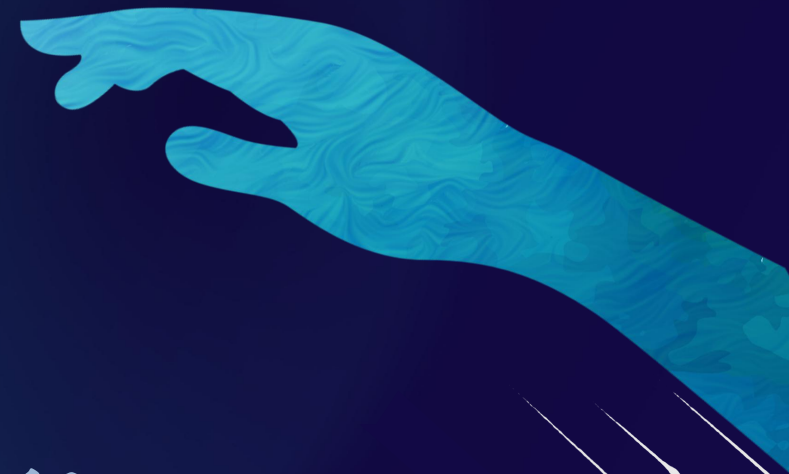
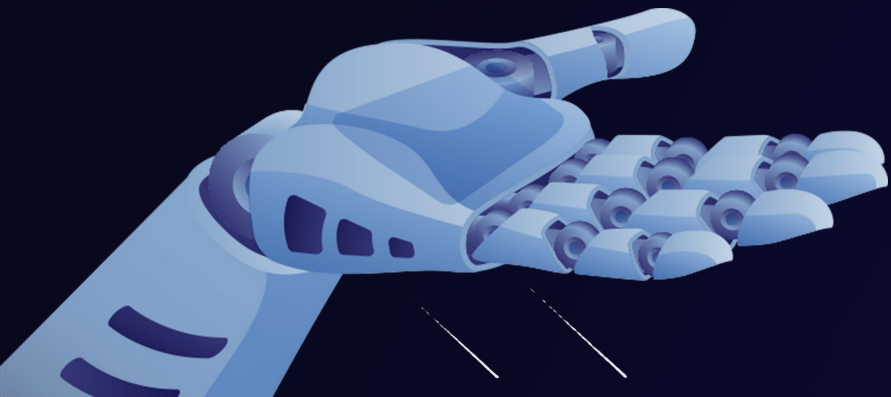


