

华为认证人工智能系列教程

HCIA-AI Solution

# 人工智能业务流程概述

## 实验指导手册

版本：1.0



华为技术有限公司

**版权所有 © 华为技术有限公司 2024。 保留一切权利。**

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

## **商标声明**



和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

## **注意**

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

**华为技术有限公司**

地址： 深圳市龙岗区坂田华为总部办公楼 邮编： 518129

网址： <https://e.huawei.com>

---

## 华为认证体系介绍

华为认证是华为公司基于“平台+生态”战略，围绕“云-管-端”协同的新ICT技术架构，打造的覆盖ICT（Information and Communications Technology，信息通信技术）全技术领域的认证体系，包含ICT基础设施认证、基础软硬件认证、云平台及云服务认证三类认证。

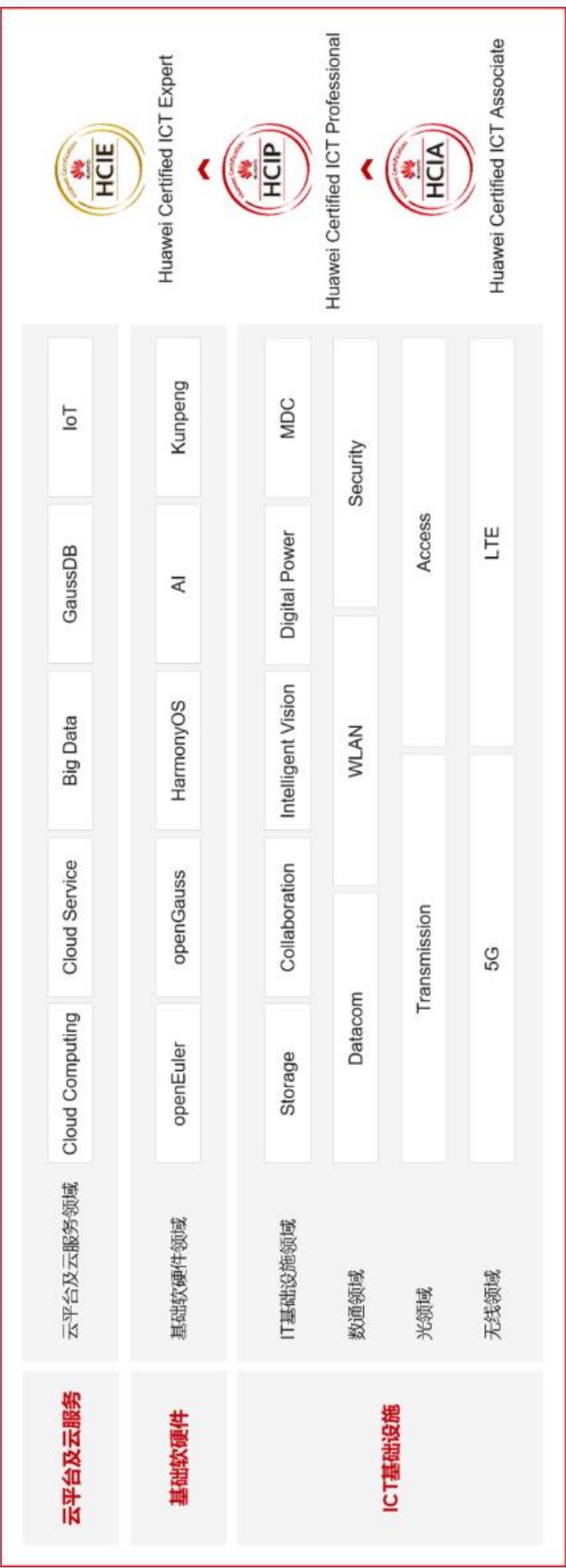
根据ICT从业者的学习和进阶需求，华为认证分为工程师级别、高级工程师级别和专家级别三个认证等级。

华为认证覆盖ICT全领域，符合ICT融合的技术趋势，致力于提供领先的人才培养体系和认证标准，培养数字化时代新型ICT人才，构建良性ICT人才生态。

HCIA-AI Solution（Huawei Certified ICT Associate-AI Solution，华为认证人工智能解决方案）主要面向华为公司办事处、代表处一线工程师，以及其他希望学习华为AI解决方案的技术人士。HCIA-AI Solution认证在内容上涵盖人工智能和算力概念、大模型业务流程、人工智能发展趋势、面向AI大模型的华为智算产品软件和硬件、大模型部署、业界大模型及应用，具备大模型简单部署、使用提示工程的能力。

华为认证协助您打开行业之窗，开启改变之门，屹立在人工智能世界的潮头浪尖！

# 华为职业认证



# 前言

---

## 简介

本书为 HCIA-AI Solution 认证培训教程，适用于准备参加 HCIA-AI Solution 考试的学员或者希望了解如何使用大模型的读者。

## 内容描述

本实验指导书共包含 3 个实验，从如何使用大模型开始，逐步介绍如何更有效的向大模型提问。

- 实验一为大模型使用实验，通过开源模型平台，体验大模型问答。
- 实验二介绍了使用典型的提示方式，使用它们向大模型提问，使大模型回答更加精准。
- 实验三介绍了两种提示框架，使用这两种框架向 GLM<sub>3</sub> 提问，观察回答结果。

## 读者知识背景

本课程为华为认证基础课程，为了更好地掌握本书内容，阅读本书的读者应首先具备以下基本条件：

- 了解大模型的基本概念。
- 了解基本的深度学习基础知识。

## 实验环境说明

实验平台：可以访问互联网的 PC。

---

# 目录

---

<b>前 言</b>	<b>4</b>
简介	4
内容描述	4
读者知识背景	4
实验环境说明	4
<b>1 初探大模型使用</b>	<b>6</b>
1.1 实验介绍	6
1.1.1 关于本实验	6
1.1.2 实验目的	6
1.1.3 实验步骤	6
1.2 思考题	7
<b>2 提示工程</b>	<b>8</b>
2.1 实验介绍	8
2.2 实验目的	8
2.3 实验步骤	8
<b>3 提示框架</b>	<b>11</b>
3.1 实验介绍	11
3.2 实验目的	11
3.3 实验步骤	11
3.4 思考题	14

---

# 1 初探大模型使用

## 1.1 实验介绍

### 1.1.1 关于本实验

本实验使用 ChatGLM 官网已部署大模型 GLM<sub>3</sub>，通过基础的提示词对其提问，使模型输出结果，帮助您了解如何使用大模型。

### 1.1.2 实验目的

- 使用大模型。

### 1.1.3 实验步骤

步骤 1 进入智谱清言官网 <https://chatglm.cn/main/detail>，在页面上端选择 GLM<sub>3</sub>。

步骤 2 在对话框输入问题。

问题 1:

中国最大的淡水湖是什么湖？

答案:

中国最大的淡水湖是鄱阳湖。鄱阳湖位于江西省北部，长江中下游的南岸，是长江最重要的调蓄湖泊之一。鄱阳湖的面积随季节变化较大，枯水期约为 900 至 1000 平方公里，而丰水期则可扩大至 4000 平方公里左右。它不仅是长江流域中重要的水源地，也是众多水生動植物的栖息地，具有极高的生态价值和生物多样性，被誉为“长江的肾脏”。

鄱阳湖的水量主要来自长江以及江西省内的赣、抚、信、饶、修五江的汇入。湖泊通过鄱阳湖出口——鄱阳湖水道与长江相连，实现水的出入。由于其特殊的地理位置和水文特征，鄱阳湖对于长江流域的水资源调节、洪水防控、生态环境维护等功能发挥着至关重要的作用。

此外，鄱阳湖还是一个重要的旅游景点，吸引了众多国内外游客前来观光。湖区的自然风光和文化遗产，如庐山、婺源等，也极大地丰富了旅游资源，对于地方经济的发展起到了推动作用。

需要指出的是，洞庭湖也曾是中国最大的淡水湖，但由于近年来当地政府实施退田还湖政策以及自然水文循环的变化，鄱阳湖的面积在某些时期已经超过了洞庭湖，成为了中国最大的淡水湖。

问题 2:

请描述一下春天樱花盛开的情景。

回答 2:

春天樱花盛开的情景是非常美丽和令人陶醉的。以下是对春天樱花盛开情景的描述：

春天的阳光温暖而明媚，春风轻拂，带来了花香和温暖的气息。走在樱花树下，您会看到密密麻麻的樱花树枝上挂满了粉色的花朵。这些花朵有的完全开放，像一朵朵粉色的小伞，有的则是半开的状态，展现出羞涩的花蕾。整棵树仿佛成了一朵巨大的樱花，给人一种如梦如幻的感觉。

当您走近樱花树，可以看到花朵的细节。樱花的花瓣柔软而薄，色彩鲜艳。花瓣边缘略带一点锯齿状，给人一种精致的感觉。花朵的中心是黄色的花蕊，像是一点点的明珠，为樱花增添了生机和活力。

樱花盛开的时候，整个樱花树林都弥漫着淡淡的花香。走在树林中，您会被花香包围，感受到一种宁静和舒适的氛围。春风吹过，花瓣随风飘落，形成一片粉色的雨。这种景象被称为“樱花雨”，给人一种浪漫和梦幻的感觉。

春天樱花盛开的时候，也是人们赏花的高峰期。许多人会和家人、朋友一起来到樱花树下，野餐、拍照，享受春天的美好。樱花盛开的情景给人们带来了快乐和幸福的感受，也让人们感受到生命的美好和希望。

