



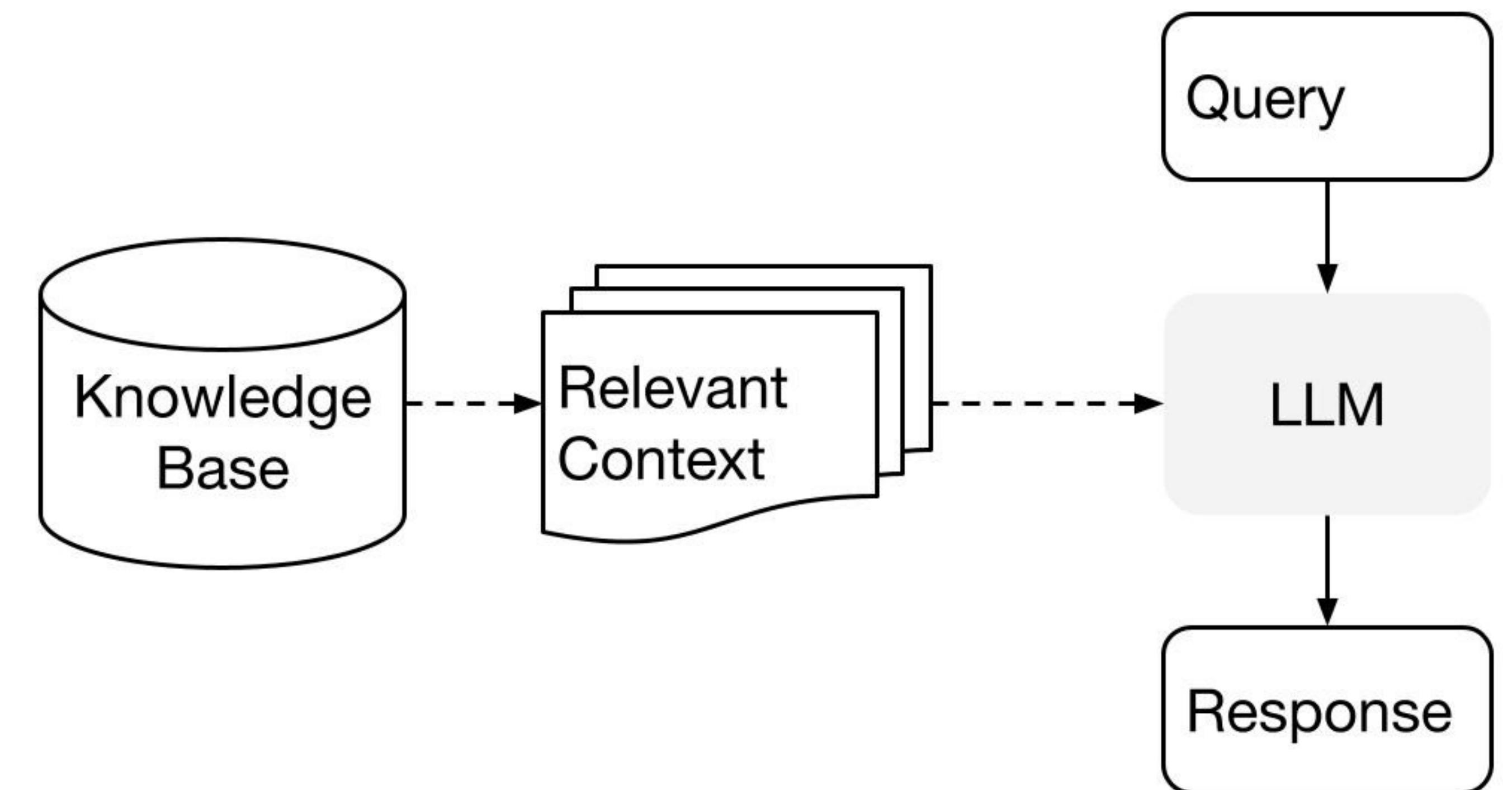
# 大模型结合 RAG 构建客服场景自动问答

齐家兴, NVIDIA 资深解决方案架构师



# RAG 基本概念

- 检索增强生成 (RAG) 从外部知识库检索事实，结合大语言模型针对用户的问题作出回答。它确保模型能够访问最新、可靠的事实，并且用户能够访问检索的来源，确保可以检查其声明的准确性。
- 开源工具包 *LangChain*, *LlamaIndex* 等提供了构建 RAG 的众多组件:
  - 文件读取 (word, pdf, md, html, ...)
  - 索引 (vectorDB, keyword, knowledge graph, ....)
  - 存储 (vector store, KV store, graph store, ....)
  - 大模型 (ChatGPT, LLaMa, ....)
  - 评估
- RAG 是一种范式，需要根据场景定制达到最好效果。



# 客服场景自动问答

## 背景介绍

- 已有的自动回复系统

- 现有的客服自动回复系统由两个部分组成：自动回复和人工客服
- 当用户提问时，系统使用关键词匹配的方法从预先准备好的问答对中寻找相关答案。如果匹配失败，系统返回若干相似的问题让用户进行选择。如果用户不满意，可以呼叫人工客服对话。
- 关键词匹配的方法虽然简单但经常失败，因为即便一个字不同也无法匹配，例如登陆 vs. 登录

- 自动问答系统的目标

- 借助大模型更精准的回答用户问题，提升服务效率，降低人工成本。

用户：登陆

客服：您是不是要咨询以下问题：

- 1, 无法登录
- 2, 手机端如何登录
- 3, 账号被盗
- 4, 从哪里下载 PC 端程序
- 5, 找回登录名

用户：找回登录名

客服：你可以参考[网页链接](#)进行找回。

使用关键词匹配进行自动回复



# 客服场景自动问答

## 数据情况

- 我们在这个场景中可以使用的数据：
- 客服部门的内部的规则介绍和常见问题， 作为知识库用来回答用户问题。文本格式与内容无规则。
- 一千条左右的用户与人工客服之间的真实对话记录。
- 400 条用户真实问题作为测试集评估效果。

访客	你好，我上次的问题解决了么？
客服	您好，请稍等
客服	让您久等了，关于您的问题这边已经反馈给工作人员进行核实了，具体结果会在3个工作日内通过@客服 私信回复您的，烦请您耐心等待下的
访客	谢谢
客服	您客气了

