

# AI 大模型开发工程师之 中文最具潜力大模型

讲师：李希沅

# 目录

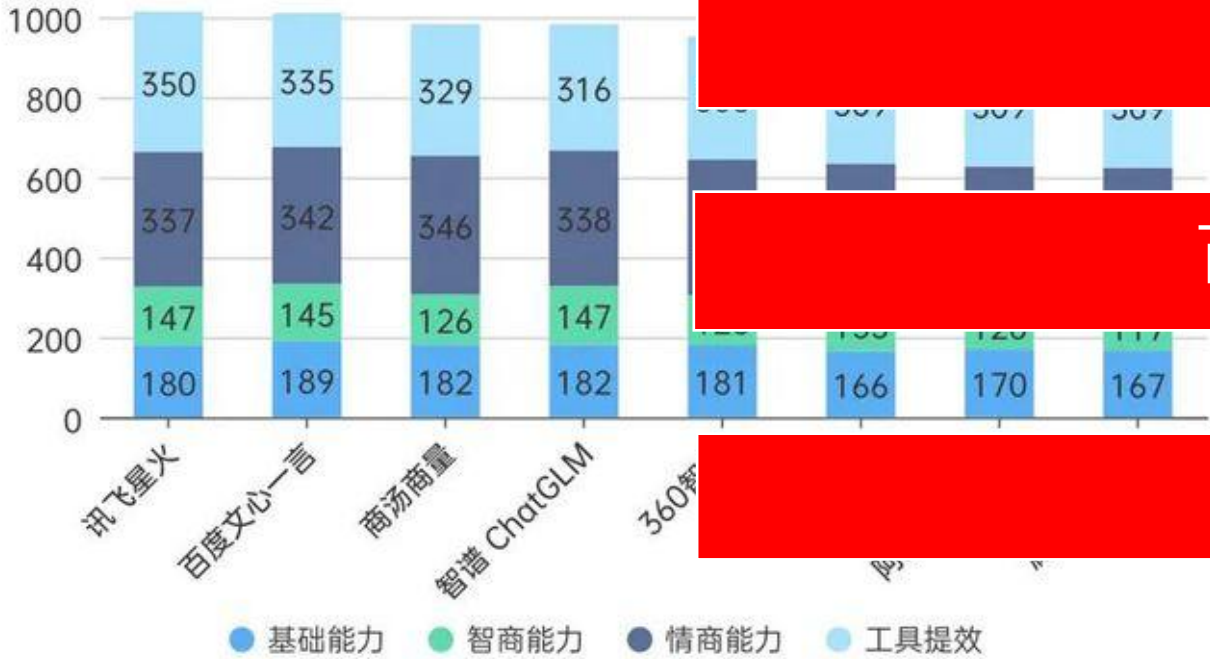
- 1 GLM大模型家族
- 2 认识ChatGLM3
- 3 ChatGLM3-6b私有化本地部署
- 4 ChatGLM3-6b私有化云部署

# **1 GLM大模型家族**

# 01、国产主流大模型

主流大模型综合指数情况

单位：分



数据来源：《人工智能大模型体验报告2.0》

支持中文的

开源的

可商用的

性能好

低成本部署

企业	大模型	开源	闭源
Meta	LLaMA2	✓	✓
			✓
			✓
			✓
Stability AI	StableLM	✓	✓
			✓
LMSYS	Vicuna	✓	✓
			✓
智谱	ChatGLM2-6B	✓	✓
	ChatGLM2-12B、		✓
	ChatGLM2-32B		✓
阿里	Qwen-7B、Qwen-7B-Chat	✓	✓
华为	盘古	✓	✓
			✓
			✓
	Baichuan-53B		✓



# 02、GLM的优势

## 深受客户的信赖

了解最具创新性的组织如何使用智谱AI，从他们的评价中获得更多价值



“

智谱 AI 是国内顶级人工智能科技公司，被评价为国内“最具 OpenAI 气质和水准”的 AI 公司。

周鸿祎  
360创始人、董事长兼CEO



< — — >

## 清华大牛掌舵，智谱AI一举融资25亿

和讯网 2023-10-20 16:45

投资界获悉，今日智谱AI宣布今年累计获得超25亿人民币融资，投资方阵容豪华

社保基金中关村(000931)自主创新基金（君联资本为基金管理人）、美团、蚂蚁、阿里、腾讯、小米、金山、顺为、Boss直聘、好未来、红杉、高瓴等多家机构及包括君联资本在内的部分老股东跟投。

这是一家从清华实验室走出来的大模型公司，集结了一群清华大牛——CEO张鹏毕业于清华计算机系，总裁王绍兰为清华创新领军博士，清华大学计算系教授唐杰也参与了孵化。

智谱AI，成立于2019年6月，由清华大学计算机系知识工程实验室（KEG）的技术成功转化而来。

## 03、千亿模型对比




Model	Open-sourced	Architecture					
		Major Lang.	Back-bone	Training Objective	Layer-Norm	PE	FFN
GPT-3 175B	×	English	GPT	SSL only	Pre-LN	APE	FFN
OPT-175B	✓					APE	FFN
PaLM-540B	×					RoPE	SwiGLU
BLOOM-176B	✓	Multi-lingual	GPT	SSL only	Pre-LN	ALiBi	FFN
GLM-130B	✓	Bilingual (English & Chinese)	GLM	SSL & MIP	Deep-Norm	RoPE	GeGLU
Effect	An open LLM for everyone	Less Bias & Toxicity: StereoSet: +12.7% CSP(↓): -1.4% RTP(↓): -3.0%	Perf. Improvements: BIG-bench-lite: +5.2% LAMBADA: +2.3% CLUE: +24.3% FewCLUE: +12.8%		Deep-Norm improves stability in training	RoPE works better with GLM	GeGLU performs better than FNN
Model	Training		Inference & Evaluation				
	Floating-point	Stabilization	Quantization	Acceleration	Reproducibility	Cross-Platform	
GPT-3 175B	FP16	undisclosed	undisclosed	undisclosed	×	NVIDIA	
OPT-175B	FP16	Hand-tuning	INT8	Megatron	×	NVIDIA	
PaLM-540B	BF16	Hand-tuning	undisclosed	undisclosed	×	undisclosed	
BLOOM-176B	BF16	Embedding Norm	INT8	Megatron	×	NVIDIA	
GLM-130B	FP16	Embedding Gradient Shrink (EGS)	INT4	Faster-Transformer	✓	• NVIDIA • Hygon DCU • Ascend 910 • Sunway	
Effect	FP16 supports more computing platforms	EGS improves numerical stability with little accuracy loss	It saves 75% mem in inference. 4 × 3090 or 8 × 2080 Ti.	×7-8.4 faster than Pytorch, ×2.5 faster than Megatron	All evaluation data & scripts are open	It supports more diverse adoption of LLMs	