总之，春天樱花盛开的情景是美丽、浪漫和令人陶醉的。它给人带来了快乐和幸福的感受，也让人们感受到生命的美好和希望。

步骤 3 请自己输入一个问题，观察模型的回答。

## 1.2 思考题

大模型的提示词格式可以是什么样的？

请观察模型的回答，是否符合您的预期。通过调整提示词，大模型的回答是否变化，是否更符合您的预期。



# 2 提示工程

## 2.1 实验介绍

本实验使用零样本提示、少样本提示、思维链提示分别向大模型提问，帮助您了解如何更有效的向大模型提问。

## 2.2 实验目的

- 使用不同的提示方式，观察大模型回答结果，了解不同提示方式，掌握提示技巧。

## 2.3 实验步骤

步骤 1 使用零样本提示向大模型提问。零样本的本质是直接向大模型提问，不给大模型任何提示，这种方式与实验 1 是相同的。

提示 1:

石头是一种特别坚硬的物体，它可以不受绝大多数外界因素干扰，保持自己自身的特性，这种品质值得我们学习。

评价:

回答 1:

这个评价富有启发性，它不仅仅是在描述石头的物理特性，更是在比喻人的品质和态度。石头之所以能够“不受绝大多数外界因素干扰，保持自己自身的特性”，是因为它的物理结构稳定，化学性质坚固。这种比喻告诉我们，人应当像石头一样坚韧不拔，保持自己的本质和原则，不被外界的干扰和诱惑所改变。这是一种很高的品质，也是我们在生活和工作中应当努力追求的目标。

提示 2:

请用一句话表达我对祖国的热爱。

回答 2:

我深深地热爱着我的祖国，为她的繁荣昌盛和人民的幸福安康感到自豪。

步骤 2 使用少样本提示向大模型提问。少样本提示是在提问之前，给出一些问题和回答的组合对，让大模型了解您期望的回答方式是如何的，以便在回答我们真正的问题时，给出更合理的回答。少样本的样本数量可以任意选择。

提示 1（提示中包含 1 个样例）：

苹果中的某些物质可能引起食用者过敏，引发皮疹等症状。

评价：负面

榴莲香甜可口，营养价值丰富。

评价：

回答 1:

正面

提示 2（提示中包含 3 个样例）：

我认为读书有益，因为读书是获取知识的最佳方式。

对读书的态度：赞同

读书会消耗大量的时间，不如通过视频学习知识速度快。

对读书的态度：反对

读书只能学到死板的知识，去实践才是最好的学习方式。

对读书的态度：反对

读书可以让读者在没有机会行万里路的情况下，开拓他们的视野，提高他们的认知水平。

对读书的态度：

回答 2:

赞同

读书是一种非常重要的学习方式，它能够帮助读者在不同的领域获取深入的知识 and 理解。尽管视频学习可能更快捷，但读书能够提供更为深入和系统的知识。此外，读书不仅仅是为了获取信息，它还能够培养读者的思维能力、批判性思维和想象力。

至于实践，它确实是学习的重要部分，但读书可以提供必要的理论基础和背景知识，使实践更加有意义和有效。因此，读书和实践相辅相成，都是学习和成长的重要途径。

步骤 3 使用思维链提示向大模型提问。思维链（CoT）提示通过引入中间推理步骤实现了复杂的推理能力，它逐步展开问题并在每个回答中保留之前问题的背景信息，链式思考可以使对话变得内容丰富和连贯。

提示 1 (Few-shot-CoT) :

李晓明有 5 个网球，他又买了两盒网球，每盒有 3 个网球。他现在有多少网球？

答：李晓明一开始有 5 个网球，新买的 2 盒网球每盒中有 3 个网球，所以新买的网球共  $2 \times 3 = 6$  个网球。因此他现在共有  $5 + 6 = 11$  个网球。答案是 11。

问：食堂有 23 个苹果，如果厨师吃掉了 20 个后，又买了 6 个。请问现在食堂有多少个苹果？

回答 1:

食堂原本有 23 个苹果，厨师吃掉了 20 个，那么剩下的苹果就是：

$23 - 20 = 3$  个苹果。

然后厨师又买了 6 个苹果，所以现在食堂有的苹果总数是：

$3$  (剩下的)  $+ 6$  (新买的)  $= 9$  个苹果。

答案是 9 个苹果。

提示 2 (Zero-shot-CoT) :

一个杂技演员可以玩杂耍 16 个球，一半是高尔夫球，其中一半的高尔夫球是蓝色的，那么蓝色的高尔夫球共有多少个？

让我们一步一步思考。

回答 2:

杂技演员可以玩杂耍 16 个球，一半是高尔夫球，那么高尔夫球有  $16 / 2 = 8$  个。其中一半的高尔夫球是蓝色的，那么蓝色的球有  $8 / 2 = 4$  个。

所以，蓝色的高尔夫球共有 4 个。

# 3 提示框架

## 3.1 实验介绍

如何更好的使用大模型，更好的向大模型提问，以便得到期望的结果，已经是使用者非常关心的问题。结合多种实验结果总结，使用者总结了一些提示框架。本实验带大家使用几种提示框架，让您可以更好的了解清晰的提示指令，会让我们得到更期望的结果。

## 3.2 实验目的

使用提示框架向大模型提问，进一步了解提示词准确、目的清晰的重要性。

## 3.3 实验步骤

### 步骤 1 APE 框架

1. Action 行动：定义要完成的工作或活动。
2. Purpose 目的：讨论意图或目标。
3. Expectation 期望：说明期望的结果。

#### 提示 1:

行动：您能为我们的环保运动鞋新产品制定一个内容营销策略吗？

目的：我们的目标是在我们的目标受众（对可持续发展充满热情的健身爱好者）中产生轰动效应，并提高他们的意识。

期望：该战略致力于推动至少 25% 的预购量增长。

#### 回答 1:

当然可以。以下是一个针对您的环保运动鞋新产品的内容营销策略：

确定目标受众：

明确您的目标受众，例如年龄、性别、地理位置、收入水平、健身偏好等。

了解他们的生活习惯、消费行为和社交媒体使用习惯。

内容营销策略：

故事讲述：制作一系列视频和博客文章，讲述您的品牌如何致力于环保，从材料选择、生产过程到包装设计，都要突出可持续性的理念。

影响者合作：与健身领域的意见领袖和影响者合作，让他们试用您的产品并分享他们的体验。选择那些已经在环保领域有影响力的个人。

互动内容：在社交媒体上发起竞赛和挑战，例如“最长跑步挑战”或“环保生活方式分享”，鼓励用户参与并分享他们的故事。

教育内容：发布关于可持续时尚、环保生活方式和运动对环境的影响的教育性内容。

社交媒体营销：

在 Facebook、Instagram、Twitter、Pinterest 等平台上创建活动专区。

使用相关的话题标签 (hashtags) 来增加话题的可见性。

运用社交媒体广告针对您的目标受众进行精准推广。

合作伙伴关系：

与其他环保品牌建立合作伙伴关系，共同推广可持续生活方式。

与健身房、跑步俱乐部等体育机构合作，举办活动或提供折扣。

SEO 和内容优化：

确保您的网站和内容针对相关的关键词进行了优化，以便在搜索引擎结果中获得更高的排名。

使用长尾关键词来吸引那些正在寻找特定类型的环保产品的用户。

电子邮件营销：

建立邮件订阅列表，定期发送更新、促销和有价值的内容。

设计引人注目的邮件模板，以吸引更多的点击和转化。

分析和调整：

使用 Google Analytics 和其他工具来跟踪您的营销活动的效果。

根据数据反馈调整您的营销策略，以提高效果。

以上是初步的策略，具体的执行和调整需要根据市场反馈和数据分析进行。希望对您有所帮助！

## 步骤 2 CARE 框架

1. Context 上下文：提供上下文信息或情况。
2. Action 行动：描述您想要做什么。
3. Result 结果：描述期望的结果。
4. Example 示例：举一个例子来说明您的观点。

提示 2：

上下文：我们的组织最近推出了一个新的服装系列。

行动：您能协助我们创建一个有针对性的广告活动，强调我们的环保承诺吗？

结果：我们期望的结果是提高产品的知名度和销量，特别是在有生态意识的消费者中。

示例：类似的成功案例中一个很好的例子是 Patagonia 的“不要买这件夹克”活动，这有效地突出了他们对可持续发展的承诺，同时提升了他们的品牌形象。

回答 2：

当然可以。以下是一个针对您的环保服装系列的广告活动建议：

活动名称：“为地球而穿”

活动目标：

提高新服装系列的知名度和销量。

突出我们品牌的环保承诺。

提升品牌形象。

活动策略：

故事讲述：制作一系列视频和博客文章，讲述我们的品牌如何致力于环保，从材料选择、生产过程到包装设计，都要突出可持续性的理念。

互动内容：在社交媒体上发起竞赛和挑战，例如“环保生活方式分享”，鼓励用户参与并分享他们的故事。

影响者合作：与生态意识强的意见领袖和影响者合作，让他们试用我们的产品并分享他们的体验。选择那些已经在环保领域有影响力的人。

教育内容：发布关于可持续时尚、环保生活方式和运动对环境的影响的教育性内容。

广告宣传：在各大社交媒体平台、网络论坛和户外广告牌上投放有针对性的广告，展示我们的环保服装系列和品牌承诺。

活动执行：

制作宣传材料：拍摄高质量的服装系列照片和视频，用于广告投放和社交媒体宣传。

社交媒体推广：在 Facebook、Instagram、Twitter、Pinterest 等平台上发布活动相关内容，使用相关的话题标签（hashtags）来增加话题的可见性。

