic configuration

Settings in this section can't be changed after the target group is created.

### Choose a target type

#### Instances

- Supports load balancing to instances within a specific VPC.
- Facilitates the use of [Amazon EC2 Auto Scaling](#) to manage and scale your EC2 capacity.

#### IP addresses

- Supports load balancing to VPC and on-premises resources.
- Facilitates routing to multiple IP addresses and network interfaces on the same instance.
- Offers flexibility with microservice based architectures, simplifying inter-application communication.
- Supports IPv6 targets, enabling end-to-end IPv6 communication, and IPv4-to-IPv6 NAT.

#### Lambda function

- Facilitates routing to a single Lambda function.
- Accessible to Application Load Balancers only.

#### Application Load Balancer

- Offers the flexibility for a Network Load Balancer to accept and route TCP requests within a specific VPC.
- Facilitates using static IP addresses and PrivateLink with an Application Load Balancer.

### Target group name

chatbotfe

A maximum of 32 alphanumeric characters including hyphens are allowed, but the name must not begin or end with a hyphen.

#### Protocol

#### Port

HTTP ▾ :

80

1-65535

## b. 注册目标端口为3000

EC2 > Target groups > chatbotfe > Register targets

### Register targets

Select instances, specify ports, and add the instances to the list of pending targets. Repeat to add additional combinations of instances and ports to the list of pending targets. Once you are satisfied with your selections, click Register pending targets.

Available instances (1/1)

Filter resources by property or value

Instance ID	Name	Status	Security groups	Zone	Public IPv4 address	Subnet ID
i-04e3a834d78983141	chatbotFEserver	Running	launch-wizard-1	us-west-2c	54.149.207.153	subnet-084270d413271ff39

1 selected

Ports for the selected instances  
Ports for routing traffic to the selected instances  
3000  
1-65535 (separate multiple ports with commas)

Review targets

## c. 创建一个application load balancer



1. ALB的VPC选择跟EC2在同一个VPC，安全组打开80端口公开访问。
2. 再配置原来EC2的安全规则，使得ALB的安全组可以访问EC2安全组的所有流量同一个安全组

A screenshot of the AWS Security Groups interface. At the top, there are filters for 'All traffic' and 'Custom'. A search bar contains the text 'sg-09d5324f1db3e3e1d'. Below the search bar is a list item with a delete button.

- d. ALB的目标组是上一步创建的目标组。

A screenshot of the AWS Load Balancer Listener configuration page. It shows a listener named 'Listener HTTP:80' with the following details:

Protocol	Port	Default action
HTTP	80	Forward to chatbotfe Target type: Instance, IPv4

Below the table, there is a 'Create target group' button. Under 'Listener tags - optional', there is a 'Add listener tag' button and a note about adding up to 50 more tags. At the bottom left, there is a 'Add listener' button.

- e. ALB的dns，即是访问chatbot的地址

## 4.4 配置管理员账户

1. 去dynamodb控制台，找到usertable.

A screenshot of the AWS DynamoDB table details page for 'BackendCdkStack-usertable46247387-N1PB6ZO2JYQ1'. The table has three items:

- BackendCdkStack-usertable46247387-N1PB6ZO2JYQ1
- chat\_user\_info
- http-crud-tutorial-items

A message at the bottom indicates '已完成. 已使用的读取容量单位: 0.5'. The '返回的项目 (0)' section shows '无项目'.

2. 点击创建项目，添加以下6个字段。

📌 **username:**yourname  
**password:**yourpassword  
**email:**[youremail@xx.com](mailto:youremail@xx.com)  
**groupname:**admin  
**status:**active  
**createtime:**2023-06-26T16:30:37.120Z

DynamoDB > Explore items: BackendCdkStack-usertable46247387-1EM3CNWDPKDKN > Create item

## Create item

You can add, remove, or edit the attributes of an item. You can nest attributes inside other attributes up to 32 levels deep. [Learn more](#)

Form

JSON view

Attributes		Add new attribute ▾
Attribute name	Value	Type
username - Partition key	admin	String
password	.....	String <span style="border: 1px solid #ccc; padding: 2px;">Remove</span>
email	admin@aws.com	String <span style="border: 1px solid #ccc; padding: 2px;">Remove</span>
groupname	admin	String <span style="border: 1px solid #ccc; padding: 2px;">Remove</span>
status	active	String <span style="border: 1px solid #ccc; padding: 2px;">Remove</span>
createtime	2023-06-26T16:30:37.120Z	String <span style="border: 1px solid #ccc; padding: 2px;">Remove</span>

Cancel Create item

访问ALB的dns或者S3托管的endpoint，进入login登陆页面。



### Sign in

Username \* —

Password \* —

Remember me

SIGN IN

SIGN UP

[Forgot password?](#)