# 第一章：大模型提示工程





# CONTENTS

1.1 大模型简介及其学习必要性

1.2 提示运用的技巧

1.3 常见的提示方法：总结、推断、转换和扩展

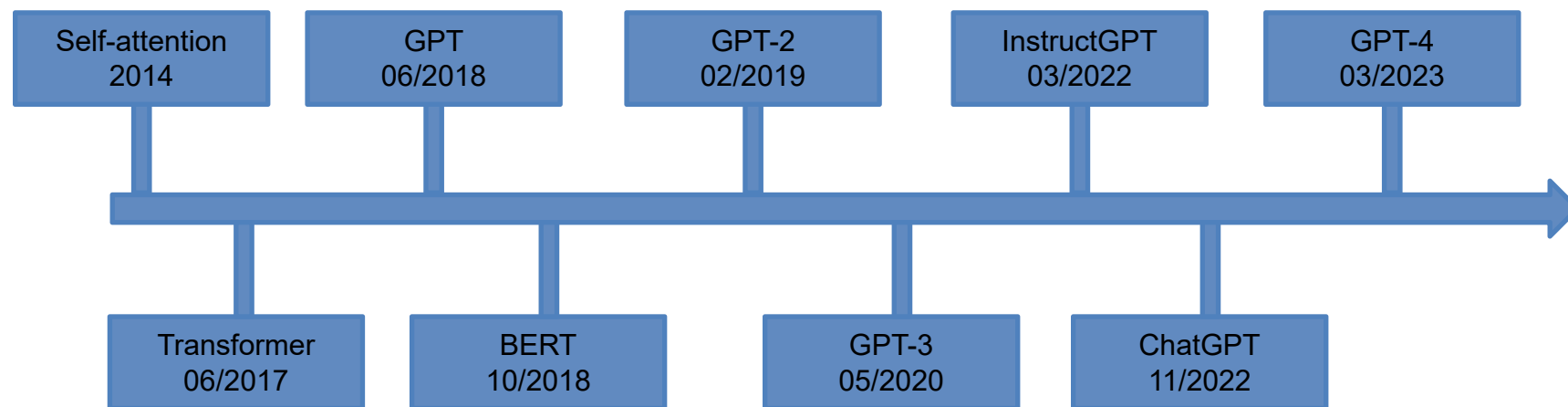
1.4 合规检查

1.5 思维链推理

1.6 提示链推理

1.7 基于大模型API构建客服机器人

## 1.1 大模型简介及其学习必要性



多模态: DALL-E

开源: Llama Qwen

专业模型: GitHub-Copilot Qwen-Math

2017年6月, Google的八位学者发明了Transformer算法, 从此为自然语言处理发展插上了腾飞的翅膀。以OpenAI为代表的机构沿用Transformer的核心路线将数据量不断增大, 将模型规模不断增大, 最终创造了接近于人类理解水平的ChatGPT。大模型的概念正是由于数据和规模巨大而来。

大模型的出现为各行各业提高了智能化水平, 比如接近于人类水平的客服机器人, 高度智能化的搜索引擎, 接近于人类水平的代码生成助手。甚至在其它领域也发挥了巨大的价值, 比如蛋白质结构预测, 该科研成果被授予了2024诺贝尔化学奖。

### Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com	Noam Shazeer* Google Brain noam@google.com	Niki Parmar* Google Research nikip@google.com	Jakob Uszkoreit* Google Research usz@google.com
--	--	---	---

Llion Jones* Google Research llion@google.com	Aidan N. Gomez* <sup>†</sup> University of Toronto aidan@cs.toronto.edu	Łukasz Kaiser* Google Brain lukaszkaier@google.com
---	---	--

Illia Polosukhin\*<sup>‡</sup>  
illia.polosukhin@gmail.com

## 1.1 大模型简介及其学习必要性

### 科研领域

**人工智能：**用于自然语言处理、图像和语音识别等领域。

**生物信息学：**分析基因组数据，预测蛋白质结构。

**药物发现：**加速新药发现与开发，提高药物疗效。

**材料科学：**设计新材料，预测材料性质，推动材料开发。

大模型算法工程师 (b... [北京·朝阳区·酒仙桥]

30-50K·16薪 5-10年 硕士 王女士 | HRBP

大模型算法 | 大模型训练

阿里淘天-大模型算法... [北京·朝阳区·望京]

35-65K·16薪 经验不限 本科 李女士 | 招聘经理

大模型评测 | 机器学习 深度学习 | NLP

### 工业领域

**智能客服：**提供实时客户支持，提高服务效率。

**推荐系统：**根据用户行为个性化推荐内容或产品。

**医疗问诊助手：**辅助医生诊断和建议治疗方案。

**智能管家：**管理家庭设备，提供个性化家庭服务。

算法工程师 [深圳·福田区·梅林]

20-30K 3-5年 本科 曾先生 | java高级工程师

算法 | 大模型

多模态内容理解算法... [深圳·南山区·科技园]

45-50K·16薪 3-5年 硕士 曾先生 | 工程师

多模态

# 1.1 大模型简介及其学习必要性

## 大模型应用程序开发

### 第一章 大模型提示工程

- 1.1 大模型简介及其学习必要性
- 1.2 提示运用的技巧
- 1.3 常见的提示方法：总结、推断、转换和扩展
- 1.4 合规检查
- 1.5 思维链推理
- 1.6 提示链推理
- 1.7 基于大模型API构建客服机器人

