AI 大模型开发工程师之
大模型微调基础

讲师：李希沅

目录

1 大模型为什么需要微调?

01、AI大模型使用阶段

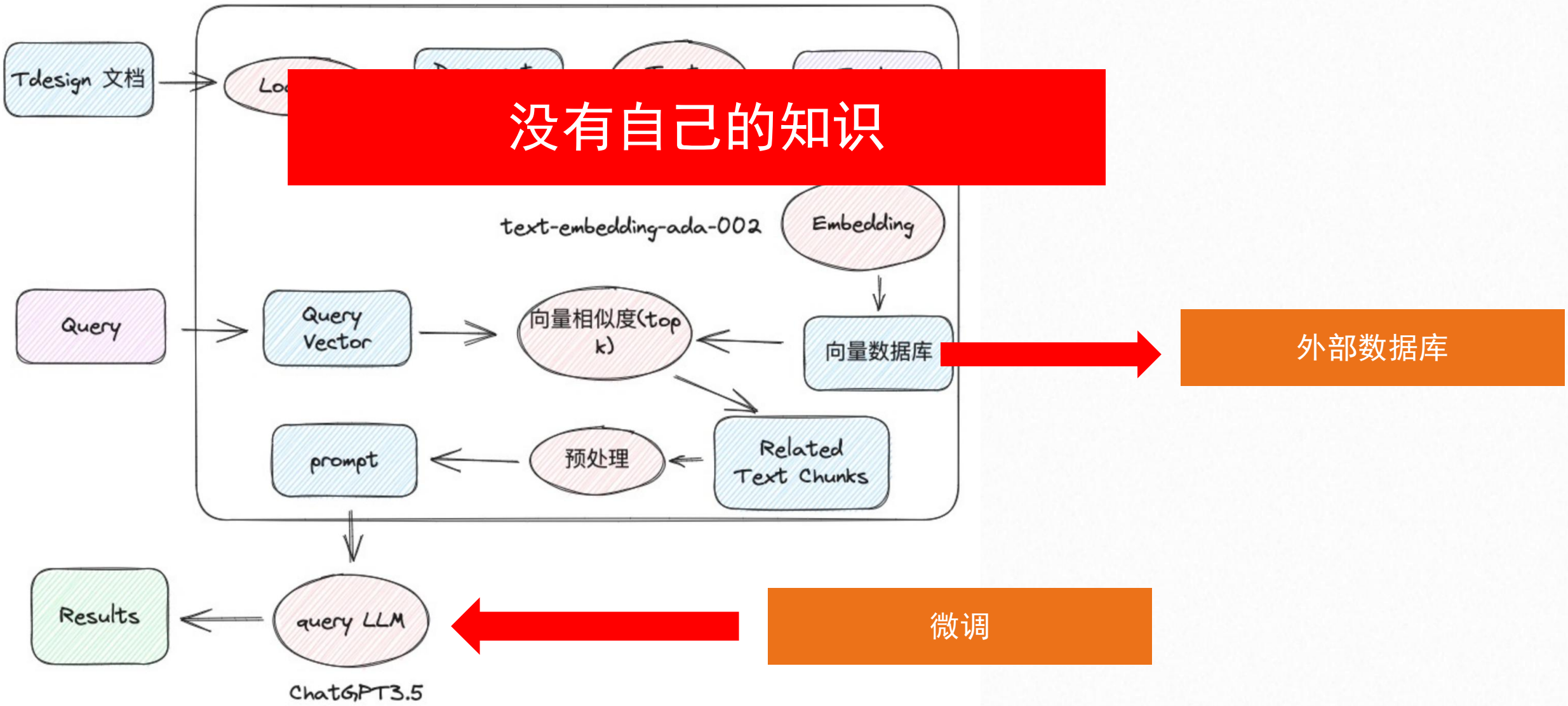
根据自然语言提示词，让大模型帮你干活，完成任务目标

借助大模型的能力，开发应用工具，比如知识库，比如自助iQuery等项目

用有监督的方式，根据我输入的数据进行训练，然后去预测答案

用无监督的方式对大量的知识对模型进行训练，然后用来预测下一个单词

02、目前大模型的缺陷


Tdesign 文档
没有自己的知识
text-embedding-ada-002
Embedding
Query
Query Vector
向量相似度(top k)
向量数据库
外部数据库
prompt
预处理
Related Text Chunks
Results
query LLM
ChatGPT3.5
微调

