大模型面试题 200 问

李博杰

为了帮助大家更好地理解《图解大模型：生成式 AI 原理与实战》，也为了方便部分有面试需求的朋友更有针对性地阅读这本书，围绕本书各章主题，译者李博杰系统梳理了大模型领域常见的面试题，其中的大多数问题可以在书中直接找到答案，部分进阶问题可以从本书的参考文献或网络上的最新论文中找到答案。希望所有的朋友都能够带着这些问题阅读本书。

第 1 章　大模型简介

Q1：Transformer 中的编码器和解码器有什么区别，只有编码器或者只有解码器的模型是否有用？

Q2：GPT 跟原始 Transformer 论文的模型架构有什么区别？

Q3：仅编码器（BERT 类）、仅解码器（GPT 类）和完整编码器 - 解码器架构各有什么优缺点？

Q4：为什么说 Transformer 的自注意力机制相对于早期 RNN 中的注意力机制是一个显著的进步？

Q5：大模型为什么有最长上下文长度的概念？为什么它是指输入和输出的总长度？

Q6：大模型的首字延迟、输入吞吐量、输出吞吐量分别是如何计算的？不同应用场景对首字延迟、输入吞吐量和输出吞吐量的需求分别是什么？

Q7：预训练和微调的两步范式为什么如此重要？基础模型通过预训练获得了哪些核心能力？微调在引导模型遵循指令、回答问题和对齐人类价值观方面起到什么作用？

Q8：Llama-3 8B 的综合能力比 Llama-1 70B 的能力还强，是如何做到的？

第 2 章　词元和嵌入

Q9：大模型的分词器和传统的中文分词有什么区别？对于一个指定的词表，一句话是不是只有一种唯一的分词方式？

Q10：为什么传统 BM25 检索对中文分词的质量很敏感，而大模型对分词器的选取不敏感？

Q11：GPT-4、Llama 等现代大模型采用的字节级 BPE 分词器相比传统的 BPE 分词器有什么优点？

Q12：国内预训练的大模型与海外模型相比，是如何做到用相对更少的词元表达中文语料的？

Q13：大模型是如何区分聊天历史中用户说的话和 AI 说的话的？

Q14：大模型做工具调用的时候，输出的工具调用参数是如何与文本回复区分开来的？

Q15：参考章节中用播放列表数据训练歌曲嵌入的案例，设计一个使用嵌入技术解决电子商务产品推荐的系统。使用什么数据作为“句子”的等价物？如何将用户行为融入嵌入模型？

Q16：word2vec 的训练过程中，负例的作用是什么？

Q17：传统的静态词嵌入（如 word2vec）与大模型产生的与上下文相关的嵌入相比，有什么区别？有了与上下文相关的嵌入，静态词嵌入还有什么价值？

Q18：与上下文相关的嵌入是如何解决一词多义问题的，如技术语境下，英文 token 可能表示词元、代币、令牌，而中文“推理”可能表示 reasoning 或 inference？

Q19：在 word2vec 等词嵌入空间中，存在 king - man + woman ≈ queen 的现象，这是为什么？大模型的词元嵌入空间是否也有类似的属性？

第 3 章　LLM 的内部机制

Q20：大模型怎么知道它的输出该结束了？

Q21：训练时如何防止模型看到未来的词元？

Q22：注意力机制是如何计算上下文各个词元之间的相关性的？每个注意力头是只关注一个词元吗？ softmax 之前为什么要除以 $\sqrt{d_k}$ ？

Q23：Q 和 K 在注意力的表达式里看起来是对称的，但 KV 缓存里为什么只有 KV，没有 Q？

Q24：如果没有 KV 缓存，推理性能会降低多少？

Q25：为什么 Transformer 中需要残差连接？

Q26：Transformer 中的 LayerNorm 跟 ResNet 中的 BatchNorm 有什么区别，为什么 Llama-3 换用了 RMSNorm？