影响者合作：与健身领域的意见领袖和影响者合作，让他们试用我们的产品并分享他们的体验。

电子邮件营销：定期向订阅者发送关于我们的环保服装系列的最新消息、优惠信息和相关文章。

线下活动：组织环保主题的线下活动，如马拉松、户外音乐节等，现场展示我们的产品，并与参与者互动交流。

活动评估：

使用 Google Analytics 和其他工具来跟踪广告投放和社交媒体宣传的效果。

收集用户反馈和评价，了解他们对我们的环保服装系列的看法。

根据数据反馈和用户反馈，调整广告策略和活动方案，以提高效果。

以上是一个初步的广告活动方案，具体的执行和调整需要根据市场反馈和数据分析进行。希望对您有所帮助！

### 3.4 思考题

1. 请使用其他对话大模型平台，使用相同问题的不同提示方式，测试各个模型的回答结果并比较哪种回答更符合您的期望，差异产生的原因是什么？
  2. 使用更多的提示框架，向同一个大模型提问，测试回答结果并比较哪种提示框架更符合您的期望。
-

华为认证 AI 系列教程

HCIA-AI Solution

大模型压缩部署

实验指导手册

版本：1.0



华为技术有限公司



**版权所有 © 华为技术有限公司 2024。 保留一切权利。**

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

## **商标声明**



和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

## **注意**

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

**华为技术有限公司**

地址： 深圳市龙岗区坂田华为总部办公楼 邮编： 518129

网址： <https://e.huawei.com>

---

## 华为认证体系介绍

华为认证是华为公司基于“平台+生态”战略，围绕“云-管-端”协同的新ICT技术架构，打造的覆盖ICT（Information and Communications Technology，信息通信技术）全技术领域的认证体系，包含ICT基础设施认证、基础软硬件认证、云平台及云服务认证三类认证。

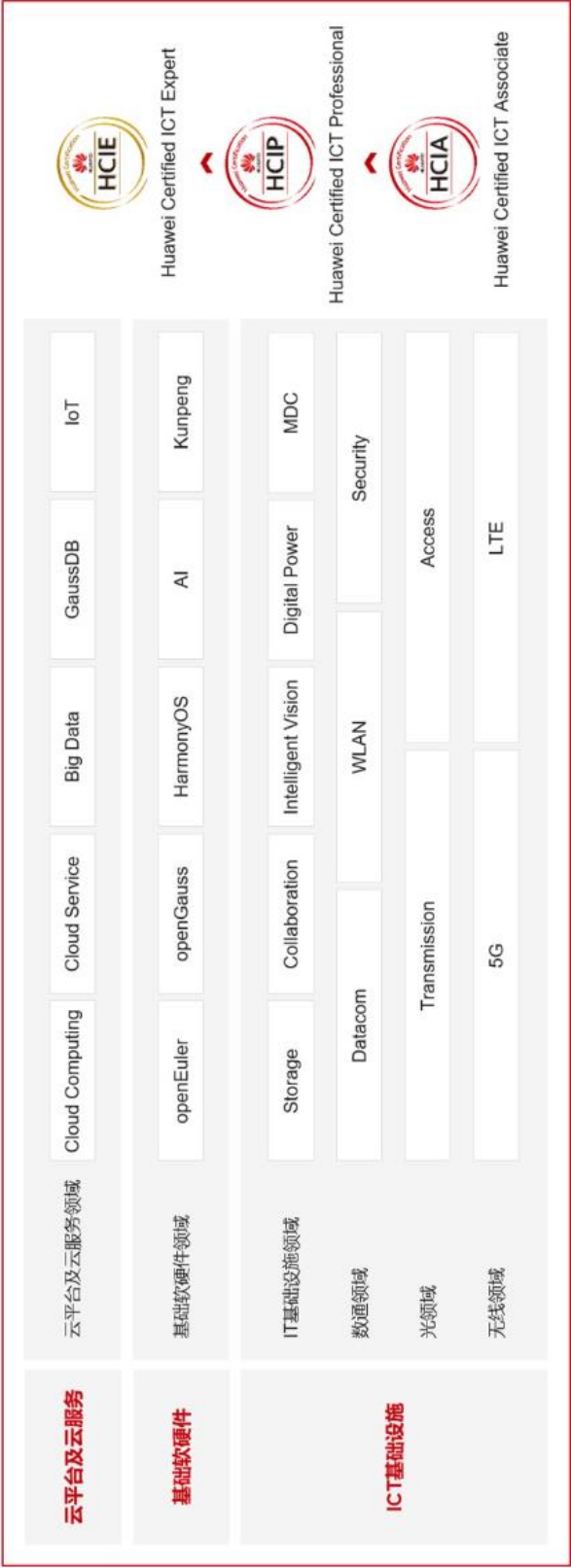
根据ICT从业者的学习和进阶需求，华为认证分为工程师级别、高级工程师级别和专家级别三个认证等级。

华为认证覆盖ICT全领域，符合ICT融合的技术趋势，致力于提供领先的人才培养体系和认证标准，培养数字化时代新型ICT人才，构建良性ICT人才生态。

HCIA-AI Solution（Huawei Certified ICT Associate-AI Solution，华为认证人工智能解决方案）主要面向华为公司办事处、代表处一线工程师，以及其他希望学习华为AI解决方案的技术人士。HCIA-AI Solution认证在内容上涵盖人工智能和算力概念、大模型业务流程、人工智能发展趋势、面向AI大模型的华为智算产品软件和硬件、大模型部署、业界大模型及应用，具备大模型简单部署、使用提示工程的能力。

华为认证协助您打开行业之窗，开启改变之门，屹立在人工智能世界的潮头浪尖！

# 华为职业认证



## 简介

本书为 HCIA-AI Solution 培训教程，适用于准备参加 HCIA-AI Solution 考试的学员或者希望了解 AI 基础知识、大模型及智算产品、华为智算硬件与解决方案的读者。

## 内容描述

本实验指导书共包含 3 个实验，从环境搭建开始，逐一介绍了如何在 Windows 电脑上安装 Miniforge、Git，然后在搭载 Intel 设备的笔记本上运行 ChatGLM<sub>3</sub> 模型。

- 实验一为实验环境搭建实验，指导读者如何在笔记本上安装 Miniforge 和 Git。
  - 实验二为大模型压缩与部署实验，包括 ChatGLM<sub>3</sub> 模型权重格式转换、压缩和部署部分。
  - 实验三为在华为云上运行实验，适用于 PC 机没有搭载 Intel 设备的学员，实验内容与前两个实验基本一致。
-

# 目录

<b>前 言</b>	<b>3</b>
简介	3
内容描述	3
<b>1 实验环境介绍</b>	<b>6</b>
1.1 实验介绍	6
1.1.1 关于本实验	6
1.1.2 实验设备介绍	6
<b>2 实验环境搭建</b>	<b>7</b>
2.1 实验思路	7
2.2 Miniforge 安装	7
2.3 Git 安装	10
<b>3 大模型压缩与部署</b>	<b>13</b>
3.1 实验思路	13
3.2 创建虚拟环境（可选）	13
3.3 项目创建	13
3.4 项目环境准备	14
3.4.1 获取项目代码（以下方式二选一即可）	14
3.4.2 获取模型权重文件	14
3.4.3 安装项目依赖	15
3.5 模型转换与压缩	15
3.6 项目运行	15
3.7 思考题	16
<b>4 （可选）在华为云运行实验</b>	<b>17</b>
4.1 实验思路	17
4.2 华为云资源购买	17
4.3 实验环境搭建	21
4.3.1 Miniforge 安装	21



4.3.2 Git 安装 ..... 22

4.4 大模型压缩与部署 ..... 22

4.4.1 创建虚拟环境（可选） ..... 22

4.4.2 项目创建 ..... 22

4.4.3 项目环境准备 ..... 23

4.4.4 模型转换与压缩 ..... 24

4.5 项目运行 ..... 24

# 1 实验环境介绍

## 1.1 实验介绍

### 1.1.1 关于本实验

Miniforge 包含 conda-forge 特有的 Conda 和 Mamba 的最小安装程序，与 Anaconda 相比占用空间更小，与 Miniconda 相比，同时由于包含 Mamba，易用性上会更好。

Git 是一个开源的分布式版本控制系统，可以有效、高速地处理从很小到非常大的项目版本管理，Git LFS 是一种开源的 Git 扩展，用于管理大文件和二进制文件，将它们存储在单独的“LFS 存储库”中，从而让 Git 存储库保持在一个可管理的规模，本实验 Git 用于下载模型权重文件。

ChatGLM<sub>3</sub> 是智谱 AI 和清华大学 KEG 实验室联合发布的对话预训练模型。ChatGLM<sub>3</sub>-6B 是 ChatGLM<sub>3</sub> 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，ChatGLM<sub>3</sub>-6B 引入了更强大的基础模型、更完整的功能支持、更全面的开源系列。

OpenVINO 是 Intel 为深度学习推理而设计的开源工具包。它可以帮助开发者优化模型，提高推理性能，减少模型的内存占用，目前仅支持在 Intel 设备上使用。

本实验参考 ChatGLM<sub>3</sub> 官方样例编写而成，原始官方链接如下（**本实验无需阅读原始链接**）：

[https://github.com/THUDM/ChatGLM3/blob/main/Intel\\_device\\_demo/opencvino\\_demo/README.md](https://github.com/THUDM/ChatGLM3/blob/main/Intel_device_demo/opencvino_demo/README.md)

