### ****Dify概述，AI应用开发框架****

Dify 是一款专为开发 AI 应用设计的开发框架。类似于 Java 中的 Spring Boot，后者被广泛应用于 Web 应用开发。使用开发框架的核心目标是提高开发效率，而 Dify 正是为此而生。

官网：https://dify.ai/zh

### ****什么是生成式？****

生成式 AI 是指通过 AI 技术生成各种形式的内容，例如文字、图片、音频或视频。这种生成能力是 AI 应用的核心特性。

### ****开放与灵活性****

Dify 的社区版是开源的，代码托管在 GitHub 上，体现了其开放性。此外，Dify 既支持在线云服务访问，也支持私有化部署，这使其在满足不同场景需求时具备高度的灵活性。

### ****Dify 能做什么？****

作为一个 AI 应用开发框架，Dify 当前支持以下四种主要应用类型：

1. **聊天助手**：基于大语言模型（LLM）构建对话式交互的智能助手。
2. **文本生成**：适用于撰写故事、文本分类、翻译等任务。
3. **Agent**：能够进行任务分解、逻辑推理和工具调用的智能助手。
4. **工作流**：通过流程编排定义灵活的 LLM 工作流。

### ****为什么需要开发聊天助手？****

虽然市场上已有 ChatGPT 和文心一言等聊天助手，但开发自有聊天助手仍然具有以下必要性：

1. **定制化需求**  
   如果需要在自己的网站或应用中集成聊天助手，仅嵌入现有产品页面（OpenAI官方的ChatGPT、文心一言等）并不现实，尤其当登录体系和现有系统不一致时。
2. **嵌入公司知识库**  
   通用聊天助手无法回答特定领域或公司内容。例如，你需要一个能解答公司产品说明书的助手，而现有助手并未接受相关培训。这时，定制开发显得尤为必要。通过Dify提供的知识库即可实现（在Dify中，知识库是一些文档的集合，可以被整体集成至一个应用中作为检索上下文使用）。

### ****为什么选择 Dify？****

**简单且快速** 是选择 Dify 的关键原因。

* **非程序员友好**  
  即使没有编程经验，也可以通过零代码和可视化方式快速构建 AI 应用。
* **程序员效率提升**  
  与 LangChain 等库相比，Dify 提供了更贴近生产需求的完整解决方案，经过精心设计和测试，相当于一套高效的脚手架。

### ****Dify 的发布方式****

Dify 允许用户完全通过零代码的方式可视化搭建 AI 应用，最终生成一个可嵌入网页或 App 的链接。此外，Dify 应用还支持 API 访问，方便灵活集成。例如，你可以开发一个自定义界面的 App，通过访问 Dify 的 API 赋予应用 AI 能力。

为了保护敏感数据，Dify 支持私有化部署，是确保数据安全的最佳选择。

### ****Dify、Coze、FastGPT对比****

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Dify** | **FastGPT** | **Coze** |
| **简介** | Dify是一个开源的平台，支持本地私有部署 | FastGPT是一个开源的平台， 提供开箱即用的数据处理、 模型调用等能力 | 字节跳动旗下的产品，Coze提供了丰富的插件系统和易用性 |
| **开源** | Dify的开源社区活跃，提供了丰富的资源 | 开源 | 不开源 |
| **定制性** | 用户可以根据自己需求进行高度定制， 并且可以在自己的服务器上进行部署， 从而更好的控制数据安全和隐私 | 可以通过Flow可视化进行工作流编排，从而实现复杂问答场景 | 相比Dify，定制性有限制 |
| **大模型支持** | 支持多种大语言模型， 如OpenAI、Claude3等， 也支持开源的本地模型 | 通过One API支持常见的大模型 | 支持少数的在线模型 |
| **便捷性** | 操作便捷 更加容易上手 可以通过零代码和可视化方式快速构建 AI 应用 | 操作便捷 | 操作便捷 |
| **功能支持** | 支持RAG 支持Agent，但不支持多Agent 支持工作流 | 擅长支持RAG场景 支持工作流 | RAG场景支持较弱 Agent功能丰富 支持工作流 |
| **面向用户** | 企业用户 程序员 | 企业用户 程序员 | C端用户 |