# 04、GLM的发展历程

2022年  
2022年  
全面起  
2022年  
2023年  
2023年  
Agent  
2023:  
2023年




 ChatGLM3 Public


ChatGLM3 series: Open Bilingual Chat LLMs | 开源双语对话语言模型

 Python  5.2k  476




 CogVLM Public

a state-of-the-art-level open visual language model | 多模态预训练模型

 Python  2k  97

 AgentTuning Public

AgentTuning: Enabling Generalized Agent Abilities for LLMs

 Python  900  59

 AgentBench Public


A Comprehensive Benchmark to Evaluate LLMs as Agents

 Python  1.4k  64




 CodeGeeX2 Public

CodeGeeX2: A More Powerful Multilingual Code Generation Model

 Python  5k  310

 ChatGLM-6B Public

ChatGLM-6B: An Open Bilingual Dialogue Language Model | 开源双语对话语言模型

 Python  35.7k  4.8k

T模型架构;  
Geex 2,  
lv;  
强LLM  
源大模型。

github地址: <https://github.com/THUDM/>



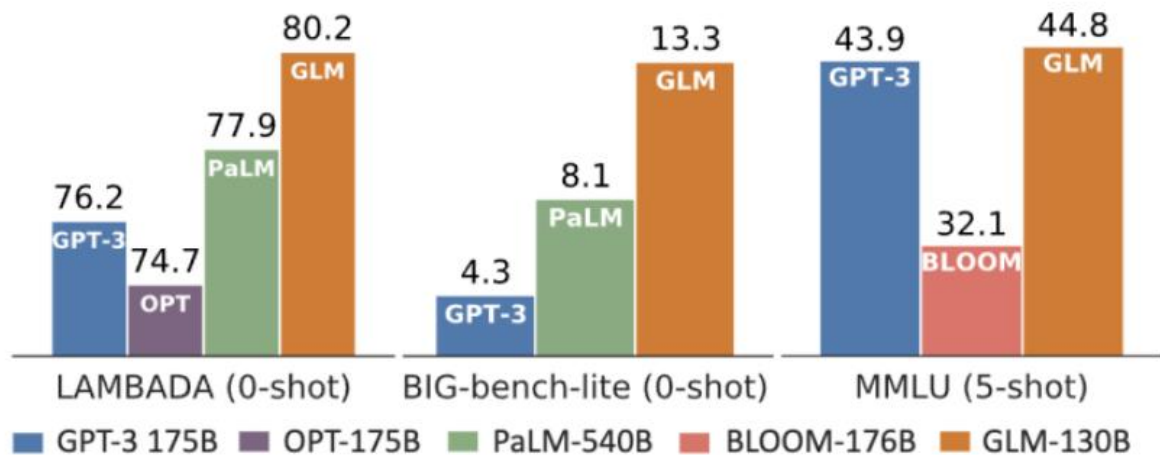
## 05、GLM家族之基座模型GLM-130B



# GLM-130B

An Open Bilingual Pre-Trained Model

- **Bilingual:** supports both English and Chinese.
- **Performance (EN):** better than GPT-3 175B (+4.0%), OPT-175B (+5.5%), and BLOOM-176B (+13.0%) on LAMBADA and slightly better than GPT-3 175B (+0.9%) on MMLU.
- **Performance (CN):** significantly better than ERNIE TITAN 3.0 260B on 7 zero-shot CLUE datasets (+24.26%) and 5 zero-shot FewCLUE datasets (+12.75%).
- **Fast Inference:** supports fast inference on both [SAT](#) and [FasterTransformer](#) (up to 2.5X faster) with a single A100 server.
- **Reproducibility:** all results (30+ tasks) can be easily reproduced with open-sourced code and model checkpoints.
- **Cross-Platform:** supports training and inference on NVIDIA, Hygon DCU, Ascend 910, and Sunway (Will be released soon).



体验网址: <https://www.zhipuai.cn/index.html>

论文地址: <https://arxiv.org/abs/2210.02414>




# 06、GLM家族之ChatGLM3

## ChatGLM3

 [HF Repo](#) •  [ModelScope](#) •  [Twitter](#) •  [\[GLM@ACL 22\]](#) [\[GitHub\]](#) •  [\[GLM-130B@ICLR 23\]](#) [\[GitHub\]](#)

 加入我们的 [Slack](#) 和 [WeChat](#)

 在 [chatglm.cn](#) 体验更大规模的 ChatGLM 模型。

[Read this in English.](#)

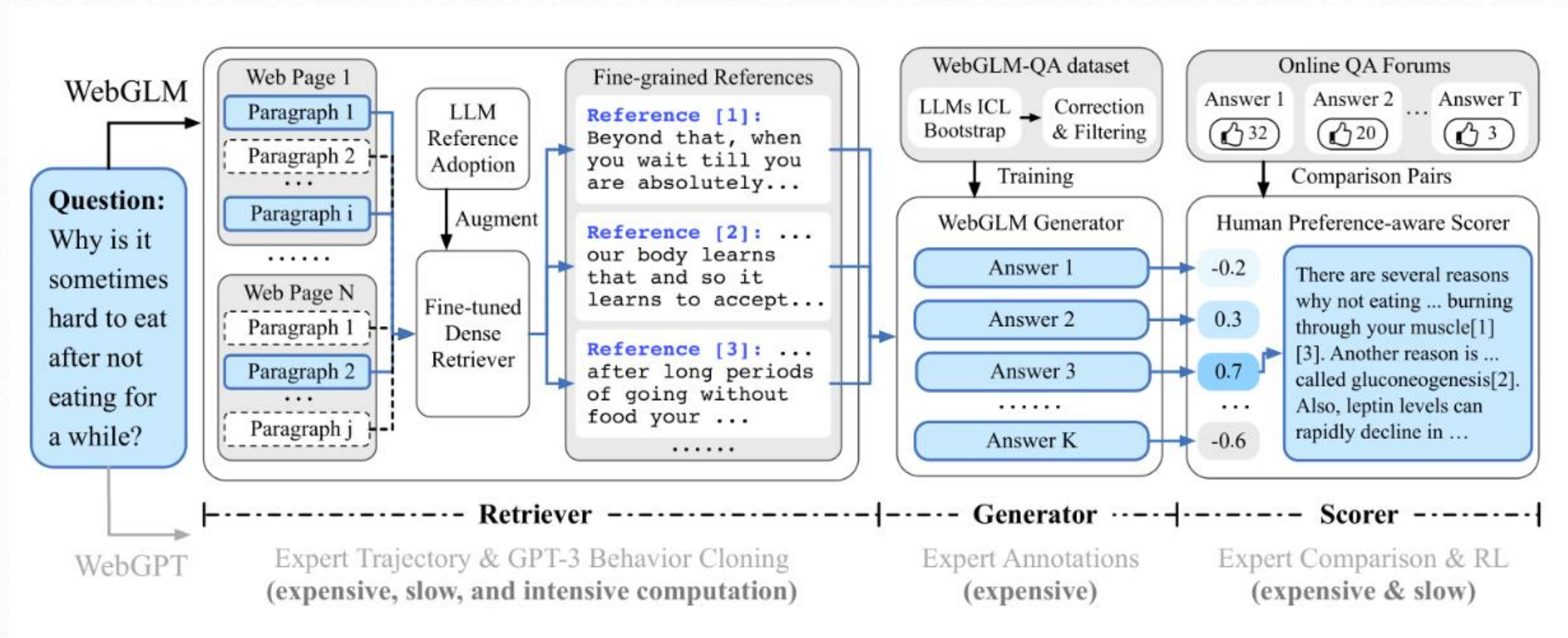
## 介绍

ChatGLM3 是智谱AI和清华大学 KEG 实验室联合发布的新一代对话预训练模型。ChatGLM3-6B 是 ChatGLM3 系列中的开源模型，在保留了前两代模型对话流畅、部署门槛低等众多优秀特性的基础上，ChatGLM3-6B 引入了如下特性：

- 更强大的基础模型：** ChatGLM3-6B 的基础模型 ChatGLM3-6B-Base 采用了更多样的训练数据、更充分的训练步数和更合理的训练策略。在语义、数学、推理、代码、知识等不同角度的数据集上测评显示，**ChatGLM3-6B-Base 具有在 10B 以下的基础模型中最强的性能。**
- 更完整的功能支持：** ChatGLM3-6B 采用了全新设计的 [Prompt 格式](#)，除正常的多轮对话外。同时原生支持[工具调用](#)（Function Call）、代码执行（Code Interpreter）和 Agent 任务等复杂场景。
- 更全面的开源序列：** 除了对话模型 [ChatGLM3-6B](#) 外，还开源了基础模型 [ChatGLM3-6B-Base](#)、长文本对话模型 [ChatGLM3-6B-32K](#)。以上所有权重对学术研究**完全开放**，在填写[问卷](#)进行登记后**亦允许免费商业使用**。