### 1.1.2 实验设备介绍

设备规格要求如下：

设备名称	设备硬件要求	OS要求
PC机	Intel CPU，内存16G以上（如果设备不满足要求，可参考实验三（4 在华为云上运行实验））	Windows10及以上

## 2 实验环境搭建

### 2.1 实验思路

- 获取 Miniforge 安装包，安装 Miniforge，测试是否安装成功；
- 获取 Git 安装包，安装 Git，测试是否安装成功。

### 2.2 Miniforge 安装

步骤 1 点击下方链接进入项目地址；

<https://github.com/conda-forge/miniforge>

步骤 2 点击对应链接下载安装包，需要注意下载安装包为 Miniforge3 而不是 Miniforge-pypy3，本实验运行环境为 Windows；

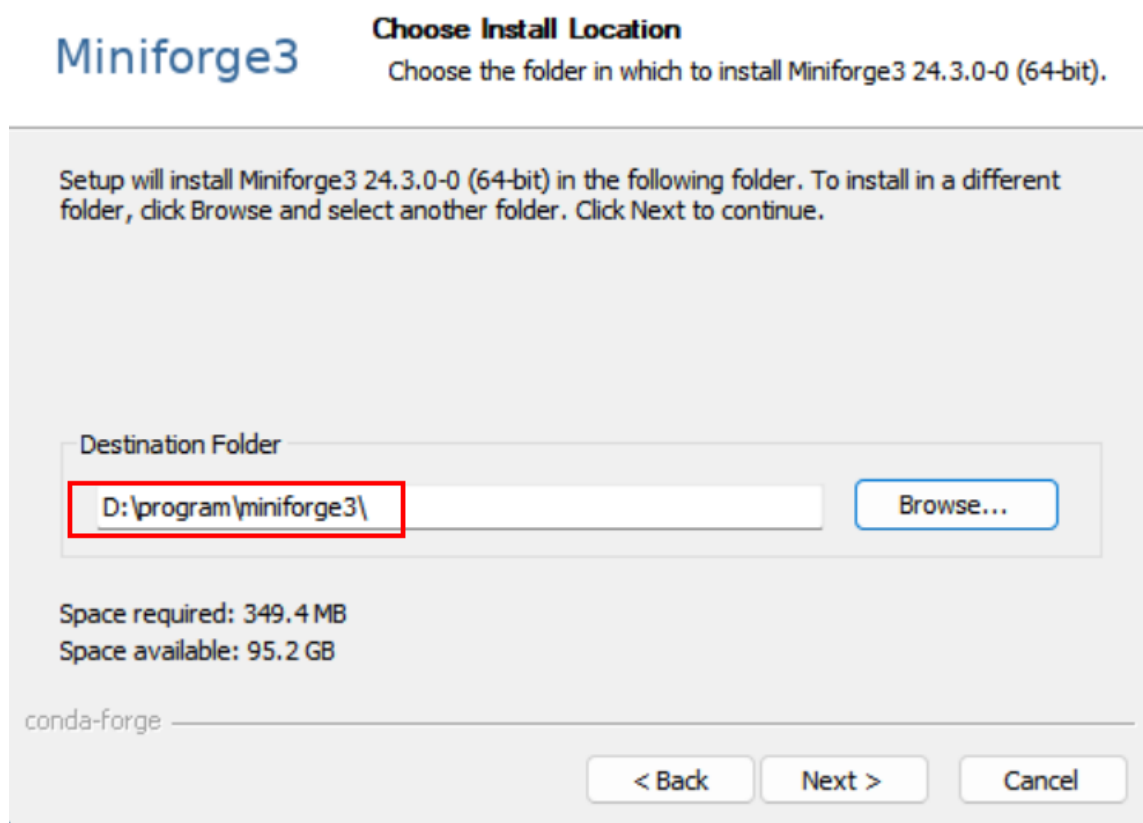
#### Miniforge3

Latest installers with Python 3.10 (\*) in the base environment:

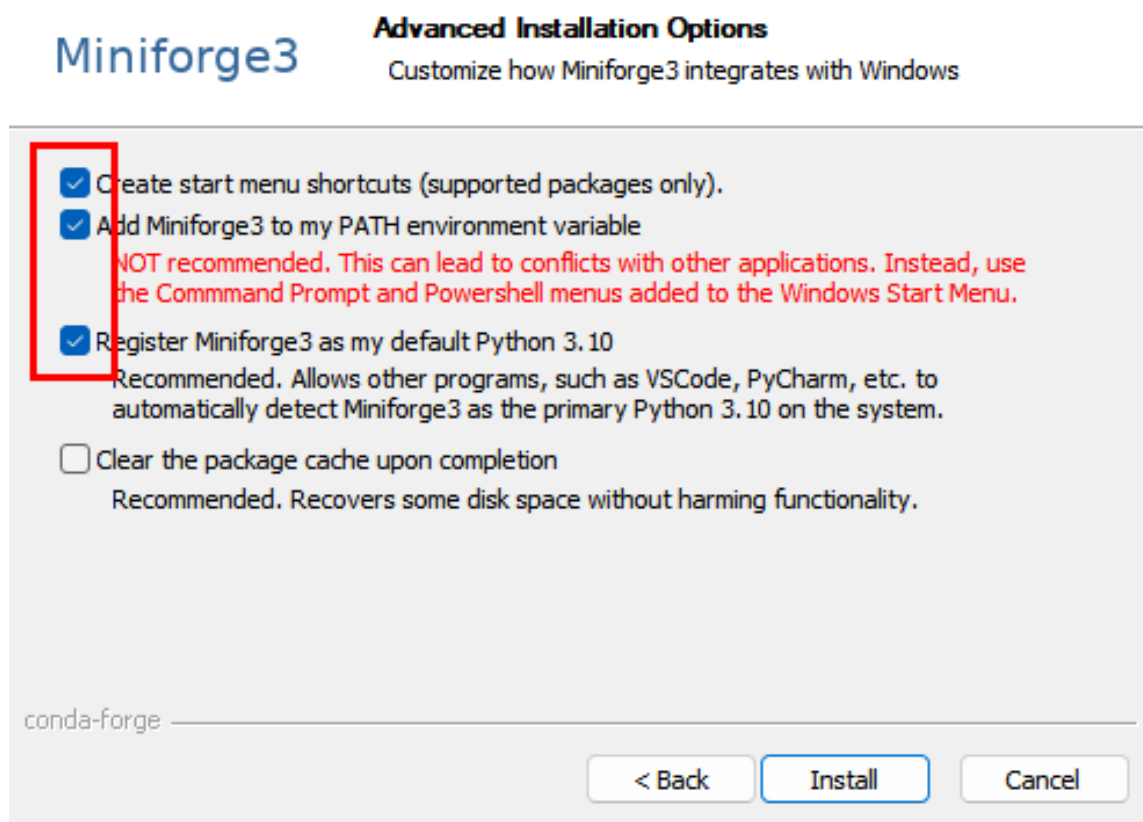
OS	Architecture	Download
Linux	x86_64 (amd64)	<a href="#">Miniforge3-Linux-x86_64</a>
Linux	aarch64 (arm64) (**)	<a href="#">Miniforge3-Linux-aarch64</a>
Linux	ppc64le (POWER8/9)	<a href="#">Miniforge3-Linux-ppc64le</a>
OS X	x86_64	<a href="#">Miniforge3-MacOSX-x86_64</a>
OS X	arm64 (Apple Silicon) (***)	<a href="#">Miniforge3-MacOSX-arm64</a>
Windows	x86_64	<a href="#">Miniforge3-Windows-x86_64</a>