Q27：Transformer 中前馈神经网络的作用是什么？注意力层中已经有 softmax 非线性层，那么前馈神经网络是否必要？

Q28：如果需要通过修改尽可能少的参数值，让模型忘记某一特定知识，应该修改注意力层还是前馈神经网络层的参数？

Q29：大模型在数学计算时，为什么经常不准确？

Q30：模型深度（层数）与宽度（隐藏维度大小）、注意力头数量、上下文长度等参数之间是如何相互影响的？如果要训练一个比当前模型参数规模大 10 倍的模型，你会如何调整这些参数？

Q31：以一个你熟悉的开源模型为例，介绍模型中每个矩阵的大小和形状。

Q32：大模型推理过程中，内存带宽和算力哪个是瓶颈？以一个你熟悉的开源模型为例，计算输入批次大小达到多少时，能够平衡利用内存带宽和算力？

Q33：从统计学角度看，Transformer 输出层假设词元符合什么分布？

Q34：给定一个支持 8K 上下文的开源模型，如何把它扩展成支持 32K 上下文的模型？上下文长度增加后对 KV 缓存会带来什么挑战？

Q35：为什么注意力机制需要多个头？ GQA、MQA 优化跟简单减少注意力头的数量相比，有什么不同？ GQA、MQA 优化的是训练阶段还是推理阶段？

Q36：Flash Attention 并不能减少计算量，为什么能实现加速？ Flash Attention 是如何实现增量计算 softmax 的？

Q37：RoPE（旋转位置嵌入）相比 Transformer 论文中的绝对位置编码有什么优点？ RoPE 在长上下文外推时会面临什么挑战？

Q38：由于训练样本长度往往小于最大上下文长度，把多个训练样本放到同一个上下文中训练时，如何避免它们互相干扰？

Q39：如何利用一个小规模的大模型提升大规模模型的推理性能，并尽量不影响大模型的推理结果？推测解码并没有减少计算量，为什么能提升推理性能？

第 4 章　文本分类

Q40：如何基于表示模型生成的嵌入向量实现文本分类？

Q41：使用嵌入向量实现分类和使用生成模型直接分类的方法相比，有什么优缺点？

Q42：如果没有标注数据，如何基于嵌入模型实现文本分类？如何优化标签描述来提高零样本分类的准确率？

Q43：书中嵌入模型 + 逻辑回归的分类方式获得了 0.85 的 F1 分数，而零样本分类方式获得了 0.78 的 F1 分数，如果有标注数据，什么情况下会选择零样本分类？

Q44：Transformer 为什么比朴素贝叶斯分类器效果好很多？朴素贝叶斯分类器的条件独立性假设有什么问题？

Q45：掩码语言建模与 BERT 的掩蔽策略相比有何不同？这种预训练方式如何帮助模型在下游的文本分类任务中获得更好的性能？

Q46：假设你有一个包含 100 万条客户评论的数据集，但只有 1000 条带有标签的数据，同时利用有标签和无标签数据，结合表示模型和生成模型的优势，构建一个分类系统？

Q47：使用生成模型进行文本分类时，以下三个提示词哪个会更有效？

- “Is the following sentence positive or negative?”
- “Classify the sentiment of this movie review as positive or negative.”
- “You are a sentiment analysis expert. Given a movie review, determine if it expresses a positive or negative opinion. Return only the label ‘positive’ or ‘negative’.”