## 4.5 使用流式输出

前端部署的websocket apigateaway用于流式输出，这时还需要后端有权限对这个apigateaway发送消息。

因此需要把后端部署的CDK wss接口参数更新一下，再执行一下后端的CDK部署代码，对wss接口授权。

## 1. 修改private-llm-qa-bot/deploy/.env文件

```
1 cd private-llm-qa-bot/deploy/  
2 vim .env
```

## 2. 加入以下3个变量：

 **use\_wss=1**

wss\_apild={apigateaway id}

如果是中国区：

wss\_resourceArn=arn:aws-cn:execute-api:{region}:{aws account id}:{apigateaway id}/\*//@connections/

非中国区：

wss\_resourceArn=arn:aws:execute-api:{region}:{aws account id}:{apigateaway id}/\*//@connections/

例如：

```
use_wss=1  
wss_apild=gje7{---sl  
wss_resourceArn=arn:aws:execute-api:us-west-2:631023274615:gje` :csl/*/*@connections/*
```

## 4. 修改保存后，再在private-llm-qa-bot/deploy 目录下执行，进行更新部署。更新部署过程加快，大约几分钟就完成

```
1 cdk synth  
2 cdk deploy --require-approval never
```

# 5. 功能介绍说明

## 5.1 基本功能介绍

AWS Chat Portal

聊天区

AWS智能问答

管理

文档库 5 提示词模板 反馈管理 Few shot示例管理 用户

对话

Input your question and press enter

自动建议

发送 新对话

Stream 使用知识库问答 多轮会话 隐藏引用 跟踪日志

更多设置 4

LLM模型 claude 最大Token数量 3000 Temperature 0.01

系统角色名 系统角色提示词 提示词模板 sso-chatbot-1102

1. 语言切换，支持中英文切换
2. 系统设置，默认不用填。如果要调用别的账户的后端，则需要配置其对应的API网关端点URL,AWS\_ACCESS\_KEY\_ID,AWS\_SECRET\_KEY,S3 Region,S3 Bucket

