# OpenWeb UI 使用指南

Open WebUI和Ollama介绍 计费说明 部署架构 RAM账号所需权限 部署流程 部署步骤

## Open WebUI和Ollama介绍

Open WebUI 是一个功能丰富且用户友好的自托管 Web 用户界面(WebUI),它被设计用于与大型语言模型 (LLMs)进行交互,特别是那些由 Ollama 或与 OpenAl API 兼容的服务所支持的模型。Open WebUI 提供了 完全离线运行的能力,这意味着用户可以在没有互联网连接的情况下与模型进行对话,这对于数据隐私和安全敏 感的应用场景尤为重要。 以下是 Open WebUI 的一些主要特点:

- 1. 直观的界面: Open WebUI 的界面受到 ChatGPT 的启发,提供了一个清晰且用户友好的聊天界面,使得 与大型语言模型的交互变得直观。
- 扩展性:这个平台是可扩展的,意味着可以通过添加新的插件或功能来定制和增强其能力,适应不同的使用 场景和需求。
- 3. 离线操作: Open WebUI 支持完全离线运行,不依赖于网络连接,适合在任何设备上使用,无论是在飞机上还是在偏远地区。
- 4. 兼容性: 它兼容多种 LLM 运行器,包括 Ollama 和 OpenAl 的 API,这使得用户可以从多个来源选择和运行不同的语言模型。
- 5. 自托管:用户可以在自己的服务器或设备上部署 Open WebUI,这为数据隐私和控制提供了更高的保障。
- 6. Markdown 和 LaTeX 支持: Open WebUI 提供了全面的 Markdown 和 LaTeX 功能, 让用户可以生成富 文本输出, 这在科学和学术交流中非常有用。
- 7. 本地 RAG 集成:检索增强生成(RAG)功能允许模型利用本地存储的数据进行更深入和具体的回答,增强 了聊天交互的功能。

Ollama 是一个开源项目,其主要目标是简化大型语言模型(LLMs)的部署和运行流程,使得用户能够在本地机 器或私有服务器上轻松运行这些模型,而无需依赖云服务。以下是 Ollama 的一些主要特点和功能:

1. 简化部署: Ollama 设计了简化的过程来在 Docker 容器中部署 LLMs,这大大降低了管理和运行这些模型的复杂性,使得非专业人员也能部署和使用。

1

- 2. 捆绑模型组件: 它将模型的权重、配置和相关数据打包成一个被称为 Modelfile 的单元,这有助于优化模型的设置和配置细节,包括 GPU 的使用情况。
- 3. 支持多种模型: Ollama 支持一系列大型语言模型,包括但不限于 Llama 2、Code Llama、Mistral 和 Gemma 等。用户可以根据自己的具体需求选择和定制模型。
- 4. 跨平台支持: Ollama 支持 macOS 和 Linux 操作系统, Windows 平台的预览版也已经发布, 这使得它在 不同操作系统上的兼容性更好。
- 5. 命令行操作: 用户可以通过简单的命令行指令启动和运行大型语言模型。例如,运行 Gemma 2B 模型只 需要执行 ollama run gemma:2b 这样的命令。
- 6. 自定义和扩展性: Ollama 的设计允许用户根据特定需求定制和创建自己的模型,这为模型的个性化使用提供了可能。

通过 Ollama, 用户可以获得以下好处:

- 隐私保护:由于模型在本地运行,因此数据不需要上传到云端,从而保护了用户的隐私。
- 成本节约:避免了云服务的费用,尤其是对于大量请求的情况。
- 响应速度:本地部署可以减少延迟,提供更快的响应时间。
- 灵活性:用户可以自由选择和配置模型,以满足特定的应用需求。

Open WebUI和Ollama做了集成,可以轻松在web界面上管理大模型,支持在线下载,Ollama支持的模型可以在这里查看https://ollama.com/search

## 计费说明

Open WebUI面板在阿里云上的费用主要涉及:

- 所选GPU云服务器的规格
- 磁盘容量
- 公网带宽 计费方式: 按量付费(小时) 或包年包月 预估费用在创建实例时可实时看到。

## 部署架构

部署架构采用ECS(云服务器)单机部署

- 1		4			ŝ				۰.		.,		i.				,				e.	,		,		1		¢.										۰.					÷	e	-				,			÷	÷			2		6	~		•	×	
1			2	ò	V	P	<i>'</i> C							×							1			ł				÷																14								1											
			•						2					2																				•																		1								1	5		
1					- 1								1									1		1	1			×																		2									1					1			
								1	-		-		-	-	-	-	-	-		-	-	-							-	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		١					1		
° 1								1		+	Y	+	í,	ŧ	ti	N :	ż.	st	A	*	n.	1		1				1									2															1	1			Ľ		1		1	۰.		2
								1		÷	^	+	1	AL5	14	κ.	×	-1	*	:17	b	1						1															1													Ľ				1			
1								1						1																																										Ľ							
								1	Č.,					ć			1				0	1		ĵ.				1					1										1									1				Ľ							
				2	1				Č.								÷.				1			Ĵ.				ļ	1																							1	1	0		Ľ							
				1				Э		2			2	2			2	1			ċ.	1		ĵ,	2			2				2	0						2				0	1	1			1				2				Ľ			0	1	1	ĉ.	
													2	1			Ĵ.	1						į.				į.				2						2	0							2						Ì.	0	1		L						0	÷.
																					Ç.																																			L							
				2																								į.																												L			į.		1		
																							2																												2					L							ι,
			2	i.				2	i.	ŝ				i.				i.			k.	9		ŝ,				i.						D			1		à.						ŝ.	s.		- 2				į.	a.			Ŀ			i.		÷.		i.
																					+			+										C				e																		Ŀ		e					
				1				4					e.											÷.			c.	ŝ						0											2							÷	÷	i.		Ŀ							
			ć	÷	÷			1	÷	÷								-			÷	÷					ċ	÷		_								÷	÷				÷		÷				2							Ŀ				÷	÷,		
~			۰.					-									÷	,					0	,				e.		E	C	25	53	Ł	19	J.								10		×						×.				Ŀ							
			ē.					-						a.			×.					1	1	x	÷			×.					÷			6									÷	÷.										1	6						
-			•					-						2									s.,	1				2					2	2					2				2			1						;				Ŀ					$\mathcal{L}$		2
- 1								•							1									ı			0																													Ŀ					<		
·								1	*	5					1		2	÷.					1	ł			1	÷									5	1					*	4	×	2	2	1			1	÷	÷.	÷.	1	1		e.		Υ.	r.		5
-								-	•					3				1						ŗ			5	ć,	1				.*				2						1		5	2		1				ε.	5			Ŀ					٠,		
·								1	*						-			-					Ċ.,	1			1	£.				1											3			a.						×.				Ŀ				1	1		
× 1			6					1																			6										9							1								2	1	×		Ľ		ŕ.					
× 1			6	1				1	•				2					7										1															*		7				1			2	1	. *		Ŀ							
·	1			*				1	۲.				-		1		÷.,				*1	. *	1	÷				÷,	-			*	*				2	1				÷.	4	100		14	- (4)	ŕ			5	•				Ľ		×.			•	~	
1			÷.	ŝ	1			1		1			ť.	1							Ť	1		Ť				ń.					4				1		1				8					1			1	1	2	1	1	Ľ		ř.	÷.				4
·			×.,		14			1	1					4														5									٢.	1					÷.,	1			1								-	Ŀ					1		1
1			1		24			- 1	1	1				1		_	-				1				1			1					1						1														1							2			
1			1	8	1									1			1	5			5			č.	2			ł.						1					1				1		1	č.		1	2			r.						1		3			
1			1						۰.					4				.*			*	.*		*			1	1				5											4									•						۰.	۲	1			1
1			1	1				1	1					1	1		1						1	1			1	1				1	1	1				1	1				1		1		1	1	1			1	1						1	1	1	1	1
1			1	2	1			۰.	1	1			5	2	-		2	2			×;	1		8	e		5	ł,	1			1		-		1	3		1			ň.,	٩.,	2		2	2	1			÷.	×.,	1	1				÷.,	÷.,	1		1	1