真实用户与客服对话记录片段

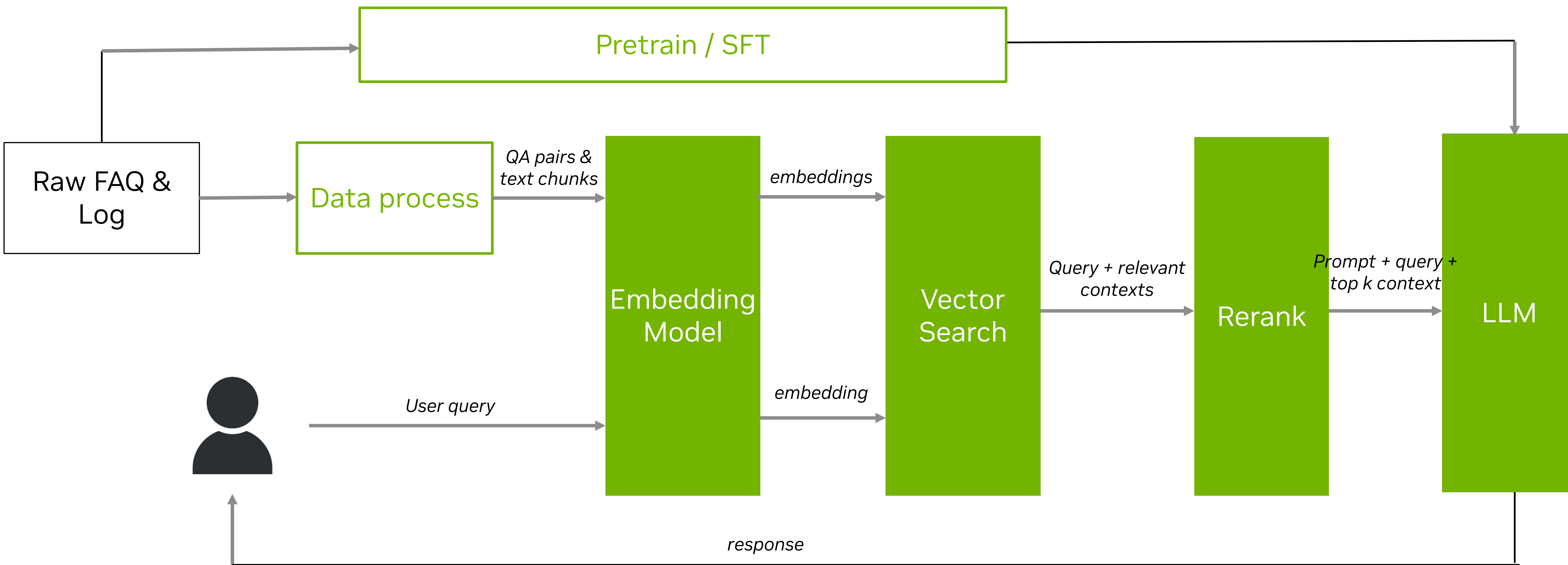
如您开启登录保护验证，请点击查看登录提示需要【登录保护验证】解决办法 为了确保您的账号安全，使用密码登录时，需完成短信或私信等方式验证，验证通过即可登录。验证时您可以选择接收短信验证码验证、私信验证、扫码验证或选择其他方式验证。 1、如验证时页面出现异常，建议您清除浏览器缓存或换个浏览器再进行尝试，如提示频繁，建议休息一段时间再进行尝试。 \*注：如选择私信验证方式，建议关注安全中心，以免出现收不到私信验证码的情况 2、如验证时提示“暂不支持该手机号验证，请更换后重试”或“系统错误，请稍后再试”，建议您使用曾经验证过此账号的手机号进行验证。

您好，登录提示【账号长期未使用，已处于保护状态】是由于账号长期未使用，存在较大安全隐患。依据《服务使用协议》，系统自动处置为保护状态所致。建议您在登录页面输入账号密码后，根据页面提示进行账号激活，或重新注册新账号使用。 操作流程如下： 一、希望继续使用该账号 1、点击【激活账号】进行账号激活，选择【希望使用该账号，申请激活】进入验证页面； 2、您可根据页面提示： ①点击【获取短信验证码】，使用手机获取验证码方式进行验证后即可登录账号； ②点击【使用其他方式验证】进行扫脸验证，通过验证后即可登录账号。如扫脸时提示“今日扫脸机会已用完，请明天再试”，建议您第二天零点之后再尝试扫脸。 ③如无法通过绑定手机或扫脸验证，可通过申请激活方式进行自助反馈，相关工作人员核实后会尽快做出回复，反馈结果您可在客服中心--服务记录中查询。 二、不想再使用原账号 如不想再继续使用原账号，可点击【不再使用该账号，注册新账号】重新注册账号即可。 如该内容对您有所帮助，烦请点击页面左下角的【有用】，期待您的反馈~

客服内部的规则介绍和常见问题片段

# 客服场景自动问答

## 系统流程





# 数据处理

## 文档切分与问答对提取

- 原始知识库文本很长，结合自动与人工方法，按语义切分成段落并提取问答对，以达到更精准检索。
- 将问答对中的问题转换为向量用于检索，进一步提升检索精度。

1.我为什么看不到XX功能？ 答：可以检查下您的app版本是否升级到了YYY或以上版本了。 2.XXX功能是干什么用的？ 答：XXX属于推送通知的功能之一，您可以通过该功能设置是否接收XXX内容推送及推送内容类型。 3.我设置推送管理/订阅，为什么会设置失败，提示达到上限？ 答：目前每个人最多可设置订阅，您可以先取消订阅其他，再订阅新的。后续我们将会根据用户反馈，扩充订阅数量上限。 4.怎么能看到我设置推送管理/订阅，集中管理这些设置？ 答：您可以通过“我”-“设置”-“推送通知设置”-“更新订阅”进入到订阅列表，在该列表内查看或管理订阅设置； 5.怎么能取消推送管理/订阅设置？ 答：您可以通过“我”-“设置”-“推送通知设置”-“更新订阅”进入到订阅列表，点击右上角“管理”按钮，进行清除订阅设置。 6.为什么我原来默认设置推送管理/订阅了某个，现在变为未设置/订阅了？ 答：结合用户反馈，该功能现已改为必须要主动设置“推送管理”才可以订阅。如果您还想订阅，辛苦再手动操作一次哦。 7.设置推送管理/订阅和特别关注有什么区别，都能及时收到推送通知吗？ 答：两种方法都可以及时收到推送通知，但设置订阅更灵活，不仅可以限制内容类型，而且可以不关注就能收到推送通知。另外订阅不限制每天的接收条数，特别关注则取决于“我”-“设置”-“推送通知设置”-“特别关注”开关设置，选择“实时通知”才会不限制每天的接收条数哦。

知识库文本样例

规则匹配

利用 LLM 抽取

人工处理

Question: 我为什么看不到XXX功能?  
Answer: 可以检查下您的app版本是否升级到了YYY或以上版本了。

Question: 管理/订阅XXX功能是干什么用的  
Answer: 订阅属于推送通知的功能之一，您可以通过该功能设置是否接收内容推送及推送内容类型。

Question: 我设置推送管理/订阅XXX，为什么会设置失败，提示达到上限？  
Answer: 目前每个人最多可设置订阅数量，您可以先取消订阅其他，再订阅新的。后续我们将会根据用户反馈，扩充订阅数量上限。

Question: 怎么能看到我设置推送管理/订阅，集中管理这些设置？  
Answer: 您可以通过“我”-“设置”-“推送通知设置”-“更新订阅”进入到订阅列表，在该列表内查看或管理订阅设置

...

提取后的问答对用于检索

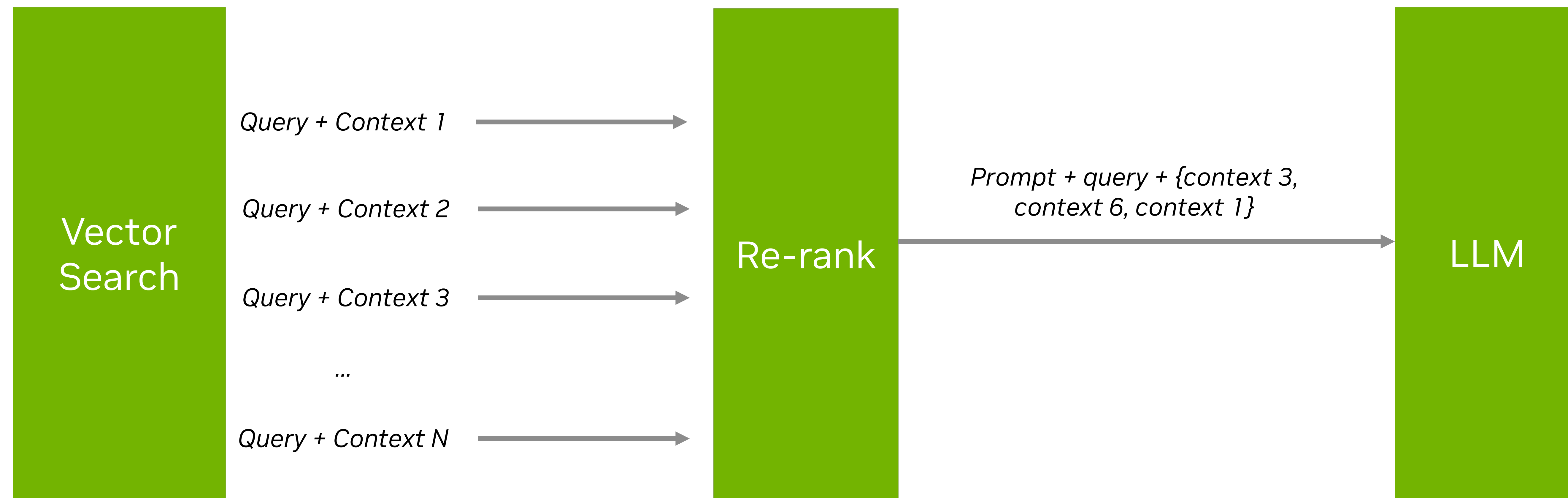
# 继续预训练 与 SFT

- 为了让 LLM 更加适应目标客服场景，我们结合继续预训练和 SFT 微调，进一步提升 LLM 回答精度。

	目的	数据
继续预训练	训练 LLM 适应客服领域的文本内容	知识库文本, 切分成 2k - 4k 的文本块。
SFT	微调 LLM 适应 RAG 根据上下文回答问题的模式	创建类似 RAG 形式的指令数据 * 指令: 用户问题 + 相关知识 * 回复: 借助 ChatGPT 生成的答案

