Comments (19)

LLM学习季Q&A-合集

1. 大模型部署运行 王昱 Yu Wang

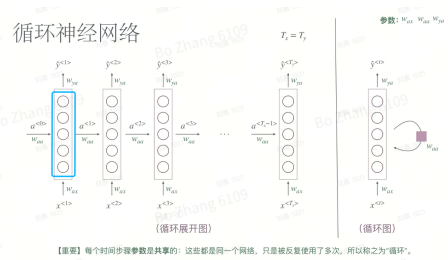
ID	提问者 Questioner (@自己哦)	问题 Question	回答人	回答 Answer
1	Yuheng Li	PyTorch模型转成ONNX后，推理性能有啥变化吗？推理结果、精度是否会有影响?	王昱 Yu W	不开启任何量化优化算法，模型转换过程成功的话，理论上推理结果和精度是不会改变的； 如果实际操作出现结果不一致问题，可能是onnx版本不兼容，或者使用了onnx不支持操作后导致转换出来的模型存在异常，需要排查一下 Yuheng Li
2	Yuheng Li	字节内部署llm服务，Laplace框架和bernard是啥区别和关系呢?	王昱 Yu W	我的理解是，这里比较类似kitex框架和tce平台的关系，前者是框架，由具体代码使用，后者是服务部署平台，管理部署机器和代码运行方式
3	Jinhui Wu	在线推理的优化策略（例如float32->float16)对推模型效果的影响，有什么评估方式么?	王昱 Yu W	目前llm推理一般使用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，需要额外验证； 可以参考正常模型训练的验证流程，在验证集上验证loss、准确率等相关指标是比较常见的做法； 对于业务场景，也可以把量化模型作为新的模型，ab验证； 另外我看到有通过kl散度等验证2个分布差异性的方法，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 Jinhui Wu
4	Zelin Yu	针对大语言模型，怎么评估训练模型的好差	王昱 Yu W	1. 训练loss稳定下降；

				2. 在不同测试集、针对各类不同任务都有不错泛化性 Zelin Yu
5	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 W	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备的，就像线上也能用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
6	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 W	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； 而vllm实际上和tensorrt-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的核心能力；使用vllm也可绑定其他更合适的服务框架来使用。 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，生态是比较重要的，如果框架不支持当前模型也没辙； 实际性能差距还需要根据实际模型比较，网上也能找到2者在特定模型上的对比，整体性能差距不大，老版本的tensorrt-llm可能在高并发有些问题，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 W	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推理引擎是竞品关系，是二选一的 Junling Du
8	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 W	这块细节不太确定，我理解是对应模型部署的需要的单位资源，和模型大小有关，不是强对应单一gpu；当然具体是否是这样可以咨询方舟同学

				Zhendong Xie
	Kai Xu	国际化的机房环境，包括TTP，可以用LLM Server做推理服务么？	Yu W	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 Kai Xu
	Junling Du	模型部署有使用ray serve来部署模型服务的吗？ray serve和哪些框架结合的比较紧密	Yu W	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
	Junling Du	多卡或多节点的模型部署一般使用什么框架和或组件？有哪些框架或组件支持根据负载自动扩缩容的功能？	Yu W	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 Laplace-Python3 使用指南； 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，可以看部署平台是否支持（如 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

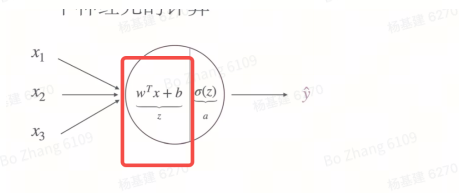
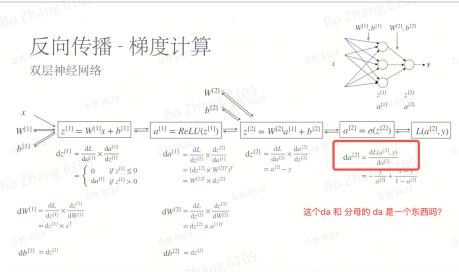
Zhen Wang