### 第二章 LangChain工具及应用

- 2.1 LangChain简介
- 2.2 模型、提示和解析器
- 2.3 记忆机制
- 2.4 模型链结构
- 2.5 代理机制
- 2.6 RAG: 访问本地文档
- 2.7 搭建基于本地知识库的问答系统

### 参考资料

- 1、Prompt Engineering for Developer. andrew ng etc.
- 2、Building Systems with the ChatGPT API. andrew ng etc.
- 3、LangChain for LLM Application Development. andrew ng etc.
- 4、LangChain Chat with Your Data. andrew ng etc.
- 5、自然语言处理：基于预训练模型的方法. 车万翔等.
- 6、Neural Network Methods for Natural Language Proce. Yoav Goldberg.
- 7、Attention Is All You Need. Ashish Vaswani etc.
- 8、BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Jacob Devlin etc.
- 9、Improving Language Understanding by Generative Pre-Training. Alec Radford etc.
- 10、Language Models are Unsupervised Multitask Learners. Alec Radford etc.
- 11、Language Models are Few-Shot Learners. Tom B. Brown etc..
- 12、Training language models to follow instructions with human feedback. Long Ouyang etc.
- 13、<https://huggingface.co/datasets/FImc/DISC-Med-SFT>. Fudan-DISC

## 大模型理论与进阶

### 第三章 自然语言处理综述

- 3.1 自然语言处理简介：概念、难点、任务和历史
- 3.2 机器学习一般过程
- 3.3 词的表示：独热表示和词嵌入表示
- 3.4 模型表示：MLP、RNN和LSTM三种神经网络结构
- 3.5 模型训练：损失函数、正则化和参数优化
- 3.6 模型评估：分类、解析、生成、回归四大问题评估方法

### 第四章 Transformer详解及翻译模型搭建

- 4.1 Transformer简介：基于自注意力机制的Seq2Seq模型
- 4.2 自注意力机制：充分考虑上下文
- 4.3 Transformer模型架构：输入、编码器、解码器和输出
- 4.4 Transformer训练过程
- 4.5 Transformer代码实操：了解大模型奠基性算法内核

### 第五章 BERT算法原理及实践案例

- 5.1 BERT简介：基于Transformer编码器的预训练模型
- 5.2 BERT预训练：掩码语言模型（MLM）和下一句预测（NSP）任务
- 5.3 BERT的微调：单句分类、句子对分类、问答、命名实体识别
- 5.4 典型微调任务实践

### 第六章 GPT模型系列理论介绍

- 6.1 GPT简介：基于Transformer解码器的预训练生成模型
- 6.2 GPT-1：预训练+微调
- 6.3 GPT-2：预训练+Zero Shot
- 6.4 GPT-3：预训练+Few Shot
- 6.5 InstructGPT：向人类对齐

### 第七章 大模型微调技术及医疗问诊机器人实战

- 7.1 大模型微调相关原理：定义、分类、作用和过程
- 7.2 模型量化和LORA技术
- 7.3 自监督模式微调
- 7.4 问答模式微调
- 7.5 多轮会话模式微调
- 7.6 医疗问诊机器人项目实战

## 1.2 提示运用的技巧

**什么是提示？** 提示即Prompt，是用户提供给模型的输入，大模型将根据Prompt返回结果（Completion）。对于大模型来说，一个合理的 Prompt 设计极大地决定了其能力的上限与下限。

**什么是提示工程？** 针对特定任务构造可以最大限度地发挥大模型能力的 Prompt 技巧。

### 提示的原则一：编写清晰、具体的指令

- 1) 使用分隔符清晰地表示输入的不同部分，例如：```, """, < >, <tag> </tag>, :
- 2) 寻求结构化的输出，比如 json 或 html。
- 3) 要求模型检查是否满足条件。