# 重排序 (Rerank)

- 大模型的回复准确率很大程度取决于召回文本与用户问题的相关度。
- 向量检索具有很高的效率，但是精度还有进一步提高的空间，因为每一条向量的计算过程是独立的。
- 我们增加了重排序 rerank 的模型，更准确的计算问题与文本之间的相关度。
- Rerank 模型使用 BERT 结构，输入是两段文本，输出是相似度。





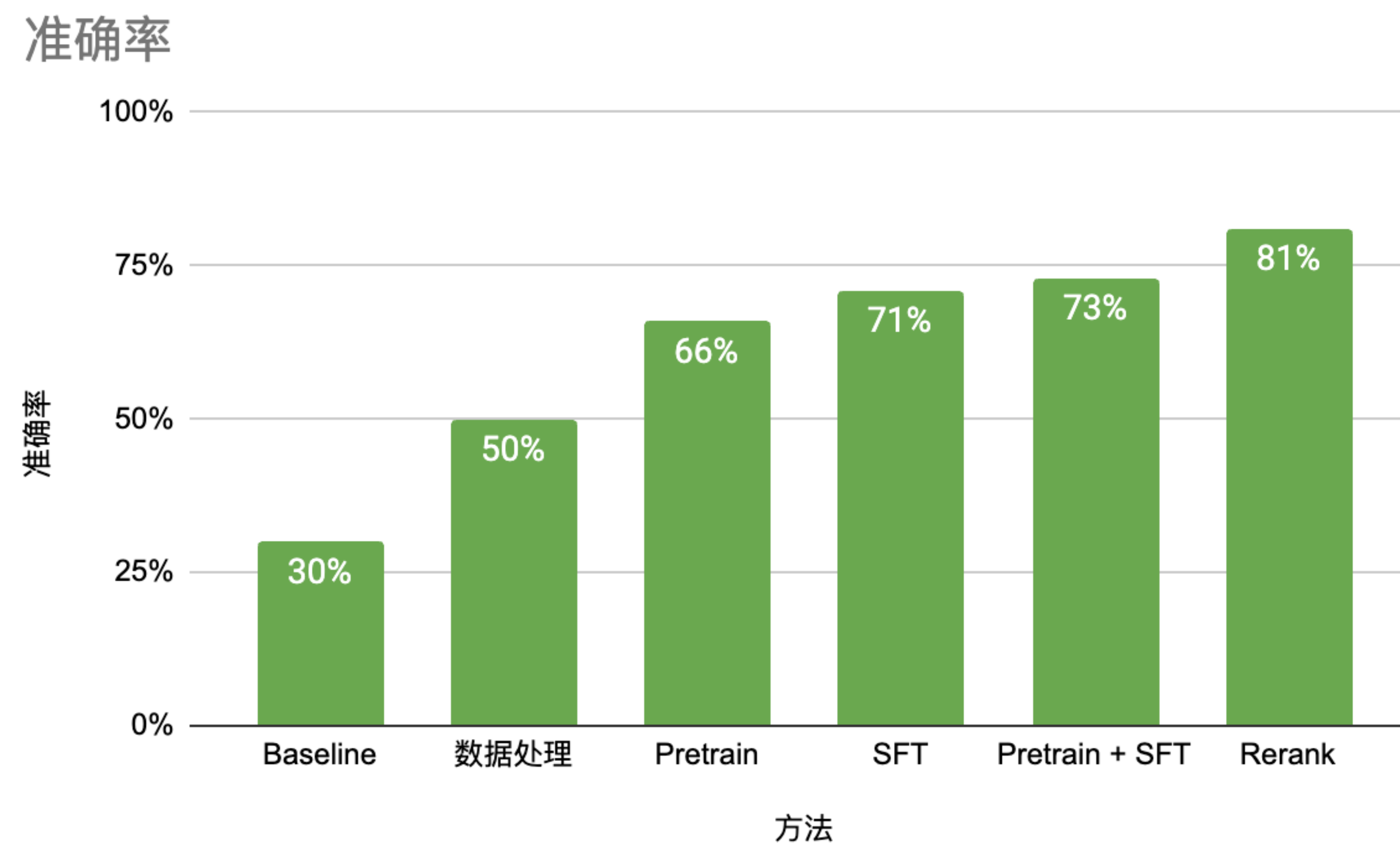
# 其他细节

- 提示工程 (Prompt Engineering)
  - 清晰的区分 context & query
  - 在 prompt 中针对场景描述额外的需求
  - 模板: 你作为 XXX 的客服, 解答用户使用 XX 时遇到的问题。以下是几条可能与用户问题相关的常见问题与答案, 你只能根据提供的材料来回答用户的问题。如果提供的材料无法回答用户问题, 请你回答: "抱歉, 无法回答您的问题! "  
以下是可能与用户问题相关的常见问题与答案:  
""  
  
<CONTEXT>  
""  
  
用户问题: <QUERY>
- 模型和超参数设置
  - 根据测试集和特定指标选择 LLM 和其他参数等。

# 效果提升

结合多种方法，提升 RAG 精度

- 在开发过程中，我们采用人工评估的方法，测评每个方法在测试集上所带来的精度提升。





# 结果对比

## 优化后精度提升

- 经过精确文档切分，继续预训练与 SFT 微调，添加 rerank 模型等优化方法，LLM 回答用户的精度有了显著提升。

问题	优化之前	优化之后
忘记密码怎么注销账号	请尝试通过以下步骤找回您的账户密码：1. 在登录界面点击“忘记密码”链接。2. 根据系统提示输入注册时使用的邮箱或手机号码，验证通过后重置密码。3. 重新登录您的账户，然后按照上述方法退出账号。	您可以尝试以下步骤找回或重置密码后再进行注销操作： 1. 在手机客户端【我】界面，点击点击右上角设置； 2. 在设置中向下滑动找到“账号与安全”选项并点击进入； 3. 点击“找回密码”按钮并根据页面提示进行操作以找回或重置您的密码。 如果您已经成功找回或重置了密码，请按照以下步骤进行注销操作： 1. 在手机客户端【我】界面，点击点击右上角设置； 2. 在设置中向下滑动找到“账号与安全”选项并点击进入； 3. 点击“注销账号”按钮并根据页面提示进行操作以完成注销流程。