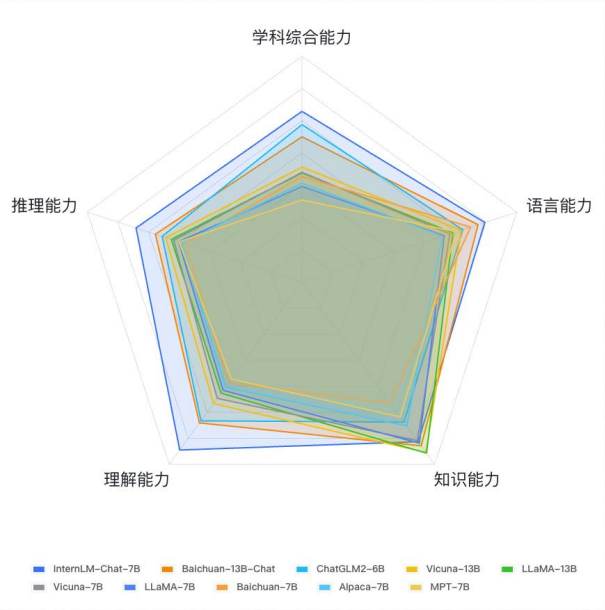
03、为什么需要微调

1、预训练成本高昂

2、企业级垂直数据分布差异

3、Prompt Engineering 推理成本

4、企业级私有数据安全



	GPT-3 (OpenAI)	Gopher (Google DeepMind)	MT-NLG (Microsoft/Nvidia)	PaLM (Google Research)
Model Parameters	175B	280B	530B	540B
FLOPs/Token/Model Parameter	6			
TPUs/Machine	4			
Peak FLOPS/TPU	275T			
FLOPS Utilization	46.20%			
Cost/Machine/Hour(1-year reserved)	$8.12			
Seconds/Hour	3600			
Training Cost/1000 Tokens	$0.0047	$0.0075	$0.0141	$0.0144
Train Tokens	300B	300B	270B	780B
Training Cost	$1,398,072	$2,236,915	$3,810,744	$11,216,529

② 大模型微调的方式有哪些?

01、大模型如何微调

- 全量微调FFT(Full Fine Tuning)

  缺点：训练成本高，灾难性遗忘

- 部分参数微调PEFT(Parameter-Efficient Fine Tuning)
- 针对的模型分为
  - 在线大模型：
    - OpenAI大模型（Fine Tunring）
  - 离线的模型：
    - LoRA、QLoRA、Adapter、Prefix-tuning、P-tuning2、Prompt-tuning

Model&Method	# Trainable Parameters	WikiSQL Acc. (%)	MNLI-m Acc. (%)	SAMSum R1/R2/RL
GPT-3 (FT)	175,255.8M	**73.8**	89.5	52.0/28.0/44.5
GPT-3 (BitFit)	14.2M	71.3	91.0	51.3/27.4/43.5
GPT-3 (PreEmbed)	3.2M	63.1	88.6	48.3/24.2/40.5
GPT-3 (PreLayer)	20.2M	70.1	89.5	50.8/27.3/43.5
GPT-3 ($Adapter^H$)	7.1M	71.9	89.8	53.0/28.9/44.8
GPT-3 ($Adapter^H$)	40.1M	73.2	**91.5**	53.2/29.0/45.1
GPT-3 (LoRA)	4.7M	73.4	**91.7**	**53.8/29.8/45.9**
GPT-3 (LoRA)	37.7M	**74.0**	**91.6**	53.4/29.2/45.1

02、掌握大模型核心三要素

关注视频号：玄姐谈AGI

助力数字化人才提升 AIGC 能力