### 提示的原则二：给模型时间去思考

- 1) 指定完成任务所需的步骤。
- 2) 指导模型在下结论之前找出一个自己的解法。

### 提示的原则三：迭代优化

- 1) 首先编写一个初始的 Prompt。
- 2) 检查 Prompt 的输出是否满足要求。
- 3) 如果不满足，根据不足的地方改进 Prompt，并再次检查输出。
- 4) 重复这个过程，直到输出满足要求。



## 1.3 常见的提示方法-总结

**提示方法-总结：**大语言模型擅长将复杂的文本信息简化，提炼出关键的观点，可以助力用户快速获取精准信息，提升工作效率。

1609年9月11日，受聘于荷兰东印度公司的英国航海家亨利·哈德逊（Henry Hudson）乘船来到今纽约州探险，并将大量土产带回荷兰，许多荷兰商人对这片土地产生了浓厚兴趣，他们来到这里寻求与当地原住民进行贸易，并建立了殖民地“新荷兰”（New Netherland）。荷兰殖民者于1614年在新荷兰建立贸易点Fort Nassau（今奥尔巴尼市附近）；于1624年建立贸易点Fort Orange（今奥尔巴尼市）；于1625年建立贸易点Fort Amsterdam（今纽约市）。

总结

1609年，亨利·哈德逊探索今纽约州，荷兰商人对土地感兴趣，建立新荷兰殖民地，包括Fort Nassau、Fort Orange和Fort Amsterdam。

## 1.3 常见的提示方法-推断

**提示方法-推断：**大语言模型擅长基于文本信息推理，得到文本输入的类别，比如情感，或者从输入文本中进行实体识别。

1609年9月11日，受聘于荷兰东印度公司的英国航海家亨利·哈德逊（Henry Hudson）乘船来到今纽约州探险，并将大量土产带回荷兰，许多荷兰商人对这片土地产生了浓厚兴趣，他们来到这里寻求与当地原住民进行贸易，并建立了殖民地“新荷兰”（New Netherland）。荷兰殖民者于1614年在新荷兰建立贸易点Fort Nassau（今奥尔巴尼市附近）；于1624年建立贸易点Fort Orange（今奥尔巴尼市）；于1625年建立贸易点Fort Amsterdam（今纽约市）。

情感判断

中性

识别年份

[1609, 1614, 1624, 1625]



## 1.3 常见的提示方法-转换

**提示方法-转换：**大语言模型擅长将文本进行不同类型的转换，比如语言翻译、风格转换、拼写纠正、语法调整、格式转换等不同的任务。

亨利·哈德逊1609年探险纽约州，带回土产，吸引荷兰商人建立新荷兰殖民地和多个贸易点。

翻译成英文

Henry Hudson explored New York in 1609, brought back goods, and attracted Dutch merchants to establish the New Netherland colony and multiple trading posts.

以海盗的风格重写

在1609年，勇敢的亨利·哈德逊，驾着他的船只，向神秘的纽约州进发。他带回了满载的战利品，引得荷兰商人们眼红，纷纷前来寻宝。他们在这片新大陆上建立了“新荷兰”殖民地和多个贸易点，正如海盗们抢占宝藏岛般，展开了一段充满冒险的征程！🚩

### 1.3 常见的提示方法-扩展

**提示方法-扩展：**大模型擅长基于已知的一段文本，进行扩展，也就是按照指令进行续写。

这个产品质量太差了，用了一周就坏了。客服态度也不好，完全不推荐购买，浪费钱。

回复  
评论

尊敬的客户，非常抱歉您对我们的产品和客服服务感到不满意。我们非常重视客户的意见和反馈，我们会立即对产品质量进行调查和改进，以确保类似问题不再发生。同时，我们会加强客服培训，提高服务质量，让每一位客户都能得到满意的解决方案。如果您愿意，我们可以为您提供退货或换货的服务，以补偿您的损失。再次感谢您的反馈，希望能够重新赢得您的信任和支持。祝您生活愉快！