# **NVIDIA 为 RAG 开发者准备的软件**



# 开箱即用的 RAG 示例

5 分钟搭建本地 RAG

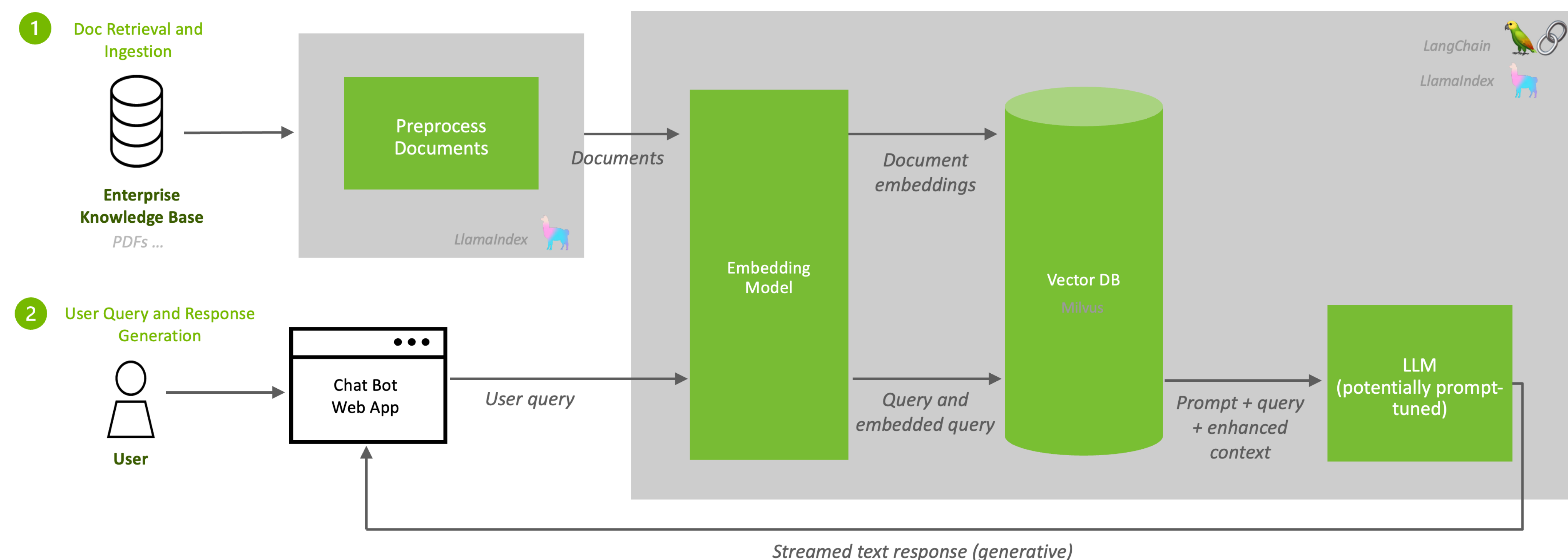
<https://github.com/NVIDIA/GenerativeAIEexamples/>