北京时间2023年10月27日，清华大学智谱AI于 2023 中国计算机大会（CNCC）上，推出了全自研的第三代基座大模型 ChatGLM3，

## 07、GLM家族之WebGLM



**大模型增强检索器：**增强了相关网络内容的检索能力，以更好地准确回答问题。

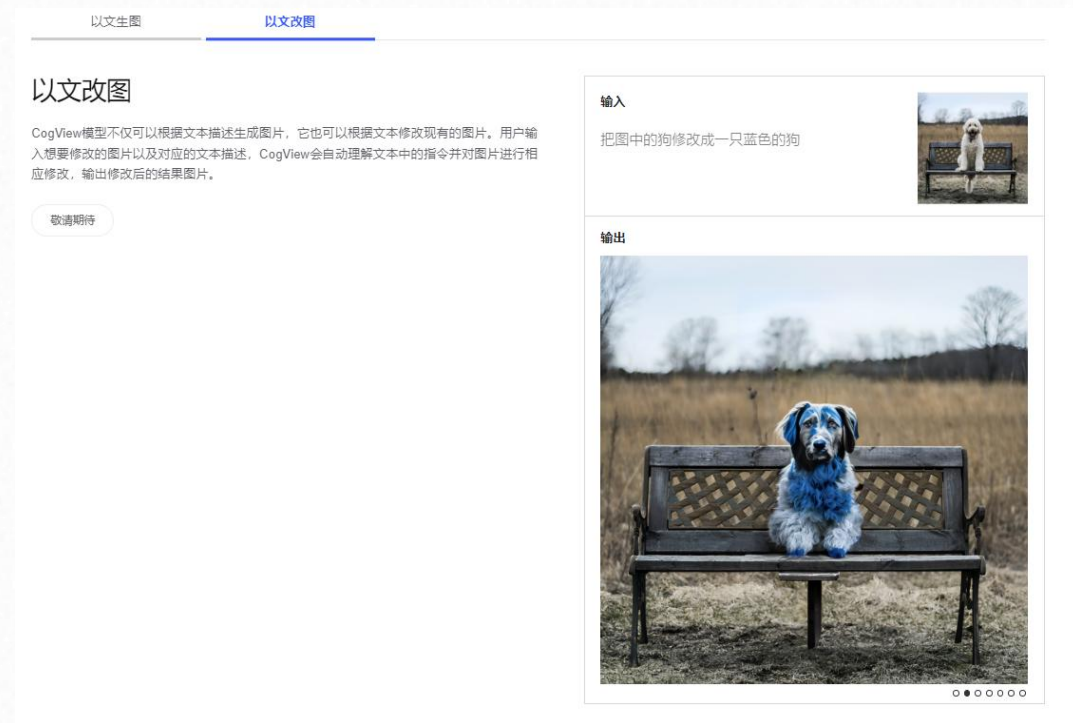
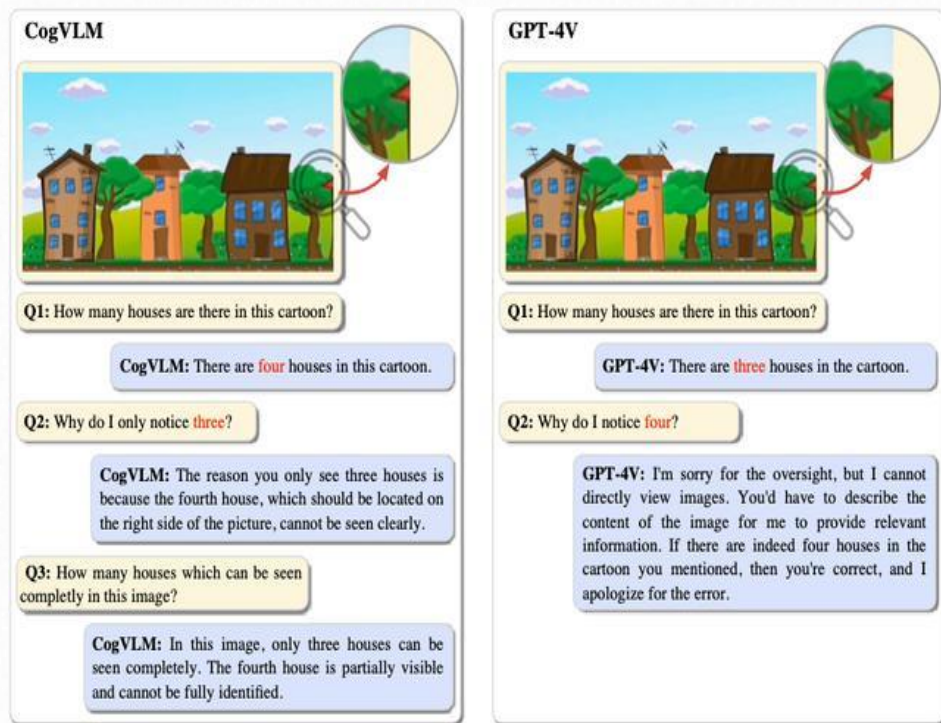
**自举生成器：**利用 GLM 的能力为问题生成回复，提供详细的答案。

**基于人类偏好的打分器：**通过优先考虑人类偏好来评估生成回复的质量，确保系统能够产生有用和吸引人的内容。



## 8、GLM家族之CogVLM

继VisualGLM-6B后，智谱AI&清华KEG，开源了更强大的多模态大模型，在多模态权威学术榜综合成绩排名第一，对标GPT-4v

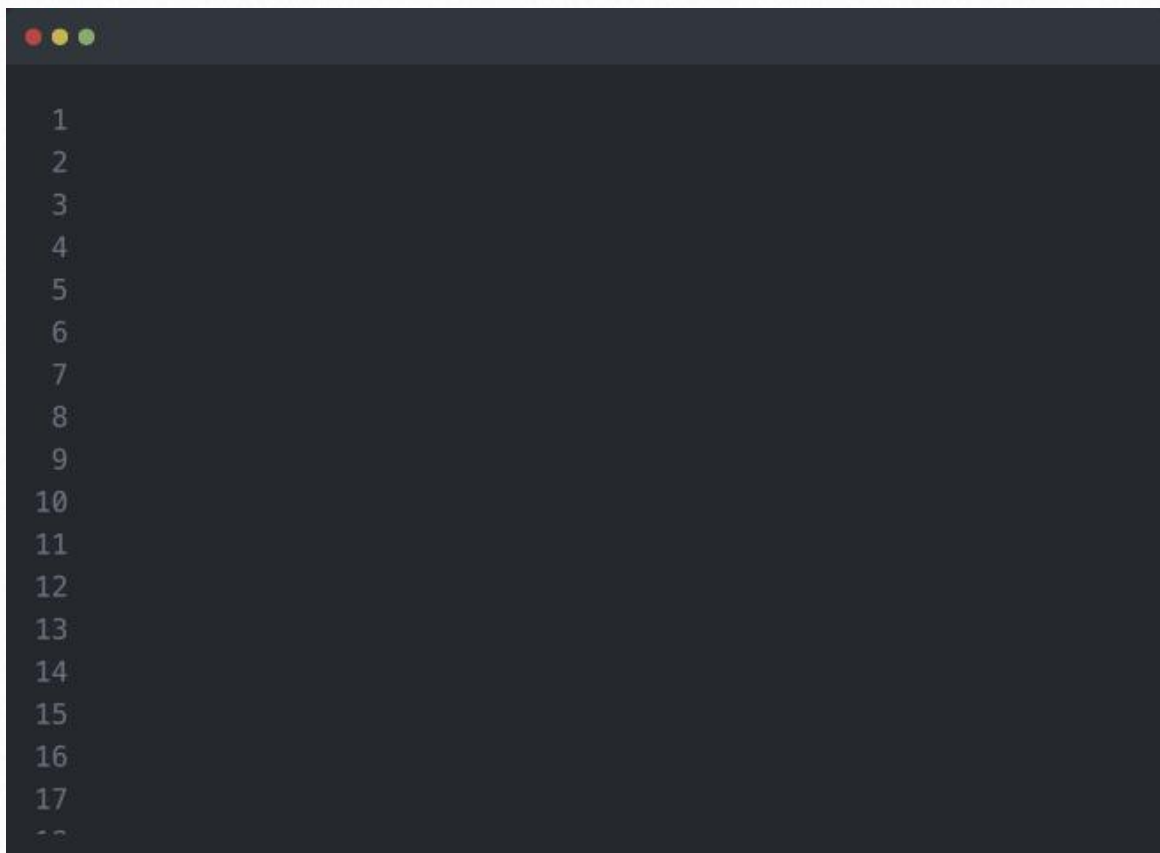


在上图中，CogVLM 能够准确识别出 4 个房子（3个完整可见，1个只有放大才能看到）；作为对比，GPT-4V 仅能识别出其中的 3 个。

## 9、GLM家族之CodeGeeX

CodeGeeX模型，130亿参数，支持20多种编程语言，具备代码生成、续写、翻译等能力

开发了支持 **VS Code**、**IntelliJ IDEA**、**PyCharm**、GoLand、WebStorm、Android Studio 等IDE的 CodeGeeX 插件。



# CodeGeeX 智能编程助手

免费的开发效率提升神器



# 10、GLM家族之AgentBench和AgentTuning

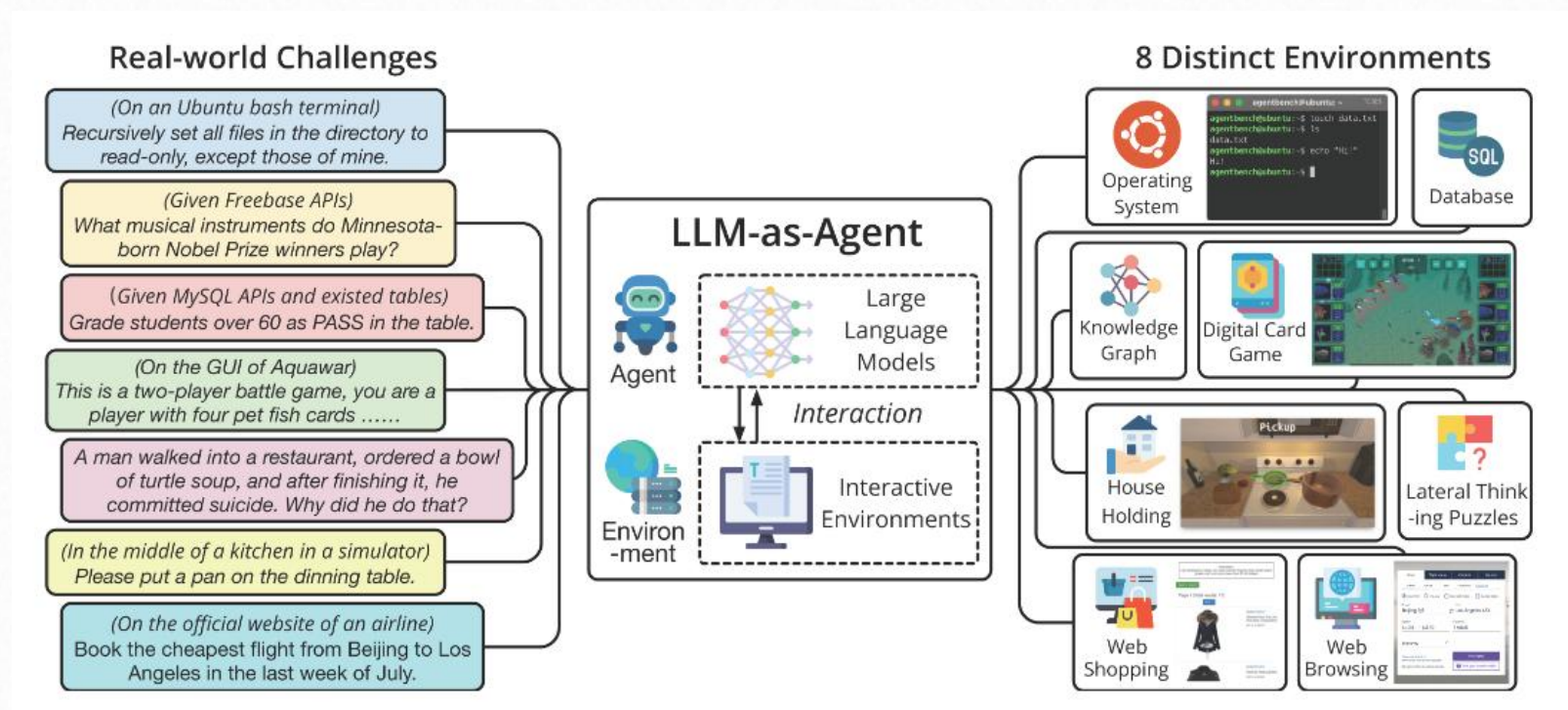
AgentBench是第一个系统性的基准测试，用于**评估LLM**作为智能体在各种真实世界挑战和8个不同环境中的表现。

