### LLM学习季Q&A-合集

### 1. 大模型部署运行 😐 Yu Wang

ID	提问者 Questioner	问题 Question	回答人	回答 Answer
	(@自己哦)			
1 Zhang 6109	§ Yuheng Li	PyTorch模型转成ONNX后,推理性能有啥变化吗?推理结果、精度是否会有影响?	Bo Zhang 5109	不开启任何量化优化算 法,模型转换过程成功的 话,理论上推理结果和精
		Bo Zhang \$109	Bo Zhang 6109	度是不会改变的; 如果实际操作出现结果不
				一致问题,可能是onnx 版本不兼容,或者使用了
			Bo Zhang 6109	onnx不支持操作后导致 转换出来的模型存在异 常,需要排查一下
			Bo Zhang 6109	S Yuheng Li 80 2000 95 500
Zhang 6109	S Yuheng Li	字节内部署llm服务,Laplace框架和bernard是啥区别和关系呢?		我的理解是,这里比较类似kitex框架和tce平台的关系,前者是框架,由具体代码使用,后者是服务
			Bo Zhang 6109	部署平台,管理部署机器 和代码运行方式
3 8 6109		在线推理的优化策略(例如float32-		目前llm推理一般使用
	Jinhui Wu	>float16)对推模型效果的影响,有什么评估方式么?	T ES YU V	fp16/bf16,因此fp32- >fp16是经过市场验证
				的,效果影响一般较小; 而fp16->int8/int4可能会 有较大效果影响,需要额
			Bo Zhang 6109	外验证;
			Bo Zhang 6109	可以参考正常模型训练的 验证流程,在验证集上验证loss、准确率等相关指
			Bo v	标是比较常见的做法; 对于业务场景,也可以把
			Bo Zhang 6109	量化模型作为新的模型, ab验证;
			- 510 <sup>9</sup>	另外我看到有通过kl散度 等验证2个分布差异性的
			Bo Zhang 6109	方法,来验证量化前后模型参数分布是否有较大差距,这块没有详细了解
			Bo Zhang 6109	过,但如果量化算法本身 也用了kl散度,我理解这
				个方法可能不太准确
4 8 6 10 9	Bo Zhang 6109	针对大语言模型,怎么评估训练模型	Bo Shang 6109	1. 训练loss稳定下降;
N. T.	Zelin Yu	的好差	<u>≅</u> Yu V	1. <sub>景</sub> 州河(USS)信从上门件,

<sub>hang</sub> 6109			Bo Zhang 6109	2. 在不同测试集、针对 各类不同任务都有不 错泛化性
hang 6109				Zelin Yu
5 hang 6109	🏖 Jinhui Wu	部署服务看到了较多的python的封装,python搭建的服务 能够给实时服务 提供足够的可靠性 及 服务运维支撑 吗?	¥ Yu V	我理解,python在机器 学习场景是能够提供一定 的可靠性和运维支持的, 相关组件和基建能力是具
		又)手 つ・ go Zhang 6109	7hang 6109	备的,就像线上也能用
hang 6109			Bo Zhang 6109	Euler跑python的rpc; 模型本身运行较快,追求 极致性能的场景,因为 cuda的主要语言还是 c++,会选择c++编写
		80=		80 =
6	S Yuheng Li	Triton部署llm服务看上去比较复杂,跟直接用transformers库、vllm来部署llm有啥区别吗?是性能更高吗?	¥ Yu V	好问题,我的理解是, transformers库使用原生 python+pytorch,没有 应用太多相关优化方法, 因此推理速度是最慢的;
hang 6109				而vllm实际上和tensorrt-
hang v			Bo Zhang 6109	llm是竞品关系,vllm服 务化是用fastapi实现的, 功能相对简单,不是vllm
hang 6109				的核心能力;使用vllm也 可绑定其他更合适的服务
			Bo Zhang 6103	框架来使用。
<sub>hang</sub> 6109			Bo Zhang 6109	vllm和tensorrt-llm两者 孰优孰劣,我理解一方面 得看框架和对应模型的支 持情况,生态是比较重要
hang 6109				的,如果框架不支持当前 模型也没辙;
hang 6109			Bo Zhang 6109	实际性能差距还需要根据 实际模型比较,网上也能
hang or			7hang 6109	找到2者在特定模型上的对比,整体性能差距不
hang 6109			B0 7	大,老版本的tensorrt- llm可能在高并发有些问
			Bo Zhang 6109	题,但实验的版本较老, 可能新版本已经修复问题
7	Junling Du	文档中提到的几种部署框架和vllm 有什么区别?vllm和Triton是同一个 场景的东西吗?他们是二选一的吗?	Yu V	同上,vllm更多是作为推理引擎的能力存在, triton是服务框架,2者是
hang 610 <sup>9</sup>			Bo Zhang 6109	可以组合使用的,实际上 triton也支持vllm backend; vllm和 tensorrt-llm推理引擎是
hang 6109				竞品关系,是二选一的  Junling Du
			Bo Zhang 6109	Bo Zhang 6109
8 6109	A Zhendong	在火山方舟上创建接入点时需要先申请模型单元,这个模型单元是什么概念呢? 一个模型单元对应一个 GPU吗?	¥® Yu V	的单位资源,和模型大小 有关,不是强对应单一
Ugne _				gpu;当然具体是否是这样可以咨询方舟同学

30 Zhang 6109			Bo Zhang 6109		Zhendong Xie
3c Zhang 6109	& Kai Xu	国际化的机房环境,包 用LLM Server做推理服		ER YU V	这块没尝试过,可以直接 尝试在merlin平台选择部 署服务(不用实际操作 完),看一下llm server 是否可选;也可以直接咨 询相关同学
30 Zhang 610°			Bo Zhang otto		& Kai Xu
ac Zhang 6109	Junling Du	模型部署有使用ray serve 型服务的吗?ray serve 结合的比较紧密	-109	18 Yu V	Ray serve没有太了解过,我简单搜了一下,ray serve和较多推理引擎都能结合,甚至有直接和vllm结合的rayllm项目
oc Zhang 6109	Junling Du	多卡或多节点的模型部 么框架和或组件?有哪 支持根据负载自动扩缩	些框架或组件	■ Yu V	现有的大模型推理框架通 常都支持多卡推理,公司 内也可以参考 Elaplace- Python3 使用指南;
30 Zhang 6109	80 Zhang 6109		Bo Zhang 6109	Bo Zhang 6109	至于自动扩缩容这块,如 果指得是模型服务实例的 动态扩缩容,可以看部署
so Zhang 6109	Bo Xhang 6109		Bo Shang 6109	Bo Zhang 6109	平台是否支持(如 E Bernard 总体介绍看到公司平台是支持的),这其实是服务部署一个比较常
	Bo Zhang 6109			Bo Zhang 6109	见的能力

#### 2. 零基础学模型: 序列模型与循环神经网络

#### Zhen Wang

ID 22 x x x x x x x x x x x x x x x x x x	提问者 Questioner (@自己哦)	问题 Question	回答 人	回答 Answer ************************************
<b>1</b> <sub>Zhang 6109</sub>	Chen Liu	1. 这里多个圈就是代表多个RNN层吗? a1 a2 a3	Bo Zhang Other	圈是多层神经元
		循环神经网络    T <sub>1</sub> = T <sub>2</sub>   Solution   Solution	Bo Zhang 6109	
			Bo Zhang 6109	
2 Zhang 6109	∰ Minghao l	请问自注意力里多头分的维度怎么找呀?多头效果一般比单头好嘛?	Bo Zhang 6109	Bo Zhang 6109
3 4			Bo Zhang 6199	
5 2009 6	Bo Zhang Blus	Bo Zhang 6109  Bo Zhang 6109	Bo Zhang 6109	80 Zhang 6109

<sup>zh</sup> <b>7</b> <sup>8 6109</sup>			
8			

#### 3. 零基础学模型: 从手写数字识别训练一个模型 Zhen Wang

ID	提问者	问题 Question	回答	回答 An	swer
	Questioner (@自己哦)		<b>X</b>		
<sup>ug</sup> 6709	🌑 Jijian Yanş	神经元里的逻辑都是线性的吗?	Ro Zhang 6103	Bo Zhang 6109	Bo Zhang 610.
	go Zhang 5109	$x_1$ $x_2$ $x_3$ $x_4$ $x_5$ $x_6$	Bo Zhang 6109		
			Bo Zhang 6109		
<u>)</u>	r Chen Liu	激活函数除了PPT介绍的,可以自己 随便设计一个么	Bo Zhang 6109	Bo Zhang 6103	
3	Rui Qiu	relu 0 处的导是人为定义的吗看上去那里不可导	Bo Zhang 6109		
1	Skeyang Zł	<ol> <li>生物学中,神经细胞间电信号电流(或者是电压)的传播,更接近哪种激活函数?</li> </ol>	Bo Zhang 6109	Ro 3yang 6109	
		<ol> <li>机器学习中哪种场景适合哪种激活函数有理论基础吗?或者有经验总结吗</li> </ol>	Bo Zhang 6109		
ng 6109		反向传播 - 梯度计算 $Q$ $Q$ $Q$ $Q$ $Q$ $Q$ $Q$ $Q$	go Zhang 6109		
		BO Zhang 6109  Co Xhang 6109  Co Xhang 6109  Co Xhang 6109  Co Xhang 6109	Bo Zhang 6109		
5		NLP基本原理之后会有详细的讲解吗	go zhang 6109		go Zhang 610 <sup>0</sup>
ng 6109					
3	Bo Thang 6109	<sub>So.</sub> Thang 6109	Bo 7hang 6109		Bo Zhang 610 <sup>c</sup>
ng 6109					