## • 特性

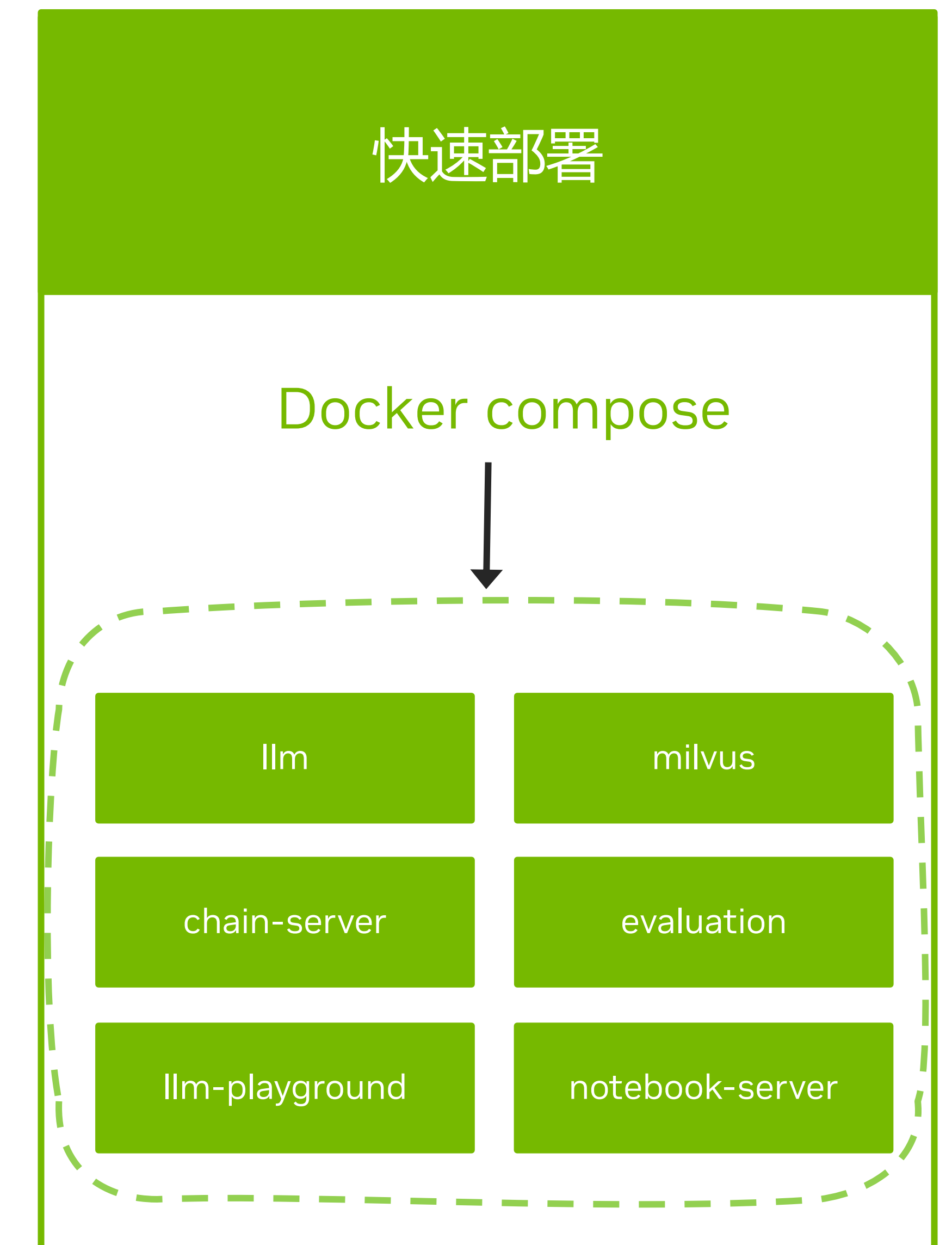
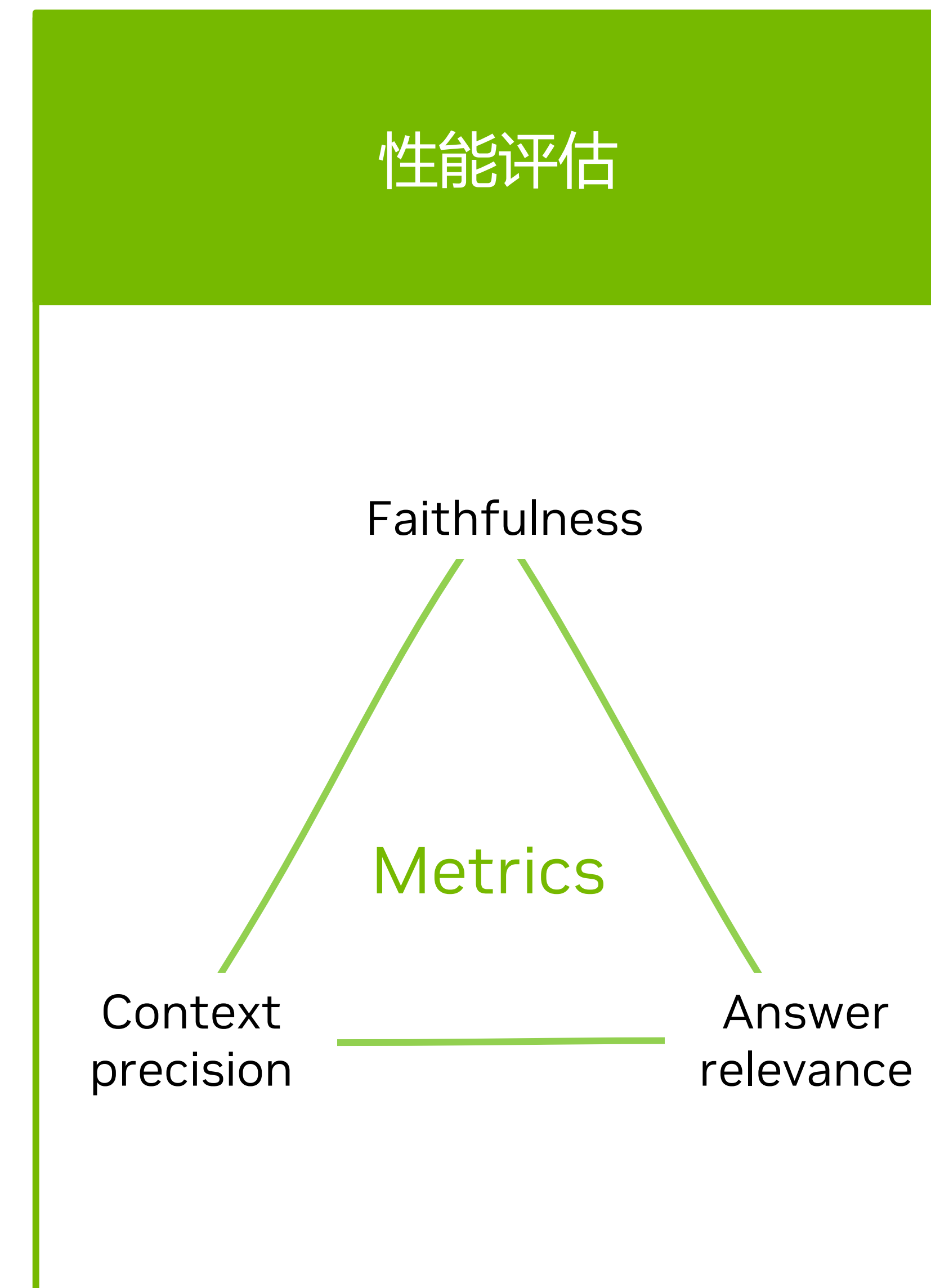
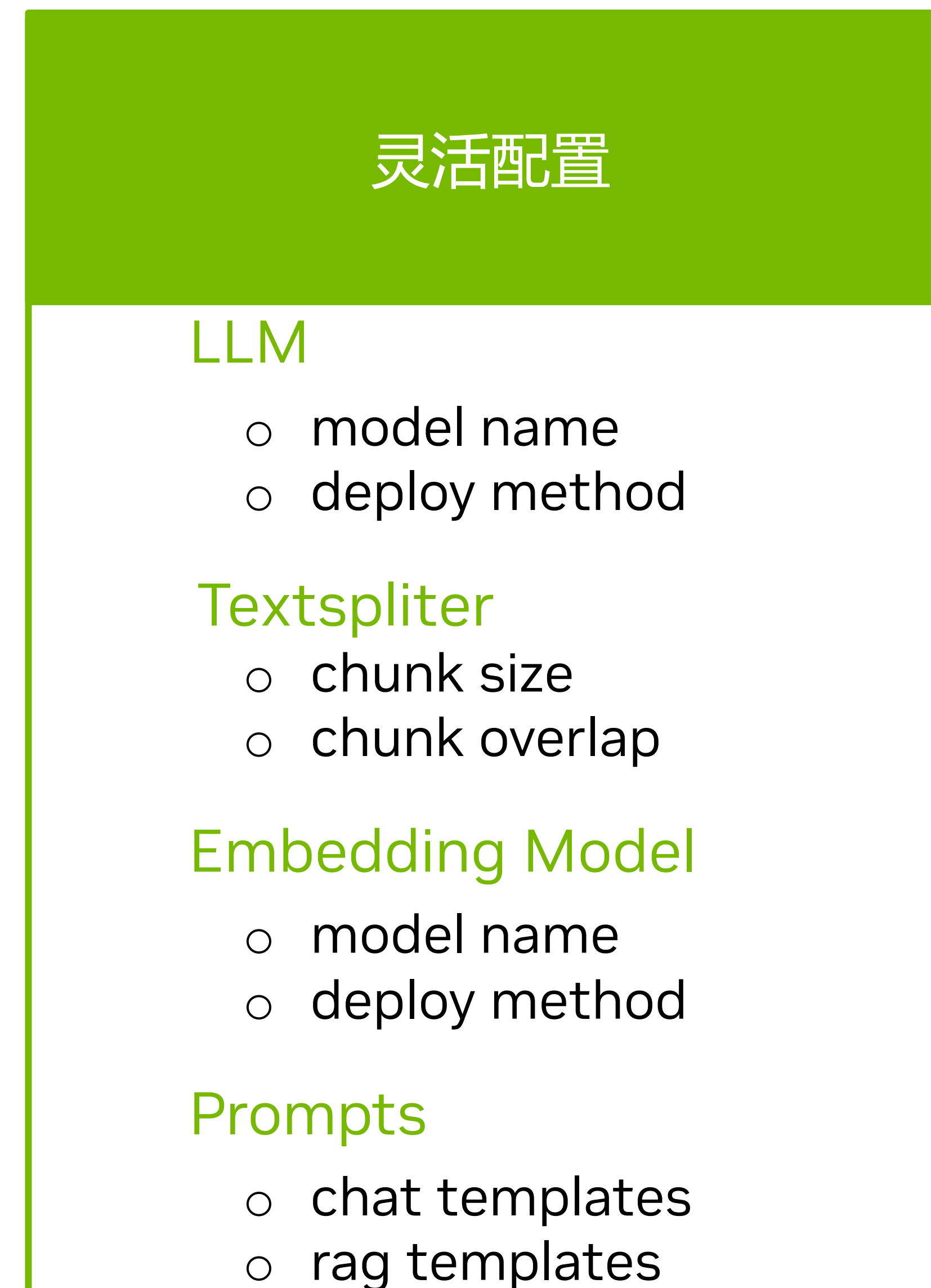
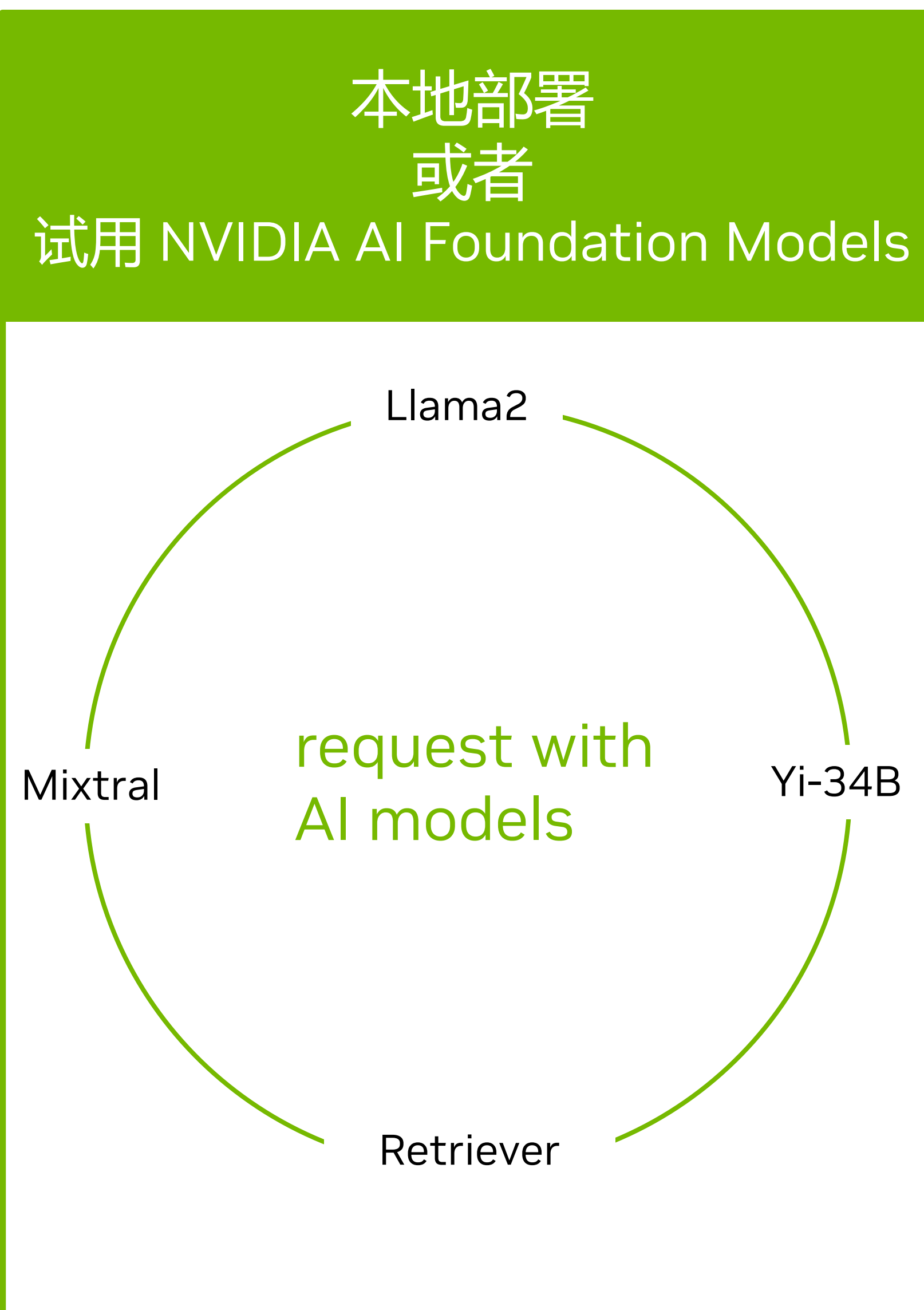
- 开源 RAG 示例，提供 notebook 方便学习
- 针对 NVIDIA GPU 全流程加速
- 借助 docker 快速部署

