



# RWKV使用指南

# 1 商品说明

RWKV(Receptance Weighted Key Value)是一种结合RNN和Transformer优势的新型大语言模型架构,由元始智能团队主导开发并开源。

本商品通过kunpeng服务器进行安装部署的RWKV-4-Pile-1B5-EngChn-test4-20230115.pth 模型

# 2 商品购买

您可以在云商店搜索"RWKV"。

其中,地域、规格、推荐配置使用默认,购买方式根据您的需求选择按需/按月/按年,短 期使用推荐按需,长期使用推荐按月/按年,确认配置后点击"立即购买"。

2.1 商品支持自定义 ECS 购买,具体见章节 3.1.1

# 2.2 使用 RFS 模板直接部署

< 立即创建资源	ŧ		
<b>1</b> 25955	2) PREE3	88892	4 REMA
* 创建方式	日和武功 在可能	現化編唱發音譜	
* 植松中源	私有機板 UR 時个市時代が最高子の市内目	<b>1. 上份情報</b> 最初,個版中心成要有 詳新	HERCE_()*IRETHERING) .
★ 欄板 URL	https://mkp-privatedata-cn	nobs.cn-north-4.myhuawe	
	0 党潜编体服务不会在1	管理资源之外的场景使用您	上计检查数据,我们不会打包的搜索进行标准,为于参数中的数据数据,并给由中交体自动提用MAAS这里包的数据参数。

必填项填写后,点击 下一步



文档名称

送除模板 2 #数配置	3 资源性设置 4 配置确认		
<b>配置参数</b> 请输入关键字语乐参数名称			
參數名称	6	类型	編述
★ ECS实例電码	۵	李符串	ECS实例的管理员来码,来码架环度要求:来码要求长度范围为6号20位,来码至少必须包含大写字母、小写字母、数字40种种字符(\@\$\%=+{(});//
* 系统盘大小	40	number	设置系统盘大小 (至少40G, 数以40G).
* 数据量大小	50	number	设置挂着的数据最大小、若不需要数据最,可设置值为0,可将展实即循环配置。默认值为50。
*版本	「清笠坪 >	字符串	造輝版本
vpc IPv4阿歐	192.168.0.0/16	字符串	职值范围 10.0.0.016 to 10.255.255.0/24, 172.16.0.0/12 to 172.31.255.0/24, or 192.168.0.0/16 to 192.168.255.0/24.
子問iPv4問設	192.168.10.0/24	字符串	必须是CIDR地区、且在VPC的CIDR说内。子列境码不能大于28。
子同的网关	192.158.10.1	字符串	子网的男子。必须是子网旋内的台法中地址
* 付養受型(不包含应用防火燎)	() 请选择 > )	字符串	prePaid 预付惠,即也平如月; postPaid 指付最,即绘图付最
订购周期美型(不包含应用防火增)	month	字符串	当chargingMode为prePaid打生效且为必谦值、职业范围:month-月,year-年
订购商期款(不包含应用防火增)	1	字符串	当chargingMode为prePaid打击攻且为应纳值、取值包面; periodType=month (周期把型为月) 打,取值为[1, 9],periodType=year (周期把型为年) 打

<u>+</u>-#

11-T (1-1



创建直接计划后,点击 确定

<	立即创建资源栈					
0						
	配置参数 亿					
	参数名称	(ii	美型	1828		
	ECS实例密码		Addition (T) L Ed	>	×	18至125位,该码至少必须包含大写字母、小写字母、数字和特殊字符(1005%^_=+(();_/?)中的三种1
	系统量大小	40	创建1代171T划			
	数据量大小	50	0 通过执行计划,可	口预选出的资源交更信息。	1	展实际情况配置。默认重为50。
	版本	v1.13.0	* 执行计划名称	executionPlan_20250324_1057_4eda		
	vpc IPv4网般	192.168.0.0/16	17.4		3	11.255.0/24, or 192.168.0.0/16 to 192.168.255.0/24.
	子网IPv4网段	192.168.10.0/24	描述	请输入对执行计划的描述	8	R
	子网的网关	192.168.10.1		0/255 //		
	付姜英型(不包含应用防火境)	postPaid				
	订购周期类型(不包含应用防火墙)	month			<u></u>	m-月,year-年
	订购周期数(不包含应用防火壤)	1	字符串	当chargingMode为prePaid时生效且为必填值,取值范围:	4: perio	odType=month (周期純型为月) 时,职值为(1, 9), periodType=year (周期純型为年) 时,职值为(1, 3)
	资源转设置					
	IAM权限委托		同语	未开曲		2019-0251 未开放
奥月	攒估: 创建执行计划 (免募) 后可获取预估展用					(上-步)((2005))(2005