### 使用Dify平台对比直接使用 ChatGPT

|  |  |  |
| --- | --- | --- |
| **角度** | **Dify** | **OpenAI 直接使用 ChatGPT** |
| **数据隐私** | Dify 支持本地私有化部署，支持调用本地训练的大模型，但是如果调用OpenAI的APIkey同样有数据泄露风险，安全性上不如国产AI百川大模型 | 数据处理全部托管在 OpenAI 云端，可能带来隐私和合规风险，尤其是在涉及严格数据保护法规的地区，部分外企有控诉过OpenAI使用泄露数据进行训练 |
| **定制化** | 可以根据具体业务场景进行模型调优、参数调整和功能扩展（workflow），能够满足复杂、个性化的需求。 | 默认模型提供通用化服务，适合大多数标准场景。虽然开箱即用，但灵活性和定制化能力相对有限，无法满足特定需求场景的优化。 |
| **成本** | Dify官方价格：https://dify.ai/pricing， 或者采用开源方案（调用OpenAI的API或者本地大模型），长期来看可以降低成本，特别是对高频使用的企业。然而，初始部署需要一定的硬件投入和技术团队支持 | 按需付费模式，初期无硬件投入，适合快速上线和临时使用的场景。然而，如果长期使用或调用频率较高，可能会导致成本累积增加。 |
| **便捷性** | 目前工作台同样依赖于网络环境，需要进项简单配置 | 无需部署，直接通过 API 或者注册即可使用，初期投入少且易于上手，适合技术能力较弱的团队或需要快速实现产品功能的企业。但其依赖于网络环境，可能对使用场景有一定限制。 |
| **性能** | 取决于调用的模型性能 | ChatGPT-4 是 OpenAI 开发的第四代大语言模型，在自然语言处理领域处于领先地位 |

### 相关关键词解释

#### 什么是RAG(检索增强生成)

**RAG**（Retrieval-Augmented Generation）是一种将信息检索与生成式模型结合的技术，用于提升生成模型的准确性和实用性。

**工作原理**：

检索（Retrieval）：

在生成内容之前，模型会从外部知识库（如文档库、数据库、网络资源等）检索相关信息，确保回答基于最新或特定领域的信息。

生成（Generation）：

利用检索到的上下文信息，生成式模型（如 GPT）根据用户的提问生成内容，保证回答更贴近真实或高质量内容。

**应用场景**：

知识问答：结合企业内知识库，为员工或客户提供精准答案。

文档生成：根据特定上下文，生成报告、摘要或策略文档。

实时更新：对依赖实时信息的应用（如新闻摘要）特别有效。

**优势**：

提高模型生成内容的准确性和可信性。

解决语言模型的“幻觉问题”（hallucination），减少无依据的回答。

#### 什么是智能体(Agent)

**智能体**（Agent）是指一种自主运行的系统或实体，能够感知环境并根据环境变化采取相应的行动来完成目标任务。智能体可以是软件系统、硬件设备或虚拟助手。

**应用场景**：

虚拟助手：如 Siri、Alexa 等，通过自然语言处理帮助用户完成任务。

机器人：工业生产中的自动化机器人，或智能家居中的清扫机器人。

自动化系统：企业流程自动化（RPA）或自动驾驶技术。

#### 什么是工作流(WorkFlow)

**工作流**（Workflow）是指一系列任务或操作的顺序化执行，用于实现某个具体目标或解决特定问题。工作流通常以规则和逻辑为基础，明确定义了任务的执行顺序、参与者及各步骤之间的关系。

#### 什么是大模型

大型语言模型，也称大语言模型、大模型(Large Language Model，LLM；Large Language Models，LLMs)。

大语言模型是一种**深度学习模型**，特别是属于自然语言处理(NLP)的领域，一般是指包含数干亿(或更多)参数的语言模型，这些参数是在大量文本数据上训练的，例如模型GPT-3，PaLM，LLaMA等，大语言模型的目的是理解和生成自然语言，通过学习大量的文本数据来预测下一个词或生成与给定文本相关的内容。

OpenAI 的 GPT (Generative Pre-trained Transformer) 系列是大语言模型的典型代表，作为目前为止公认最强的 GPT-4 架构，它已经被训练在数十亿的单词上。从实际应用表现来看，大语言模型具备回答各种问题、编写文章、编程、翻译等能力，如果深究其原理，LLM建立在Transformers架构之上，并在很大程度上扩展了模型的大小、预训练数据和总计算量。