## • 组件

- **LLM**: 支持众多社区开源模型
- **LLM Backend**: 借助 NVIDIA TensorRT-LLM 推理优化，使用 Triton 推理服务器部署。
- **Vector DB**: Milvus-GPU
- **Embedding Model**: e5-large-v2



# RAG 示例特性






# RAG 示例 WebUI

## 使用步骤

- 上传知识库文件
- 选择基于知识库问答
- 与大模型对话

 NVIDIA LLM Playground

Converse Knowledge Base ⚙️

**Converse**

💬 Yi-34B

☐ Use knowledge base

Enter text and press ENTER

提交

清除

Clear history

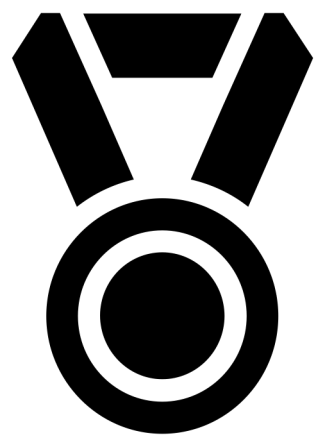
Show Context

# NVIDIA AI Foundation Models

<https://catalog.ngc.nvidia.com/>

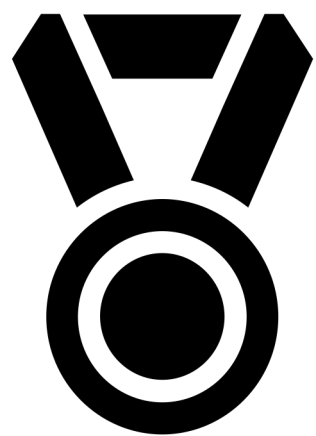


LLM



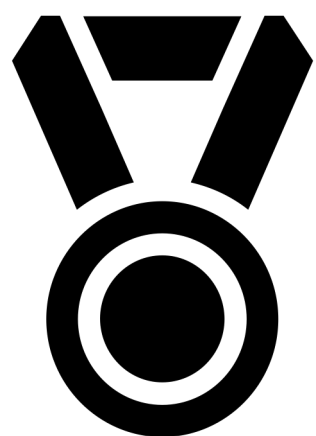
## 试用社区开源大模型

- 众多著名模型可供试用
  - Llama
  - Mixtral



## 统一的使用界面

- 通过 API 调用模型
- 在请求中指定模型名称选择指定模型



## 体验 NVIDIA GPU 的加速效果

- 使用 NVIDIA Triton 推理服务器部署
- 企业级服务保障 API 一致性



### Llama 2 70B

Text Generation

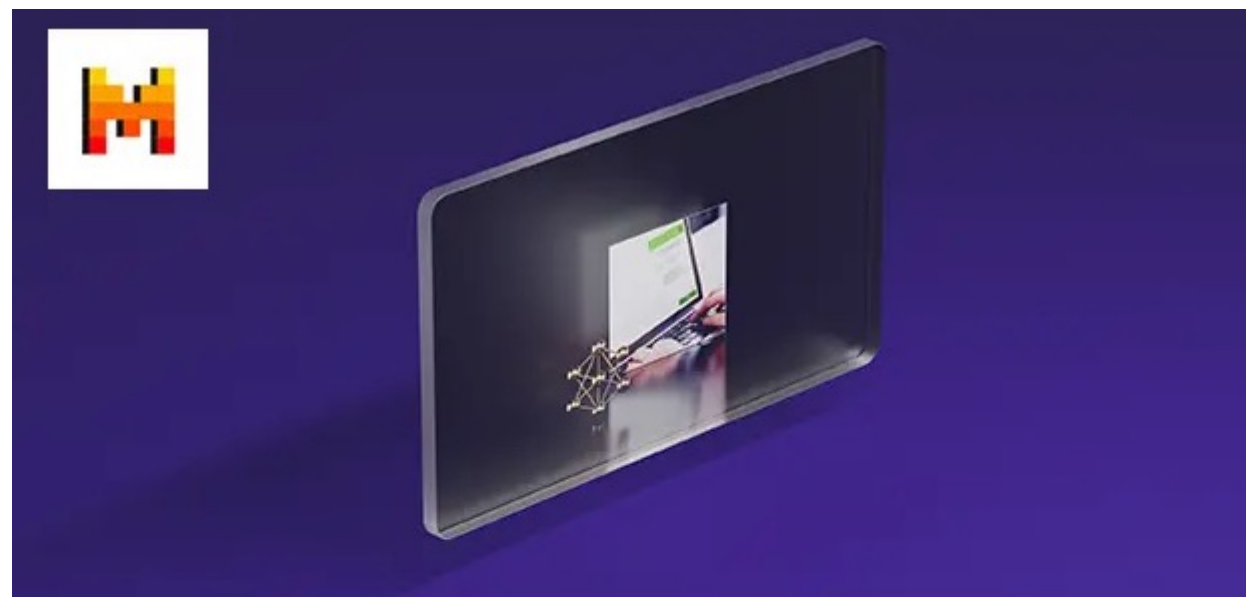
Llama 2 is a large language AI model capable of generating text and code in response to prompts.



### Llama 2 13B

Text Generation

Llama 2 is a large language AI model capable of generating text and code in response to prompts.