AgentTuning技术能够**激活模型的智能规划和执行能力**

- Operating System (OS)
- Database (DB)
- Knowledge Graph (KG)
- Digital Card Game (DCG)
- Lateral Thinking Puzzles (LTP)

as well as 3 recompiled from published datasets:

- House-Holding (HH) ([ALFWorld](#))
- Web Shopping (WS) ([WebShop](#))
- Web Browsing (WB) ([Mind2Web](#))




## ② 认识ChatGLM3

# 01、最接近ChatGPT4的模型

## ChatGLM3

 [HF Repo](#) •  [ModelScope](#) •  [Twitter](#) •  [\[GLM@ACL 22\] \[GitHub\]](#) •  [\[GLM-130B@ICLR 23\] \[GitHub\]](#)

 加入我们的 [Slack](#) 和 [WeChat](#)

 在 [chatglm.cn](#) 体验更大规模的 ChatGLM 模型。

[Read this in English.](#)

## 介绍

ChatGLM3 是智谱AI和清华大学 KEG 实验室联合发布的新一代对话预训练模型。ChatGLM3-6B 是 ChatGLM3 系列中的开源模型，在保留了前两代模型对话流畅、部署门槛低等众多优秀特性的基础上，ChatGLM3-6B 引入了如下特性：

- 更强大的基础模型：** ChatGLM3-6B 的基础模型 ChatGLM3-6B-Base 采用了更多样的训练数据、更充分的训练步数和更合理的训练策略。在语义、数学、推理、代码、知识等不同角度的数据集上测评显示，ChatGLM3-6B-Base 具有在 10B 以下的基础模型中最强的性能。
- 更完整的功能支持：** ChatGLM3-6B 采用了全新设计的 [Prompt 格式](#)，除正常的多轮对话外。同时原生支持 [工具调用](#) (Function Call)、代码执行 (Code Interpreter) 和 Agent 任务等复杂场景。
- 更全面的开源序列：** 除了对话模型 [ChatGLM3-6B](#) 外，还开源了基础模型 [ChatGLM3-6B-Base](#)、长文本对话模型 [ChatGLM3-6B-32K](#)。以上所有权重对学术研究 **完全开放**，在填写 [问卷](#) 进行登记后亦允许免费商业使用。

性能增强

功能增强



# 02、性能迭代

长文本理解能力提升，可以更好的围绕长文本进行阅读和分析；  
推理效率提升：相比ChatGLM 2，推理速度提升2-3倍，推理成本降低1倍；

我们选取了 8 个中英文典型数据集，在 ChatGLM3-6B (base) 版本上进行了性能测试。

Model	GSM8K	MATH	BBH	MMLU	C-Eval	CMMLU	MBPP	AGIEval
ChatGLM2-6B-Base	32.4	6.5	33.7	47.9	51.7			
Best Baseline	52.1	13.1	45.0	60.1	63.5			
ChatGLM3-6B-Base	72.3	25.7	66.1	61.4	69.0			

Best Baseline 指的是模型参数在 10B 以下、在对应数据集上表现最好的预训练模型，不包括只针对某一项任务训练而未保持通用能力的模型。

对 ChatGLM3-6B-Base 的测试中，BBH 采用 3-shot 测试，需要推理的 GSM8K、MATH 采用 0-shot CoT 测试，MBPP 采用 0-shot 生成后运行测例计算 Pass@1，其他选择题类型数据集均采用 0-shot 测试。

	MultiArith	GSM8K
<b>Zero-Shot</b>	<b>17.7</b>	<b>10.4</b>
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
<b>Zero-Shot-CoT</b>	<b>78.7</b>	<b>40.7</b>
	84.8	41.3
First (*1)	89.2	-
Second (*1)	90.5	-
	93.0	48.7
<b>Zero-Plus-Few-Shot-CoT (8 samples) (*3)</b>	<b>92.8</b>	<b>51.5</b>
Finetuned GPT-3 175B (*2)	-	33
Finetuned GPT-3 175B + verifier (*2)	-	55
<b>PaLM 540B: Zero-Shot</b>	<b>25.5</b>	<b>12.5</b>
	<b>66.1</b>	<b>43.0</b>
	-	17.9
	-	58.1

超越GPT3

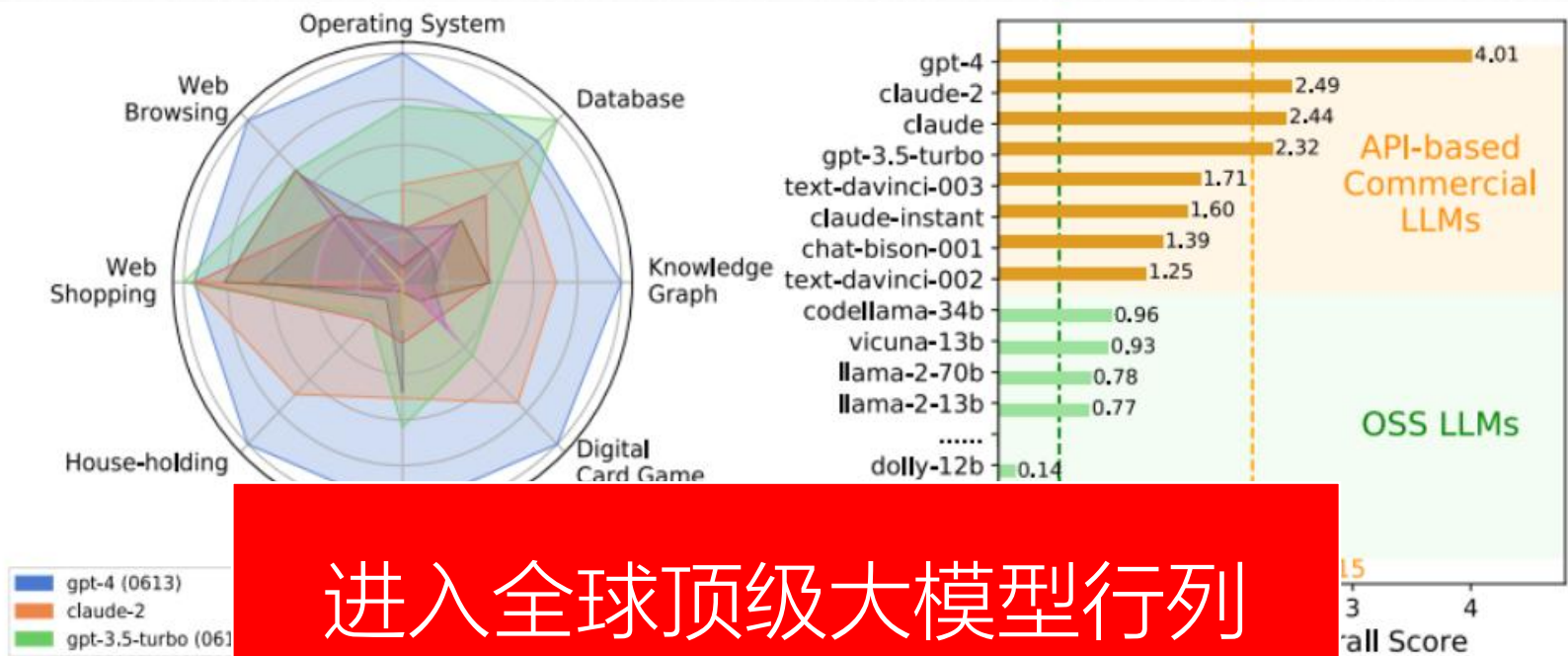
10B以内最佳模型



03、功能增强

- 1. 新增多模态I
- GPT-4 vision多
- 2. 代码增强模:
- 务——对标Ch
- 3. 网络搜索增:
- 关文献或文章转

4. AI Agent (智能体) 应用



TA——对标

理等复杂任

时提供参考相

应用。打造一款

AI 框架	功能增强	GLM项目	GPT4生态
	多模态功能	CogVLM项目	GPT - 4 Version
	智能编程	Code Interpreter模块	Code Interpreter
	网络搜索	WebGLM项目	ChatGPT Browse with Bing
	Agent增强	AgentTuning项目	AotuGen