设置

API网关端点URL

OPENAI API KEY

AWS\_ACCESS\_KEY\_ID

AWS\_SECRET\_KEY

S3 Region

S3 Bucket

S3 Object prefix ai-content/

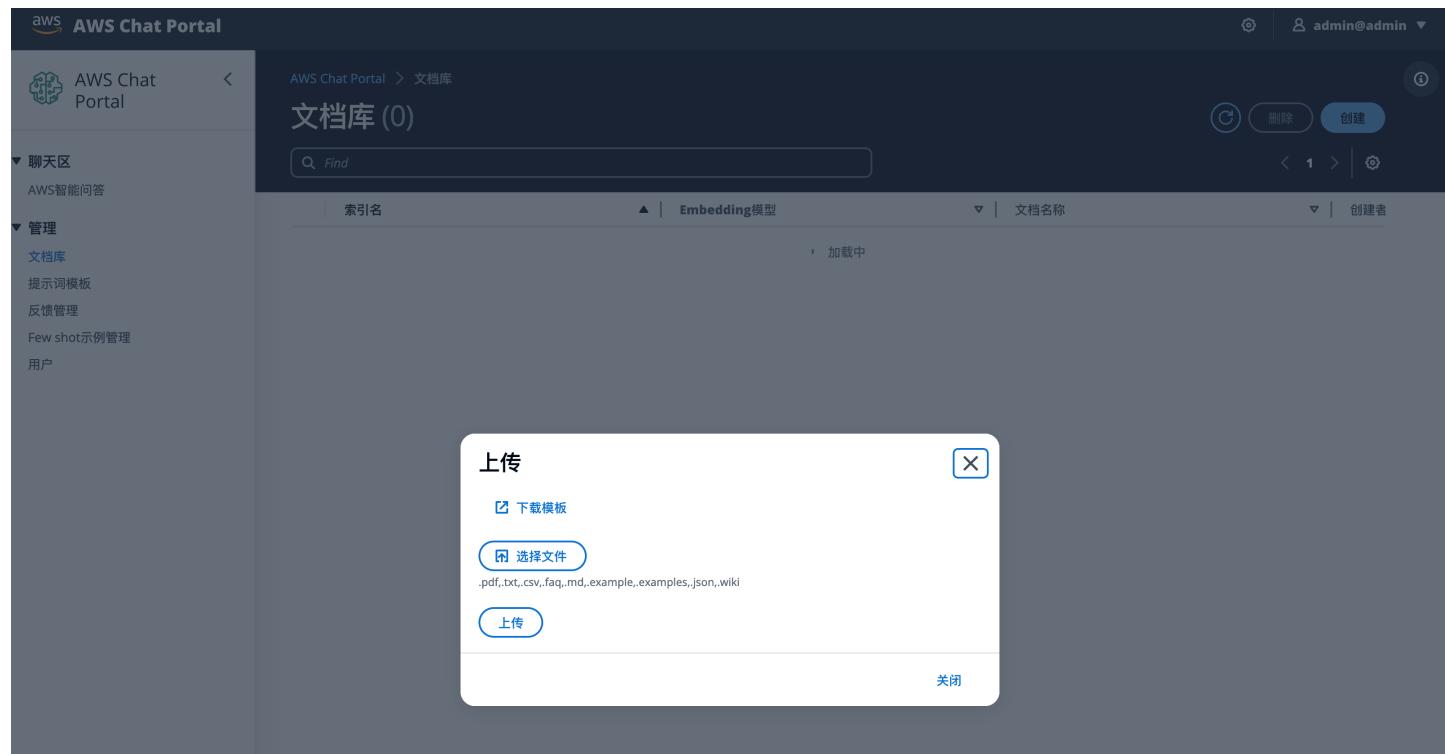
关闭 确认

3. 控制开启流式输出，是否启用知识库，多轮会话，显示引用文档，输出追踪日志功能。
4. 展开更多设置，可以选择模型，提示词模板，模型参数等。

## 5.2 上传知识

管理->文档库， 点击创建， 支持txt, pdf等类型， 也可以下载csv模板上传faq类型知识。

也选中之后删除知识。



## 5.3 管理提示词模板

管理->提示词模板， 可以创建， 删除， 修改提示词模板， 注意例子中{}的是保留字段， 不能自行修改， 具体解释可以见[说明](#)



## 添加模板

## 聊天区

AWS智能问答

## 管理

文档库

提示词模板

反馈管理

Few shot示例管理

用户

## 模板名称

Required

## 模板

Keywords: {system\_role\_prompt},{question},{role\_bot},{chat\_history},{context} [使用说明](#)

```

1 {system_role_prompt} {role_bot}, 请根据反括号中的资料提取相关信息, 回答用户的各种
    问题
2 ``
3 {chat_history}
4 {context}
5 ``
6 用户:{question}
7 ▼ {role_bot}:

```

Python

undefined: 0

△ undefined: 0



## 预览

, 请根据反括号中的资料提取相关信息, 回答用户的各种问题  
``

{chat\_history}  
{context}

``

用户:{question}

## 备注

Optional

取消

确认

## 5.4 反馈管理

用户可以对chatbot回答进行反馈，纠错。

## 对话



ODCR预留实例怎么计费?

AN

只要交付到客户账户里, 无论是否预留实例是否运行, 都按等同的按需费率计费。如果客户没有使用预留, 将在客户的EC2账单中显示为未使用的预留。如果客户运行的实例属性与预留匹配, 则客户只需要为该实例付费, 不需要为预留付费。


纠正 

具体的反馈在**反馈管理**中能够查看, 管理员角色用户, 可以双击单元格对反馈内容进行修改, 或者点击注入知识库。这样将会把这条内容直接注入到知识库中, 形成新的知识。生效时间大约1分钟。

AWS Chat Portal > 反馈管理

## 反馈管理 (1+)

Filter by text, property or value

Question	Original Answer	New Answer	Status	Timestamp
erik chen 是负责什么的	根据公开报道,Erik Chen是一位年轻的...	事实上 ✓ ✗	injected	2023-11-05 11:40:...

也可以用于逐条创建FAQ类型的知识

AWS Chat Portal > 反馈管理

## 反馈管理 (1+)

Filter by text, property or value

Question	Original Answer	New Answer	Status	Timestamp
erik chen 是负责什么的	根据公开报道,Erik Chen是一位年轻的...	事实上 ✓ ✗	injected	2023-11-05 11:40:...

### 创建新的FAQ

问题  
Your question

答案  
Your answer

取消 提交

## 5.5 Fewshot 示例管理

跟文档库功能类似，上传的是意图识别的示例，具体模板可以参考[这里的文件](#)

AWS Chat Portal

## Few shot示例管理 (5)

Find

索引名	Embedding模型	文档名称	创建者
chatbot-example-index	paraphrase-m-base-v2-2023-08-31-05-2...	ai-content/aws_service_owner.example	s3event
chatbot-example-index	paraphrase-m-base-v2-2023-08-31-05-2...	ai-content/emotion.example	s3event
chatbot-example-index	paraphrase-m-base-v2-2023-08-31-05-2...	ai-content/aws_service_status.example	s3event
chatbot-example-index	paraphrase-m-base-v2-2023-08-31-05-2...	ai-content/conversations.example	s3event
chatbot-example-index	paraphrase-m-base-v2-2023-08-31-05-2...	ai-content/admin/aws_faq.example	admin