步骤 3 下载完成后双击安装包进行安装，安装路径建议**只包含英文**，不然可能会出现问題；

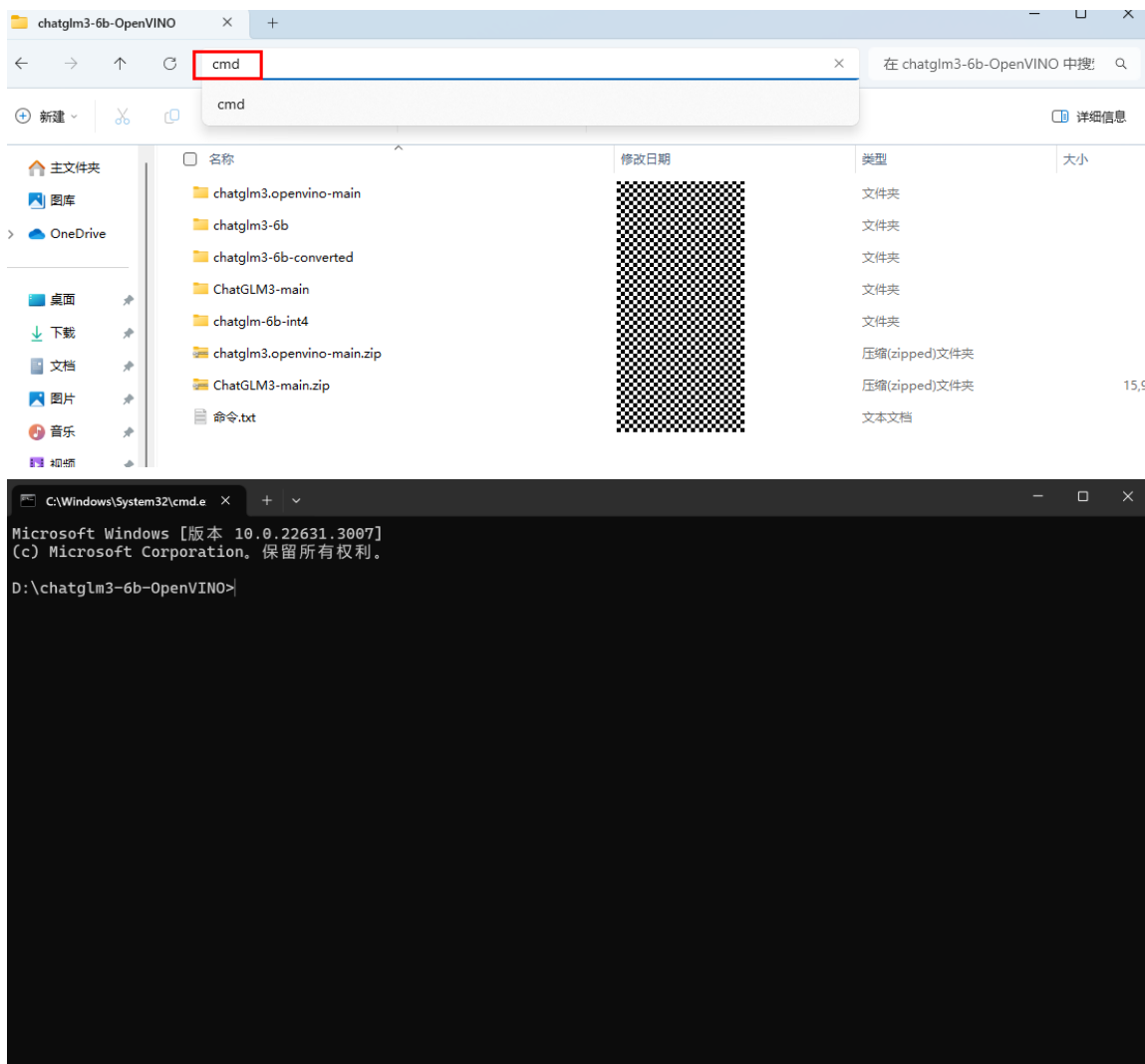




步骤 4 安装过程中下面三个打勾，不然后续需要手动配置环境变量；



步骤 5 安装完成后打开资源管理器，随后在地址栏输入 cmd 打开命令行窗口；



步骤 6 输入 mamba，如果有回显则表示软件安装成功；

```
C:\Windows\System32\cmd.e  X  +  v

-V, --version          Show the conda version number and exit.

commands:
The following built-in and plugins subcommands are available.

COMMAND
activate              Activate a conda environment.
clean                 Remove unused packages and caches.
compare               Compare packages between conda environments.
config                Modify configuration values in .condarc.
create                Create a new conda environment from a list of specified packages.
deactivate             Deactivate the current active conda environment.
doctor                Display a health report for your environment.
env                   See 'conda env --help'.
info                  Display information about current conda install.
init                  Initialize conda for shell interaction.
install                Install a list of packages into a specified conda environment.
list                  List installed packages in a conda environment.
notices                Retrieve latest channel notifications.
package               Create low-level conda packages. (EXPERIMENTAL)
remove (uninstall)    Remove a list of packages from a specified conda environment.
rename                Rename an existing environment.
repoquery              Advanced search for repodata.
repoquery             Query repositories using mamba.
run                   Run an executable in a conda environment.
search                Search for packages and display associated information using the MatchSpec format.
update (upgrade)       Update conda packages to the latest compatible version.

D:\chatglm3-6b-OpenVINO>
```

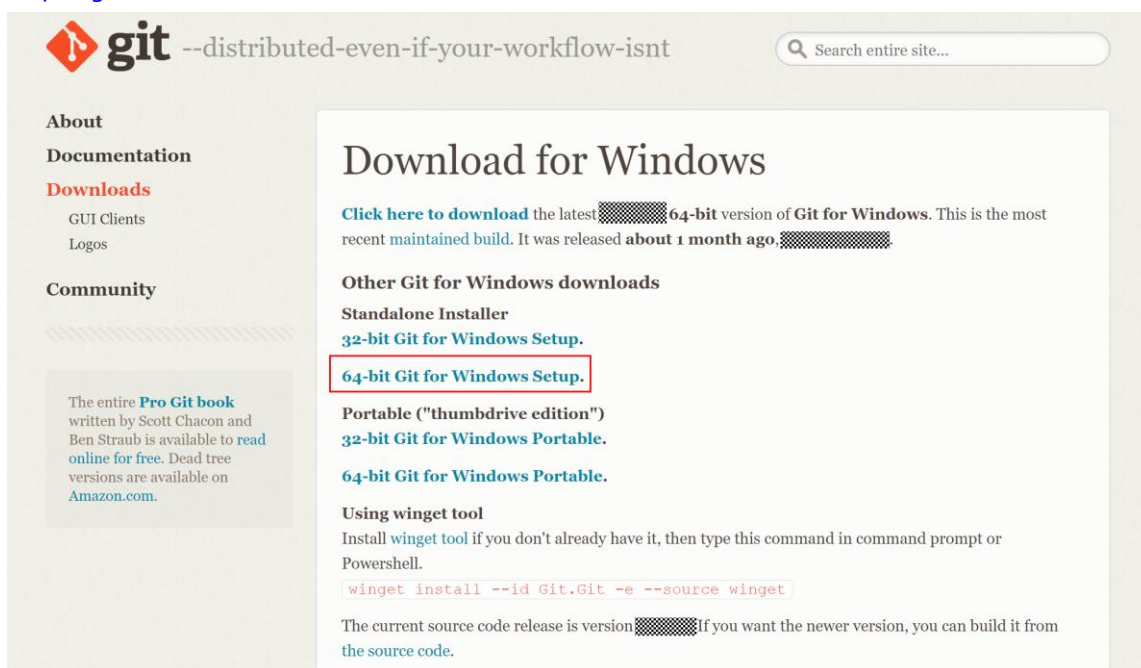
步骤7 输入以下命令更换 PyPI 源，在使用 pip 安装软件包时速度更快：

```
python -m pip install -i https://pypi.tuna.tsinghua.edu.cn/simple --upgrade pip
pip config set global.index-url https://pypi.tuna.tsinghua.edu.cn/simple
```