#### 什么是ChatGPT

GPT，全称"Generative Pre-training Transformer"，是一个由OpenAI开发的自然语言处理(NLP)的模型。它的主要目标是理解和生成人类的自然语言。通过对大规模文本数据进行预训练，GPT模型能学习到语言的各种模式，如语法、句法、一词多义等，以及一些基础的世界知识。

总的来说，它通过预训练和生成技术，以及Transformer的自注意力机制，来理解和生成人类的自然语言。

#### GPT与大模型的关系

GPT-3、GPT-4、GPT-4o 是具体的大模型的例子。当提到“大模型”时，通常是指具有大量参数的机器学习模型。

#### ChatGPT、GPT与大模型三者的关系

ChatGPT是一种特定的GPT应用，GPT是一种大模型，而大模型是一类具有大量参数的深度学习模型。

#### 开源大模型举例

\* **LLaMa**：LLaMa是Meta AI公司开源的一组大规模语言模型，参数范围从7B到65B。它们在多达14,000亿tokens的语料上进行了训练。其中，LLaMA-13B在大部分基准测评上超过了GPT3(175B)，6B可在个人GPU上使用，13B时性能相当于GPT3(175B)，训练复杂度高于ChatGLM。

\* **ChatGLM**:ChatGLM-6B是清华大学知识工程和数据挖掘小组发布的一个开源的、支持中英双语的对话语言模型，基于General Language Model (GLM)架构，具有62亿参数。根据智谱AI公司的GLM-130B修改而来，结合模型量化技术，可以在消费级的显卡上进行本地部署(INT4量化级别下最低只需6GB显存)。

\* **Alpaca**：斯坦福大学根据LLAMA7B模型训练得到，性能接近GPT3.5，测试中发现中文支持较差。

\* **MOSS**：复旦大学团队开发，是一个支持中英双语和多种插件的开源对话语言模型，moss-moon系列模型具有160亿参数，在FP16精度下可在单张A100/A800或两张3090显卡运行，在INT4/8精度下可在单张3090显卡运行。MOSS基座语言模型在约七千亿中英文以及代码单词上预训练得到，后续经过对话指令微调、插件增强学习和人类偏好训练具备多轮对话能力及使用多种插件的能力，同样，经测试对中文支持一般。

\* **miniGPT4**：沙特国安大学的华人团队，未公布参数，使用4块A100(80G显存)GPU训练得到，本地部署最低需要12G显存，具备多模态能力。即除了识别文字外，还可识别图像，可以根据图像输出文字。

\* **GPT4ALL**：基于LLAMA7B模型微调得到，训练数据采用了OpenAlGPT-3.5-Turbo模型创建的对话语料和其他语料，GPT-3.5-TurboAPI总花费500刀，全部语料均已开源，本地部署需要至少16G显存，缺乏中文训练语料，因此对中文支持较差。

#### OpenAI API

OpenAI API 就是 OpenAI 为第三方开发者提供的接口，开发者可以将 OpenAI API 集成到自己的产品中，实现用 AI 生成文字的效果。在产品中不需要训练自己的模型，只需把用户输入的句子发送到 OpenAI 的 API，API 就会返回按照自定义指令润色之后的句子，在网站上展示。

#### OpenAI付费方式的区别

OpenAI 旗下各产品简单的分为两种：

**\* ChatGPT Plus**：先付后用，20 美元/月，开通时绑定信用卡并付款，之后每月自动从信用卡扣费(也可以通过 App Store 开通，从 App Store 余额或绑定的支付方式中扣费)。

**\* OpenAI API**：先用后付，按量付费，绑定信用卡后，每月月初自动从信用卡按上个月用量扣费。

#### 其他

\* **OpenAI** 一个研究 AI 技术的机构

\* **GPT** (Generative Pre-trained Transformer)OpenAI 发布的一种人工智能语言模型

\* **GPT-3.5** GPT 模型的 3.5 版本

\* **GPT-4** GPT 模型的 4 版本，相比 GPT-3.5 模型拥有更强的推理能力

\* **ChatGPT** OpenAI 基于 GPT 模型发布的 AI 聊天产品

\* **ChatGPT Plus** ChatGPT 的付费版本，可以理解为 ChatGPT 的高级会员

\* **OpenAI API** OpenAI 向开发者提供的接口服务，方便开发者在自己的产品(App 或网站)中使用 OpenAI 的模型实现 AI 相关的功能

\* **DALL-E OpenAI** 发布的一种文本转图像模型

\* **Whisper** OpenAI 发布的一种语音识别模型