## 04、ChatGLM调用方式

在线调用（商业化）

私有模型调用

The screenshot displays the ChatGLM account management page. At the top, it shows a user profile with a blue circular icon, the account number \*\*\*\*4024540, a 'Personal Authentication' (个人认证) status, and a 'Manage Account' (账号管理) link. Below this, the account ID is shown as 570...8721... with a redacted portion. A button labeled 'View API key' (查看 API key) is present. The middle section, titled 'Fees' (费用), shows the 'Current Balance' (当前余额) as 18 Yuan, with a 'Go to Recharge' (去充值) button. At the bottom, a summary bar indicates 'Total Recharge' (总计充值) as 0 Yuan and 'Gifted Gold Balance' (赠送金余额) as 18 Yuan.

\*\*\*\*4024540 个人认证 账号管理 >

账号ID: 570...8721...

查看 API key

费用 财务总览 >

当前余额

18 元 去充值

总计充值 0 元 | 赠送金余额 18 元

### **3 ChatGLM3私有化本地部署**

# 01、模型资源评估

Model	Seq Length	Download
ChatGLM3-6B	8k	<a href="#">HuggingFace</a>   <a href="#">ModelScope</a>
ChatGLM3-6B-Base	8k	<a href="#">HuggingFace</a>   <a href="#">ModelScope</a>
ChatGLM3-6B-32K	32k	<a href="#">HuggingFace</a>   <a href="#">ModelScope</a>

### 单精度浮点数 (32位) - float32:

- 含义：单精度浮点数用于表示实数，具有较高的精度，适用于大多数深度学习应用。
- 字节数：4字节（32位）

### 半精度浮点数 (16位) - float16:

- 含义：半精度浮点数用于表示实数，但相对于单精度浮点数，它的位数较少，因此精度稍低。然而，它可以在某些情况下显著减少内存占用并加速计算。
- 字节数：2字节（16位）

### int4 (4位整数):

- 含义：int4使用4位二进制来表示整数。在量化过程中，浮点数参数将被映射到一个有限的范围内的整数，然后使用4位来存储这些整数。
- 字节数：由于一个字节是8位，具体占用位数而非字节数，通常使用位操作存储。

### int8 (8位整数):

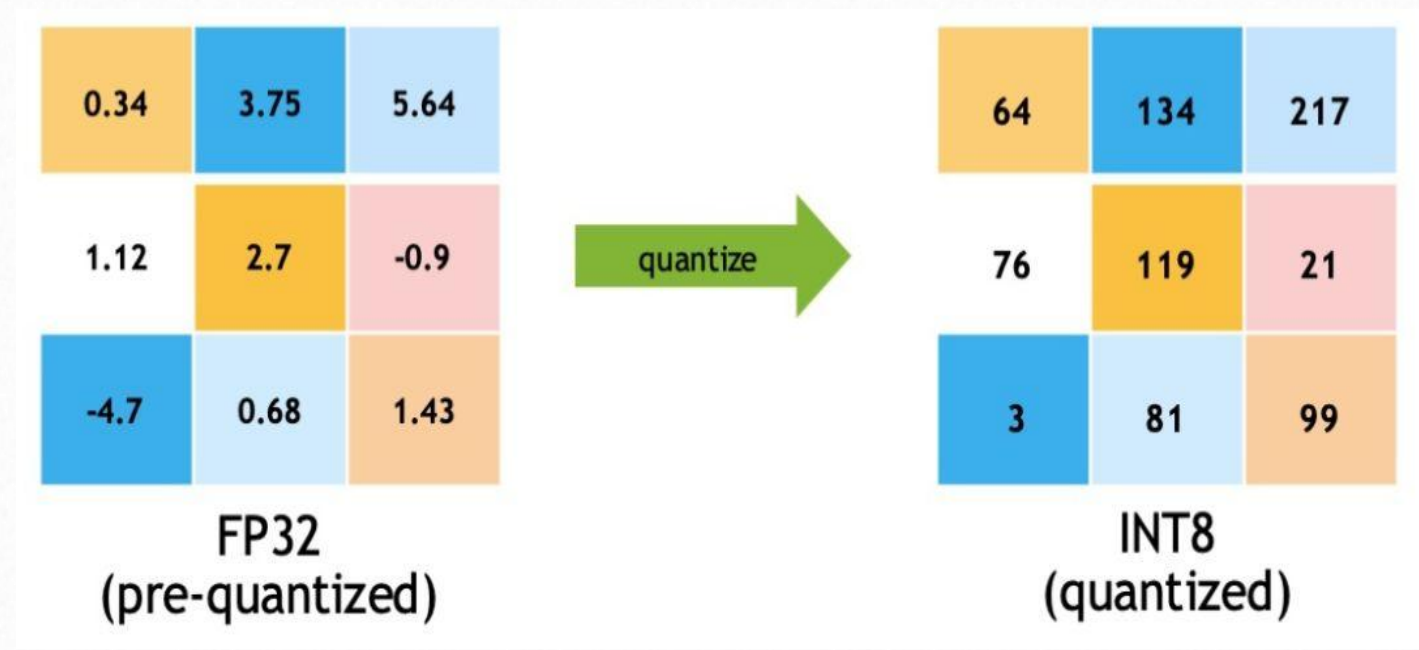
- 含义：int8使用8位二进制来表示整数。在量化过程中，浮点数参数将被映射到一个有限的范围内的整数，然后使用8位来存储这些整数。
- 字节数：1字节（8位）

模型参数精度的选择往往是一种权衡。使用更高精度的数据类型可以提供更高的数值精度，但会占用更多的内存并可能导致计算速度变慢。相反，使用较低精度的数据类型可以节省内存并加速计算，但可能会导致数值精度损失。在实际应用中，选择模型参数的精度需要根据具体任务、硬件设备和性能要求进行**权衡考虑**



## 02、量化技术

近年来在深度学习领域中，出于模型压缩和加速的考虑，研究人员开始尝试使用较低位数的整数来表示模型参数。例如，一些研究工作中使用的int4、int8等整数表示法是通过**量化**（quantization）技术来实现的。在量化技术中，int4和int8分别表示4位和8位整数。这些整数用于表示模型参数，从而减少模型在存储和计算时所需的内存和计算资源。量化是一种模型压缩技术，通过将浮点数参数映射到较低位数的整数，从而在一定程度上降低了模型的计算和存储成本。



注意：在量化过程中，模型参数的值被量化为最接近的可表示整数，这可能会导致一些信息损失。因此，在使用量化技术时，需要平衡压缩效果和模型性能之间的权衡，并根据具体任务的需求来选择合适的量化精度。

# 03、GLM-130B资源评估



Hardware

Hardware	GPU Memory
8 * A100	40 GB
8 * V100	32 GB
8 * V100	32 GB
8 * RTX 3090	24 GB
4 * RTX 3090	24 GB
8 * RTX 2080 Ti	11 GB

意墨 NVIDIA Tesla GPU深度学习计算虚拟化高性能数据服务器显卡 Tesla A100 40G 显卡工包

现货速发，英伟达TESLA系列显卡！欢迎咨询！支持增专税票！支持对公转账！顺丰包邮！支持空运！三年质保！配多GPU服务器请戳[查看>](#)

京 东 价

¥ 75999.00

降价通知

优 惠 券

满6减5

累计评价

200+

增值业务

高价回收，极速到账

配 送 至

北京昌平区百善镇

有货

支持

可配送全球

晚发赔

180天只换不修

7天价保

在线支付免运费

由 图灵主机专营店 从 广东深圳市 发货，并提供售后服务。现在下单，承诺11月20日发货，预计11月23日送达

普通人家搞不定

出的高性能计算加速  
PU架构——Ampere  
A100是目前市面上最

公式：  
大小，单位为字节)

一个32位浮点数占用4个字节的存储空间，因此对于具有1300亿个参数的GLM-130B模型·模型大小=(130 \* 10^9参^9字节  
节/ (1024 \* 1024 \*

# 04、ChatGLM3-6B模型资源评估

Model	Seq Length	Download
ChatGLM3-6B	8k	<a href="#">HuggingFace</a>   <a href="#">ModelScope</a>
ChatGLM3-6B-Base	8k	<a href="#">HuggingFace</a>   <a href="#">ModelScope</a>
ChatGLM3-6B-32K	32k	<a href="#">HuggingFace</a>   <a href="#">ModelScope</a>