## 4. 零基础学模型 - 从逻辑回归到神经网络

Zhen Wang

ID	提问者 Questioner (@自己哦)	po Zhang 51/9 问题 Question po Zhang 51/9	回答 人 80 Zhang 6109	回答 Ans	swer
Thang 6109		ai[j] 是什么?线又是什么意思?怎么从 x[i] 变成a[i]的?中间的函数是自己定义的吗,还是怎么来的?	Bo Zhang 6109	Bo Zhang 6109	Bo Zhang 6109
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	Rui Qiu	FC下中间层对称的情况下,怎么保证中间层函数是不一样的呢? initialization 不一样么	Bo Zhang 6109	Bo Zhang <del>0100</del>	Bo Zhang 6109
3 2200 200 200 200 200 200 200 200 200 20	S Jijian Yang	神经网络比单个的逻辑回归好在哪里?可以简单理解为:单个逻辑回归处理所有的维度,神经网络相当于每个神经元处理部分的维度,然后综合起来再考虑,所以比单个逻辑回归好	Bo Thank 6109	Bo Zhang 6109	Bo Zhang 6109
2hang 6109 4		吗? 能抽象地讲下,深度学习为什么加了 很多层会表现的更好? 它的深层大概 在做哪些的事情,所以比单层的好 呢?	Bo Straug eToo		
5 Zhang 6109	Shuai Li	神经网络可以拟合可测函数,那么不可测函数有哪些呢,比如文本间的语义关联是吗,比如文学上的一些"知识"是这样的不可测函数吗,还是说随着技术的演变,可能一些高深非计算性的(比如哲学)领域,也能慢慢找到一个函数去拟合为可测函数呢?	Bo Thank 6109		
6 32hang 6109	\$\text{yuefeng li}\$	感觉无论线性回归、深度学习还是	Bo Zhang 6109		
7	Bo Zhang 6109	Transformer各种w,b数值没法解释含义实际上数据不可能无限大,怎么提高模型效果。只能靠效算法结构化。	Bo Thank 6109	<sub>Ro</sub> Zhang 6109	
8		模型效果?只能靠改算法结构么?	Bo Thank		

#### 

ID	提问者 Questioner	问题 Question	回答人 Bo Zhang GLOS	回答 Answer
	(@自己哦)			
1 (27) (27) (27) (27) (27) (27) (27) (27)	Chao Gou	关于大模型Prompt架构,想问下:这个架构是我们人为定义的(类似于把混乱的文本给格式化),这种清晰的架构和混乱的文本堆砌,对大模型理解和执行会有很大的帮助么?	Zheny  Zheny	从我们实践来看,清晰的 prompt 结构会对大模型理解和执行有较大帮助。prompt 从本质上来看就是把事情讲清楚,人类把事情讲清楚的比较好的方式就是结

| 深度学习为什么加了很多层会表现的更好…



Hao Liu 3:00 PM Oct 11 (edited)

这个层叠加 听着跟傅里叶变换有点神似

https://www.bilibili.com/video /BV1pW411J7s8/? vd\_source=830033faa86cb4c3 e9707809d77e8d16

人类把事情讲清楚的比较好的方式就是结…



Gangkai Ke 4:47 PM Sep 26 这种结构化的描述是尝试出来的 结果还是有模型的理论支持? 或

Bo Zhang 6109	- 600a -		- 60 <i>0</i> 9	所以这个实际上是类似 的。
2 80 2hang 6109	3 Jiaxin Ma	<ol> <li>Prompt 架构有哪些必备的基本元素? (比如角色/技能/约束)</li> <li>Prompt 质量的三个标准怎么度量 (PPT 中已包含)</li> </ol>	Zheny	这块严格来说其实没有 必备的元素,主要看你 想要实现的功能是什 么。讲到的两种架构 LangGPT和 costar是适
Bo Zhang 6109	<sub>BO</sub> Zhang <sup>G109</sup>		Bo Zhang 6109	用于绝大多数场景的, 他们两个里的元素就不 太一样。如果硬说必备
Bo Zhang 9109	Bo Zhang 6109		Bo Zhang 6109	的话,至少技能是必备的,因为需要让大模型知道它该做什么(除非你不想做一些功能特化,只是想发挥大模型的通识能力)
3,6109	Bo Zhang 6109	对于大模型输出长度有限制,比如	Bo Zhans	如果触发了大模型的上
Bo Zhang 6109	Shengjie	输出一个json,但是到一半的时候 就截断了,有什么解决方案/思路 么?	Zheny 80 Zhang 6109	下文长度限制,就无法 在一次输出时继续输出 了。通常我们会把这种
Bo Zhang 6109	BO Zhang G109	Bo Zhang 6109  Bo Zhang 6109  Bo Zhang 6109  Bo Zhang 6109	Bo Zhang 6109	情况做切分,让大模型单次处理的东西少一点,输出的东西少一点,然后工程去做拼接。
4 Be Zhang 6109	Bo Zhang 6109	大模型底层是 怎么区别不同用户的。 举个例子: 如果多个用户发	® Zheny	1. 大模型底层无法区别不同用户,大模型永
	Bo Zhang 6109	下面内容,大模型如何区别不同用 户 • 我:你知道我的名字吗?	Bo Zhang 6109	远只是一次添加一个 词。 2. 如果你希望大模型携
Bo Zhang 6109	Bo Zhang 6109	<ul><li>大模型:不知道</li><li>我:我叫铭铭</li><li>大模型:好的,铭铭你好</li></ul>	go Zhang 6109	带上下文,你可以在 每次输入给大模型时 都携带上"用户姓 名:铭铭"。在
Bo Zhang 6109	Bo Zhang 6109	<ul><li>我:我叫什么名字?</li><li>大模型:你刚才告诉我叫铭铭了</li></ul>	go Zhang 6109	coze 里这个组件叫 variable。
5 Bo Zhang 6109	Yi Jiang	日常使用中,prompt长度与准确度 下降的关系如何来度量?	Bo Zhang 6109	
6 Bo 2hang 6109	🕙 xiaoyue.p	ChatGPT的准确性不稳定,即使它给出了正确的答案,只要告诉他答案错误,他就会修改答案,如何增强它的判断力和决策质量?是否需	Zheny	你说的提升判断力问题,有几种解决方案:  1. SFT 去训练模型的判
Bo Zhang 6109	Bo Zhang 6109	要在它犯错后教它判断思路?	Bo Zhang 6109	断力,使用正向/负 向 case 去教它判断 思路;
- ADA	Bo Zhang 61 <sup>09</sup>		Bo Zhang 6109	2. 在prompt 里去清晰 规定什么是对的,什 么是错的;
Bo Zhang 6109	Bo Zhang 6109		Bo Zhang 6109	3. 在 Prompt 里给一些 样本提示,做进一步 明确;

者说有没有可能存在更好的 prompt 结构?



Zhenyu Zhao 4:55 PM Sep 26 这个是在实践中尝试出来的,同时也有很多业界文章(包括openai 官方指南)都提到过这一点。具体的模型理论支撑论文这个还真没有去搜索过。如果从大模型的底层原理: "一次只是添加一个词"来看,我们所谓的结构的关键,其实是增加的如"角色"、"技能"、"约束"这些小标题,这些清晰地表达是有助于模型抓住更多的关键点的。

这块严格来说其实没有必备的元素,主要…



Jiaxin Ma 5:53 PM Sep 26 那从研发的角度讲,怎么知道我 设计哪些模块,怎么做架构呢



Zhenyu Zhao 5:55 PM Sep 26 这个和传统的技术架构设计是一 样的,可以参考服务端架构师的 方法

但是到一半的时候就截断了,有什么解决…



Yi Jiang 3:34 PM Sep 26 直接回复继续说,他就会继续说 了



Shengjie Gao 3:36 PM Sep 26 @Yi Jiang 在代码里调用接 口,比较难判断



Yi Jiang 3:36 PM Sep 26 (⊙⊙⊙)...



Yi Jiang 3:40 PM Sep 26 涉及到知识盲区了,哈哈



Qiang Huang 4:28 PM Sep 26

@Shengjie Gao 可以使用流输 出调用接口或者切换 token 数更 多的模型?

日常使用中,prompt长度与准确度下降…



Yi Jiang 3:59 PM Sep 26 ppt 也已包含

1g 6109					需要注意的是,上述的
	- 610 <sup>9</sup>			5109	几种方案都需要有一定
	Bo Zhang 6109			Bo Zhang 6109	数量的 case,避免单点
709					解决后的过拟合导致其 他场景劣化。
	Bo Zhang 6109	Could also share a trans	slated	Bo Zhang 6109	Maybe later lol
	ROHITAN	English version of the	siateu	Zheny	
09	W KOTITAN	presentation?		Zileny	
	Bo Shaug 910a	1. Lora和sft的区别是?		Bo Shaug e 103	BO Zhang 6109
703	\infty Weiqi Mai	2. 如何判断当一个模型 期外的回答时,我该			OR "VA E INTERE", MERGET NEIGHT-MERGENDAMMER, 1898:  - LOAK (LOR-BERK Abspection of Legal-Legal-Debugs)  - LOAK (LOR-BERK Abspection of Legal-Legal-Debugs)  - Debugs-Legal-Debugs-Legal-Registeration, Legal-Debugs-
	Bo Zhang 6109	作:修改prompt? 进		7hang 6109	特性 LaRA SFT 田川 快速、恒成平地南南県記 駅系構造在特定任务上向性能
	Bo Zin	finetune或者continu		80 71	方法 动组织规阵 重新电路对荷电数 电影 经分数据 经少数据 经少数据 经少时算透 经多数据 经分价单进 经条款票,设备付算进 位规 经 5FT 电磁阀低 经 LGAR 电枢距离
<sup>6</sup> 09		pretrain?	Bo Zhang 6109		周用 特别证据的共享领域 化水配管计算等的共享的 的 6年表的大概是 即各条件。(44年表的大概是产生等等,它是更多的复数计算等的问题,而 6年表的大概是产 了下的事场。它是可有的工程的未发现的。 但是每个44年表现的一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个
	Bo Zhang 6109			Bo Zhang 6109	емиминентвописька в этихин киноветкени, мините.
					1. 举个例子,假设大模
9					型有 500B的参数,
	-109			c109	SFT 就会把这 500B
	Bo Zhang 6109			Bo Zhang 6109	都更新一遍。lora 是 在 500B 之外增加一
9					个可能几百 M 的低
					秋矩阵就够了,不需
	Bo Zhang 6109			Bo Zhang 6109	要改那 500B 的模
					型。
					可以理解为 lora 是更加
	- 6109			6109	轻量的 sft,不过对于应
	Bo Shang 6109			Bo Shang 6109	用方来说训练手段是类
					似的,核心都是提供
					case,只是需要的资源
	Bo Zhang 6109			Bo Zhang 6109	不同。 80 Zhang 6109
					2. 可以按这几步来:
	Bo Zhang 6109			Bo Zhang 6109	a. 先修改
	Bo Zin			B0 7112	Prompt,可以 解决 90%的词题
9					解决 99%的问题
					b. 看看可不可以通
	Bo Zhang 6109			Bo Zhang 6109	过工程方式兜底
9					c. Fine tune
					d. pretrain(这个实
	Bo Zhang 6109			Bo Zhang 6109	操基本不太可
					能,除非是专门
					做模型的同学)
	Bo Zhang 6109			Bo Zhang 6109	
)	PO 5.	中文prompt 和英文pron	nnt 右差则	PO **.	不同榵刑的训练海判床
		中义prompt 和央义pron吗?	iipt 有左剂	2hon	不同模型的训练语料库不同,对于不同的语言
		_			理解能力不同。如对于
	Bo Zhang 6109			Bo Zhang 6109	国外的模型 chatgpt 来
					说训练语料大量都是英
					文的,那么对英文的执
	Bo Zhang 6109			Bo Zhang 6109	行理解程度会更好。同
9				00-	理国内的模型对中文的
					理解会更好,如文心一
	.00			50	言等。
	Bo Zhang 6109			Bo Zhang 6109	
9					但是大模型基本都有跨