第 5 章　文本聚类和主题建模

Q48：有了强大的生成式大模型，嵌入模型还有什么用？请举一个适合嵌入模型但不适合生成模型的例子。（提示：推荐系统）

Q49：给定大量的文档，如何把它们聚类成几簇，并总结出每一簇的主题？

Q50：词袋法和文档嵌入在实现原理上有什么区别？词袋法是不是一无是处了？

Q51：BERTopic 中的 c-TF-IDF 与传统 TF-IDF 有何不同？这种差异如何帮助改进主题表示的质量？

Q52：LDA、BTM、NMF、BERTopic、Top2Vec 等主题模型有什么优缺点？对长文档、短文档、高质量需求的垂直领域分别应使用何种模型？

Q53：基于质心的和基于密度的文本聚类算法有什么优缺点？

Q54：为什么在主题建模流程中，将聚类和主题表示这两个步骤分开处理是有益的？

Q55：在一个主题建模项目中，你发现生成的主题中有大量重叠的关键词，如何使用本章介绍的技术来改进主题之间的区分度？

Q56：在使用 BERTopic 时，如果很大比例的文档被归类为离群值，这可能是什么原因导致的？如何调整聚类参数？

Q57：在新闻或社交媒体推荐系统中，主题往往随时间快速演化，如何检测新兴主题？

Q58：如何构建一个内容平台的推荐系统，冷启动时通过文本聚类和主题建模提供推荐，有一定量用户交互数据后又能利用这些数据提升推荐效果？

第 6 章　提示工程

Q59：针对翻译类任务、创意写作类任务、头脑风暴类任务，`temperature` 和 `top_p` 分别该怎么设置？如何验证你选择的参数设置是否最优？

Q60：为什么一些模型把温度设置成 0，输出的内容仍然有一定的不确定性？（提示：推测解码）

Q61：对于指定的大模型，如何通过提示词减少其幻觉？

Q62：一个专业的提示词模板应该由哪几部分构成？为什么提示词中需要描述角色定义？

Q63：对于一个复杂的提示词，如何测试其中哪些部分是有用的，哪些部分是无用的？

Q64：如何设计提示词模板，尽量防止提示词注入？如何在系统层面检测提示词注入攻击？

Q65：如果把用户信息放在系统提示词中，但在对话轮数较多后，大模型经常忘记用户信息，如何解决？

Q66：如何让 ChatGPT 输出它自己的系统提示词？

Q67：在没有推理模型之前，如何让模型先思考后回答？思维链、自洽性、思维树等几种技术有什么优缺点？

Q68：在创意写作任务中，如何让模型生成多个可能输出，再从中选取一个最好的？

Q69：如果需要模型遵循指定的格式输出，提示词应该怎么写？

Q70：如何保证模型的输出一定是合法的 JSON 格式？（提示：限制采样）

Q71：将大模型用于分类任务时，如何保证其输出一定是几个类别之一，不会输出无关内容？（提示：限制采样）

Q72：如果做一个学英语的应用，如何保证它说的话一定在指定的词汇表中，绝不会出现超纲的生词？（提示：限制采样）

第 7 章　高级文本生成技术与工具

Q73：如果我们需要生成小说的标题、角色描述和故事梗概，单次模型调用生成效果不佳时，如何分步生成？

Q74：如果用户跟模型对话轮次过多，超出了模型的上下文限制，但又希望尽可能保留用户的对话信息，该怎么办？

Q75：在角色扮演场景中，用户跟模型对话轮次过多后（但没有超过上下文限制），模型经常没有注意到过去对话中发生过的关键事件，怎么办？

Q76：用户跟模型对话轮数较多后，处理输入词元的预填延迟升高，应该如何解决？（提示：持久化 KV 缓存）

Q77：如何编写一个智能体（agent），让它像 OpenAI Deep Research 一样，能够自主思考下一步该搜索什么关键词，浏览哪个网页？

Q78：如何编写一个智能体，帮助用户规划一次包含机票预订、酒店安排和景点游览的旅行？需要配置哪些工具？如何确保系统在面对不完整或矛盾信息时仍能提供合理建议？

Q79：如果单一智能体的提示词过长，导致性能下降，如何将其拆分为多个智能体，并在合适的时机调用不同的智能体？不同智能体间如何进行有效的上下文传递和结果整合？

Q80：不同基础模型在不同任务上的表现不同，如何基于任务特性自动选择最合适的模型？