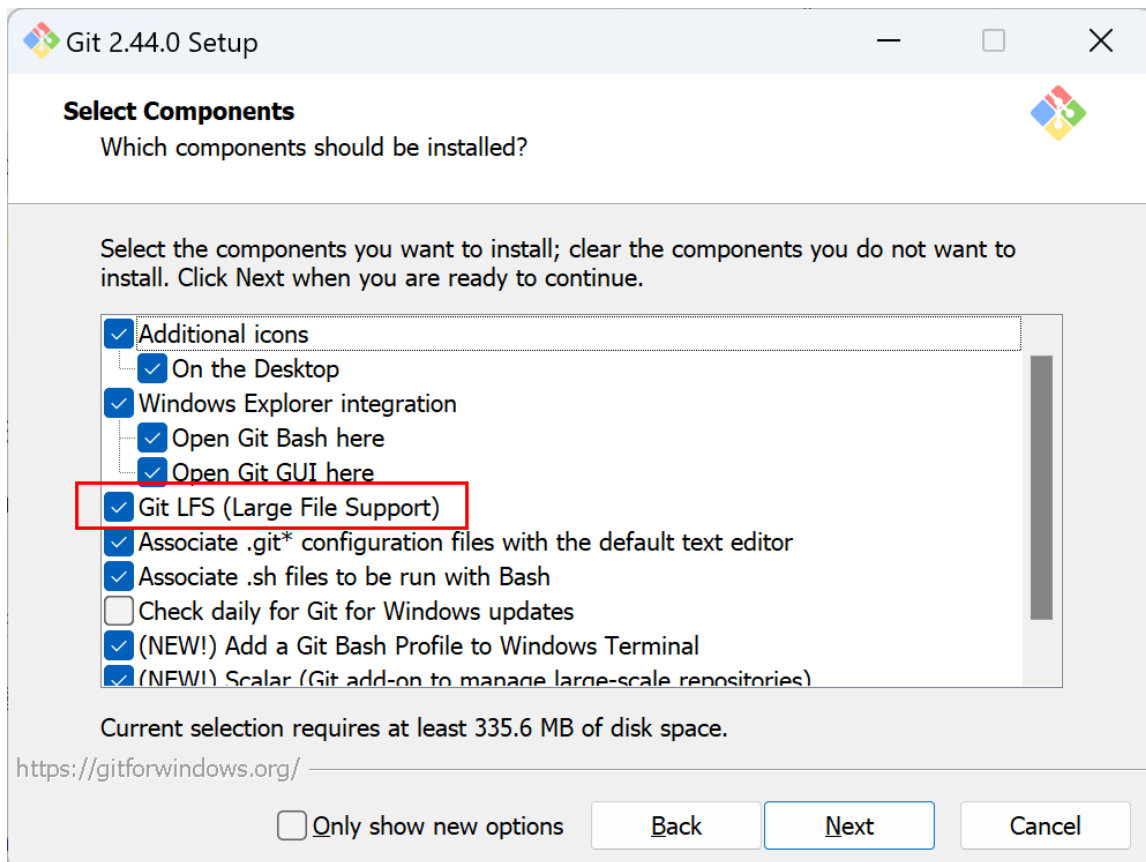
## 2.3 Git 安装

步骤1 点击下方链接进入软件下载页面，选择 64 位版本下载；

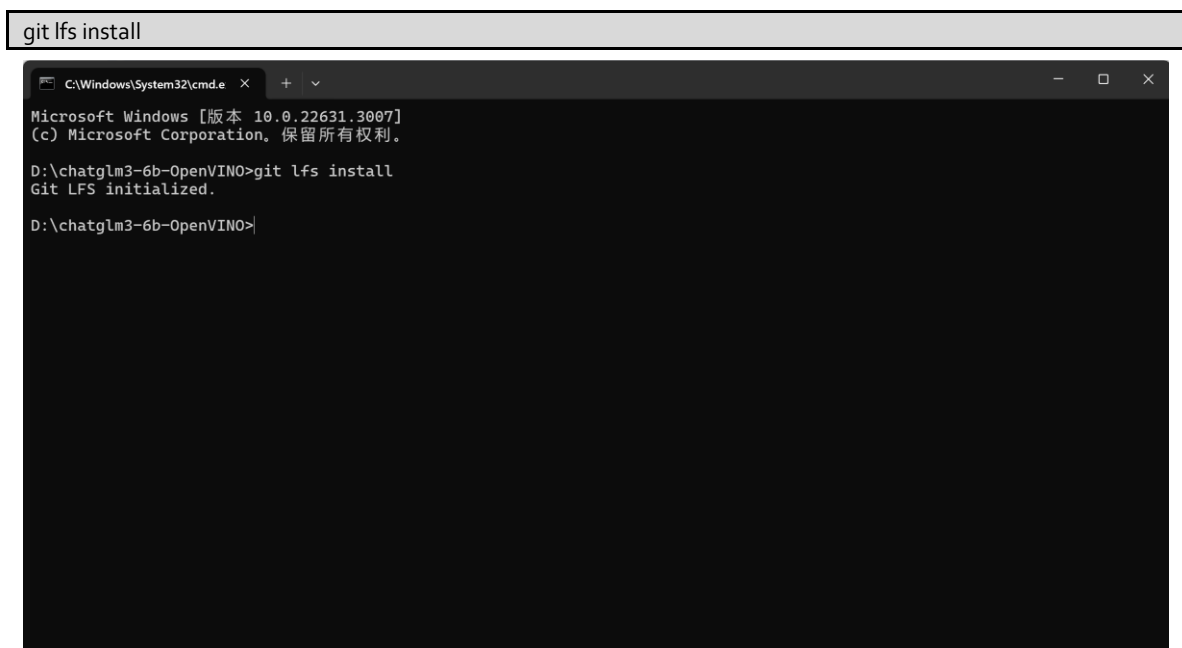
<https://git-scm.com/download/win>



步骤 2 下载完成后双击安装包进行安装，注意在 Git LFS 前面打钩；



步骤 3 安装路径建议全英文，其余选项默认即可，安装完成后在命令行窗口输入以下命令，有回显表示 Git 和 Git LFS 安装成功；





# 3 大模型压缩与部署

## 3.1 实验思路

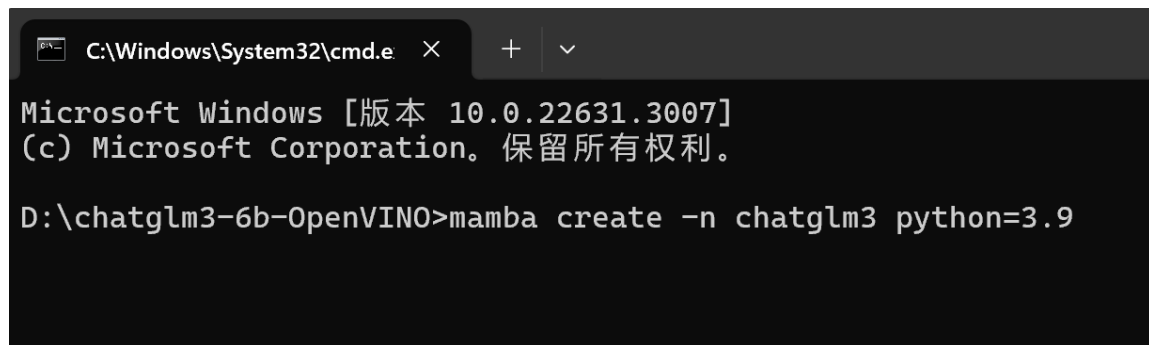
- PC 上配置项目运行环境，用于模型转换和推理；
- 获取模型原始权重和推理代码，完成模型转换后开启推理服务；
- 与模型对话，体验 ChatGLM3-6B。

## 3.2 创建虚拟环境（可选）

推荐为每个项目创建一个虚拟环境，防止软件包冲突。

步骤 1 在命令行窗口输入以下命令创建虚拟环境，红色部分为虚拟环境名称可自定义；

```
mamba create -n chatglm3 python=3.9
```



```
C:\Windows\System32\cmd.e  X  +  v  
Microsoft Windows [版本 10.0.22631.3007]  
(c) Microsoft Corporation. 保留所有权利。  
D:\chatglm3-6b-OpenVINO>mamba create -n chatglm3 python=3.9
```

步骤 2 输入回车，然后输入 Y 确认创建即可。

## 3.3 项目创建

步骤 1 新建一个空文件夹用于存放项目代码，**路径不能包含中文**。

## 3.4 项目环境准备

### 3.4.1 获取项目代码（以下方式二选一即可）

#### 3.4.1.1 方式一（推荐）

点击下方链接下载代码包，解压后放入**项目文件夹**中即可。

<https://certification-data.obs.cn-north-4.myhuaweicloud.com/CHS/HCIA-AI%20Solution/V1.0/chapter6/chatglm3.openvino.zip>

#### 3.4.1.2 方式二

在**项目文件夹**中进入命令行窗口，输入以下命令获取项目代码：

```
git clone https://github.com/OpenVINO-dev-contest/chatglm3.openvino.git
```

### 3.4.2 获取模型权重文件（以下方式二选一即可）

#### 3.4.2.1 方式一（推荐）

步骤 1 在**项目文件夹**中进入命令行窗口，输入以下命令安装 ModelScope Library：

```
pip install modelscope
```

步骤 2 输入以下命令进入 Python 环境，然后执行 Python 代码下载模型权重：

```
Python
from modelscope import snapshot_download
model_dir = snapshot_download('ZhipuAI/chatglm3-6b', cache_dir='D:\chatglm3-6b', revision='master')
```

步骤 3 下载完成后输入以下命令退出 Python 环境：

```
exit()
```

```
Downloading: 100%|
Downloading: 100%|
>>> exit()
(base) [root@ecs-d992 chatglm3]#
```

#### 3.4.2.2 方式二

步骤 1 在**项目文件夹**中进入命令行窗口，输入以下命令下载模型权重，这部分时间会比较久：

```
git lfs install --skip-smudge
git clone https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git
cd chatglm3-6b
git lfs pull
```

### 3.4.3 安装项目依赖

步骤 1 输入以下命令进入到代码文件夹中（需要根据命令行实际打开路径进行调整），然后激活虚拟环境：

```
#进入代码文件夹
cd chatglm3.openvino
#激活虚拟环境，红色部分为前面创建的虚拟环境名称
activate chatglm3
```

步骤 2 输入以下命令安装项目依赖：

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

## 3.5 模型转换与压缩

步骤 1 不同的框架对模型格式要求不同，输入以下命令将 Huggingface 模型转换为 OpenVINO IR 模型：

```
python convert.py --model_id THUDM/chatglm3-6b --output {your_path}/chatglm3-6b
```

--model\_id - 模型所在**目录**的路径（绝对路径，不能包含中文）。

--output - 转换后模型保存的**目录**路径（不能包含中文）。

步骤 2 （可选）不量化的情况下，运行模型需要 12G 左右内存，如果内存空间不足或想获得更快的输出，输入以下命令进行模型量化：

```
python quantize.py --model_path {your_path}/chatglm3-6b --precision int4 --output {your_path}/chatglm3-6b-int4
```

--model\_path - OpenVINO IR 模型所在**目录**的路径。

--precision - 量化精度：int8 或 int4。

--output - 保存模型的**目录**路径。

## 3.6 项目运行

步骤 1 输入以下命令运行 ChatGLM3 机器人：

```
python chat.py --model_path {your_path}/chatglm3-6b --max_sequence_length 4096 --device GPU
```

--model\_path - OpenVINO IR 模型所在**目录**的路径。

--max\_sequence\_length - 输出标记的最大大小。



`--device` - 运行推理的设备，有 Intel 核显即可选择 GPU，速度比 CPU 快。

效果如下图：



```
(openvino_chatglm3) D:\chatglm3-6b-OpenVINO\chatglm3.openvino-main\chatglm3.openvino-main>python chat.py --model_path D:\chatglm3-6b-OpenVINO\chatglm3-6b-int4 --max_sequence_length 4096 --device GPU
INFO:nncf:NNCF initialized successfully. Supported frameworks detected: torch, onnx, openvino
Setting eos_token is not supported, use the default one.
Setting pad_token is not supported, use the default one.
Setting unk_token is not supported, use the default one.
====Compiling model====
The argument 'trust_remote_code' is to be used along with export=True. It will be ignored.
Compiling the model to GPU ...
====Starting conversation====
用户：你好
ChatGLM3-6B-OpenVINO: 我是人工智能助手，很高兴为您服务。请问有什么问题我可以帮您解答吗？

用户：老抽和生抽有什么区别？
ChatGLM3-6B-OpenVINO:
老抽和生抽都是我国烹饪中常用的酱油种类，它们的主要区别在于颜色和味道：

1. 颜色：生抽的颜色较深，呈棕红色，而老抽的颜色较浅，呈暗红色。
2. 味道：生抽的口感较鲜，有鲜美的滋味，而老抽的味道较浓，带有一定的甜味。
3. 使用场景：生抽通常用于炒菜、拌面等烹饪过程中，需要上色和提味；而老抽主要用于炖煮、红烧等需要调色的菜肴中，使菜肴颜色更深、口感更浓郁。

总的来说，生抽更适合用于烹饪过程，而老抽则更多地用于调味和增色。

用户：|
```