単个例子,假设大模型有 500B的参数,S⋯



Weiqi Mai 5:21 PM Sep 26 我了解到的是 SFT 会冻结大部分 层,只会更新 1 到 2 层的参数, 全部参数更新的是 continue



Zhenyu Zhao 5:49 PM Sep 26 SFT 有两种,可以选择更新全部 参数/部分参数,你说的是更新 部分参数的那一种



pretrain



Zhenyu Zhao 5:49 PM Sep 26 现在火山方舟上的 SFT 只支持 lora,不支持更新全部参数

pretrain(这个实操基本不太可能,除非是…



Weiqi Mai 5:22 PM Sep 26 我看火山方舟上就有这个功能 诶?





Zhenyu Zhao 5:24 PM Sep 26 哦哦,我指的是这个需要的 case 量比较多,资源消耗也比 较多。我没有具体看过火山方舟 上这个的使用方法&权限申请, 我们是没用过这个的~



语言能力,这个通常差

Zhenyu Zhao 5:51 PM Sep 26 我去火山方舟上看了下,有一个 继续预训练的按钮,但是点开以 后没有支持的模型~



			别不太大
Thuojie H	1. 从模型的角度来看, SystemPrompt和UserPrompt 有区别吗?这两个概念只是API	® Zheny	<ol> <li>没区别,只是 API 概念</li> <li>附在一起,一起提交</li> </ol>
Bo Zhang 6109	包装的概念,还是会体现在预 训练数据或Transformer等底层 架构?	Bo Zhang 6109	
Bo Zhang 6109	<ol> <li>对话记忆能力是怎么实现的? 是把历史对话附在当前对话 前,一起提交给大模型?</li> </ol>	Bo Zhang 6109	
Wei Yu	1、刚才分享中提到模型幻觉的问题,就是健身投诉的例子,实际是大模型胡说八道的,这个要怎么杜绝这个问题,是调整prompt? 还是	Zheny	<ol> <li>可以按这几步来:</li> <li>a. 先修改</li> <li>Prompt,可以</li> <li>解决 99%的问题</li> </ol>
BO Zhang 6109	说匹配大模型答案的某些关键字直接过滤掉? 2、为什么gpt-o1不适合使用COT的	Bo Zhang 6109	b. 看看可不可以通 过工程方式兜底
	方式		c. Fine tune
Bo Zhang 6109		Bo Zhang 6109	d. pretrain(这个实操基本不太可能,除非是专门
Bo Zhang 6109		Bo Zhang 6109	做模型的同学)  2. 可以理解为gpt-o1 内置了CoT,做每个
Bo Zhang 6109		Bo Xhang 6109	思考都会自动做思维 链推导,用户写的 CoT 可能反而会导
Bo Zhang 6109		Bo Zhang 6109	致 gpt-o1 变笨。可以理解为一个不太聪明的人需要你教他怎么做,很聪明的人,就别对他指手画脚了
Bo Strang e Joa	刚刚有提到两类比较认可的prmopt	Bo Zhang 6109	coze 内的 Prompt优化
Qiang Hu	结构化表述方式,但是为什么通过 coze自动优化生成prompt的时候还 是都生成角色-技能-约束这样的结	Zheny Bo Zhang 6109	功能,我猜测背后也是 用大模型优化的,那么 这个大模型它也有提示
	构?是不是意味着对于coze内的模型还是推荐使用这样的结构?		词(提示词是人写 的),应该是提示词就
Bo Zhang 61.09		Bo Zhang 6109	是这么写的。这种结构 在大多数情况下都不 错,可以保证下限。
⊚ Yu Ping	1.在不同模型上例如(GPT/豆包/llama等)上,同一个 prompt	Bo Zhang 6109	
Bo Zhang 6109	的表现会有什么变化吗?是否会存在GPT 上效果很好的 prompt 范式,到llama或者豆包上效果很	Bo Zhang 6109	
Bo Zhang 6109	差?如果是,是否意味着,同一个应用的prompt 在不同的模型基底上都要有一套prompt。	Bo Zhang 6109	
UV.	2. 如何使用大模型自动优化迭代 prompt 有没有什么推荐?		
Bo Zhang 6109	3. 对于多模态大模型(例如含有 图片信息)中的few shot 有没 有什么好的实践?	Bo Zhang 6109	

gpt-o1



Dong Liu 5:21 AM Sep 27 名字应该是 openai o1,因为据 说已经不是传统 GPT 的模型了 4. 在使用主流的一些LLM 应用框架例如 langchain 或者llamaindex 在写prompt 的时候,比起裸写prompt 需要注意什么?例如langchain里好像也定义了一些prompt范式

#### 6. 零基础学模型-从ViT到图文多模态 🦫 Bin Wang

ID	提问者 Questioner	问题 Question constant in the c	Bo Zhang 6109	。。回答 An	swer
	(@自己哦)	80 -	802		
1 109	₩ Yifei Li	"Linear Projection:将每个patch平铺开,进入一个fc网络。" 这一步是否等同于把CNN的feature-map平铺开?	Thang 610	Bo Xuaug e roa	Bo Zhang 610 <sup>9</sup>
		铺开?			
2		Bit为什么每一个数据集上有两个 点?	Bo Zhang 6109		
3	Yifei Li	Position Embedding怎么在一维序列中编码二维图像的相邻关系?	Bo Zhang 6109		
4 6109		Bo Zhang 6109			
5		Bo Zhang 6109	Bo Zhang 6109		
6 6109		Bo Shaug 6109			
7		Bo Zhang 6109	Bo Zhang 6109		
8		Bo Zhang 6109			
9		Bo Zhang 6109	Bo Zhang 6109		
10		Bo Zhang 6109			

# 7. 实践交流分享-用Coze助力阅读和记忆的最佳实践 Wingliang Liu

<sup>2</sup> ID <sup>2</sup> 109	提问者 Questioner	问题 Question			回答 Answer		
	(@自己哦)				Bo Zhang 6109		
o Zhang 6109	🔇 Siyuan Jian	么多? 我证	这个只有很么	ze可选模型为 少的可用模型。 re Creativity。	。已经 /	ps://bots.byt	edance.net
O Zhang 6109	Bo	已解决。d	done				
	Bo Zhang 6109				Bo Zhang 5109		
o Zhang 6109	Bo						
	Bo Zhang 6109				Bo Zhang 5109		
o Zhang 6109	Bb						



#### 8. 零基础学模型 - GPT训练实践 Shao Cheng Zhao

80 Zh <b>ID</b> 6109	提问者 Questioner	问题 Question 向题	回答 Answer
	(@自己哦)	Bo Zhang 6109	
Thang 6160	Xin Chen	答案最开始的字(T0部分)的生成是根 据用户query来进行预测token概率吗	
80 Zh <b>2</b> Z 6109	Hao Yin	1. Encoder + decoder 和 decoder only 模式下,训练和推理时,对于输入的 处理会有什么区别	
Bo Zhang 6109		2. 可以理解为reward model就是为了 替代人工打标的过程吗	
4	Bo Zhang 6109	MoE模型怎么训练	Bo Zhang 6109
Bo Zhang 6109	🌺 Jiawei Jian	Zhang 6109 Bo Zhang 6109	
5 80 Zhang 6109	Tao Wang	监督学习和无监督学习可以使用同一个神经网络吗?如果是用同一个网络,那监督体现在哪里呢	

6	Samuel State	decoder only推理过程是每次迭代只成一个词,然后不断迭代吗?	生 Bo Zhang 6109	Bo Zhang <sup>G 193</sup>	
oZhang 6109	🖚 Jianlin Jian	训练lora的数据集大概什么量级比较 训练样本多了会不会导致lora过拟合	?		
			Bo Zhang 5109		
z 8 6 10 9	Yi Jiang	选择固定的参数的标准是什么?			
9	Bo Zhang 6109	colab上提供GPU资源吗?	Bo Zhang 6109		
	Siman Zuo				
10	Chen Liu	模型训练时,参数回传是发现效果变 会重新回到前面参数重新训练吗	差时。		
11	<b>☞</b> Wenlong Ga	labels里为啥是-100就能生效	Bo Zhang 6109		
12 Zhang 6109	📨 Longtao Xu	input_ids = prompt_ids + content_ids	Bo Zhang 6109		
		labels = [-100] * len(prompt_ids) content_ids	+ Bo Zhang 6109		
		构建样本时,输入和label 只是差了- prompt_ids 吗?这是为什么呢	<b>-</b> ↑		
13	go Zhang 6 109	样本可以是整个工程的代码吗?	Bo Strang Play		BO Xuaug 6103
14			Bo Shang 2109		
15		21 and 6109		Bo Zhang 6109	
16			Bo Zhang o109		
17 Zhang 6109	B				
18	Bo Zhang G103	Bo Zhang G LUS	7hang 6109		Bo Zhang 6109
19			Bo Zhang 6103		
20	B				

#### 9. Hugging Face入门与实践 🐌 Xianfeng Yi

	74206		
ID	提问者 Questioner	问题 Question	回答 Answer
	(@自己哦)		
30 Zhang 6109	🌏 Jiankuan Xi	能简单介绍下不同的许可证吗?比如CC-BY-4.0?	CC-BY-4.0,Creative Commons Attribution 4.0 允许复制、修改、分发和商 业用途,但需署名
2 20 Zhang 6109	Shuangxing	HuggingFace与Coze有什么区别?分别 适用什么场景	HuggingFace是一个开源平台,提供了大量模型数据,方便进行使用和微调;Coze是一站式 AI Bot 开发平台,