ID	提问者 Questioner (@自己哦)	问题 Question	回答人	回答 Answer
1	Chen Liu	1. 这里多个圈就是代表多个RNN层吗？a1 a2 a3 		圈是多层神经元
2	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？		
3				
4				
5				
6				

7				
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 Questioner (@自己哦)	问题 Question	回答人	回答 Answer
1	Jijian Yang	神经元里的逻辑都是线性的吗？ 		
2	Chen Liu	激活函数除了PPT介绍的，可以自己随便设计一个么		
3	Rui Qiu	relu 0 处的导是人为定义的吗看上去那里不可导		
4	Keyang Zh	1. 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？ 2. 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		
5				
6		NLP基本原理之后会有详细的讲解吗		
7				
8				
9				
10				

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)	问题 Question	回答人	回答 Answer
1	Yuchi Wan	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?		
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		
3	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吗?		
4	Jijian Yang	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?		
5	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?		
6	yuefeng li ，	有没有从根本上可解释的神经网络热门研究? 感觉无论线性回归、深度学习还是Transformer各种w，b数值没法解释含义		
7		实际上数据不可能无限大，怎么提高模型效果？只能靠改算法结构么?		
8				

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=830033faa86cb4c3e9707809d77e8d16

5. Prompt Engineering 指北 Zhenyu Zhao Xiaochen Luo

ID	提问者 Questioner (@自己哦)	问题 Question	回答人	回答 Answer
1	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	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构化表达/金字塔原理，

人类把事情讲清楚的比较好的方式就是结…

刚铠 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

				所以这个实际上是类似的。
2	Jiaxin Ma	1. Prompt 架构有哪些必备的基本元素？（比如角色/技能/约束...） 2. ~~Prompt 质量的三个标准怎么度量~~（PPT 中已包含）	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	Shengjie	对于大模型输出长度有限制，比如输出一个json，但是到一半的时候就截断了，有什么解决方案/思路么？	Zheny	如果触发了大模型的上下文长度限制，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		大模型底层是 怎么区别不同用户的。 举个例子： 如果多个用户发下面内容，大模型如何区别不同用户 • 我：你知道我的名字吗？ • 大模型：不知道 • 我：我叫铭铭 • 大模型：好的，铭铭你好 • 我：我叫什么名字？ • 大模型：你刚才告诉我叫铭铭了	Zheny	1. 大模型底层无法区别不同用户，大模型永远只是一次添加一个词。 2. 如果你希望大模型携带上下文，你可以在每次输入给大模型时都携带上“用户姓名：铭铭”。在 coze 里这个组件叫 variable。
5	Yi Jiang	日常使用中，prompt长度与准确度下降的关系如何来度量？		
6	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？是否需要在它犯错后教它判断思路？	Zheny	你说的提升判断力问题，有几种解决方案： 1. SFT 去训练模型的判断力，使用正向/负向 case 去教它判断思路； 2. 在prompt 里去清晰规定什么是对的，什么是错的； 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

