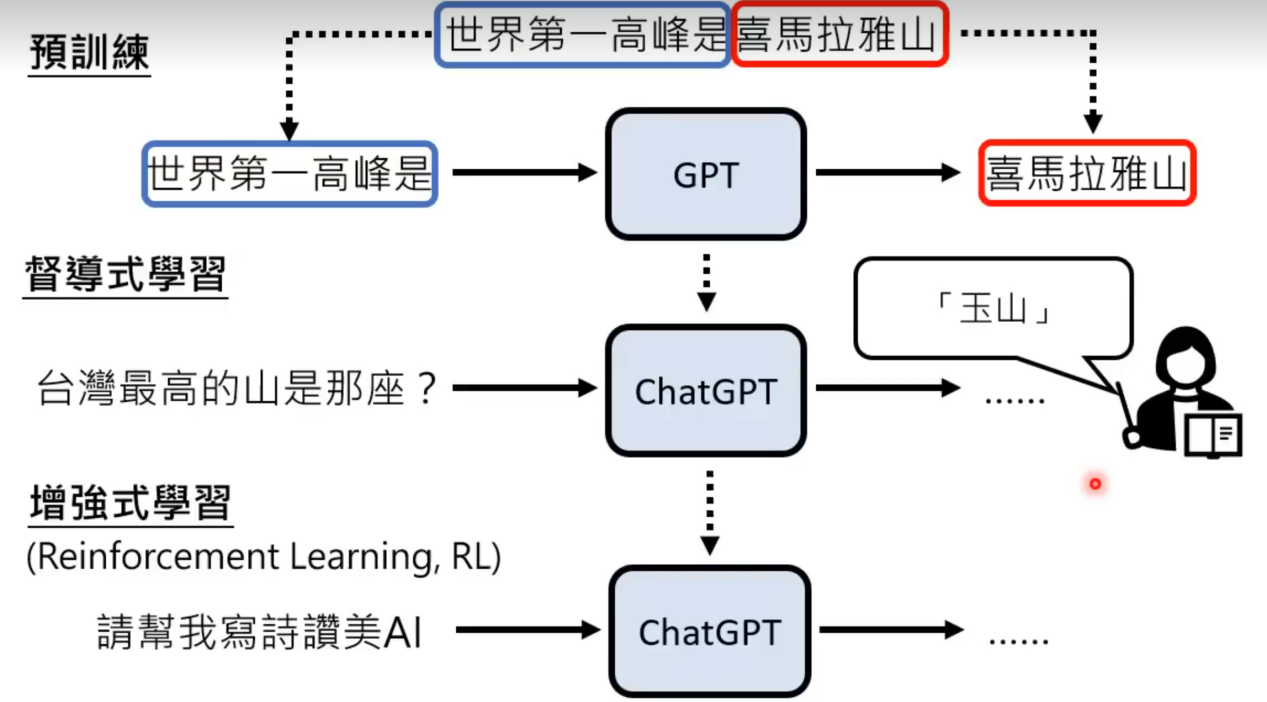
# 大语言模型

1. ChatGPT原理剖析：
2. ：预训练也叫自督导式学习/基石模型
3. ：



1. ：催眠——prompting
2. 生成式学习的两种策略：各个击破、一次到位
3. 中文的token就是字、英文的token是word piece
4. New Bing是有搜寻网路的，何时搜寻是机器自己决定的