## 1.4 合规检查

当前的大模型是由互联网海量的数据训练而来，其中不乏一些暴力、种族歧视等方面的内容，如果要将大模型投入生产应用，必然需要对其合规性进行审查，并禁止回答类似的不合规的问题。

虽然主流的大模型在对齐阶段已在合规性方面做了很多工作，但是通过一些特定的手段对用户的内容进行专门的审查也是很有必要。本小节我们学习OpenAI提供的合规审查api Moderation。

请教我如何杀掉一个人？

Moderation

```
{
  "flagged": true,
  "categories": {
    "sexual": false,
    "hate": false,
    "harassment": false,
    "self-harm": false,
    "sexual/minors": false,
    "hate/threatening": false,
    "violence/graphic": false,
    "self-harm/intent": false,
    "self-harm/instructions": false,
    "harassment/threatening": false,
    "violence": true
  },
  "category_scores": {
    "sexual": 8.19847991806455e-06,
    "hate": 0.00039588319486938417,
    "harassment": 0.006893481593579054,
    "self-harm": 0.00023407311527989805,
    "sexual/minors": 4.6891163947293535e-06,
    "hate/threatening": 2.1002895664423704e-05,
    "violence/graphic": 5.6011267588473856e-05,
    "self-harm/intent": 6.489454972324893e-05,
    "self-harm/instructions": 1.9671333575388417e-06,
    "harassment/threatening": 0.019682535901665688,
    "violence": 0.9210704565048218
  }
}
```

## 1.5 思维链推理

思维链推理是指大模型在回答问题前先进行一步步的思考，并最终给出答案，和人类的思维相似，如果不加思索就回答一个复杂的问题，与经过一步步推理深思熟虑而来的答案肯定相差甚远。

为了利用大模型的思维链推理能力，可以使用思维链提示。思维链提示是一种引导大模型进行逐步推理的 Prompt 设计技巧。它通过在 Prompt 中设置系统消息，要求语言模型在给出最终结论之前，先明确各个推理步骤。

### 系统消息

请按照以下步骤回答客户的提问。客户的提问将以{delimiter}分隔。

步骤 1:{delimiter} 首先确定用户是否正在询问有关特定产品或产品的问题。

步骤 2:{delimiter} 如果用户询问特定产品，请确认产品是否在以下列表中。所有可用产品：

.....

步骤 3:{delimiter} 如果消息中包含上述列表中的产品，请列出用户在消息中做出的任何假设，\ 例如手机X手机Y大，或者手机Z有2年保修期。

步骤 4:{delimiter} 如果用户做出了任何假设，请根据产品信息确定假设是否正确。

步骤 5:{delimiter} 如果用户有任何错误的假设，请先礼貌地纠正客户的错误假设（如果适用）。\ 只提及或引用可用产品列表中的产品，因为这是商店销售的唯一三款产品。以友好的口吻回答客户。

使用以下格式回答问题：

步骤 1: {delimiter} <步骤 1 的推理>

步骤 2: {delimiter} <步骤 2 的推理>

步骤 3: {delimiter} <步骤 3 的推理>

步骤 4: {delimiter} <步骤 4 的推理>

回复客户: {delimiter} <回复客户的内容>

请确保每个步骤上面的回答中中使用 {delimiter} 对步骤和步骤的推理进行分隔。

### 用户提问

请问华为手机比小米手机贵吗？



步骤 1: ====用户正在询问有关华为手机和小米手机价格的比较。

步骤 2: ====确认华为和小米手机都在产品列表中。

步骤 3: ====用户假设华为手机比小米手机贵。  
步骤 4: ====根据产品信息，华为MateX的价格为9999，小米Mi11 Ultra的价格为5999，因此小米手机更便宜。

回复客户: ==== 华为手机比小米手机贵，华为MateX的价格为9999，而小米Mi11 Ultra的价格为5999。

## 1.5 思维链推理

内心独白：很多时候用户并不需要看到大模型的思维过程，为了有更好的用户体验，可以将思维过程隐藏起来，仅展示最后的结果。

### 系统消息

请按照以下步骤回答客户的提问。客户的提问将以{delimiter}分隔。

步骤 1：确定用户是否正在询问有关特定产品或产品的问题。

步骤 2：如果用户询问特定产品，请确认产品是否在以下列表中。所有可用产品：

.....