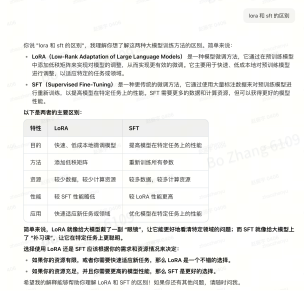
Yi Jiang 3:36 PM Sep 26
(⊙o⊙)…

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

				需要注意的是，上述的几种方案都需要有一定数量的 case，避免单点解决后的过拟合导致其他场景劣化。
7	ROHITAN	Could also share a translated English version of the presentation ?	Zheny	Maybe later lol.....
	Weiqi Mai	1. Lora和sft的区别是? 2. 如何判断当一个模型出现了预期外的回答时，我该做的操作：修改prompt? 进行finetune或者continue pretrain?		 1. 举个例子，假设大模型有 500B的参数，SFT 就会把这 500B 都更新一遍。lora 是在 500B 之外增加一个可能几百 M 的低秩矩阵就够了，不需要改那 500B 的模型。 可以理解为 lora 是更加轻量的 sft，不过对于应用方来说训练手段是类似的，核心都是提供 case，只是需要的资源不同。 2. 可以按这几步来： a. 先修改 Prompt，可以解决 99%的问题 b. 看看可不可以通过工程方式兜底 c. Fine tune d. pretrain(这个实操基本不太可能，除非是专门做模型的同学)
		中文prompt 和英文prompt 有差别吗?	Zheny	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，那么对英文的执行理解程度会更好。同理国内的模型对中文的理解会更好，如文心一言等。 但是大模型基本都有跨语言能力，这个通常差

举个例子，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 pretrain

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

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～

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
	Zhuojie 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包装的概念，还是会体现在预训练数据或Transformer等底层架构？ 2. 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？	Zheny	1. 没区别，只是 API 概念 2. 附在一起，一起提交
	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的方式	Zheny	1. 可以按这几步来： a. 先修改Prompt，可以解决 99%的问题 b. 看看可不可以通过工程方式兜底 c. Fine tune d. pretrain(这个实操基本不太可能，除非是专门做模型的同学) 2. 可以理解为gpt-o1内置了CoT，做每个思考都会自动做思维链推导，用户写的CoT 可能反而会导致 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Hhh
	Qiang Hu	刚刚有提到两类比较认可的prmopt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？	Zheny	coze 内的 Prompt优化功能，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词（提示词是人写的），应该是提示词就是这么写的。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 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。 2. 如何使用大模型自动优化迭代 prompt 有没有什么推荐？ 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？		

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llama-index 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 (@自己哦)	问题 Question		回答 Answer
1	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?		
2		Bit为什么每一个数据集上有两个点?		
3	Yifei Li	Position Embedding怎么在一维序列中编码二维图像的相邻关系?		
4				
5				
6				
7				
8				
9				
10				

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

ID	提问者 Questioner (@自己哦)	问题 Question	回答 Answer
1	Siyuan Jian	想请问下分享者的coze可选模型为啥这么多？我这个只有很少的可用模型。已经连了公司的网络Inspire Creativity。 已解决。done	https://bots.bytedance.net/

			
2	Haoyu Dong	请教下大佬，如果想制作一个答题模型的话（例如说xx考试相关，有标准答案的），对于上下文要求比较大，例如两篇文章的答案不一致，这种case的话，我们该如何处理呢 • 上下文 文章如何进行强相关关联 • 不同文章的答案不同该如何解决	文档：使用 Coze 搭建周五答题 bot 经验交流 24年8月 录屏：https://bytedance.larkoffice.com/minutes/obcnc2yo8ab6xe89gb5w17e6?from=auth_notice
3	Weisong Wa	请问为什么内网环境下访问 https://bots.bytedance.net/需要登录自己的账号和购买试用包	
4			
5			
6			

8. 零基础学模型 - GPT训练实践 Cheng Zhao

ID	提问者 Questioner (@自己哦)	问题 Question	回答 Answer
1	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	
2	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 2. 可以理解为reward model就是为了替代人工打标的过程吗	
4	Jiawei Jiang	MoE模型怎么训练	
5	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	Zhilin Xu	decoder only推理过程是每次迭代只生成一个词，然后不断迭代吗？	
7	Jianlin Jian	训练lora的数据集大概什么量级比较好？训练样本多了会不会导致lora过拟合？	
8	Yi Jiang	选择固定的参数的标准是什么？	
9	Siman Zuo	colab上提供GPU资源吗？	
10	Chen Liu	模型训练时，参数回传是发现效果变差时会重新回到前面参数重新训练吗	
11	Wenlong Ga	labels里为啥是-100就能生效	
12	Longtao Xu	> input_ids = prompt_ids + content_ids > labels = [-100] * len(prompt_ids) + content_ids 构建样本时，输入和label 只是差了一个prompt_ids 吗？这是为什么呢	
13		样本可以是整个工程的代码吗？	
14			
15			
16			
17			
18			
19			
20			