## 3.7 思考题

不同的硬件设备推理模型时，对模型回答准确性有影响吗？

如何调整提示词，可以让大模型的回答更符合您的期望？

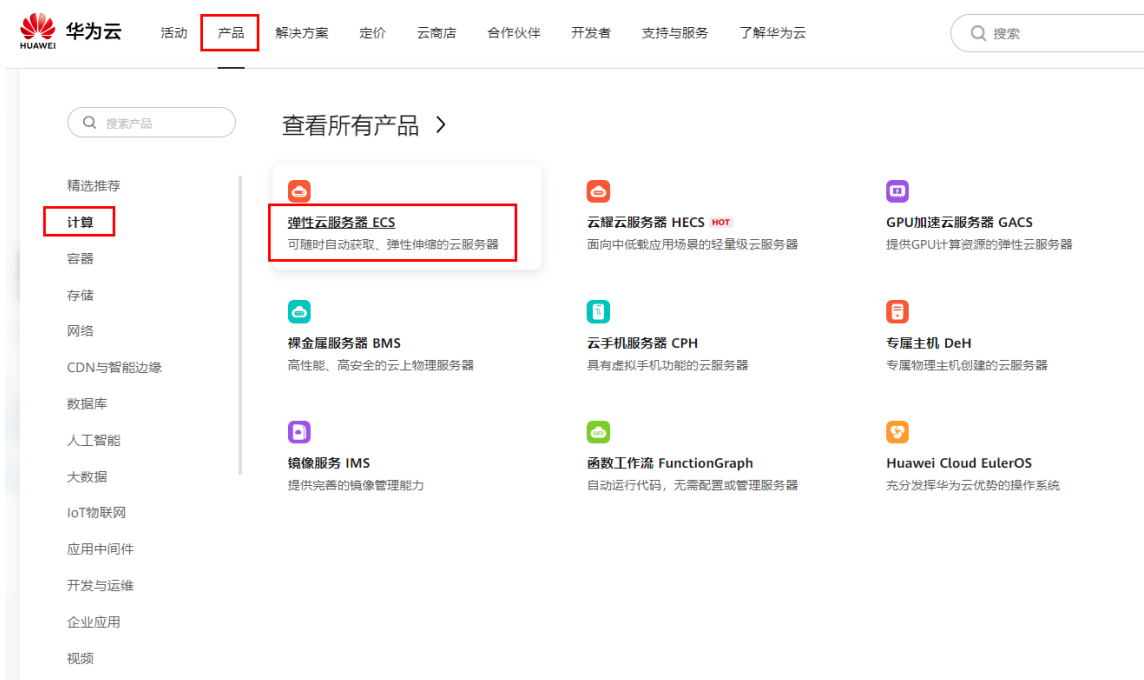
# 4 在华为云运行实验（可选）

## 4.1 实验思路

- 购买华为云 ECS；
- ECS 上完成环境部署，与实验一（2 实验环境搭建）基本一致，命令上有部分区别；
- ECS 完成大模型的压缩和部署，与实验二（3 大模型压缩与部署）基本一致，命令上有部分区别。

## 4.2 华为云资源购买

步骤 1 登录华为云，找到弹性云服务器产品；



步骤 2 点击购买，进入产品购买界面；



步骤3 选择**按需计费**，X86 计算，内存优化型（实验对内存容量要求较高）；



步骤4 选择内存为 32GiB 的实例，系统选择 EulerOS，然后点击下一步；

全选
m7
am7
m6
m3ne

收起帮助 ^

可应对大型内存数据集和高网络场景。适用于内存要求高，数据量大并且数据访问量大的场景，同时要求快速的数据交换和处理。

实例类型	规格名称	vCPUs	内存	CPU
<input type="radio"/> 内存优化型m7	m7.large.8	2vCPUs	16GiB	Intel Ice Lake 3.0GHz
<input checked="" type="radio"/> 内存优化型m7	m7.xlarge.8	4vCPUs	32GiB	Intel Ice Lake 3.0GHz
<input type="radio"/> 内存优化型m7	m7.2xlarge.8	8vCPUs	64GiB	Intel Ice Lake 3.0GHz
<input type="radio"/> 内存优化型m7	m7.3xlarge.8	12vCPUs	96GiB	Intel Ice Lake 3.0GHz
<input type="radio"/> 内存优化型m7	m7.4xlarge.8	16vCPUs	128GiB	Intel Ice Lake 3.0GHz
<input type="radio"/> 内存优化型m7	m7.6xlarge.8	24vCPUs	192GiB	Intel Ice Lake 3.0GHz
<input type="radio"/> 内存优化型m7	m7.8xlarge.8	32vCPUs	256GiB	Intel Ice Lake 3.0GHz

当前规格 **内存优化型 | m7.xlarge.8 | 4vCPUs | 32GiB**

您当前选择规格，仅支持使用SCSI磁盘模式挂载磁盘，不支持使用VBD磁盘模式挂载磁盘。磁盘标识为wwn号。

公共镜像
私有镜像
共享镜像
市场镜像

Huawei Cloud EulerOS
Huawei Cloud EulerOS 2.0 标准版 64位(40GiB)

☒ 免费试用一个月主机安全基础防护
 ☐ 购买高阶防护

步骤5 网络设置界面选择现在购买弹性公网IP，**按流量计费**，带宽选择 300Mbit/s，实验流量基本为入流量，不计费，其余默认即可，最后完成实例购买；

弹性公网IP
 ☒ 现在购买
 ☐ 使用已有
 ☐ 暂不购买

线路
 全动态BGP
静态BGP

不低于99.95%可用性保障

公网带宽
 

按带宽计费  
流量较大或较稳定的场景

**按流量计费**  
流量小或流量波动较大场景

加入共享带宽  
多业务流量错峰分布场景

指定带宽上限，按实际使用的出公网流量计费，与使用时间无关。

带宽大小
 

5 10 20 50 100 自定义

300

 带宽范围：1-300 Mbit/s

免费开启DDoS基础防护

释放行为
 ☐ 随实例释放

对于设置了随实例释放的弹性公网IP，将在删除云服务器同时执行删除。

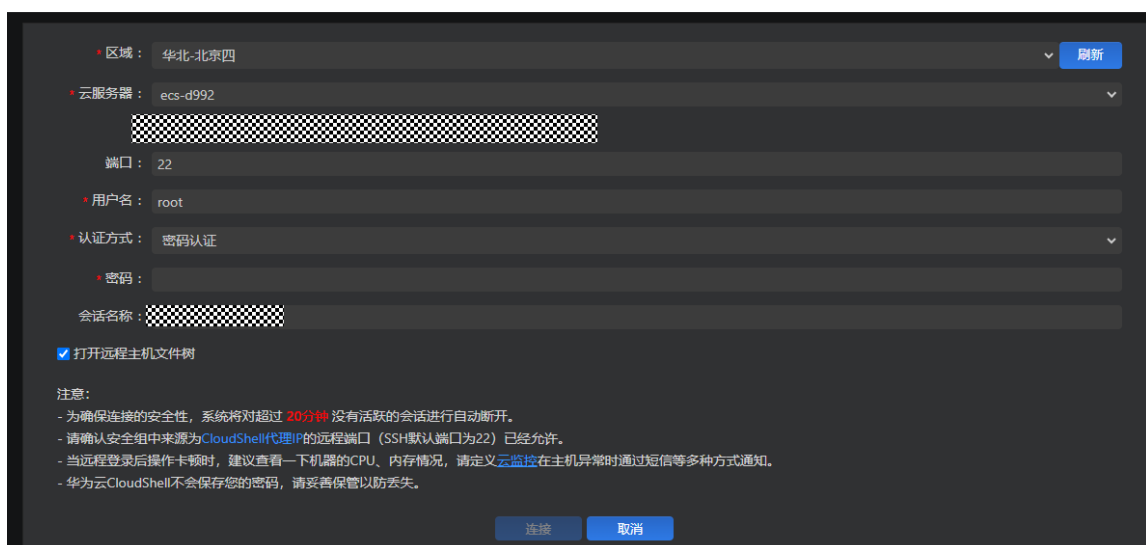
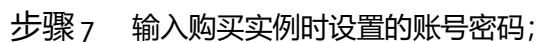
买量
 

1

 台
 配置费用 **¥1.4878/小时** + 弹性公网IP流量费用 **¥0.80/GB**

步骤6 实例购买成功后可通过 CloudShell 登录，如果电脑安装了 Xshell 等软件也可以使用；

名称ID	监控	安全	状态	可用区	规格名称	操作系统	IP地址	计费模式	企业项目	标签	操作
ecs-0992			运行中	可用区7	4vCPUs   32GiB   m7.xlarge.8 Huawei Cloud EulerOS 2.0 标准版	Linux		按量计费	default		<span>远程登录</span>



## 4.3 实验环境搭建

### 4.3.1 Miniforge 安装

步骤 1 输入以下命令下载 Miniforge 安装包（Miniforge 与 Mambaforge 已经合并，releases 界面的名称为 MambaForge）：

```
wget https://github.com/conda-forge/miniforge/releases/download/24.3.0-0/Mambaforge-24.3.0-0-Linux-x86_64.sh
```