Q81：如果一个工具的调用时间较长，如何让智能体在等待工具调用返回前能够持续与用户交互或调用其他工具，并在工具调用返回时及时做出下一步动作？

Q82：对于角色扮演场景下的持续对话任务，如何缓存角色设定和历史对话，降低输入词元的成本和延迟？

Q83：智能体如何处理记忆中的时间信息，例如“昨天讨论的问题”？如何在用户长时间不回复时，主动询问用户？

Q84：多个智能体在同一房间里讨论时，如何防止多个智能体互相抢话，又避免冷场？

Q85：支持实时语音的智能体如何既保持低延迟，又避免与用户抢话？

Q86：支持语音输入的智能体，如何用非声学方法，通过语义理解用户是在对旁边人说话还是对它说话？

Q87：PTQ 和 QAT 量化方法的区别是什么，有什么优缺点？

第 8 章　语义搜索与 RAG

Q88：在 RAG 中，为什么要把文档划分成多个块进行索引？如何解决文档分块后，内容上下文缺失的问题？如何处理跨片段的依赖关系？

Q89：如果发现向量相似度检索的匹配效果不佳，除了更换嵌入模型，还有哪些办法？

Q90：向量相似度检索不能实现关键词的精确匹配，传统关键词检索不能匹配语义相近的词，如何解决这对矛盾？

Q91：向量相似度检索已经是根据语义相似度匹配，为什么还需要重排序模型？

Q92：为什么要在向量相似度检索前，对用户输入的话进行改写？

Q93：RAG 系统检索的文档可能包含冲突信息或过时数据，如何在生成回答时防止被这些信息误导？

Q94：如何使检索模块能够从生成模块获得反馈并动态调整检索策略，例如给不同的文档标注可信度？

Q95：如何提升 RAG 系统的可解释性，包括清晰标注生成内容的来源，以及量化展示系统对回答的确信度？

Q96：智能体如何把处理企业任务的经验总结到知识库中，并在后续任务中引入知识库中的经验？如何保证经验不断积累，而不是简单用新的经验覆盖已有的经验？

Q97：如果需要根据一本长篇小说的内容回答问题，小说长度远远超出上下文限制，应该如何综合利用摘要总结和 RAG 技术，使其能同时回答故事梗概和故事细节？

Q98：如何将 RAG 系统从纯文本扩展到多模态，支持检索图像、视频、图文并茂的文档等多模态信息，并在生成回答时以多模态形式呈现，例如包含原始文档中的图表和视频？

Q99：如果需要设计一个 AI 智能伴侣，每天记录用户说过的所有话、做过的所有事，持续多个月，如何在需要的时候快速检索出相关的记忆，让 AI 能够根据记忆回答问题？综合对话历史窗口化、摘要总结、RAG 等技术。

第 9 章　多模态大模型

Q100：为什么 ViT 不能简单地像处理文本词元那样，为每个图像块分配一个唯一的、离散

的 ID，而是必须采用线性投影生成连续的嵌入向量？

Q101：在 CLIP 训练过程中，为什么需要同时最大化匹配图文对的相似度和最小化非匹配对的相似度？

Q102：BLIP-2 采用了冻结预训练 ViT 和 LLM，仅训练 Q-Former 的策略。这种设计的核心动机和优势是什么？

Q103：BLIP-2 是如何连接预训练的图片编码器和预训练 LLM 的？为何不直接将视觉编码器的输出连接到语言模型，而要引入 Q-Former 这一中间层结构？

Q104：将多模态特征映射到文本特征空间时不可避免会产生信息损失，交叉注意力、Q-Former 和线性映射等方法的信息保留能力有什么区别？

Q105：BLIP-2 模型的图像 - 文本对比学习、图像 - 文本匹配、基于图像的文本生成三个任务分别是什么作用？与今天的 Qwen-VL 等多模态模型有什么区别？

Q106：基于已经预训练好的模态编码器、模态解码器、文本大模型做多模态模型，多模态预训练和多模态微调两个阶段分别需要什么数据，需要冻结模型的哪些参数？

Q107：CLIP 和 BLIP-2 在处理图像时，都会将其预处理成固定尺寸。如何处理长宽差异巨大的图像？

Q108：在 BLIP-2 实现视觉问答（VQA）时，模型是如何同时处理输入的图像和文本问题的？

Q109：以一个你熟悉的开源多模态模型为例，输入一张 512×512 的图片和一个 100 词元的问题，其首字延迟大约是多少，其中模态编码器、Q-Former 和 LLM 部分各占多少？