点击 部署



基本信息 资源 输出 事件 模板 負	机行计划				
22					【请输入关键字 Q
执行计划名称IID	秋志	奏用预结 ③	创建时间	描述	操作
executionPlan_20250324_1057_4eda 18a03c49-7e20-4b60-b8ca-689e5c63f2e7	创建成功,侍部署	重著集用的组	2025/03/24 10:58:08 GMT+08:00		部署

如下图 "Apply required resource success."即为资源创建完成

			1	■ 「 諸 個 長 個 子 一 前 個 人 天 個 子
inifi 🖕	中林英臣 公	the training	资源名称:类型	关联资源ID
5/03/24 11:00:06 GMT+08:00	88	Apply required resource success	Ē	-
5/03/24 11:00:01 GMT+08:00	生成完成	module.ecs.huaweicioud_compute_instance.ecs[0]: Creation complete after 57s [id-aa08d528.dc2e.409- b388.fb333ece8b44]	ecs ECS	aa08d928-dc2e-4019-b388-fb333ece8b44
5/03/24 11:00:01 GMT+08:00	支更倡要	Apply completel Resources: 8 added, 0 changed, 0 destroyed.	2	-
5/03/24 10:59:54 GMT+08:00	正在生成	module ecs.huaweicloud_compute_instance.ecs[0]: Still creating[50s elapsed]	ecs ECS	æ
5/03/24 10:59:44 GMT+08:00	正在生成	module ecs huaveicloud_compute_instance eci(0): Stil creating (40s elapsed)	ecs ECS	ан (т. т. т
5/03/24 10:59:34 GMT+08:00	正在生成	module.ecs.huaweicloud_compute_instance.ecs(0): Still creating(30s elapsed)	ecs ECS	-
5/03/24 10:59:24 GMT+08:00	正在生成	module ecs.huaweicloud_compute_instance ecs[0]: Still creating [20s elapsed]	ecs ECS	
5/03/24 10:59:14 GMT+08:00	正在生成	module ecs.huawelcloud_compute_instance ecs[0] SIII creating_ [10s elispsed]	ecs ECS	-
5/03/24 10:59:04 GMT+08:00	生成完成	module.vpc.husweicloud_vpc_subnet.subnet: Creation complete after 9s [id=4b3eceef.a475.4a7d-9e7c. 000003b63763]	subnet Subnet	4b3eceef-a475-4a7d-9e7c-000003b63763
503774 ID 50 04 OMT-08 00	TRAC	madula are humaniclassi compute incluses accilit Prantina	ecs	1.00

# 3 商品资源配置

商品支持ECS控制台配置,下面对资源配置的方式进行介绍。

## 3.1 ECS 控制台配置

## 3.1.1 准备工作

在使用ECS控制台配置前,需要您提前配置好安全组规则。

### 安全组规则的配置如下:

- 入方向规则放通CloudShell连接实例使用的端口22,以便在控制台登录调试。
- 出方向规则一键放通

## 3.1.2 创建 ECS

前提工作准备好后,选择ECS控制台配置跳转到购买ECS页面,ECS资源的配置如下图所



文档密级

### 示:

其砂砂黑			
<b>圣</b> 吨化 <b>旦</b>			
计费模式 ②			
包年/包月 <sup>芭</sup> 按需计费 竞优	介计费		
按需计费实例不支持备案。了解备案限制 🕑			
区域 ⑦			
<ul> <li>华北-北京四</li> <li>✓ 位 推</li> </ul>	荐区域 华北-北京四 华南-广州	华东-上海一 🍈 华北-乌兰察布一	前 西南-贵阳一
云服务器创建后无法更改区域;不同区域之间	内网互不相通, 请就近选择靠近您业务	的区域,减少网络时延。如何选择区域 📿	
可用区 ⑦			
随机分配 可用区1 可用	区2 可用区3 可用区7	随机至多可用区	
立例			
规格类型选型 业务场景选型			
CPU架构 ⑦			
x86计算			
VOON SIT			
实例筛选 ⑦			
请选择vCPUs V请	青选择内存 V	青输入规格名称模糊搜索	Q
<u> </u>	化型 鲲鹏招高1/0型		