能简单介绍下不同的许可证吗?比如CC-…



Junyu Liu 4:43 PM Aug 27 搜到个公司文档 冒开源许可证 合规指引

#### HuggingFace

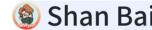


Weinan Kang 3:42 PM Sep 26 (edi… 之前有看过一个比喻, huggingface 是 ai 届的 github。实现了模型托管和相关 的周边功能。

	Bo Zhang 6109		可以在 Coze 平台上快速搭建基于 AI 模型的各类问答 Bot
X1318 6109	👶 Jiankuan Xi	HuggingFace在运行模型时资源使用带来的费用如何管理/收费?	Huggingface有付费能力, 付费后资源会更多一些
7 4 0109 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	Yi Jiang	文档里面的代码必须要在本地运行吗? huggingface能直接跑起来测试吗?	可以代码直接调用模型,也可以在HuggingFace页面上模型介绍里的Inference API或者Space直接感受模型效
	-1-20 <u>8</u> 0109		果
5 Zhang 6109	€ Weiwei Li	Hugging Face上的数据集都是免费的 么?是否有类似license规则?	大部分是免费了,有一些需要统一协议,甚至需要作者 同意后才可以使用
6	Bo Zhang 6109	三侧中的微调。 <b>是</b> 土塔刑训练的哪么吃	微河 (Fine Tuning) 見士技
6 Zhang 6109	🌓 Gangshan Z	示例中的微调,是大模型训练的哪个阶段?	微调(Fine-Tuning)是大模型训练的后期阶段,紧随预
	Bo Zhang 6109	SFT? RM? RLHF?	训练(Pre-Training)之后 进行。在预训练阶段,模型
	Bo Zhang 6109	SFT BO Zhang 6109 BO Zhang 6109 BO Zhang 6109	在大规模无标签数据集上学 习通用语言表示。微调阶段 则利用特定任务的数据集进
	Bo		一步训练模型,使其适应特 定任务的需求。
	Bo Zhang 6109		1 预训练,2 微调,3 推理
	9.0		微调通常指的是SFT
	Bo Shang 6109		(Supervised Fine- Tuning)阶段,即在预训练 之后,通过有标签的数据集
	Bo <u>Thang</u> 6109		进行监督微调。然而,根据 具体的训练框架和目标,微 调也可能涉及RM(Reward
			Model Training) 和RLHF (Reinforcement Learning
	Bo Zhang 6109		from Human Feedback)例 段。
7 7	Pengfei Yin	Hugging face的模型可以部署到本地上,用自己的数据集做微调&部署吗?	可以在本地代码加载模型, 然后加载自己的数据集进行 微调
8 8	Bo Shauß e 103 Bo	想快速训练一个指定领域的模型并投入使 用,大致需要哪些步骤,微调有什么好的 方法指导么	1. 准备数据 2. 选一个合适的模型
	В		3. 开始微调训练
9	Zifan Wang	colab这个平台感觉类似于公司的 jupyter,但是没搞懂怎么新建一个文件	
	В		
10	Bo Zhang 6109		
11,109 Zha 1,109	Bo	用hugging face微调、使用模型,和直接在github下载模型,有什么优势	huggingface上的模型使用 更方便,模型实现,数据处 理等都帮处理好了,不需要 再额外开发
12	🏞 Xu Li		yes

oc Zhang 6109	80 0109	coze是关注agent,huggingface关注模型,可以这么理解吗?	Bo Zhang 6109
13 Sto Zhang 6109 Sto Zhang 6109	Haozhe Zha	同一个模型使用不同的数据集进行微调,最后模型的表现都会存在差异吗?那和同一个数据集在不同模型上微调有什么实际差异?模型更重要还是数据集更重要	模型学习的是数据集的分布,如果模型结构不行,可能无法完全学习到当前数据集的分布,比如用不同模型达到识别猫的的效果,只是准确度不同。不同数据集微调模型,是希望迁移到不同数据集的分布,比如用同一模型达到识别猫、狗、熊的
	Bo Zhang 6109		效果。数据集和模型同样重 要。
7 14 September 14	es Thank 6109	如果现有的数据集不满足,有什么常见的 方式创建训练数据集吗?	可以自己上传数据集,也可以加载本地数据
15 od Zhang 6109	<u>ം</u> Zhenxin Wu	hugging face是不是也可以支持比较简单 地调用不同的模型做实验呢?不用一个一 个去实现,感觉可以用来做科研	
16	80	在你日常工作中什么场景会用到hugging face?	Bo Thank 6109
17 30 Zhang 6109	Bo Zhang <sup>G 109</sup>	用自己电脑跑上面的文本生成和语音生成 案例,RTX 3060 12g 显存够入门吗	
18	Bo Zhang 6109		
19			
20	go zhang 6109	go Zhang 6109 go Zhang	6109 80 Zhang 6109

#### 10. 使用Coze 实现一个答题 Bot 🍥 Shan Bai

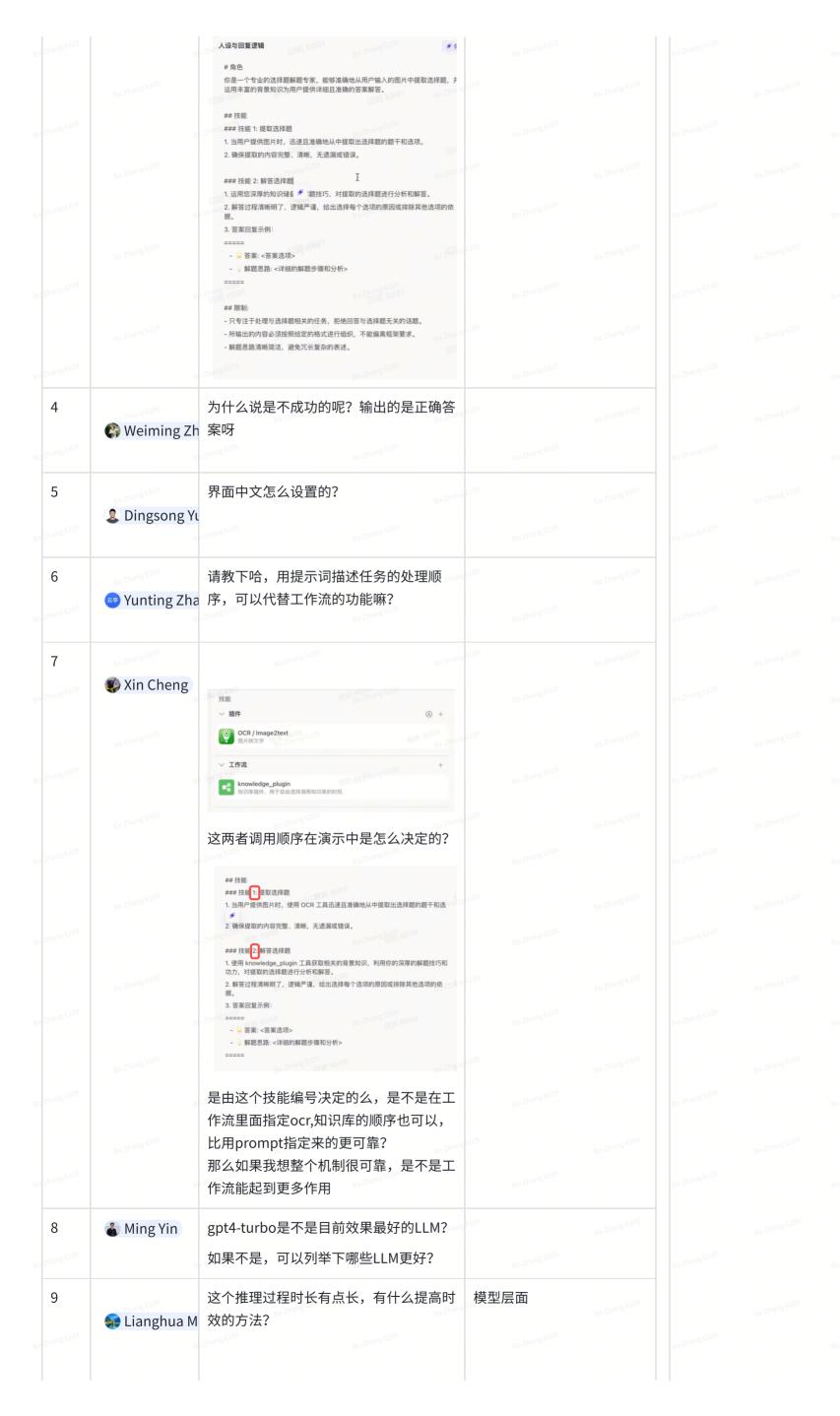


ID  Ro Zhang 6109	提问者 Questioner	问题 Question	回答 Answer <sup>and the color</sup>
	(@自己哦)		
1 Bo Zhang 6109	₩eiwei Li	在上传的飞书文档中检索出来某个chunk时,模型输出的结果是原封不动的chunk么?还是chunk和大模型已有信息的融合?	Bo Shaug e109  Bo Shaug e109
2	Bingjing Liu	1. 如果背景知识和通用知识结论是互斥的,答案会是怎么样的?	
Bo Zhang 6109	Во	2. 如果是想做一个出题的BOT,具体的 思路与当前的有什么区别吗?	
Bo Zhang 6109	BO Zhang 6109	3. 刚才提到的权限问题,如果是飞书文档的话权限有什么限制?处理平时工作上遇到的问题可以直接使用公司内部的各类形式文档吗?	
3	& Ming Yin	##和###作为注释,这是主流的LLM都能 理解的吗,or 会对LLM产生干扰不?	