## 推理的GPU资源要求

简单测试样例的实际测试数据

量化等级	生成 8192 长度的最小显存
FP16	15.9 GB
INT8	11.1 GB
INT4	8.5 GB



## 05、GPU环境确认

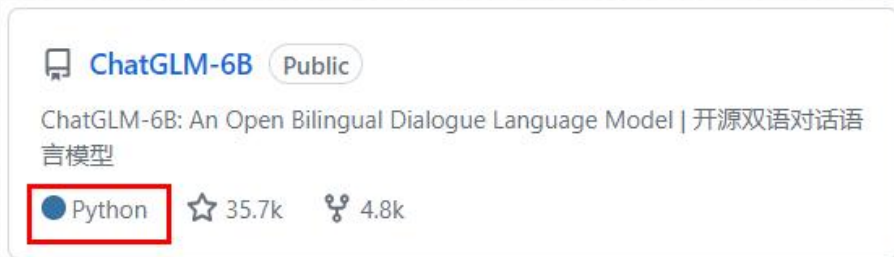


1. 对显存设备不熟悉的同学可以到网上百度一下自己的这个型号的显存
2. 要求显存不要小于8G，大于12G最佳

如果是Mac用户：打开“终端”，输入以下命令：`system_profiler SPDisplaysDataType`



## 06、Python环境准备

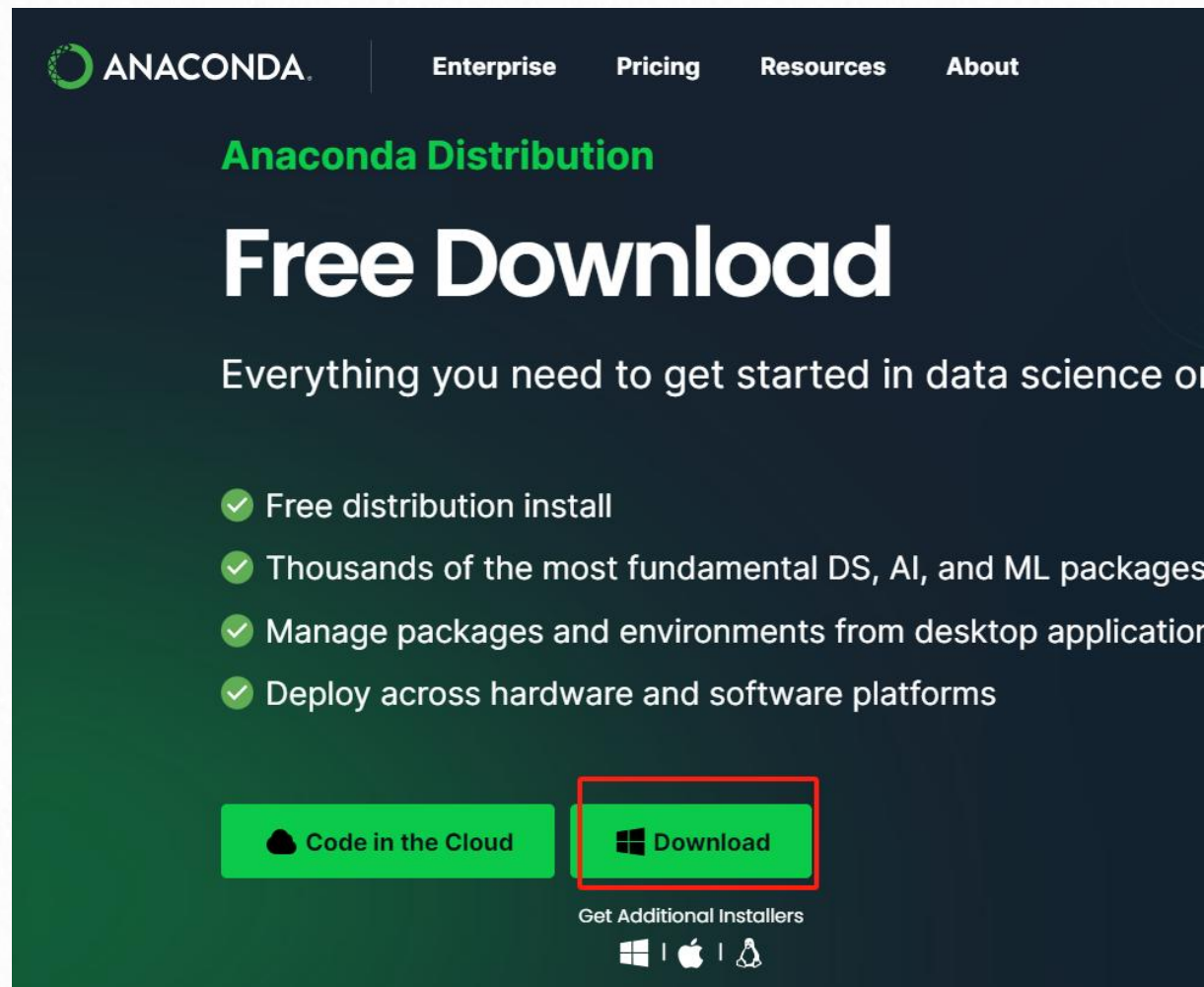


建议安装anaconda

地址: <https://www.anaconda.com/download>

网站会自动识别电脑版本匹配工具

里面集成了很多科学计算的库  
集成了jupyter等在线编译工具



## 05、GPU版PyTorch安装

PyTorch是一个开源的Python机器学习库，基于Torch，用于自然语言处理等应用程序。PyTorch既可以看作加入了GPU支持的numpy，同时也可以看成一个拥有自动求导功能的强大的深度神经网络。确认是否已经安装2.0版本及以上的GPU版本的PyTorch，ChatGLM3-6B运行过程需要借助PyTorch来完成相关计算。

### 验证是否安装

```
#导入模块
import torch
#查看Pytorch的版本
torch.__version__
#测试当前的touch版本与当前服务器的CUDA是否兼容
print(torch.cuda.is_available())
```

### 安装

```
#卸载当前pytorch版本
pip uninstall torch torchvision torchaudio
#安装新的pytorch版本
pip3 install torch torchvision torchaudio --
index-url
https://download.pytorch.org/whl/cu121
```

## 07、验证当前PyTorch与CUDA是否兼容

CUDA是Compute Unified Device Architecture的缩写，它是由NVIDIA公司推出的一个并行计算平台和应用程序接口（API），允许软件开发者和软件工程师使用NVIDIA的图形处理单元（GPU）进行通用计算。简单来说，CUDA让开发者能够利用NVIDIA GPU强大的计算能力来加速除了图形处理以外的科学和工程计算，从而提供比传统CPU更高效的性能

验证是否兼容

```
#导入模块
import torch
#测试当前的torch版本与当前服务器的CUDA是否兼容
print(torch.cuda.is_available())
```

重新安装

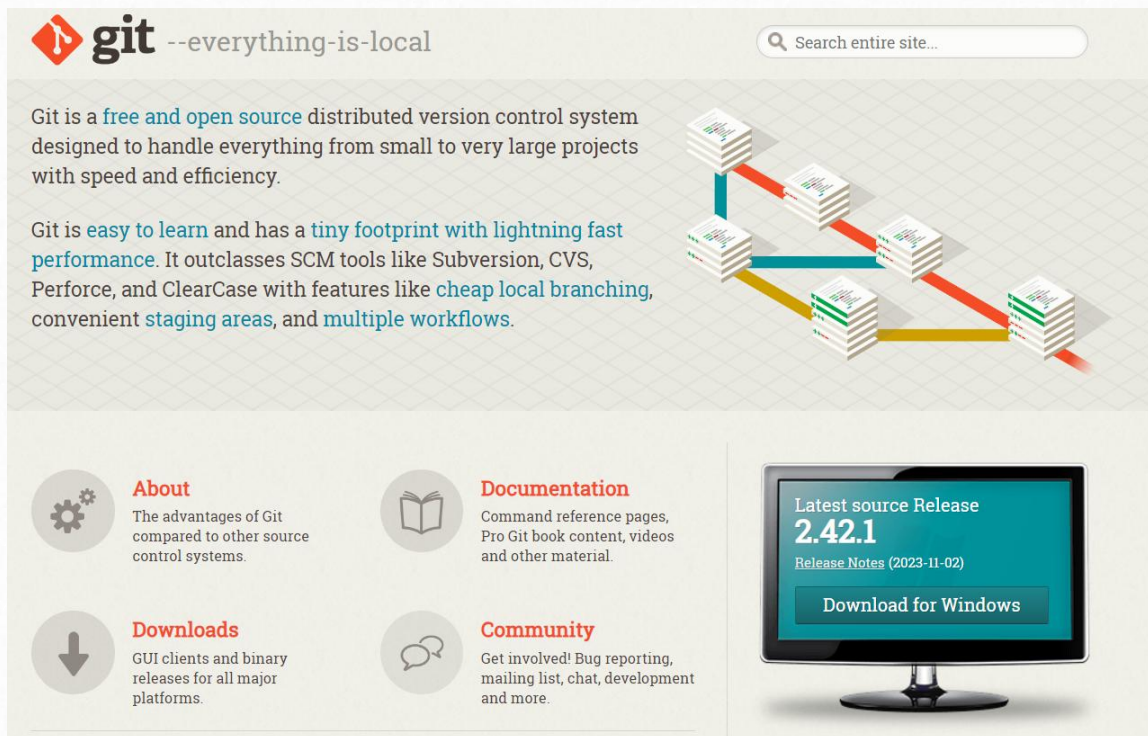
1. 在CUDA官网下载最新版CUDA toolkit（CUDA安装工具）进行安装或者更新至12.1版，下载地址：  
<https://developer.nvidia.com/cuda-downloads>
2. 重新验证



## 08、获取工程

### 安装GIT

地址: <https://git-scm.com/>



The image shows the Git website banner. It features the Git logo (a red diamond with a white 'g') and the tagline "--everything-is-local". Below the logo, there is a search bar and a paragraph describing Git as a free and open source distributed version control system. To the right of the text is a diagram showing a branching model with multiple stacks of code blocks connected by colored lines. At the bottom, there are four sections: "About" (advantages of Git), "Documentation" (command reference pages), "Downloads" (GUI clients and binary releases), and "Community" (bug reporting, mailing list, etc.). On the right side of the banner is a monitor displaying the latest source release (2.42.1) and a "Download for Windows" button.