步骤 3：如果消息中包含上述列表中的产品，请列出用户在消息中做出的任何假设，

例如手机X手机Y大，或者手机Z有2年保修期。

步骤 4：如果用户做出了任何假设，请根据产品信息确定假设是否正确。

步骤 5：如果用户有任何错误的假设，请先礼貌地纠正客户的错误假设（如果适用）。

只提及或引用可用产品列表中的产品，因为这是商店销售的唯一三款产品。以友好的口吻回答客户。

使用以下格式回答问题：

步骤 1：<步骤 1 的推理>

步骤 2：<步骤 2 的推理>

步骤 3：<步骤 3 的推理>

步骤 4：<步骤 4 的推理>

回复客户：<回复客户的内容>

请确保每个步骤上面的回答中中使用 {delimiter} 对步骤和步骤的推理进行分隔。

### 用户提问

请推荐几个合适得电视给我，谢谢！



步骤 1: ====用户正在询问有关特定产品的问题。

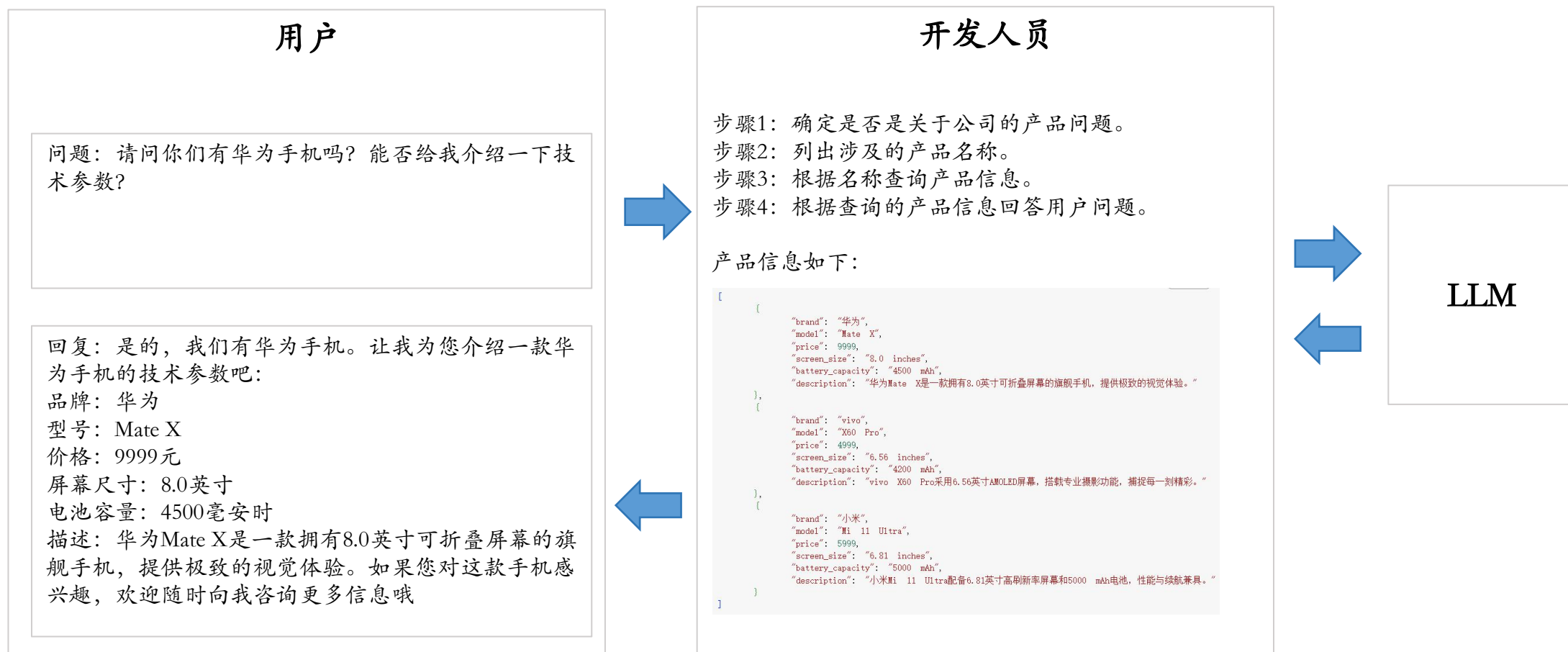
步骤 2: ====根据可用产品列表，我们只有手机产品的信息，没有电视产品的信息。