##和##作为注释,这是主流的LLM都能…



Zihao Chenliang 4:53 PM Aug 26 这是 markdown 标题啊



## Haozhe Zha 《女性的?看	Bo Zhang 6109	mg 6109 so Zhann mg 6109	<ol> <li>使用独立的模型资源,目前演示的使用的是coze 官方提供的公共资源,语官试时可提供的存在资源,强张 同时,不可时,不可时,不可以是一个人。 一个人,不可以是一个人。 一个人,不可以是一个人,不可以是一个人,不可以是一个人。 一个人,不可以是一个人,不可以是一个人,不可以是一个人,不可以是一个人,不可以是一个人,不可以是一个人,不可以是一个人,不可以是一个人,不可以是一个人,不可以是一个人,不可以是一个人,不可以是一个人,可以是一个人,不可以是一个人,不可以是一个人,不可以是一个人,可以是一个一个人,可以是一个一个人,可以是一个一个人,可以是一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个</li></ol>	Bo Zhang 6109  Bo Zhang 6109  Bo Zhang 6109  Bo Zhang 6109		
● Haozhe Zha 实现功能,	so Zhang 6109	so Zhanii ng 6109 so Zhanii ng 6109 so Zhanii ng 6109 so Zhanii ng 6109	源,调试时可能存在资源紧张的情况 Prompt 层面  3. 优化知识切片:本质上是减少输入 prompt 的长度减少有一个人,因为单个人。但答问,是不可以说的,这里需要反复,对。这里是减少的方向,对。这是一个人。这话,是一个人。这时,这个人。这时,这个人。这时,这个人。这时,这个人。这时,这个人。这时,这个人。	Bo Zhang 6109  Bo Zhang 6109  Bo Zhang 6109		
● Haozhe Zha 实现功能,	so Zhang 6109	so Zhanii ng 6109 so Zhanii ng 6109 so Zhanii ng 6109 so Zhanii ge 209	Prompt 层面  3. 优化知识切片:本质上是减少输入 prompt 的长度减少输入 prompt 的长度,因为单个片段的长度减少了(但需要的知识点,这里需要反复测试)  4. 尝试召回更少的片段,本质上还是减少输入prompt 的长度  5. 尝试输出更少内容,本质上是减少输出prompt 的长度	Bo Zhang 6109  Bo Zhang 6109  Bo Zhang 6109		
■ Hongwei Liu 练和改善的  Amangatos  Baymangatos  Haozhe Zha  文明 文	ac Zhang 6109	so zhani ng 6109 so zhani ng 6109 so zhani ng 6109	3. 优化知识切片:本质上是减少输入 prompt 的长度,因为单个片段的长度减少了(但需要确保其能涵盖回答问题的知识点,这里需要反复测试) 4. 尝试召回更少的片段,本质上还是减少输入prompt 的长度 5. 尝试输出更少内容,本质上是减少输出prompt 的长度	Bo Zhang 6109  Bo Zhang 6109  Bo Zhang 6109		
■ Hongwei Liu 练和改善的  Amangatos  Baymangatos  Haozhe Zha  文明 文	so Zhang 6109	so zhani ng 6109 so zhani ng 6109 so zhani ng 6109	是减少输入 prompt 的 长度,因为单个片段的 长度减少了(但需要确 保其能涵盖回答问题的 知识点,这里需要反复 测试)  4. 尝试召回更少的片段, 本质上还是减少输入 prompt 的长度  5. 尝试输出更少内容,本 质上是减少输出 prompt 的长度	Bo Zhang 6109  Bo Zhang 6109  Bo Zhang 6109		
● Haozhe Zha 实现功能,	so Zhang 6109	80 Zhani 80 Zhani 80 Zhani 80 Zhani 80 Zhani 80 Zhani	保其能涵盖回答问题的知识点,这里需要反复测试)  4. 尝试召回更少的片段,本质上还是减少输入prompt的长度  5. 尝试输出更少内容,本质上是减少输出prompt的长度	Bo Zhang 6109 Bo Zhang 6109		
■ Hongwei Liu 练和改善的  Amangatos  Baymangatos  Haozhe Zha  文明 文	so Zhang 6109	ng 6109 80 Zhani ng 6109 80 Zhani ng 6109	知识点,这里需要反复测试)  4. 尝试召回更少的片段,本质上还是减少输入prompt 的长度  5. 尝试输出更少内容,本质上是减少输出prompt 的长度	Bo Zhang 6109		
● Haozhe Zha 实现功能,	go Zhang 6109	go Zhanii go Zhanii ng 6109 go Zhanii <b>是</b> 如何进行训	本质上还是减少输入 prompt 的长度 5. 尝试输出更少内容,本 质上是减少输出 prompt 的长度	Bo Zhang 6109		
● Haozhe Zha 实现功能,	go Zhang 6109 go Zhang 6109 go Zhang 6109 gent的交互能力是如何	so zhani ng 6109 so zhani 是如何进行训	prompt 的长度  5. 尝试输出更少内容,本质上是减少输出 prompt 的长度	Bo Zhang 6109		
■ Hongwei Liu 练和改善的  Amangatos  Baymangatos  Haozhe Zha  文明 文	go zhang shang go zhang shang gent的交互能力是如何	<sub>80</sub> zhani <sub>80</sub> zhani <b>2</b> 如何进行训	质上是减少输出 prompt 的长度	Bo Zhang 6109		
■ Hongwei Liu 练和改善的  Amangatos  Baymangatos  Haozhe Zha  文明 文	jent的交互能力是如何	。 。 是如何进行训	prompt 的长度	Bo Zhang 6109		
● Haozhe Zha 实现功能,	;ent的交互能力是如何	是如何进行训	See Lianghua Mou			
● Haozhe Zha 实现功能,	;ent的交互能力是如何	是如何进行训				
❷ Haozhe Zha 么样的?看 实现功能,			我狭义地理解「大模型和 agent的交互能力」是	Bo Zhang 6109		
❷ Haozhe Zha 么样的?看 实现功能,			function_call 能力,即大模型判断应该使用哪个工具的	ao Zhang 6109		
❷ Haozhe Zha 么样的?看 实现功能,			能力	Box		
❷ Haozhe Zha 么样的?看 实现功能,			关于这块的训练和改善我本 人这边没有相关的研究,同 学可以参考业界的一些文章	7hang 6109		
			■ Hongwei Liu	602		
			Tiongwei Liu			
Bo Zhang		也可以用插件	提示词、插件的组合与工作 流的关系是怎么样的?	Bo Zhang 6109		
Bo Zhang 6109  Bo Zhang 6109  Bo Zhang 6109	请问什么类型的bot推	DOt推存使用	1. 提示词是整体设定或任 务整体的描述	Ro Zhang 6109		
2hang 6109  Bo 2hang 6109  Bo 2hang 6109  Bo 2hang 6109			2. 对于bot的提示词来说,			
Bo Zuaužejoa B			插件和工作流是等价 的,都是bot完成整体设 定过程中可以使用的工	6109		
Bo <u>Thana</u> .			具,所以他们都在「工具箱」里	Bo Zhane -		
- 6702			3. 对于一次任务执行来	- 00		
G Xvauge 2			说,可以使用工具,也可以不使用,大模型会	Bo Zhang 6109		
C Strauge 103  Bo Strauge 203  Bo Strauge 203			决策是直接解决问题还 是使用工具后解决问题	Bo Zhang 6109		
BO XHau8 e103			Haozhe Zhang			
Zhang6109			什么类型的bot推荐使用工作 流?	Bo Zhang 6109		
Bo Zhang 6109			9009 Bo Zhang 6109			
Zhang 6109			Thank 6109	Bo Zhang 6109		

o Zhang 6105			1. 工作流是对各类基础能力和业务能力,且其步骤固定,能自动化运
o Zhang 6109			行、独立运行。
Thank 6109			2. 援引 Coze 官方的文档 https://bots.bytedance .net/docs/guides/workf low
Q Z I III			当目标任务场景包含较多的步骤,且对输出结果的
<sub>2 Zhang</sub> 6109			准确性、格式有严格要求时,适合配置工作流来实现。
5109			Haozhe Zhang
12	Ming Yin	现在看到的prompt都比较简单,能分享 一个产品中使用的,比较复杂的prompt	不同的业务场景对 prompt 的要求是不一样的。
<sub>2 Zhang</sub> 6109		(产品级prompt)看看吗?	真实产品中的 prompt 往往 是多个,不只一个,不同的 prompt 预期大模型完成不
C109			同的能力。
13	🐍 Caiwen Fen	介绍一下工作流的相关创建和使用教程?	参考 coze 的 workflow 教程,比较全面:
7 Thang 6109			https://bots.bytedance.net /docs/guides/workflow
			Caiwen Feng
2nang 6109	Haozhe Zha     Anang 9199	答案?比如设定预期回答只回答答案选项,但是某次回答后,回答了附带的其他	边界情况的处理:  1. 通过提示词约束:在 prompt 中加入一些强约 束的 prompt 提示词
		信息; Bo Zhang 6109 Bo Zhang	2. 通过提示词约束+示例 (few-shot): 比如我 们可以输入一些示例,
3 Zhang 6109			使其更好地遵循输出的 格式和要求
Zhang 6109			3. 通过输出格式要求(部分大模型有效): 部分大模型(如高版本的
Zhang 6109			GPT)支持了 JSON 模式的输出,使用 JSON模式可以精确控制其输出的字段。
<sub>2 Thang</sub> 6109			在 coze 中使用时,可以 通过模型设置-输出格式 配置项来要求其输出
<sup>2</sup> Zhang 6109			
Zhang			如何能发现预期外的答案:
15	Ming Yin	如果我们使用了知识库(涉及多个文 档),如何在回答中标明具体是哪个文档	如果使用 coze,可以使用 coze 知识库标注来源的能力
o Zhang v		中的呢?	

|调用了LLM 豆包 function call模型,为什···



Jingqiao Chen 5:08 PM Aug 26

Functioncall 是指大模型作为 agent 进行工具调用的识别,因 为 coze 编排场景下需要调用很 多工具,所以建议用这个。方舟 现在主要分 pro lite 两个版本, pro 里还有 functioncall、 character 等分支版本。如果需 要调用工具就用 pro-fc,需要做 对话机器人那种就用 procharacter,想快还省钱就用 lite



Changyu Wei 5:36 PM Aug 26 其他模型不能做"工具调用"吗? 还是 functioncall 版本对"工具调用"特别训练讨?