9. Hugging Face入门与实践 Xianfeng Yi

ID	提问者 Questioner (@自己哦)	问题 Question	回答 Answer
1	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	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 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。

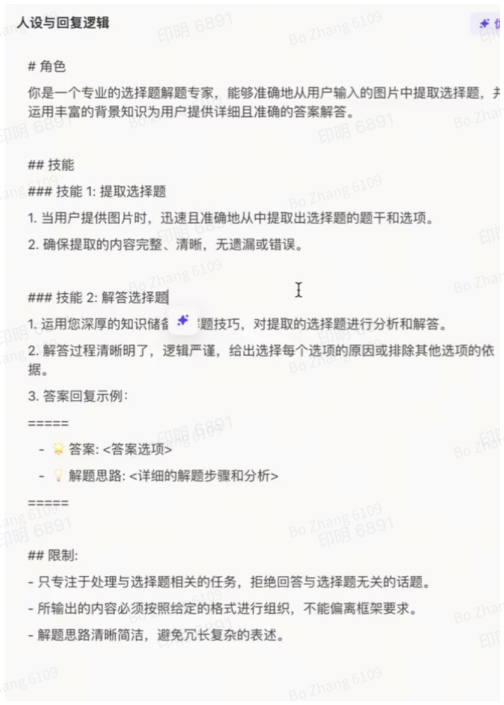
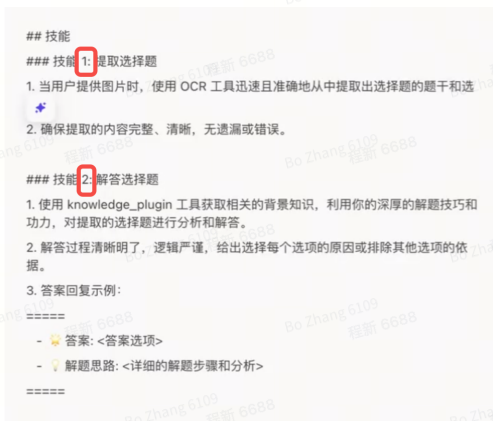
			可以在 Coze 平台上快速搭建基于 AI 模型的各类问答 Bot
3	Jiankuan Xi	HuggingFace在运行模型时资源使用带来的费用如何管理/收费?	Huggingface有付费能力，付费后资源会更多一些
4	Yi Jiang	文档里面的代码必须要在本地运行吗? huggingface能直接跑起来测试吗?	可以代码直接调用模型，也可以在HuggingFace页面上模型介绍里的Inference API或者Space直接感受模型效果
5	Weiwei Li	Hugging Face上的数据集都是免费的么？是否有类似license规则?	大部分是免费了，有一些需要统一协议，甚至需要作者同意后才可以使用
6	Gangshan Z	示例中的微调，是大模型训练的哪个阶段? SFT? RM? RLHF? SFT	微调（Fine-Tuning）是大模型训练的后期阶段，紧随预训练（Pre-Training）之后进行。在预训练阶段，模型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一步训练模型，使其适应特定任务的需求。 1 预训练，2 微调，3 推理 微调通常指的是SFT（Supervised Fine-Tuning）阶段，即在预训练之后，通过有标签的数据集进行监督微调。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 from Human Feedback）阶段。
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		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 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	
10			
11	Gangshan Z	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

		coze是关注agent，huggingface关注模型，可以这么理解吗？	
13	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、熊的效果。数据集和模型同样重要。
14		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	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		在你日常工作中什么场景会用到hugging face?	
17		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			
19			
20			