## RAM账号所需权限

权限策略名称	备注
AliyunECSFullAccess	管理云服务器服务(ECS)的权限
AliyunVPCFullAccess	管理专有网络(VPC)的权限
AliyunROSFullAccess	管理资源编排服务(ROS)的权限
AliyunComputeNestUserFullAccess	管理计算巢服务(ComputeNest)的用户侧权限



## 部署步骤

1. 单击部署链接,进入服务实例部署界面,根据界面提示,填写参数完成部署。

### 2. 参数填写完成后可以看到对应询价明细,确认参数后点击下一步:确认订单

15.4%	4977 ( 101771)				<u> </u>		
责类型配置 ^							
四大	按量付费	包年包月					
和置 ~							
列类型	<b>猜逸</b> 选择vCPU	▼ 选择内存	▼ 报索实例规格	٩	<u> </u>		
	架构 异构计算 GPU / FPGA / NPU	J					
	分类 A10加速 V100加速	T4加速 GRID虚拟化					
	空间提择按: acs on 7i acs on 7a acs	an7s ess an7 ess an7s ess an6j	已进程格: ers on7i-	cBn1 2vlarne V			
	201001018 000-911 (000-911 0/000)	gin stees gin tees gin itees gind.		ood intruside V	ALCOHOLD AN ALCOHOLD	AL 2010 10	A NUMBER OF
	和時間時一〇	关例规范	VCPU + MMP +	GPU/FPGA @	父埠信王朔/晉羽	李考D(他 ①	- 处理的至今
	● GPU 计算型 gn7i	ecs.gn7i-c8g1.2xlarge	8 vCPU 30 GiB	A10	2.9 GHz/3.5 GHz	¥ 9.5326/85	Platinum 8369B
	〇 GPU 计算型 gn7i	ecs.gn7i-c16g1.4xlarge	16 vCPU 60 GiB	1 * NVIDIA A10	2.9 GHz/3.5 GHz	¥ 10.0934/85	Intel Xeon(Ice Lake) Platinum 8369B
	〇 GPU 计算型 gn7i	ecs.gn7i-4x.8xlarge	32 vCPU 128 GIB	4 * NVIDIA A10	2.9 GHz/3.5 GHz		Intel Xeon(Ice Lake) Platinum 8369B
	〇 GPU 计算型 gn7i	ecs.gn7i-2x.8xlarge	32 vCPU 128 GIB	2 * NVIDIA A10	2.9 GHz/3.5 GHz	-	Intel Xeon(Ice Lake) Platinum 8369B
				1 * KIUIDIA			Intel Veneriles Lakel
							1 2

#### 3. 确认订单完成后同意服务协议并点击立即创建

4. 等待部署完成后就可以开始使用服务,进入服务实例详情点击Address访问。

既觉	资源	事件	监控	法律管理	這規項	日志管理	升级历史	费用统计	番份与恢复	
¢ Intentite	¢.	иникан О	E#15	<b>天间条</b> 0	电影件	*	例安全事件.		服务资源 ECS 1	
被用 verAddras	aa 🗇	http://120	9.26.133.205.8	1080						
H使用 NerAddrau H信息	81 ©	http://120	3.26.133.205.8	1080						
使用 irAddra 信息	aa ()	Http:()120	1.26.133.205 a	8080					服务实务名称	11-699dc 708x3234eb2a728
D使用 NerAddrei H放息 服送	aa ()	Htp:@120 ● CBF#	1.26.133.205-8	1080					服务实件名称 创建和34	11-69#51:768x73334eb2#728 2024#117月19⊟ 10:42:16
的使用 rverAddrei N信息 5 1号述 1号记	aa ()	Http://120	).26.133.2053 1,919日 10:50	59					服务实件名称 创建90间 标签	si-69est:788x3334eb2a728 20248111月19⊟ 10-42:16 ©

### 5.使用服务

0		
	<b>R</b> 2 0 W W	
	豆束 Open webUI	
	电子邮箱	
	输入您的电子邮箱	
	密码	
	输入您的密码	
	登录	
	没有账号? 注册	