### Mixtral 8x7B Instruct

Text Generation

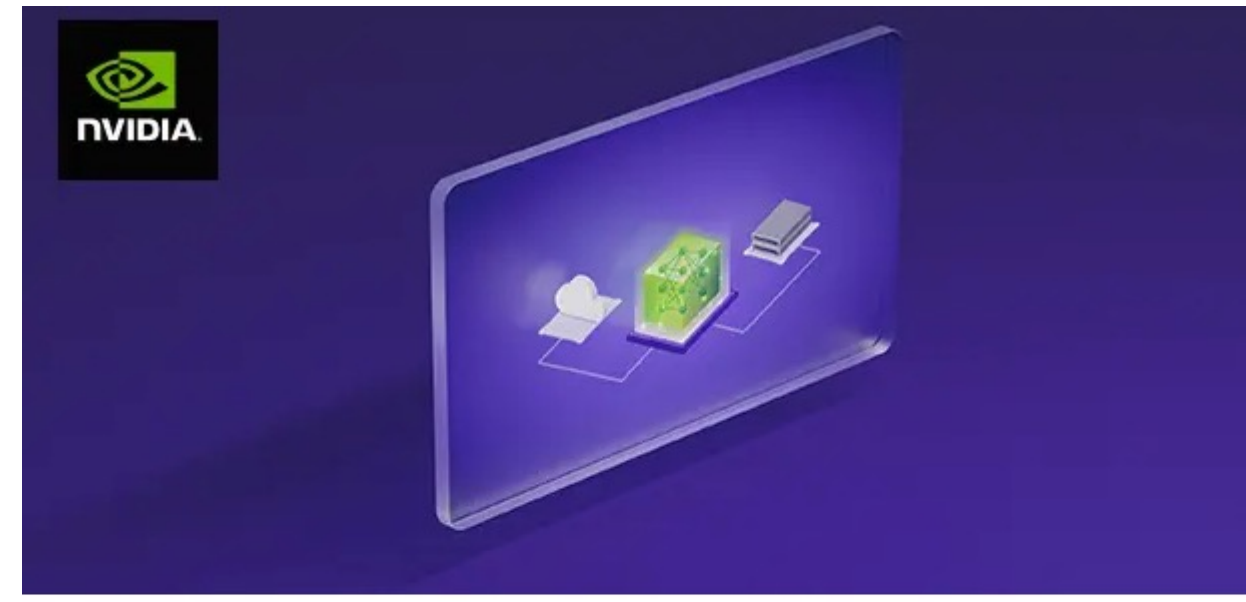
Mixtral 8x7B Instruct is a language model that can follow instructions, complete requests, and generate creative text...



### Yi-34B

Text and Code Generation

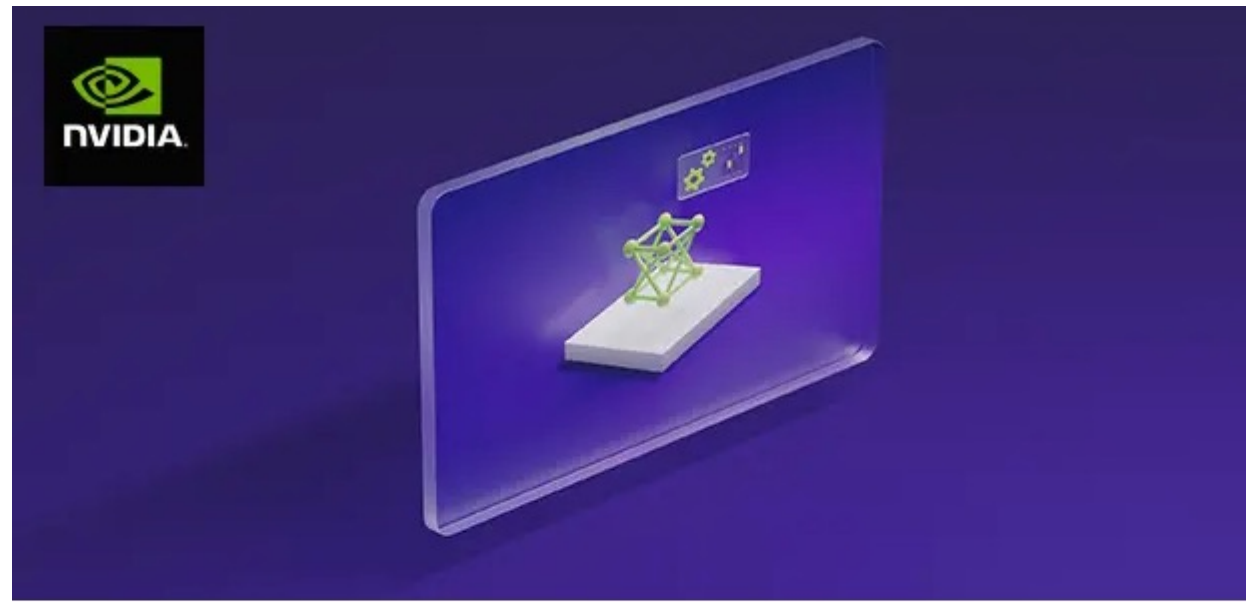
The Yi-34B is a large language model trained from scratch by developers at 01.AI. Yi-34B has been finetuned for...



### NV-Llama2-70B-RLHF

Text Generation

NV-Llama2-70B-RLHF-Chat is a 70 billion parameter generative language model instruct-tuned on LLama2-70B model. It...



### Nemotron-3-8B-QA

Text Generation

Nemotron-3-8B-QA is a 8 billion parameter generative language model based on the Nemotron-3-8B base model. The model...



# 总结

- 针对客服自动问答场景的 RAG 算法流程。
- 多种优化手段提升问答精度，包括 数据清洗提升召回精度，预训练与 SFT 训练大模型，增加重排序模型等。
- 开发过程中，不断迭代测试来选定最合适的模型与参数。
- NVIDIA 提供了 TensorRT-LLM, Triton 推理服务器等软件能加速大模型的推理和部署。
- NVIDIA 提供了一个开源 RAG 示例，方便开发者快速部署，定制化开发。



# 您将获得 NVIDIA 深度学习培训中心（DLI）大语言模型课程 75 折优惠码

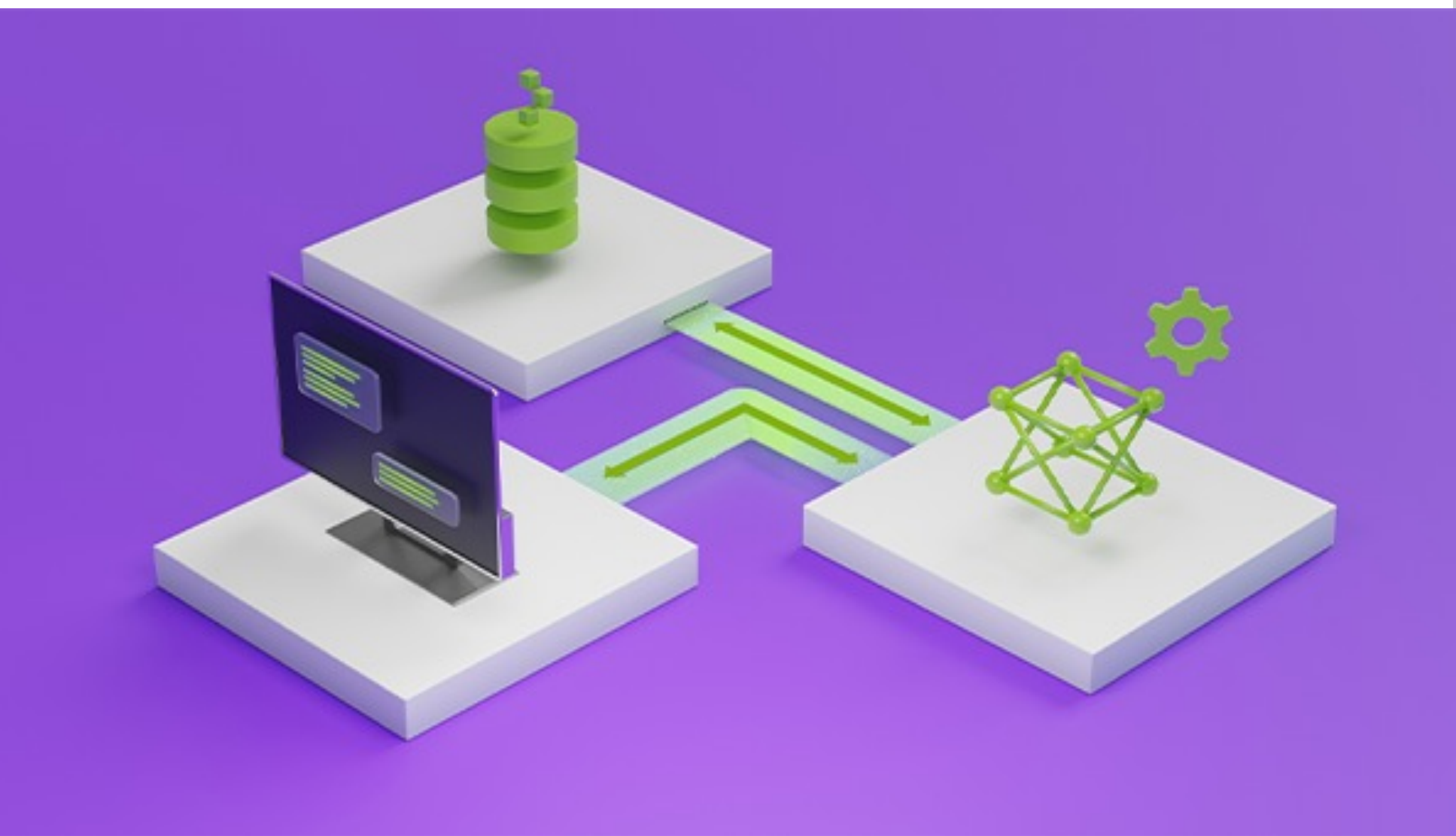
实战开发技能 | 实时讲师授课 | 实验用 GPU | 实名 NVIDIA 证书