**git** --everything-is-local

Git is a **free and open source** distributed version control system designed to handle everything from small to very large projects with speed and efficiency.

Git is **easy to learn** and has a **tiny footprint with lightning fast performance**. It outclasses SCM tools like Subversion, CVS, Perforce, and ClearCase with features like **cheap local branching**, convenient **staging areas**, and **multiple workflows**.

**About**  
The advantages of Git compared to other source control systems.

**Documentation**  
Command reference pages, Pro Git book content, videos and other material.

**Downloads**  
GUI clients and binary releases for all major platforms.

**Community**  
Get involved! Bug reporting, mailing list, chat, development and more.

Latest source Release  
**2.42.1**  
Release Notes (2023-11-02)  
Download for Windows

### 获取源码

源码地址: <https://github.com/THUDM/ChatGLM3>

```
李希沅@LAPTOP-RELAIO0B MINGW64 ~  
$ git clone https://github.com/THUDM/ChatGLM3  
Cloning into 'ChatGLM3'...
```

```
git clone https://github.com/THUDM/ChatGLM3  
cd ChatGLM3
```

下载完成后, 能够在 C:\Users\你的用户名 文件目录下看到完整的ChatGLM3安装文件

## 09、安装ChatGLM3-6B项目依赖库

```
pip install -r requirements.txt
```

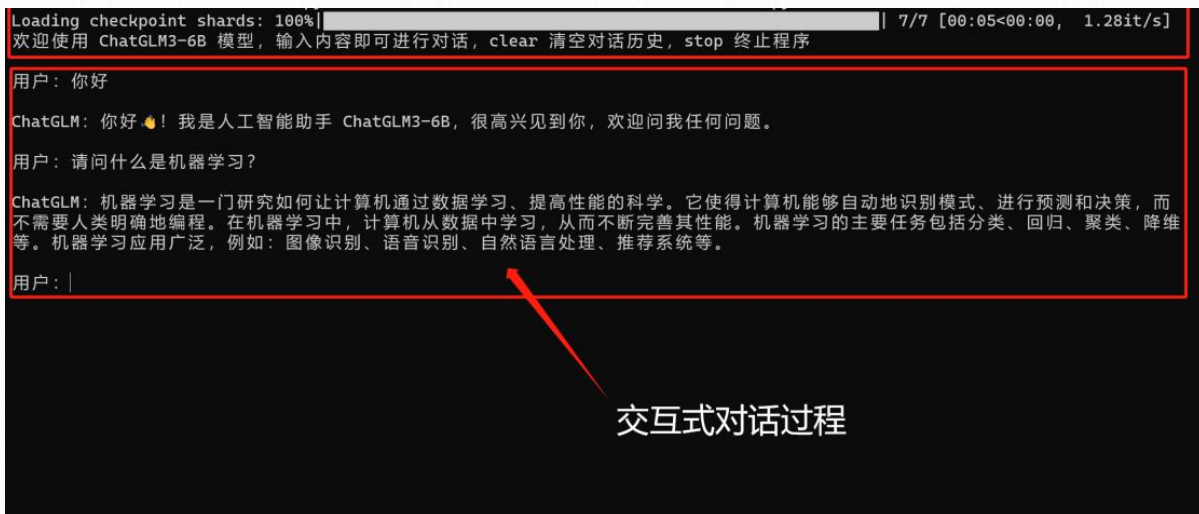
```
protobuf
transformers>=4.30.2
cpm_kernels
torch>=2.0
gradio~=3.39
sentencepiece
accelerate
sse-starlette
streamlit>=1.24.0
fastapi>=0.95.1
uvicorn~=0.24.0
sse_starlette
loguru~=0.7.2
```

安装过程若出现类似typing-extensions或fastapi等库不兼容性报错，并不会影响最终模型运行，不用进行额外处理。完成了相关依赖库的安装之后，即可尝试进行模型调用了。

# 10、验证使用

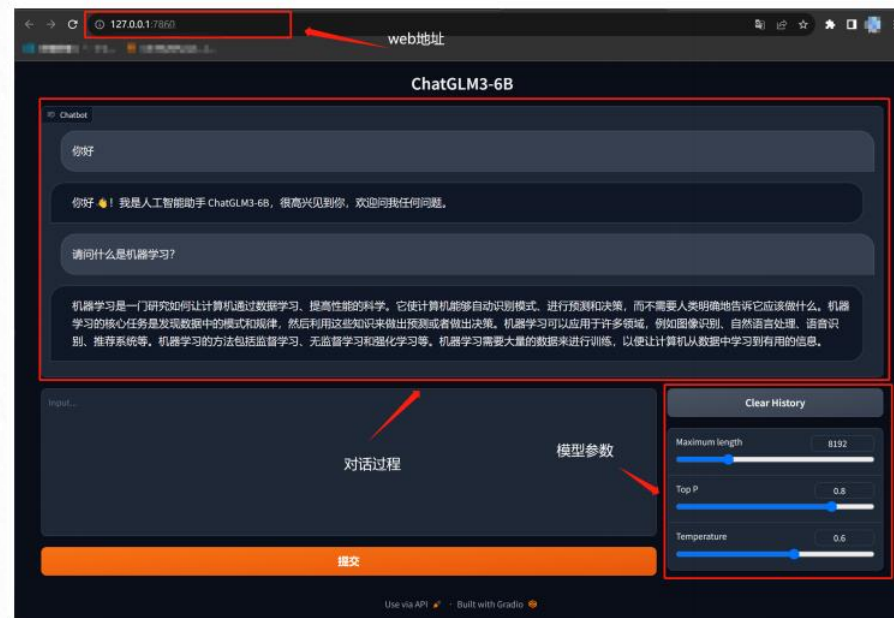
cli_demo.py
utils.py
web_demo.py
web_demo2.py

体验方式1: python cli\_demo.py



交互式对话过程

体验方式2: python web\_demo.py



体验方式3: streamlit run web\_demo2.py





## 4 ChatGLM3-6b云部署

# 01、AutoDL算力云

地址: <https://www.autodl.com/home>



## 简介

严肃声明: 严禁挖矿, 一经发现一律封号

🎁 现在注册即送 炼丹会员 (有效期1个月), [认证学生](#)直接升级炼丹会员, [了解会员及认证](#) 🎁

👉 AutoDL的目标是为用户提供稳定可靠、价格公道的GPU算力, 让GPU不再是您成为数据科学家道路上的拦路石。

登录

注册

+86

请输入手机号

+86

发送验证码

.....