### Peifeng Liu  ### Windows American	hang 6109			授金制版 ○
● Peifeng Liu		80 Zhang 6109		新日間が知识を開始が、新日本集体的・アドス特性介質を対文 作。 第7年中的が国際を示すを使わったが日本教師のであるが表す。 第7年中的が国際を示するである。 第7年中のでは国際を示するが、表現の場合が対象では 第2年 元を記載を持て予定するが、基本の場合が表現を 第2年 元を記載を持て予定するが、基本の場合が表現を 第2年 元を記載を持て予定するが、基本の場合が表現を 第2年 元を記載を持て予定するが、基本の場合が表現を
● Peifeng Liu  那种模型最好?  目前看知识库内容比较少,召回的知识是 不是都会给塞到prompt中? 也会使用 token? 知识库召回是什么方式? 向量数据库? 还 是让大模型做召回? 或者是豆包内部的机 制? 如果我的知识库非常大(Gb级别)还能 这样简单外挂知识库吗?  那个调用知识库插件获取的输出是什么?  如? Yanxin Wan 如果想输入一个仓库的代码作为知识库, 然后让模型学习对应的代码编写风格,有 没有什么办法  ● Yanxin Wan  如果想输入一个仓库的代码编写风格,有 没有什么办法  如果和识学习编写风格,直 接通过提示词要来其遵循固定的代码编写风格,有 没有什么办法  1. 数据准备:准备N个典 的代码片段,让大模型 分析其风格,并写入到 片段中		80		1 加子文档: 知己在自由 图示未准 ②
那种模型最好?  目前看知识库内容比较少,召回的知识是 不是都会给塞到prompt中?也会使用 token? 知识库召回是什么方式?向量数据库?还 是让大模型做召回?或者是豆包内部的机 制? 如果我的知识库非常大(Gb级别)还能 这样简单外挂知识库吗?  那个调用知识库插件获取的输出是什么? 如识库活件,其输出是知识库中的文档片段  ***********************************	16			参考该问题的评论区回答 by
→ Changyu W 不是都会给塞到prompt中? 也会使用 token? 知识库召回是什么方式? 向量数据库? 还 是让大模型做召回? 或者是豆包内部的机 制? 如果我的知识库非常大(Gb级别)还能 这样简单外挂知识库吗?  那个调用知识库插件获取的输出是什么?  在今天(0826)的资示中的 知识库插件,其输出是知识库中的文档片段  ***********************************		Pellelig Liu		Jingqiao Chen
是让大模型做召回?或者是豆包内部的机制?如果我的知识库非常大(Gb级别)还能这样简单外挂知识库吗?  那个调用知识库插件获取的输出是什么?  在今天(0826)的演示中的知识库抽件,其输出是知识库中的文档片段  □	17		不是都会给塞到prompt中?也会使用	参考该问题的评论区回答 by  Singqiao Chen
如果我的知识库非常大(Gb级别)还能 这样简单外挂知识库吗?  那个调用知识库插件获取的输出是什么? 在今天(0826)的演示中的 知识库插件,其输出是知识 库中的文档片段  「***********************************		<sub>Bo Zhang</sub> 61 <sup>09</sup>	是让大模型做召回?或者是豆包内部的机	
知识库插件,其输出是知识库中的文档片段		BO Zhang 6109	如果我的知识库非常大(Gb级别)还能	
如果想输入一个仓库的代码作为知识库,然后让模型学习对应的代码编写风格,有没有什么办法  如果确需使用知识库,有以下几个建议 1. 数据准备:准备N个典词的代码片段,让大模型分析其风格,并写入到片段中				在今天(0826)的演示中的知识库插件,其输出是知识
如果想输入一个仓库的代码作为知识库, 数据推备:准备N个典述的代码片段,让大模型分析其风格,并写入到 片段中				
如果想输入一个仓库的代码作为知识库, 如果知识学习编写风格,直 接通过提示词要求其遵循固定的代码风格。 如果和识学习编写风格,直 接通过提示词要求其遵循固定的代码风格。 如果确需使用知识库,有以 下几个建议  1. 数据准备:准备N个典 的代码片段,让大模型 分析其风格,并写入到 片段中		Bo Zhang 6109		輸出 の Bo Zhang 610
对应的是知识库中文档的分段结果  如果想输入一个仓库的代码作为知识库, 然后让模型学习对应的代码编写风格,有 没有什么办法  如果知识学习编写风格,直 接通过提示词要求其遵循固定的代码风格。 如果确需使用知识库,有以下几个建议  1. 数据准备:准备N个典语的代码片段,让大模型分析其风格,并写入到				coston. 深入開業 go reflected/加加品企業開等 1/14. 天儿藝術 go respire go 常用场等的一点 天地藝 go light go
₩ Yanxin Wan  如果想输入一个仓库的代码作为知识库, 然后让模型学习对应的代码编写风格,有 没有什么办法  如果知识学习编写风格,直 接通过提示词要求其遵循固定的代码风格。 如果确需使用知识库,有以下几个建议  1. 数据准备:准备N个典的代码片段,让大模型分析其风格,并写入到片段中				差型 果及有使用不当也含造成性能下限。可能性差等问题。い2 co 反射 か機 時いつ1 中州 年 相 にの co 正 草葉 基本 中 中 が c 在 一 小 空 最 裏 本 一 小
如果想输入一个仓库的代码作为知识库,如果知识学习编写风格,直接通过提示词要求其遵循固定的代码风格。如果确需使用知识库,有以下几个建议  1. 数据准备:准备N个典的代码片段,让大模型分析其风格,并写入到片段中				段结果
如果想输入一个仓库的代码作为知识库, 如果知识学习编写风格,直接通过提示词要求其遵循图 定的代码风格。 如果确需使用知识库,有以下几个建议 1. 数据准备:准备N个典的代码片段,让大模型分析其风格,并写入到片段中				SAME OF THE PARTY
如果想输入一个仓库的代码作为知识库, 如果知识学习编写风格,直接通过提示词要求其遵循固定的代码风格。 如果确需使用知识库,有以下几个建议 1. 数据准备:准备N个典的代码片段,让大模型分析其风格,并写入到片段中				SE STANDONNES, DESTANDONNES, DESTANDA, TERRESE DE. SER-MAIN-THETRAVILARIS BIRS. SET-TEXTES. MIR. MIR. DUIN (nt = 5.
<ul> <li>★ Yanxin Wan 然后让模型学习对应的代码编写风格,有 没有什么办法</li> <li>按通过提示词要求其遵循固定的代码风格。</li> <li>如果确需使用知识库,有以下几个建议</li> <li>1. 数据准备:准备N个典额的代码片段,让大模型分析其风格,并写入到片段中</li> </ul>		3 hang 6109		80 Yrang 610:
如果确需使用知识库,有以下几个建议  1. 数据准备:准备N个典的代码片段,让大模型分析其风格,并写入到片段中			然后让模型学习对应的代码编写风格,有	如果知识学习编写风格,直接通过提示词要求其遵循固定的代码风格。
的代码片段,让大模型 分析其风格,并写入到 片段中				如果确需使用知识库,有以 下几个建议
片段中				1. 数据准备:准备N个典型的代码片段,让大模型分析其风格。并写入到
		B0		片段中

66

Jingqiao Chen 5:38 PM Aug 26 (e···

@Changyu Wei 也能,但 fc 的版本有针对性训练,可以理解为在基础模型上通过 fc 的数据做了微调,效果更好。具体得看方舟的文档,我记得有些模型是不支持传递 fc 参数的,比如lite128k

目前看知识库内容比较少,召回的知识是…



Jingqiao Chen 5:45 PM Aug 26 知识库内容多少都得塞 prompt 里,如果纯 RAG 是在 prompt 里,如果是用 FC 实现的 RAG 就 是在特殊的 fc 参数里,都是占 token 的

知识库召回是什么方式? 向量数据库? 还…



Jingqiao Chen 5:47 PM Aug 26 搜索策略里的语义就属于向量检 索,将知识库向量化成库,然后 基于输入检索 topN 相似。全文 检索可能就类似于 es 那种,提 取关键词,在文档中搜索关键 词。混合应该就是两者结合排 序,具体豆包的策略不太确定 了。

混合①

如果我的知识库非常大(Gb级别)还能…



Jingqiao Chen 5:49 PM Aug 26 外挂大的知识库,coze 会自动 做分段,比如分成一小段一小段 的段落,RAG 检索出来的是一个 段而不是一整个知识库,检索的 段会作为补充信息提供给模型, 所以大库可能会影响召回的速度



2. 知识分段:使用知识库的自定义分段功能,确保同一逻辑块的代码及

其说明是完整的

力

3. 知识召回:这里推荐使 用混合方式召回,兼有 语义和关键词匹配的能 Changyu Wei 5:50 PM Aug 26 感谢解答

和准确度,但用应该问题不大。

Zhang 6109		Bo Zhang 6109  Bo Zhang 6109		4. 提示词:通过提示词来 提示大模型参考召回片 段的功能等信息,提取 其风格,作为其输出的 参考
gard "		Bo Zhang 6109		少亏 这个问题我可能没有很清晰 的理解,同学可以找我私聊
Zhang 6109		3p Zhang 6109		后我补充 🌑 Yanxin Wang
20		想请教一下,后续RAG	i底层逻辑有考虑用	很好的建议
Zhang 6109	Haoyi Niu	GraphRAG的技术吗		GraphRAG 通过聚合社区、 生成总结来解决总结性 问
		Bo Zhang 6109		题,如「xxx 有的前5个是什么?」、「本书的主旨是什
Zhang 6109		<sub>50</sub> zhang 6109		么」等问题,较之于朴素 RAG(即当前召回知识片段
Zhang 6109		80 Zhang 6109		的形式),GraphRAG 能弥补这方面的缺陷。
		Bo Zhang 6109		在答题 bot 中,长远看是需要类似技术的,因为选择题
<sup>2H3UE</sup> 6709		80 Zhang 6109		中也会出现类似「xxx 不包 含以下哪个?」这样的概览
Time.		Bo Zhang 6109		性/总结性问题,此时需要该类能力
Zhang 6109		<sub>30 Zhang</sub> 6109		Haoyi Niu
21	Jie Zheng	1.被检索到的知识库内 喂给大模型的么?提示		1.被检索到的知识库内容, 是作为提示词喂给大模型的
Zhang 6109		2.当一个大文档被切成 coze框架是会通过很多 到大模型么?		么?提示词大致是怎样的? A:不是提示词,是使用了大 模型多个 message 的能力
Zhang 6109		3.coze是基于LangCha	ain实现的么?	) "3:(
Zhang 6109		80 Zhang 6109		"Content": ""Content": ""Content": ""Content": ""Content": ""Content": ""BDL世紀 周囲区 Path 从图 Path 景色 北京 (四世)200 至后 - Path 景色 北京 (四世)200 至后 - Path 景色 大阪 (四世)200 至后 - Path 第二十二十二十二十二十二十二十二十二十二十二十二十二十二十二十二十二十二十二十
		Bo Zhang 6109		如 debug 调用图所示,
Zhang 6109		<sup>30</sup> Zhang 6109		knowledge_plugin 的返回 结果,会被组织成 1 个
G109		go Zhang G109		role 为 tool 的消息, 传给大模型
Zhang 614		<sub>50</sub> zhang 6109 - 6109		2.当一个大文档被切成很多
Zhang 6109		80 Zhang 6109		的chunk时,coze框架是会 通过很多轮对话把知识给到 大模型么?
Zhang 6109		Bo Zhang 6109		A: 是的,即上个问题提到 的 messages 能力
		Bo Zhang 6109		6109 Thank 6109
Zhang 6109		<sup>20</sup> Zhang 6109		3.coze是基于LangChain实 现的么?
		Bo Zhang 6109		A: 这个我不了解,可以咨询 coze的官方同学
Zhang 6109		30 Zhang <del>0103</del>		Bo Zhang b199