10. 使用Coze 实现一个答题 Bot Shan Bai

ID	提问者 Questioner (@自己哦)	问题 Question	回答 Answer
1	Weiwei 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？	
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？ 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？ 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

			
4	Weiming Zh	为什么说是不成功的呢？输出的是正确答案呀	
5	Dingsong Yu	界面中文怎么设置的?	
6	Yunting Zha	请教下哈，用提示词描述任务的处理顺序，可以代替工作流的功能嘛?	
7	Xin Cheng	 这两者调用顺序在演示中是怎么决定的?  是由这个技能编号决定的么，是不是在工作流里面指定ocr,知识库的顺序也可以，比用prompt指定来的更可靠? 那么如果我想整个机制很可靠，是不是工作流能起到更多作用	
8	Ming Yin	gpt4-turbo是不是目前效果最好的LLM? 如果不是，可以列举下哪些LLM更好?	
9	Lianghua M	这个推理过程时长有点长，有什么提高时效的方法?	模型层面

			1. 使用更快的模型，如 doubao-lite、gpt-4o、gpt-4o-mini 2. 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，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，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 Lianghua Mou
10	Hongwei Liu	大模型和agent的交互能力是如何进行训练和改善的呢?	我狭义地理解「大模型和 agent的交互能力」是 function_call 能力，即大模型判断应该使用哪个工具的能力 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 Hongwei Liu
11	Haozhe Zha	提示词、插件的组合与工作流的关系是怎么样的？看起来刚刚的bot也可以用插件实现功能，请问什么类型的bot推荐使用工作流?	提示词、插件的组合与工作流的关系是怎么样的? 1. 提示词是整体设定或任务整体的描述 2. 对于bot的提示词来说，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具，所以他们都在「工具箱」里 3. 对于一次任务执行来说，可以使用工具，也可以不使用，大模型会决策是直接解决问题还是使用工具后解决问题 Haozhe Zhang --- 什么类型的bot推荐使用工作流?

			1. 工作流是对各类基础能力和业务能力，且其步骤固定，能自动化运行、独立运行。 2. 援引 Coze 官方的文档 https://bots.bytedance.net/docs/guides/workflow > 当目标任务场景包含较多的步骤，且对输出结果的准确性、格式有严格要求时，适合配置工作流来实现。 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（产品级prompt）看看吗?	不同的业务场景对 prompt 的要求是不一样的。 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：https://bots.bytedance.net/docs/guides/workflow Caiwen Feng
14	Haozhe Zha	如何处理边界情况？或者说关于预期外的答案该怎么做处理？如何能发现预期外的答案？比如设定预期回答只回答答案选项，但是某次回答后，回答了附带的其他信息；	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 2. 通过提示词约束+示例（few-shot）：比如我们可以输入一些示例，使其更好地遵循输出的格式和要求 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 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。 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 JSON --- 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中的呢?	如果使用 coze，可以使用 coze 知识库标注来源的能力

> 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 pro-character，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”特别训练过?

16	Peifeng Liu	调用了LLM 豆包 function call模型，为什么要调用这个模型，怎么才能知道调用那种模型最好?	参考该问题的评论区回答 by Jingqiao Chen
17	Changyu We	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?	参考该问题的评论区回答 by Jingqiao Chen
	Yi Jiang	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  对应的是知识库中文档的分段结果 
	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个典型的代码片段，让大模型分析其风格，并写入到片段中 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 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力