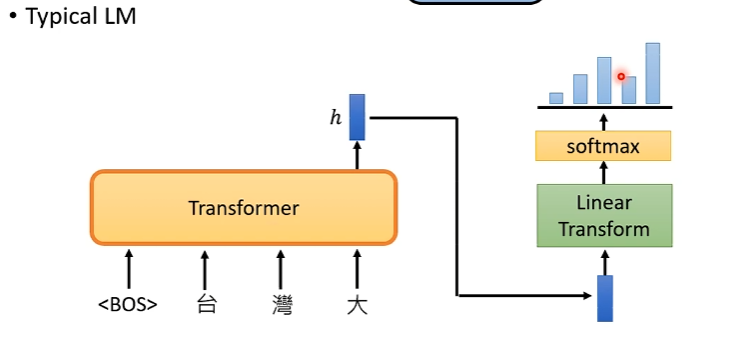
WebGPT：使用搜寻引擎的GPT （使用搜寻引擎也是文字接龙）

Toolformer：使用多重不同的工具（使用工具就是文字接龙）

1. GPT做的文字接龙，BERT做的文字填空
2. 大型语言模型两个期待，一个专才一个通才
3. 成为专才的好处：专才在单一任务上有机会赢过通才
4. BERT是一个语言模型，经过微调和加外褂成为一个专才——在语言模型的本身加上一个Adapter
5. In-context Learning:透过范例来学习。通过例子让机器知道作什么样的任务，但是也会进行范例学习，越大的模型受数据错误率影响越大
6. Instructing-tuning（eg：T0）:透过题目叙述来工作。先搜集大量自然语言的任务，然后改写成指令。训练有了这样的任务训练的时候就不用这个任务
7. Chain of Thought（CoT）Prompting：有推论