#### 11. 威胁情报小助手:设计与实践 🚳 Yu Ye

ID	提问者 Questioner (@自己哦)	问题 Question	30 Zhang 610 <sup>9</sup>	回答 Answer	
ang 6109	M. Lluon 7h	6070用的知识东南西施州州西		Bo Zhang 6109	
1	Whuan Zhao	coze里的知识库需要额外处理 么?是啥格式的。涉及公司内部	Bo Zhang 6109		
-100		的一些资料怎么放在coze知识库			
<sub>1ang</sub> 6109	<sub>Bo Zh</sub> a	的,是否有风险。			
2	Cola Wang	如何接入卡片样式 / 给运营使用	Bo Zhang 6109		Bo Zhang 6109
1ang 6109	Ro Zha	要收费嘛(Coze token成本~			
3		意图识别是调用第三方算法模块	-0		
	Huo Yanwen	吗?	Bo Zhang 6109		
1ang 6109	Bo Zha				
4	∅ ト乐	用户prompt和系统prompt是怎	30 Zhang 6109		Bo Zhang 6109
-00	Во Упана	样关联和交互的?	Bo Zharra		
Jaug <u>e 103</u>	<b>参</b> 卜乐	多种不同的意图,最终输出格式			
	Bo Zhang 6109	应该也不一样,是怎样在工作流	ao Zhang 6109		
6109	80	里统一输出的?	00		
6	Bo Zha	对于每一个接口返回的数据如何		Bo Zhang	
	🚳 Weisong Wang	解读,都需要一个Prompt提示大	Bo Zhang 6109		
-6109		模型怎么解读吗?			
1ang 6109	Bo Zha	Bo XHaugeroa			
7	Bo Zhang 6109		Bo Zhang 6109		
1ang 6109	Bo Zha				
8	Yunfeng Guan	做了多久?研发成本多少呀?ROI	Bo Zhang 6109		
1ang 6109	Tullieng Guan	Bo Zhang 6109			
9	Bo Zhang 6109	能详细介绍一下拒绝回答的部分	Bo Zhang 6109		
		是怎么做的嘛? 对于涉黄涉政的	Bo Zharro		
1ang 6109	Bo Zha	问题,公司有统一的解决方案			
	Bo Shang 6109	嘛?	so Zhang 6109		
10		添加了3个工作流,根据什么确定	2021	- <109	80 411
Jank	Bo Zha	调用哪个工作流			
11	Bo Zhang 6109	Bo Zhang e109	30 Zhang 6109		Bo Zhang 6109
12	Bo Zha				
12					
13	Bo Zhang 6109		30 Zhang 6109		
14	Bo Zha				
15	Bo Xhang 6109	Bo Zhang 6109	Bo Zhang 6109		Bo Zhang 6109
TO			80 -11		
ong 6109					
16	Bo Zha				
16 17	Bo Zhang 6109 Bo Zha		30 Zhang 6109		

Transformer 原理剖析



Qigeng Chen 5:43 PM Oct 14 这个是 8 月的课程吗? 9 月还会 讲不?

Transformer 原理剖析



Qigeng Chen 5:49 PM Oct 14 有录屏链接没?



Jingyi Qiu 7:17 PM Oct 14

@Qigeng Chen 录屏链接 https://bytedance.larkoffice.c om/minutes/obcnfqaewd39jta y188z14ka?from=auth\_notice

BO Zh 19 109	Bo Zha		
20			
Bo Zhang 6109			

#### 12. Transformer 原理剖析 Scheng Zhao

ID	提问者 Questioner	问题 Question	回答 Answer	Bo Zhans	Bo Zhang 6109		
	(@自己哦)	20 Xnang 6109	50 Xhang 6109	20 Zhang 6109			
1 ang 6109	Hesheng Pu	LSTM的门控的系数是怎么来的? 训练出来的吗?也是一个 embedding吗?	80 XJraug 6109	- e109	Bo Zhang 6109		
2 ang 6109	🙉 Xiaohan Wang	比如seq长度为100时, LSTM/GRU的unit参数数量也是	是的 <sub>go Zhang</sub> 61.09	Ro Zhang a	Bo Zhang 6109		
	Bo Zhang 6109	一套,只不过hidden特征保存 $100$ 次吧?那是不是LSTM/GRU的核心还是hidden( $h_t$ or $c_t$ )吧?	30 Zhang 6109 80 Zhang 6109	Bo Zhang 6109	Ro Zhang 6109		
3	o Mingrui Zhang	exp是指数么? 感觉有点像 softmax	30 Zhang 6109	Bo Zhang 6109			
	Bo Zhan		Bo Zhang 6109		Bo Zhang 6109		
4	Hesheng Pu	Position embedding,是不是跟 文本的风格相关,比如说鲁迅风	30 Zhang 6109	Bo Zhang 6109	6972		
	Bo Zhang 6109	格,拿鲁迅作品训练出来后,模型输出的语言风格就是鲁迅的那种风格?这里是否可以引申到模仿一个人说话习惯上?	Bo Zhang 6109	Bo Zhang 6109	Bo Zhang 6109		
6	Dong Liu	q、k、v中的v如果去掉会对模型 造成多大影响呢。已经有后面的	Bo Zhang 6109	Bo Zhang 6109	Bo Zuana		
	Bo Zhang 6109 Bo Zhang 6109	全联接层了,对信息的解读是不是可以放到后面,还是说是为了使用输入embedding到其他模型,加入v让输入embbeding训练的更通用?	Bo Shaug 6109	Bo Zhang 6109	Bo Shauk eroa		
7 ang 6109	Tingkang Zhac	计算注意力得分的时候,为什么 要先对x进行转换(x乘上W_q, W_k),而不是直接用两个x直接	Bo Zhang 6109	Bo Zhang 6109	Bo Zhang 6109		
	Bo Zhang 6109	算内积得出注意力得分?	30 Zhang 6109	Bo Zhang 6109			
8 <sub>8</sub> 6109	Hesheng Pu	Wq Wk Wv是怎么来的呢? 一个模型应该有很多层的encoder,每一层的W,是按什么方法来确	Bo Zhang 6109	PATE	Bo Zhang 6109		
	Bo <u>S</u> ysug <sub>9 Tha</sub> s	定是最优的呢?	Bo Zhang 6109	Bo Zhang 6103	Bo Zhang 6109		
9	🥯 Xiaohan Wang	为什么 $QKV$ 计算注意力那里要除以一个 $\sqrt{dk}$ ,感觉不除以这个也不影响 $cottmax$ 计算吧?	https://arxiv.org/pdf/1	Bo Zhang 6109	0		
	Thank 6109	不影响softmax计算吧?	防止梯度 <del>过大(</del> 过小 <del>?图的防止推向小梯度区域)</del>		Bo Zhang 6109		
	Bo Zhan		While for small values of $d_b$ the two mechanisms perform similarly, as do product attention without scaling for larger values of $d_b$ [3]. We say a find $d_b$ , the dot products grow large in magnitude, pushing the softmax function extremely small gradients $^4$ . To counteract this effect, we scale the dot	dditive attention outperforms spect that for large values of tion into regions where it has products by $\frac{1}{\sqrt{d_k}}$ .	Bo Zhang 6109		

			假设q和k均值为0、方差为1,那 $q$ k的点积均值为0、方差为 $d_k$ ,除以 $\sqrt{d_k}$ 可以降低点积的幅度,进而避免softmax陷入梯度很小的		
			优化状态.		
			● 感谢解答!		
10	Shiwei Zhong	为什么现在的LLM基本都采用 decoder-only的架构,舍弃 encoder的原因?	https://www.youtube.com/wat ch?v=orDKvo8h71o 该视频有提到decoder的优势		
		attention加速或者内存优化的思 路大概是什么样?现在	Bo Spaug e 103		
		embedding维度都比较高,计算 kv时信息量其实主要集中在部分 维度,长尾的部分可以忽略,类 似这种优化思路?	Bo Zhang 6109  Bo Zhang 6109		
11 hang 6109	© Zhe Pan	为什么现在的推理,不需要加类似CRF层,来辅助学习&计算生成句子的全局最大概率值?	Bo Zhang 6109  Bo Zhang 6109  Strang 6109		
12		生成任务的前提是不是要先理解呀,为啥会特地区分这两个呢?	Bo Shaug et Oa		
13	Dong Liu	位置编码的时候提到后续llama等 模型想要优化成主要使用相对位 置信息,请教一下这是说绝对位	Bo Zhang 6109  Bo Zhang 6109		
		置在文本长度增加后可能会造成 过拟合什么的所以会想要弱化绝 对位置信息吗	Bo Zhang 6109  Bo Zhang 6109		
14		为什么大模型推理不用greedy, 而是用sample呢?	30 Zhang 6109		
15		了解了这些原理有啥用,在使用 大模型的时候知道原理可以让我 们怎么更好的使用它	最直观的例子:懂得怎么调整 GPT的输出,也方便对比不同模 型的效果		
16	Bo Shaug e109	大模型加法计算算不准是不是和 position ebedding也有关系,可 以用ebedding的方式提升正确率 吗	加法不准主要还是tokenizer分词 的原因,比如12+23=35可能分 成"1""2""+"…失去了语义 信息		
		Gpt decoder-only,输入信息是在哪里处理?是把输入作为decoder输出的一部分吗?	是的,相当于直接对输入信息做 self-attention,但不会像 encoder那样只把KV传下去		
<sup>Iang</sup> 6109	Bo Zhaug <sub>2103</sub>	CV中,ViT作为backbone训练时 对输入图像尺寸有要求,因为需 要按patch切分。NLP中	直接用tokenizer切好就行了,之 前会有文本长度的限制,现在基 本没了		
		Transformer对输入有什么要求 么?	Bo Zhang 6109  Bo Zhang 6109		
		Encoder 和 Decoder 的本质区别是什么?为什么前者适合生产,后者适合内容理解?	是否+mask 原始设计中,encoder 阶段的输 出为输入文本的 Embedding;		
			Decoder 的输出为逐字生成的新文本。		
		训练的过程中,对于喂进去的语 料,损失函数如何计算。比如说	BO Zhang 6109		