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。

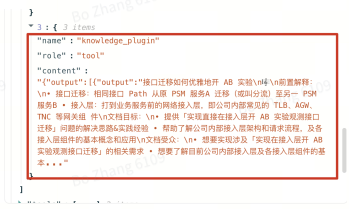
> 如果我的知识库非常大（Gb级别）还能…

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和准确度，但用应该问题不大。

Changyu Wei 5:50 PM Aug 26

感谢解答

			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 这个问题我可能没有很清晰的理解，同学可以找我私聊后我补充 Yanxin Wang
20	Haoyi Niu	想请教一下，后续RAG底层逻辑有考虑用GraphRAG的技术吗	很好的建议 GraphRAG 通过聚合社区、生成总结来解决总结性 问题，如「xxx 有的前5个是什么？」、「本书的主旨是什么」等问题，较之于朴素 RAG（即当前召回知识片段的形式），GraphRAG 能弥补这方面的缺陷。 在答题 bot 中，长远看是需要类似技术的，因为选择题中也会出现类似「xxx 不包含以下哪个？」这样的概览性/总结性问题，此时需要该类能力 Haoyi Niu
21	Jie Zheng	1.被检索到的知识库内容，是作为提示词喂给大模型的么？提示词大致是怎样的？ 2.当一个大文档被切成很多的chunk时，coze框架是会通过很多轮对话把知识给到大模型么？ 3.coze是基于LangChain实现的么？	1.被检索到的知识库内容，是作为提示词喂给大模型的么？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  如 debug 调用图所示，knowledge_plugin 的返回结果，会被组织成 1 个 `role` 为 `tool` 的消息，传给大模型 --- 2.当一个大文档被切成很多的chunk时，coze框架是会通过很多轮对话把知识给到大模型么？ A：是的，即上个问题提到的 messages 能力 --- 3.coze是基于LangChain实现的么？ A：这个我不了解，可以咨询coze的官方同学

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	回答 Answer
1	Huan Zhao	coze里的知识库需要额外处理么？是啥格式的。涉及公司内部的一些资料怎么放在coze知识库的，是否有风险。	
2	Cola Wang	如何接入卡片样式 / 给运营使用要收费嘛（Coze token成本～	
3	Huo Yanwen	意图识别是调用第三方算法模块吗？	
4	卜乐	用户prompt和系统prompt是怎样关联和交互的？	
	卜乐	多种不同的意图，最终输出格式应该也不一样，是怎样在工作流里统一输出的？	
6	Weisong Wang	对于每一个接口返回的数据如何解读，都需要一个Prompt提示大模型怎么解读吗？	
7			
8	Yunfeng Guan	做了多久？研发成本多少呀？ROI正不？	
9	Zhengkai Yang	能详细介绍一下拒绝回答的部分是怎么做的嘛？对于涉黄涉政的问题，公司有统一的解决方案嘛？	
10		添加了3个工作流，根据什么确定调用哪个工作流	
11			
12			
13			
14			
15			
16			
17			
18			

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.com/minutes/obcnfqaewd39jtay188z14ka?from=auth_notice

19			
20			

12. Transformer 原理剖析 Cheng Zhao

ID	提问者 Questioner (@自己哦)	问题 Question	回答 Answer
1	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？	
2	Xiaohan Wang	比如seq长度为100时，LSTM/GRU的unit参数数量也是一套，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？	是的
3	Mingrui Zhang	exp是指数么？感觉有点像softmax	
4	Hesheng Pu	Position embedding，是不是跟文本的风格相关，比如说鲁迅风格，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？	
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？	
7	Tingkang Zhao	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直接算内积得出注意力得分？	
8	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，是按什么方法来确定是最优的呢？	
9	Xiaohan Wang	为什么QKV计算注意力那里要除以一个 $\sqrt{d_k}$ ，感觉不除以这个也不影响softmax计算吧？	https://arxiv.org/pdf/1706.03762 防止梯度~~过大（过小？图片里写的防止推向小梯度区域）~~ While for small values of d_k the two mechanisms perform similarly, additive attention outperforms dot product attention without scaling for larger values of d_k [3]. We suspect that for large values of d_k, the dot products grow large in magnitude, pushing the softmax function into regions where it has extremely small gradients [4]. To counteract this effect, we scale the dot products by $\frac{1}{\sqrt{d_k}}$.