步骤 2 输入以下命令给安装包添加执行权限，然后执行安装：

```
chmod +x Mambaforge-24.3.0-0-Linux-x86_64.sh
./Mambaforge-24.3.0-0-Linux-x86_64.sh
```

步骤 3 安装过程中需要根据提示按回车，输入 yes，最后初始化部分默认 no，需注意输入 yes；

```
agreement.
Please, press ENTER to continue
>>>

Do you accept the license terms? [yes|no]
>>> yes

conda config --set auto_activate_base false

You can undo this by running `conda init --reverse $SHELL`? [yes|no]
[no] >>> yes
```

步骤 4 安装成功后断开连接，重新连接 ECS，这时可以发现 root 前面多了（base），输入 mamba，有回显表示安装成功；

```
(base) [root@ecs-d992 ~]# mamba
usage: mamba [-h] [-v] [--no-plugins] [-V] COMMAND ...

conda is a tool for managing and deploying applications, environments and packages.

options:
  -h, --help            Show this help message and exit.
  -v, --verbose          Can be used multiple times. Once for detailed output, twice for INFO logging, thrice for DEBUG logging, four times for TRACE logging.
  --no-plugins          Disable all plugins that are not built into conda.
  -V, --version          Show the conda version number and exit.

commands:
  The following built-in and plugins subcommands are available.

COMMAND
activate      Activate a conda environment.
clean         Remove unused packages and caches.
compare       Compare packages between conda environments.
config        Modify configuration values in .condarc.
create        Create a new conda environment from a list of specified packages.
deactivate    Deactivate the current active conda environment.
doctor        Display a health report for your environment.
export        Export a given environment
info          Display information about current conda install.
init          Initialize conda for shell interaction.
install       Install a list of packages into a specified conda environment.
list          List installed packages in a conda environment.
notices       Retrieve latest channel notifications.
package       Create low-level conda packages. (EXPERIMENTAL)
remove (uninstall) Remove a list of packages from a specified conda environment.
rename        Rename an existing environment.
repoquery     Advanced search for repodata.
repoquery     Query repositories using mamba.
run           Run an executable in a conda environment.
search        Search for packages and display associated information using the MatchSpec format.
update (upgrade) Update conda packages to the latest compatible version.
```

步骤 5 输入以下命令更换 PyPI 源：

```
python -m pip install -i https://pypi.tuna.tsinghua.edu.cn/simple --upgrade pip
```

```
pip config set global.index-url https://pypi.tuna.tsinghua.edu.cn/simple
```

## 4.3.2 Git 安装

步骤 1 输入以下命令安装 Git：

```
yum makecache  
yum install git -y
```

步骤 2 输入以下命令下载 Git Lfs 安装包并解压：

```
wget https://github.com/git-lfs/git-lfs/releases/download/v3.5.1/git-lfs-linux-amd64-v3.5.1.tar.gz  
tar -zxvf git-lfs-linux-amd64-v3.5.1.tar.gz
```

步骤 3 输入以下命令安装 Git Lfs：

```
cd git-lfs-3.5.1/  
./install.sh
```

步骤 4 输入以下命令，有回显表面 Git Lfs 安装成功：

```
git lfs install
```

```
(base) [root@ecs-d992 git-lfs-3.5.1]# git lfs install  
Git LFS initialized.
```

## 4.4 大模型压缩与部署

### 4.4.1 创建虚拟环境（可选）

步骤 1 输入以下命令创建 Python 虚拟环境，防止依赖包冲突：

```
mamba create -n chatglm3 python=3.9
```

### 4.4.2 项目创建

步骤 1 输入以下命令回到家目录，然后创建项目目录，用于存放模型权重和项目代码：

```
cd ~  
mkdir chatglm3
```

## 4.4.3 项目环境准备

### 4.4.3.1 获取项目代码（以下方式二选一即可）

#### 4.4.3.1.1 方式一（推荐）

步骤 1 输入以下命令进入项目目录，然后下载和解压实验代码：

```
cd chatglm3
wget https://certification-data.obs.cn-north-4.myhuaweicloud.com/CHS/HCIA-
AI%20Solution/V1.0/chapter6/chatglm3.openvino.zip
unzip chatglm3.openvino.zip
```

#### 4.4.3.1.2 方式二

步骤 1 输入以下命令获取原始项目代码：

```
git clone https://github.com/OpenVINO-dev-contest/chatglm3.openvino.git
```

### 4.4.3.2 获取模型权重文件（以下方式二选一即可）

#### 4.4.3.2.1 方式一（推荐）

步骤 1 输入以下命令安装 ModelScope Library：

```
pip install modelscope
```

步骤 2 输入以下命令进入 Python 环境，然后执行 Python 代码下载模型权重：

```
Python
from modelscope import snapshot_download
model_dir = snapshot_download('ZhipuAI/chatglm3-6b', cache_dir='/root/chatglm3/chatglm3-6b', revision='master')
```

步骤 3 下载完成后输入以下命令退出 Python 环境：

```
exit()
```

```
Downloading: 100%|
Downloading: 100%|
>>> exit()
(base) [root@ecs-d992 chatglm3]#
```

#### 4.4.3.2.2 方式二

步骤 1 输入以下命令下载模型权重，这部分时间会比较久：

```
git lfs install --skip-smudge
git clone https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git
cd chatglm3-6b
git lfs pull
```



### 4.4.3.3 安装项目依赖

步骤 1 输入以下命令进入到代码目录：

```
cd ~
cd chatglm3/chatglm3.opengvino
```

步骤 2 输入以下命令进入到虚拟环境并安装依赖，红色为前面自己设置的环境变量名称：

```
mamba activate chatglm3
pip install wheel setuptools
pip install -r requirements.txt
```

### 4.4.4 模型转换与压缩

步骤 1 不同的框架对模型格式要求不同，输入以下命令将 Huggingface 模型转换为 OpenVINO IR 模型：

```
python convert.py --model_id /root/chatglm3/chatglm3-6b/ZipuAI/chatglm3-6b/ --output
/root/chatglm3/chatglm3-6b-converted
```

--model\_id - 模型所在目录的路径（绝对路径，不能包含中文）。

--output - 转换后模型保存的目录路径（不能包含中文）。

步骤 2 （可选）不量化的情况下，运行模型需要 12G 左右内存，如果内存空间不足或想获得更快的输出，输入以下命令进行模型量化：

```
python quantize.py --model_path /root/chatglm3/chatglm3-6b-converted/ --precision int4 --output
/root/chatglm3/chatglm3-6b-int4
```

--model\_path - OpenVINO IR 模型所在目录的路径。

--precision - 量化精度：int8 或 int4。

--output - 保存模型的目录路径。

## 4.5 项目运行

步骤 1 输入以下命令运行 ChatGLM3 机器人：

```
python chat.py --model_path /root/chatglm3/chatglm3-6b-int4 --max_sequence_length 4096 --device CPU
```

--model\_path - OpenVINO IR 模型所在目录的路径。

--max\_sequence\_length - 输出标记的最大大小。

--device - 运行推理的设备，ECS 上只能选 CPU。

效果如下图：

```
C:\Windows\System32\cmd.exe X + v

(openvino_chatglm3) D:\chatglm3-6b-OpenVINO\chatglm3.openvino-main\chatglm3.openvino-main>python chat.py --model_path D:\chatglm3-6b-OpenVINO\chatglm3-6b-int4 --max_sequence_length 4096 --device GPU
INFO:nnf:NNCF initialized successfully. Supported frameworks detected: torch, onnx, openvino
Setting eos_token is not supported, use the default one.
Setting pad_token is not supported, use the default one.
Setting unk_token is not supported, use the default one.
====Compiling model====
The argument 'trust_remote_code' is to be used along with export=True. It will be ignored.
Compiling the model to GPU ...
====Starting conversation====
用户：你好
ChatGLM3-6B-OpenVINO: 我是人工智能助手，很高兴为您服务。请问有什么问题我可以帮您解答吗？

用户：老抽和生抽有什么区别？
ChatGLM3-6B-OpenVINO:
老抽和生抽都是我国烹饪中常用的酱油种类，它们的主要区别在于颜色和味道：

1. 颜色：生抽的颜色较深，呈棕红色，而老抽的颜色较浅，呈暗红色。
2. 味道：生抽的口感较鲜，有鲜美的滋味，而老抽的味道较浓，带有一定的甜味。
3. 使用场景：生抽通常用于炒菜、拌面等烹饪过程中，需要上色和提味；而老抽主要用于炖煮、红烧等需要调色的菜肴中，使菜肴颜色更深、口感更浓郁。

总的来说，生抽更适合用于烹饪过程，而老抽则更多地用于调味和增色。

用户：|
```

## 4.6 资源释放（很重要）

实验完成之后需要及时释放资源，避免持续扣费。

步骤 1 进入 ECS 控制台，点击更多，删除；



步骤 2 勾选删除公网 IP 和删除数据盘，释放所有计费资源。

## 删除



① 删除设置 ———— ② 资源明细 ———— ③ 完成

以下1台云服务器，可直接删除

⚠ 已开启云备份的云服务器在删除后，云备份会保留并继续计费，您可以前往[云备份页面](#)进行删除操作。



名称 ▾	ID ▾
ecs-06f3	

删除方式 ☒ 立即删除 ☐ 定时删除

是否同步删除关联资源

⚠ 未勾选的弹性公网IP和数据盘不会随云服务器同步删除，会继续计费。

☒ 删除云服务器绑定的弹性公网IP地址 ☒ 删除云服务器挂载的数据盘

下一步

取消