Least-to-most prompting：复杂的问题拆分

12.一般的LM（语言模型）:h为一个向量，对h进行分类

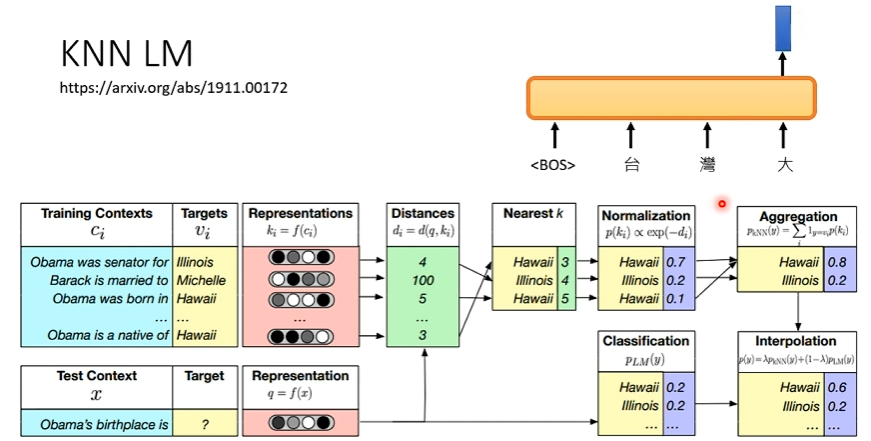


13.KNN LM（太花时间：计算Representations花费的）

Representations：参数通过Transformer等得出来的

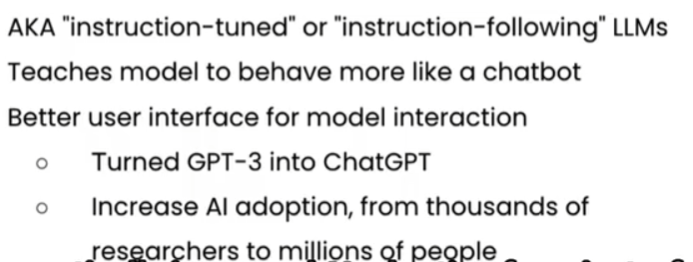
Aggregation：最后输出的概率

Classification：一般LM经过Linear Transform得到的



# **大模型微调相关**

1. **指令微调——Instruction Finetuning**
2. 什么是指令微调



1. 指令微调的数据集

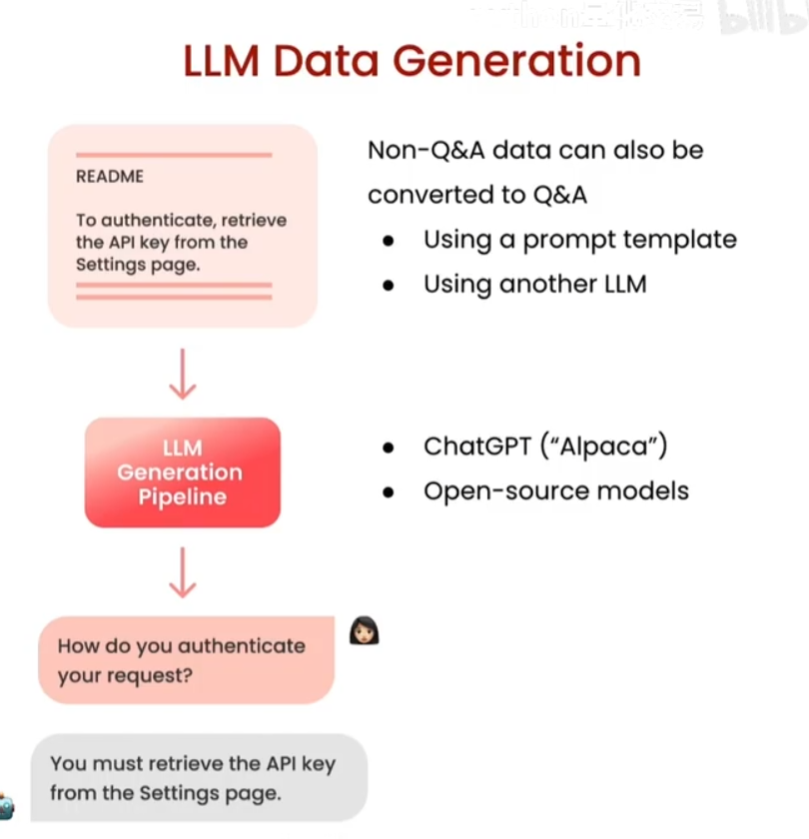
可以使用现有的易得的数据

如果没有数据——可以使用提示模板将自己的数据转换为更具问答形式或指令跟随形式的 内容

——可以使用另一个LLM来帮你完成

——使用ChatGPT

——在不同开源模型上使用管道



1. 微调的步骤概述

数据准备——训练——评估——数据准备....

1. **微调方法：LoRA、Prefix Tuning、P-Tuning、Prompt Tuning、AdaLoRA**

（1）Bert变体

RoBERTa，即BERT预训练方法的鲁棒优化

* 模型训练时间更长，batch size更大，训练数据更多
* 删除了NSP任务
* 使用动态MASK

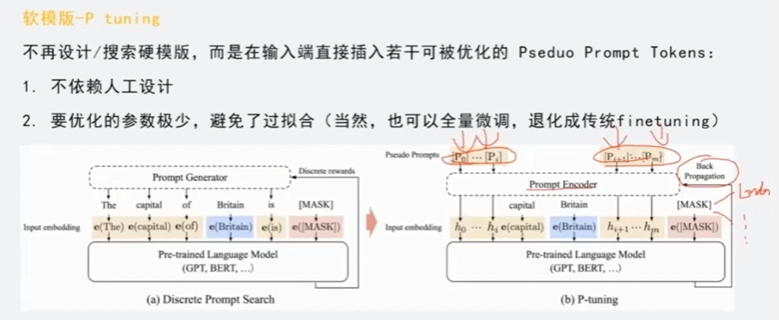
DeBERTa，相比于RoBERTa的更强模型

* 解耦注意力机制
* 虚拟的对抗方法，通过扰动归一化词嵌入来提高训练稳定性

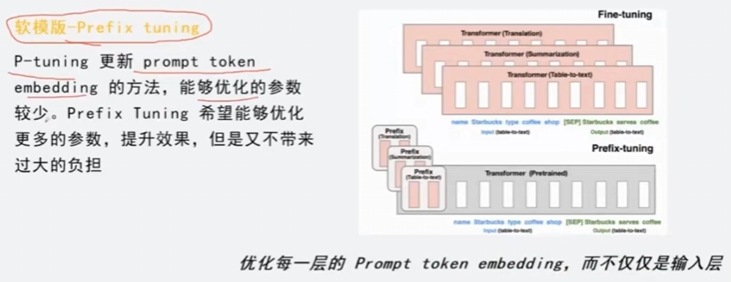
（2）Prompt Tuning



1. P-Tuning



1. Prefix Tuning



1. LoRA



LoRA权重就是A和B

1. **LangChain**
2. 模型：基础的语言模型

提示（prompt）：创建输入，用来给模型传递消息的一种方式

解析器：接受模型的输出，并将输出结果解析成更结构化的格式

1. 记忆存储管理

方案一：ConversationBufferMemory



方案二：ConversationBufferWindowMemory——保留最后若干轮（k轮）对话消息



方案三：ConversationTokenBufferMemory——限制保存在记忆存储的令牌数量





方案四：ConversationSummaryBufferMemory——在记忆中存储历史消息的摘要