HUAWEI	文档名称

#### 操作系统

镜像 ⑦					
公共镜像和	A有镜像	市场镜像			
C Huawei Cloud EulerOS	CentOS	Ubuntu	EulerOS	O	<b>OS</b> KylinOS
UnionTechOS	openEuler				
Huawei Cloud EulerOS	S 2.0 64bit for kAi2p with	HDK 23.0.1 and CANN	• • Q		

#### 存储与备份

系统盘 ⑦				
磁盘类型	系统盘大小(GiB)			
通用型SSD ~	- 40 +			
IOPS上限2,280, IOPS <u>突发上限</u> 8,000	高级设置			

#### ⊕ 増加一块数据盘

您还可以挂载 23 块磁盘 (云硬盘)

#### 一 开启备份

开启备份		CSDN @p_xcn
云服务器名称	ect-kette         一 先序覆宽           所天多公云服务器时, 艾特自动增加部方面最合成高有能至义规则命名。         ⑦	
描述		
登录凭证	在時 他的时 创造市业量	
	· · · · · · · · · · · · · · · · · · ·	
密钥灯	-2008- • C #MEMPOT ()	
云聲份	使用云面合质的,则称实面合在被声,在包裹是开放服装器产生的面合副本的容器。	
	夏在购买 使用已有 \$\$7585次 ⑦	
	最创可以解创造在服务器制度时代成数据。为了常约数据要全、强引强议定由用导创。	
云服外器组 (可违)	losses 0	
	—通信#25版品書編目- ▼ C	
	<b>和徽云昭外福山</b>	
No.45.10.19		
menoral-sc		
实例目定又数据注入	LUCKSUS UCHTS RUHLA/	
	echo footxxx ( dipassivid bash /home/int sh •	
购买量 — 1	+ * EEEEMI <u>¥0.3988.</u> ////.+ i#to:/invite.mii ¥0.80.os ③	上一步下一步。确认配置

### 值得注意的是:

- VPC您可以自行创建
- 安全组选择3.1.1章节中配置的安全组
- 弹性公网IP选择现在购买, 推荐选择"按流量计费",带宽大小可设置为5Mbit/s •
- 高级配置需要在高级选项支持注入自定义数据,所以登录凭证不能选择"密码",选 25-4-21 华为保密信息,未经授权禁止扩散 第5页,共7页 • 2025-4-21



择创建后设置

● 其余默认或按规则填写即可。

# 4 商品使用

4.1 RWKV 使用

### 4.1.1 激活环境

RWKV模型主要的通途包含文本生成、对话系统等,测试使用的模型文件主要是用来和机

器人进行聊天,会获得对应的回答。

登录到服务器上运行conda activate py39

在/opt/ChatRWKV下执行python v2/app.py 运行之后先选择语言,然后选择要使用的处理

器,CPU推理就选择CPU,最后选择需要进行推理的模型。

<pre>(base) [root@rwkv ~]# conda acti (pv39) [root@rwkv ~]# cd (opt/Ch</pre>	vate py39 atRWKV				
(py39) [root@rwkv ChatRWKV]# pyt	hon v2/app.	ру			
Please choose the language. 0 = English 1 = 简体中文 Waiting for the language ref (et	c. 0): 1				
Please choose the mode. 0 = GPU 1 = CPU Waiting for the mode ref (etc. 0	): 1				
Please choose the model. 0 = .cache 1 = RWKV-4-Pile-1B5-EngChn-test4 Waiting for the model ref (etc.	-20230115.p 0): 1	th			
Loading ChatRWKV - Chinese - cpu	- fp32 - Q	A_PROMP	T False		
RWKV_JIT_ON 1					
Loading model - /opt/models/RWKV	-4-Pile-1B5	-EngChn	-test4-2023011	15	
blocks.0.ln1.weight	float32	cpu	2048		
blocks.0.ln1.bias	float32	сри	2048		
blocks.0.ln2.weight	float32	сри	2048		
blocks.0.ln2.bias	float32	сри	2048		
blocks.0.att.time_decay	float32	cpu	2048		



文档名称

# 4.1.2 推理结果

#### Run prompt... ### prompt ###

l The following is a verbose and detailed conversation between an AI assistant called Bot, and a human user called User. Bot is intelligent, knowledgeable, wise and polite. ser: wat is lhc ot: LHC is a high-energy particle collider, built by CERN, and completed in 2008. They used it to confirm the existence of the Higgs boson in 2012. ser: wat is nlp ser: wat is nlp ot: the natural language processing language of smart phones, and its use on the internet. It is the main form of communication for modern day computers and smartphones

# 4.2 参考文档

<u>RWKV参考文档</u>