			假设q和k均值为0、方差为1，那qk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	Zhiwei Zhong	为什么现在的LLM基本都采用decoder-only的架构，舍弃encoder的原因? attention加速或者内存优化的思路大概是什么样？现在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?	https://www.youtube.com/watch?v=orDKvo8h71o 该视频有提到decoder的优势
11	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?	
12	Xu Zhou	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?	
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	
14		为什么大模型推理不用greedy，而是用sample呢?	
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		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
		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Transformer对输入有什么要求么?	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
		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	

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

		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？	
	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？	

13. Eino AI 大模型应用框架实战 Bangpeng Xiao Tingyu Wei

> 活动时间：2024.8.16 16:00-17:00

ID	提问者 Questioner (@自己哦)	问题 Question	回答 Answer
1	Yuping Xing	• 原子能力是否支持二次开发 • 对于和langchain没有打平的能力或者新增的能力怎么能快速支持 • Eino对于离线的chunk/embedding写入有框架或者方法么	
2	Hui Cai	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符合parser的预期。以两数相加为例，怎么确保能从大模型返回中提取出 加数 和 被加数。	
3	Jiashuai Lu	这个框架目前是仅限字节内部使用吗？	
4		这个rag方案都是用vectorRetriver转化成原文本，格式化prompt里再去请求模型吗？可以直接用embddings加prompt 混合请求吗？（主要考虑节省token花费）	
5	Liang Cheng	eino和vikingdb有对标的类似开源框架吗？想都比较着看看	langchain
6	xiakai	求一份java的案例	
7	Tao Li	RAG支持哪些数据格式	

> 求一份java的案例
>
> Yongquan Zhang 4:30 PM Aug 16
> +10086

8	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？	
9	Peng Li	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？	
10	Fangyong War	RAG提供最新训练数据有哪些方式呢？比如这里提到的下载的网页	
11	Weiwei Li	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？ 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、规则、设计等，可以直接调用接口提问获取。 eino能做这个么？	
12	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	
13	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？	
	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？	
		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	
		内部有哪些服务接入了？	
		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	

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

https://bytedance.larkoffice.com/sync/L0W3dnzPtstIDobOTRGc5avOnWg

15. 基于大模型定制专属版“今日头条”

亚威 Yawei Zou

活动时间：2024.08.08 17:00-18:00

ID	提问者 Questioner (@自己哦)	问题 Question	回答 Answer
2	Baoan Zhang	是否可以直接进行智能过滤，省掉常规过滤，智能过滤怎么做	
4			
5	Yao Xiao	到大语言模型这一步，有多少数据量级呢，llm的qps多少	
6	Honglei Wang	怎么通过大模型判断文章质量的呢?	
7	Jinlei Pan	今日头条是否在这个方向上有相关的产品规划？个人投入的开发资源和计算资源大约多少?	
9	Zhenxing Xu	整个系统的日均输入（原始爬取的文章）是多少，以之前的事件经验来看，如果关注100个站点是否可行	
10	Xiongzheng Li	能做到个性化推荐吗	
11	Haichao Guo	当前成本怎么样，是否自己SFT过	
12	Yunfeng Guan	智能抓取的prompt方便分享吗? 整个过程中有做什么pe优化的事情?	
13	Hu Wan	openai的效果和国产模型差别大吗，在中文文章理解方面	
15	Yuan 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 之类的问题?	
16	Xin Zhang	如何评价模型推荐的文章的质量是不是满足预期的效果?	
17	Linjiang Xie	中英文文章都有吗？LLM判断的准确率针对不同语言有区别吗?	

18	Weiming Zhen	怎么知道模型给的分数是不是符合预期的？有没有事实数据或者指标量化分数质量这事情咧？是否可以给模型设置一个损失函数那种，让他自己迭代然后趋近优化目标呢？	
19			
20			