回复客户: ====非常抱歉，我们目前只提供手机产品的信息，没有电视产品的推荐。如果您需要任何关于手机的建议或信息，请随时告诉我们。感谢您的理解！

## 1.6 提示链推理

提示链推理是将复杂任务分解为多个简单Prompt的策略。与一次将一系列步骤融合到一个Prompt的思维链推理不同，它的复杂度更小，成本更低。

如果将思维链推理比喻成将一个任务的详细操作步骤告诉一个人来实施的话，提示链推理可以比喻成首先把任务的第一步告诉一个人来实施，完成之后根据情况在告诉他第二步，直到完成任务。





## 1.7 基于大模型API构建客服机器人

在各行各业，客户服务都是极其重要的部分，包括售前咨询和售后处理，得益于如今大模型强大的理解和对话能力，搭建性能卓越的客服机器人变得容易。本小节实现一个完整的客服机器人搭建过程。

### 用户

question: 请问你们有华为手机吗？能否给我介绍一下技术参数？

answer: 是的，我们有华为手机。让我为您介绍一款华为手机的技术参数吧：

品牌：华为

型号：Mate X

价格：9999元

屏幕尺寸：8.0英寸

电池容量：4500毫安时

描述：华为Mate X是一款拥有8.0英寸可折叠屏幕的旗舰手机，提供极致的视觉体验。如果您对这款手机感兴趣，欢迎随时向我咨询更多信息哦

### 开发人员

步骤1: 对输入进行合规检查

步骤2: 确定是否是关于公司的产品问题

步骤3: 列出涉及的产品名称

步骤4: 根据名称查询产品信息

步骤5: 根据查询的产品信息回答用户问题

步骤6: 对输出结果进行合规检查

步骤7: 对输出结果进行有效性检查

产品信息如下：

```
[
  {
    "brand": "华为",
    "model": "Mate X",
    "price": 9999,
    "screen_size": "8.0 inches",
    "battery_capacity": "4500 mAh",
    "description": "华为Mate X是一款拥有8.0英寸可折叠屏幕的旗舰手机，提供极致的视觉体验。"
  },
  {
    "brand": "vivo",
    "model": "X60 Pro",
    "price": 4999,
    "screen_size": "6.56 inches",
    "battery_capacity": "4200 mAh",
    "description": "vivo X60 Pro采用6.56英寸AMOLED屏幕，搭载专业摄影功能，捕捉每一刻精彩。"
  },
  {
    "brand": "小米",
    "model": "Mi 11 Ultra",
    "price": 5999,
    "screen_size": "6.81 inches",
    "battery_capacity": "5000 mAh",
    "description": "小米Mi 11 Ultra配备6.81英寸高刷新率屏幕和5000 mAh电池，性能与续航兼具。"
  }
]
```

LLM



感谢您的观看