- ✓ 听取 China AI Day 任一演讲，即可获取 DLI 75 折优惠码
- ✓ 会后您将收到优惠码专属邮件
- ✓ 任选右侧一门公开课使用优惠码（仅限一门）
- ✓ 开启 LLM 实战之旅

## 构建基于大语言模型（LLM）的应用

4 月 18 日

利用大模型开源生态系统  
快速开发基于预训练大语言模型的应用

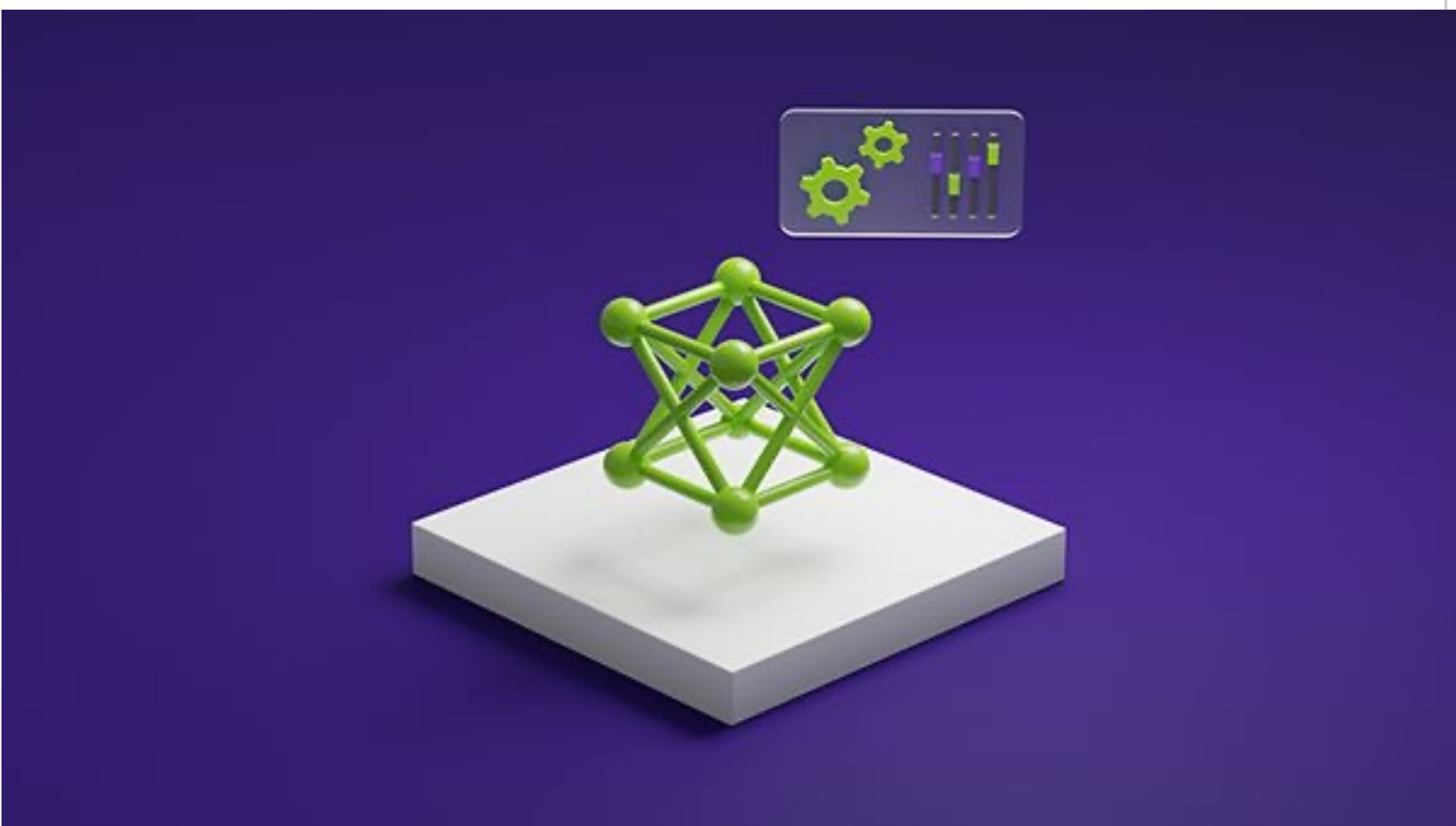


基础

## 高效定制大语言模型（LLM）

5 月 16 日

学习提示工程和各类高效参数微调方法  
对预训练 LLM 模型进行定制以适应特定应用



进阶



大语言模型（LLM）、生成式 AI 系列近期公开课：

4 月 11 日 [深度学习基础 —— 理论与实践入门](#)

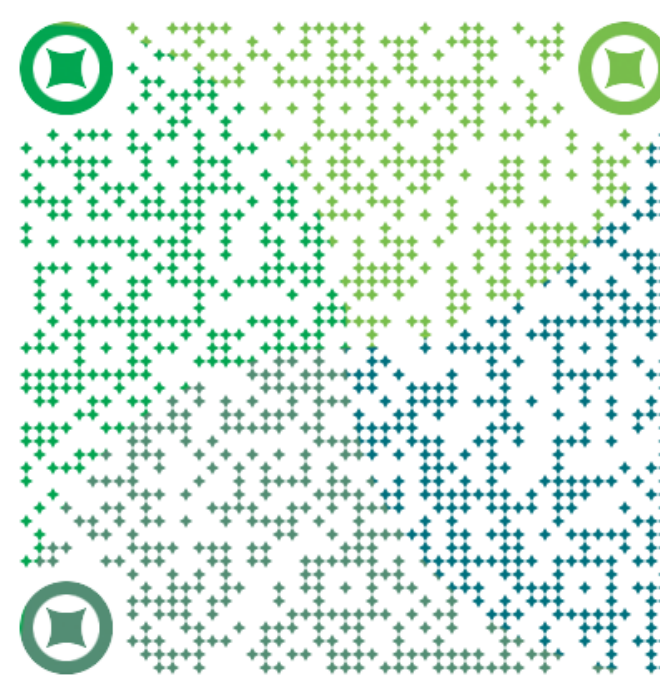
4 月 18 日 [构建基于大语言模型 \(LLM\) 的应用](#)

5 月 16 日 [高效定制大语言模型 \(LLM\)](#)

5 月 23 日 [构建基于扩散模型的生成式 AI 应用](#)

6 月 13 日 [构建基于 Transformer 的自然语言处理应用](#)

6 月 27 日 [模型并行 —— 构建和部署大型神经网络](#)



扫描二维码查看课程详情页

课程咨询，微信联系NVIDIALearn