## 5.6 用户管理

管理员可以在这里添加其他账户

首页 > 用户 > 添加用户

## 添加用户

用户名

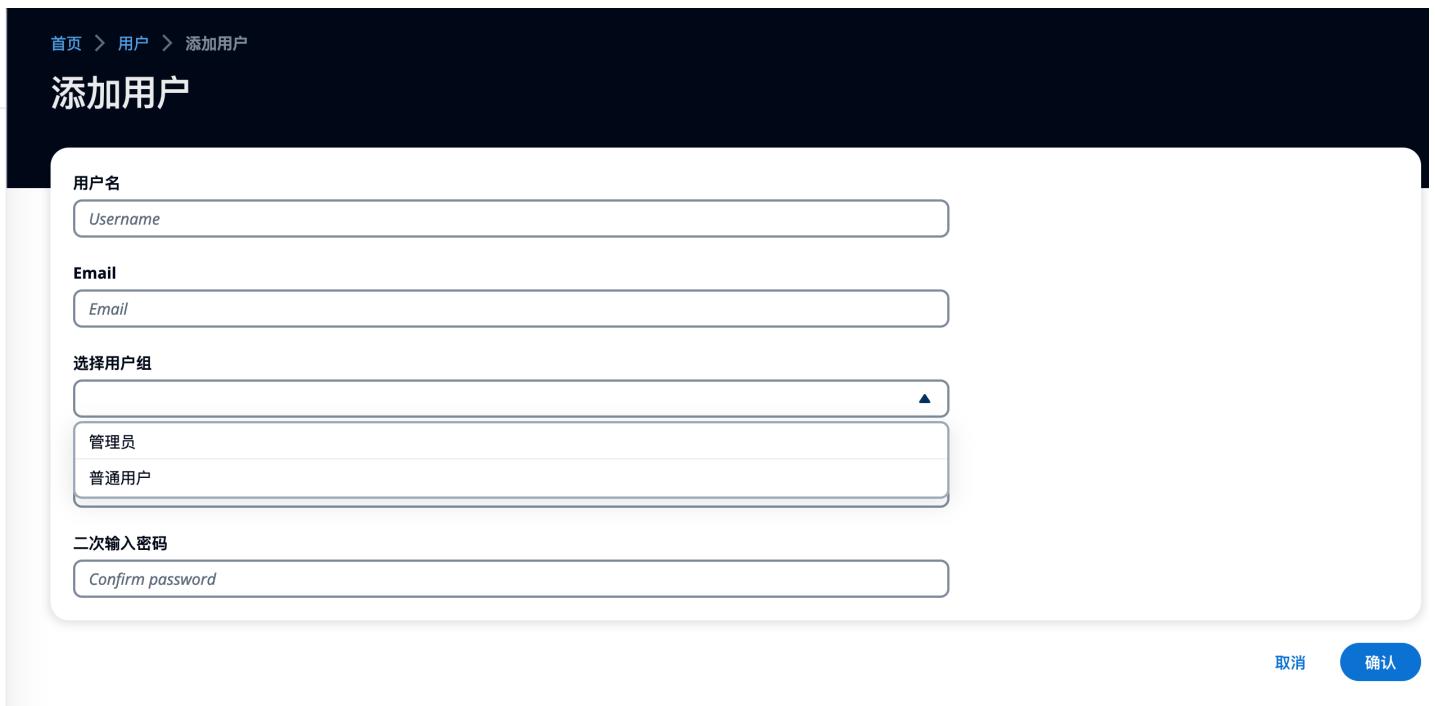
Email

选择用户组

管理员  
普通用户

二次输入密码

取消 确认

A screenshot of a user addition form. At the top, there's a breadcrumb navigation: 首页 > 用户 > 添加用户. Below it is a title 添 加用 户. The form consists of several input fields: '用户名' (Username) with placeholder 'Username', 'Email' with placeholder 'Email', '选择用户组' (Select User Group) with a dropdown arrow icon, and '二次输入密码' (Confirm Password) with placeholder 'Confirm password'. At the bottom right are two buttons: '取消' (Cancel) and '确认' (Confirm).

6. Demo视频

[https://www.bilibili.com/video/BV1HN4y1D7vy/?vd\\_source=2cb87d8dd3ca4ea778f5468be12405b3](https://www.bilibili.com/video/BV1HN4y1D7vy/?vd_source=2cb87d8dd3ca4ea778f5468be12405b3)