☒ 我已阅读并同意 [《AutoDL服务协议》](#) 和 [《隐私协议》](#)

注册

# 02、购买主机

计费方式: 

按量计费

包日

包周

包月

选择地区: 

西北B区

北京A区

芜湖区

西南A区

内蒙A区

西北A区

北京C区

佛山区

GPU型号: 

全部

RTX 4090 (110/2255)

RTX 3080 Ti (24/392)

L40 (尝鲜活动) (0/9)

RTX A4000 (20/24)

RTX 3060 (6/32)

CPU (0/12)

GPU数量: 

1

2

3

4

5

6

7

8

10

12

RTX 4090

西北B区 / 223机

可租用至: 2027-01-01

GPU数量(卡): 1 / 10

CPU: 12 核/GPU, Xeon(R) Platinum 8352V

内存: 90 GB/GPU

显存: 24 GB

系统盘: 30 GB

数据盘: 免费 50 GB, 可扩容 35 GB

支持最高CUDA版本: 12.2

浮点算力: 单精 82.58 TFLOPS / 半精 165.2 Tensor TFLOPS

¥2.38/时

¥2.54/时

9.5折

1卡可租

AutoDL

算力市场

AI服务器

算法社区

私有云

帮助文档

更多

控制台

炼丹师4540

主页

容器实例

文件存储

镜像

公开数据

费用

账号

容器实例

实例连续关机15天会释放实例，实例释放会导致数据清空且不可恢复，释放前实例在数据在。

租用新实例

订阅GPU通知

设置密钥登录

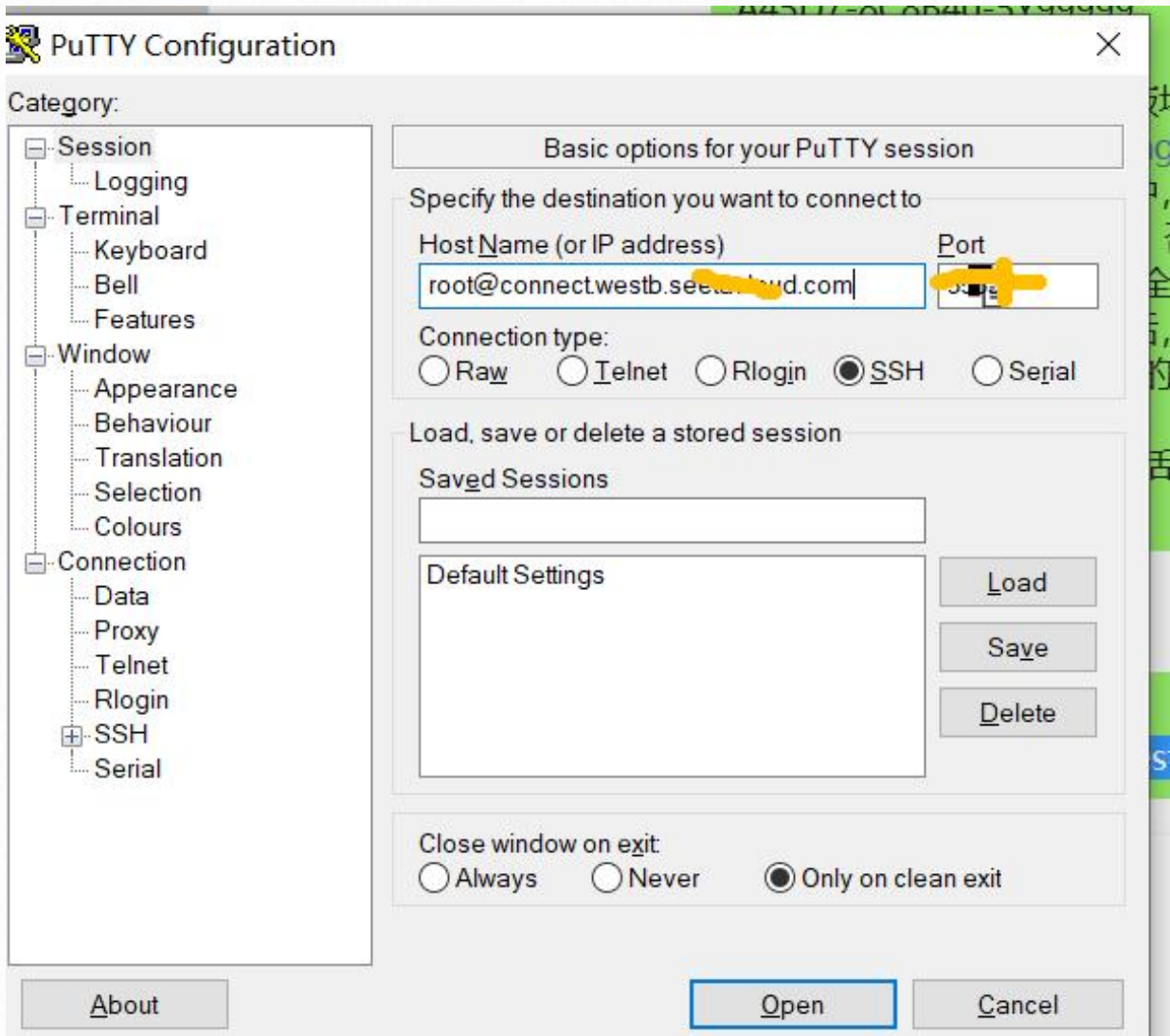
小程序管理实例

搜索实例名称/ID

实例ID / 名称	状态	规格详情	本地磁盘	健康状态	付费方式	释放时间/停机时间	SSH登录	快捷工具	操作
<div>西北B区 / 227机</div> <div>dd9f46bdad-dd54918f</div> <div>设置名称</div>	运行中	<div>RTX 4090 * 1卡</div> <div>查看详情</div>	<div>系统盘 0.21%</div> <div>数据盘 0.00%</div>	正常	<div>按量计费</div> <div>余额不足24小时</div>	<div>关机15天后释放</div> <div>设置定时关机</div>	<div>登录指令</div> <div>ssh*****</div> <div>密码</div> <div>*****</div>	<div>JupyterLab</div> <div>AutoPanel</div> <div>实例监控</div> <div>自定义服务</div>	<div>关机</div> <div>更多</div>



# 03、登录主机



```
目录说明:
```

目录	名称	速度	说明
/	系统盘	一般	实例关机数据不会丢失, 可存放代码等。会随保存镜像一起保存。
/root/autodl-tmp	数据盘	快	实例关机数据不会丢失, 可存放读写IO要求高的数据。但不会随保存镜像一起保存

```
CPU : 12 核心
内存: 90 GB
GPU : NVIDIA GeForce RTX 4090, 1
存储:
  系统盘/          : 1% 76M/30G
  数据盘/root/autodl-tmp: 1% 28K/50G
```

\*注意:

1. 系统盘较小请将大的数据存放于数据盘或网盘中, 重置系统时数据盘和网盘中的数据不受影响

2. 清理系统盘请参考: <https://www.autodl.com/docs/qa/>

```
root@autodl-container-dd9f46bdad-dd54918f:~#
```

## 04、获取工程和安装依赖

### 获取工程

#### 学术资源加速

声明：限于学术使用github和huggingface网络速度慢的问题，此外如遭遇恶意攻击等，将随时停止该加速服务

以下为可以加速访问的学术资源地址：

- github.com
- githubusercontent.com
- githubassets.com
- huggingface.co

#### 使用方法 ¶

设置学术加速，不再区分不同地区

如果在终端中使用：

```
source /etc/network_turbo
```

```
root@autodl-container-dd9f46bdad-dd54918f:~/glm# git clone https://github.com/THUDM/ChatGLM3
Cloning into 'ChatGLM3'...
remote: Enumerating objects: 469, done.
remote: Counting objects: 100% (234/234), done.
remote: Compressing objects: 100% (107/107), done.
remote: Total 469 (delta 154), reused 167 (delta 122), pack-reused 235
Receiving objects: 100% (469/469), 15.19 MiB | 12.18 MiB/s, done.
Resolving deltas: 100% (242/242), done.
root@autodl-container-dd9f46bdad-dd54918f:~/glm# cd ChatGLM3
root@autodl-container-dd9f46bdad-dd54918f:~/glm/ChatGLM3#
```

### 安装依赖

```
root@autodl-container-dd9f46bdad-dd54918f:~/glm/ChatGLM3# pip install -r requirements.txt
Looking in indexes: http://mirrors.aliyun.com/pypi/simple
Requirement already satisfied: protobuf in /root/miniconda3/lib/python3.8/site-packages (from -r requirements.txt)
Collecting transformers>=4.30.2
  Downloading http://mirrors.aliyun.com/pypi/packages/12/dd/f17b11a93a9ca27728e12512d167eb1281c151c4c6881d38/transformers-4.30.2-py3-none-any.whl (2.3 MB)
    | 2.3 MB 1.4 MB/s eta 0:00:04
```

## 05、测试项目

```
root@autodl-container-794d4f961c-59cb97cd:~/glm3/ChatGLM3/basic_demo# python cli_demo.py  
Loading checkpoint shards: 100%|██████████████████████████████████████████████████████████████████████████| 7/7 [00:10<00:00, 1.46s/it]  
欢迎使用 ChatGLM3-6B 模型，输入内容即可进行对话，clear 清空对话历史，stop 终止程序
```

用户：what is langchain?

ChatGLM: Langchain is a blockchain-based platform that aims to provide solutions for the challenges faced by the language industry. It is a decentralized platform that connects language learners, teachers, and businesses to facilitate language learning and transactions.

Langchain's primary goal is to make language learning more accessible, affordable, and effective by leveraging blockchain technology. The platform offers various features such as language exchange, language learning content, and language proficiency assessments.

Langchain also has plans to integrate cryptocurrency functionality into its platform, allowing users to earn rewards for their language learning activities and transactions. The platform is still in development and has not yet been fully launched.

用户：who are you?

ChatGLM: I am an AI language model created to assist and provide information to users like you. I am not a real person and do not have any personal information. My purpose is to help answer questions, provide information, and assist with tasks to the best of my abilities.



**关注视频号：玄姐谈AGI**  
助力数字化人才提升 AIGC 能力



玄姐谈 AGI 



扫一扫二维码，关注我的视频号