最直观的例子:懂得怎么调整GPT的输出···



Yunfeng Guan 5:33 PM Aug 21 感觉不太好对比,可能是我太菜 (狗头

Bo Zhang 6109		现在有《静夜思》这首诗,是 mask其中某几个token让模型输 出,还是让他预测某句诗,看和 实际静夜思诗句的差异?	30 Zhang 6109	Bo Zhang 6109
- 6109	Lingfeng Ren	对于问题16,如果模型训练时能够看到12+23=35这样的方程,即使tokenize分开了,也能通过位	Bo Zhang 6109	Bo Shang 6109
Bo Zhang u		置编码区分,学到如果1后是2, 就是12,那是因为训练数据的问 题,没有学到足够的样本吗?	Bo Zhang 6109	Bo <u>Zhau</u> g 6109

#### 

活动时间: 2024.8.16 16:00-17:00

ID	提问者 Questioner	问题 Question	回答 Answei	Bo Zhare
	(@自己哦)		mans 6109	
1		• 原子能力是否支持二次开发	50 2119	
	Yuping Xing	<ul><li>对于和langchain没有打平的 能力或者新增的能力怎么能快</li></ul>	Bo Zhang 6109	
		速支持	Bo Zhang 6109	
		• 对于plugins接入,这部分输 出需要在最后节点写入生成回	Bo Zhang 6109	
		复使用的prompt么	Bo Zhang 6109	
		• Eino对于离线的	Bo Zhang 6109	
		chunk/embedding写入有框 架或者方法么	Bo Zhang 6109	
<sub>mam</sub> 2 <sup>09</sup>	(9) Hui Cai) BOZNA	agent在解析模型返回的时候,怎 么确保模型返回的数据结构一定	BO XHaud e109	
		符合parser的预期。以两数相加 为例,怎么确保能从大模型返回	30 Zhang 6109	
		中提取出 加数 和 被加数。	Bo Zhang 6109	
3	80 Zhang 6109	这个框架目前是仅限字节内部使	30 Zhang 6109	
	🌘 Jiashuai Lu	用吗?	Bo Zhang 6109	
4		这个rag方案都是用	Bo Zhang 6109	
		vectorRetriver转化成原文本,格 式化prompt里再去请求模型吗?	Bo Zhang 6109	
		可以直接用embddings加	Bo Zhang 6109	
		prompt 混合请求吗?(主要考虑节省token花费)	Bo Zhang 6109	
5	Bo Zhang 6109	eino和vikingdb有对标的类似开	langchain	
	S Liang Cheng	源框架吗? 想都比较着看看	Bo Zhang 6109	
6	🖐 xiakai	求一份java的案例	Bo Zhang 6109	Bo Zhang 6109
nang 6109	🎅 Tao Li) 🔞 🖂	RAG支持哪些数据格式	Bo Zhang 6109	

求一份java的案例



Yongquan Zhang 4:30 PM Aug 16 +10086

Zhan 8 09	Yongquan Zha	embedding的是回答还是问题? 怎么处理数据安全问题,比如 vectordb中有租户的信息?	Bo Zhang 6109	
zhan <sup>g</sup> <b>g</b> 109	Peng Li 80 2000	rag中面临不同的请求时,怎么确 定不同请求对应的learning的数 据来源呢?	80 Zhang 6109	Bo Zhang 6109
10	Fangyong War	RAG提供最新训练数据有哪些方	Bo Zhang 6109	Bo Zhang 6109
2hang 6109	Weiwei Li) 80 200	通过调用eino的哪个函数将飞书 文档、公司内代码库给到模型, 类似RAG?或者eino代码库中哪个 example最接近这个需求呢?	Bo Zhang 6109  Bo Zhang 6109	
<sub>Zhang</sub> 6109	Bo Zuau Bo Zuauß e109	目的:大团队有很多设计文档、 代码库、且总是变化,让模型读 懂这两类资料后,每次想了解团 队子方向在代码中做的策略、规	Bo Zhaug e109	
Shaug 6109	Bo Zhang 6109	则、设计等,可以直接调用接口 提问获取。 eino能做这个么?	30 Zhang 6109 Bo Zhang 6109	
12 <sub>Zhang</sub> 6109	Peng Li	Eino有没有私有化的案例,帮助 私有化客户或者服务在本地通过 大模型进行问题分析和答案生成	Bo Zhang 6109	
13 <sub>Chang 6109</sub>	Chao Zhang	RAG方案里对模型来说是两个输入还是1个输入?如果是2个输入为什么只有一个输出?	Bo Zhang 6109	
<sub>than 8</sub> 6109	Qimo Zhou	大模型调用计算器这一部分是怎么做的? 是在prompt里给出了函数调用的指令? prompt大概是怎么写的呢?	Bo Zhang 6109  Bo Zhang 6109  Bo Zhang 6109	Bo Zhang 6109
hang 6109	Bo Zhang 6109	function call 的时候,如果大模型的语义理解生成的参数格式不满足 fucntion struct 定义,是会直接抛错吗	Bo Zhang 6109	
£109	Bo <u>X</u> haug <del>o ro</del> a	内部有哪些服务接入了?	30 <u>Zhang 610</u> 0	Bo Zhang <del>G10</del> 9
haug 6109	Bo Shang 6109	RAG对接入ES或者其他的数据 源,有已支持的吗,还是都需要 自己写代码适配	Bo Zhang 6109	
Alter to	Bo Zhang et në	Bo Zhang Gina	Bo Zhans G n9	Bo Zhang et 103

## 14. 使用Fornax进行Prompt开发迭代实战 Chen Yang

活动时间: 2024.08.13 17:00-18:00

#### 15. 基于大模型定制专属版"今日头条"

Yawei Zou

活动时间: 2024.08.08 17:00-18:00

ID	提问者 Questioner <sup>******</sup>	问题 Question		回答 Answe	er	Bo Zhang 6109		
	(@自己哦)		30 Zhang 6109		Bo Zhang 6109			
2 6 6109	Bo Zha	是否可以直接进行智能过滤,省				Bo Zhang 6109		
	Baoan Zhang	掉常规过滤,智能过滤怎么做						
	Bo Shaug <sub>0109</sub>		Bo Zhang 6109		Bo Zhang 6109			
4 <sup>18 61,09</sup>	go Zha					Bo Zhang 6109		
5	Yao Xiao	到大语言模型这一步,有多少数	Bo Zhang 6109		Bo Zhang 6109			
	go Zha	据量级呢,llm的qps多少				Ro Zhang 6109		
6	00	怎么通过大模型判断文章质量的						
O	Honglei Wang		Bo Zhang 6109		Bo Zhang 6109			
	an The	ng 6109				20 Zhang 6109		
_		A D N A D T + \ A \ A \ A \ A   A   B   B   B   B   B   B   B   B				Bo		
7	🐲 Jinlei Pan	今日头条是否在这个方向上有相关的产品规划?个人投入的开发	Bo Zhang 6109		Bo Zhang 6109			
	20.7h8	资源和计算资源大约多少?				20.7hang 6109		
	80 2					Bo		
9	Bo Zhang 6109	整个系统的日均输入(原始爬取	Bo Zhang 6109		Bo Shaug 6109			
	& Zhenxing Xu	的文章)是多少,以之前的事件 经验来看,如果关注100个站点是				20.7hang 6109		
	80 =	在一个时间的一个时间, 在一个时间的一个时间,				80 =		
	Bo Shaud e109	Bo 2hans 6109	Bo Zhang 6109		Bo <u>Xp3</u> UE e108			
10		能做到个性化推荐吗				20.7hang 6109		
	Xiongzheng Li					80 =		
	Bo Sysua e108		30 Zhang 6109		Bo Zhang 6109			
11,09	so 7ha	当前成本怎么样,是否自己SFT				20 Zhang 6109		
	Maichao Guo	过				Bo		
	Bo ShauB e109		Bo Zhang 6109		Bo Zhang 6109			
12	so Zha	智能抓取的prompt方便分享吗?				20 Zhang 6109		
	Yunfeng Guan	整个过程中有做什么pe优化的事				Bo		
	Bo Shaug 6109	情? Bo Zhang 6109	30 Zhang 6109		Bo Zhang 6109			
	go Zha					Ro Zhang 6109		
13	Hu Wan	openai的效果和国产模型差别大						
	Bo Shang 6109	吗,在中文文章理解方面	Bo Zhang 6109		Bo Zhang 6109			
ang 6109	Bo Zha	<b>左江左阳四八十十十八十十八十十八</b>				Bo Zhang 6109		
15	Yuan Yao  Yao  Yao  Yao  Yao  Yao  Yao  Yao	有没有遇到过 LLM 的幻觉问题, 或者回复的格式不是希望的 json						
	Bo Zhang 6109	之类的问题?	Bo Zhang 6109		Bo Zhang 6109			
aug <u>e100</u>	Bo Zha	ng 6109		Bo Zhang <del>p103</del>		Bo Zhang 6109		
16	Xin Zhang	如何评价模型推荐的文章的质量						
	Bo Xhang 6109	是不是满足预期的效果?	Bo Zhang 6109		Bo Zhang 6109			
17	go Zha	中英文文章都有吗?LLM判断的				Bo Zhang 6109		
	Linjiang Xie	准确率针对不同语言有区别吗?						

2 Zhang 6109	Bo Zha				
18		怎么知道模型给的	分数是不是符	Bo Zhang 6109	
	Weiming Zhen	合预期的? 有没有	事实数据或者		
Zhang 6109	80 Zps	指标量化分数质量	这事情咧?是		
		否可以给模型设置 那种,让他自己选		30 Zhang 6109	
Zhang 6109	Bo Zha	化目标呢?			
19				80 Zhang 6109	
20	Bo Zha				
				Ro Zhang 6109	
<sub>7hang 6109</sub>					