Q110：能够操作计算机图形界面的多模态大模型每步操作的延迟通常需要几秒，延迟的构成是什么？

Q111：人类对不熟悉的界面操作较慢，但对熟悉的界面操作很快。如何让多模态模型像人类一样快速操作熟悉的界面？

Q112：现有一个能力较弱的多模态模型和一个能力较强的文本模型（如 DeepSeek-R1），如何结合两者的能力，回答多模态问题？

Q113：如果一个垂直领域（如医学）的图文对训练数据极为有限，如何为该领域构建多模态大模型？

Q114：如何构建一个 AI 照片助手，能够索引用户的上万张照片，根据用户的查询高效地检索到相关照片？

Q115：端到端语音模型中，语音是如何转换成词元表示的？

Q116：端到端语音模型是如何实现在工具调用进行过程中，继续与用户实时语音交互的？工具调用的结果与用户的语音输入在模型的上下文中如何区分？

Q117：图像生成模型（如 Stable Diffusion）与图像理解模型（如 CLIP、BLIP-2）在技术路线上有什么异同？为什么扩散模型在推理时需要噪声，而自回归模型不需要？

第 10 章　构建文本嵌入模型

Q118：为什么通过对比（相似 / 不相似样本）学习通常比仅学习相似样本能更有效地捕捉文本的语义或特定任务特征？

Q119：如何生成负例以提升模型性能？如何构建高质量的难负例？

Q120：双编码器和交叉编码器有什么区别？假设你需要构建一个大规模语义搜索引擎，你会优先选择哪种架构来计算查询与文档的相似度，为什么？如果任务变为对少量候选对进行精确重排序，你的选择会改变吗？

Q121：多负例排序损失（MNR）、余弦相似度损失和 softmax 损失在训练嵌入模型时有哪些优缺点？在什么场景下，余弦相似度损失可能比 MNR 损失更合适？

Q122：为什么 TSDAE 选择使用特殊词元而非平均池化作为句子表征？

Q123：相比有监督方法，TSDAE 这类无监督预训练方法在处理领域外数据或进行领域适配时有何优缺点？

Q124：MTEB 相比基础的语义相似度测试（STSB）有哪些改进？其中包括哪些类别的嵌入任务？

Q125：如何根据用户偏好反馈数据，持续提升 RAG 系统的重排序模型性能？

Q126：如果一个 RAG 系统没有人类用户，仅供 AI agent 使用，如何自动收集 AI agent 的反馈，持续提升 RAG 系统的重排序模型性能？

Q127：如果要构建一个类似 Google 图片搜索的文本嵌入模型，根据输入图片找到相似图片，应该如何训练？

Q128：如果要构建一个非自然语言垂直领域（如氨基酸序列、集成电路设计）的语义搜索系统，但该领域标注数据极少，应该如何训练嵌入模型？

Q129：随着新数据和新概念的不断产生，如何检测何时需要更新文本嵌入模型，实现增量的持续学习？

第 11 章　为分类任务微调表示模型

Q130：在微调任务中，应该冻结哪些层的权重？微调编码器前几层、编码器后几层、前馈神经网络层有什么区别？

Q131：如果有标注的训练数据很少，如何扩增训练数据的数量？（提示：SetFit）

Q132：SetFit 在训练分类头之前，会先利用对比学习微调 Sentence Transformer。为什么这个微调步骤对于在极少标注样本下取得高性能至关重要？

Q133：相比直接使用一个冻结的通用 Sentence Transformer 提取嵌入向量再训练分类器，SetFit 的对比学习微调方法能让嵌入向量学习到哪些更适用于下游分类任务的特性？

Q134：在继续预训练时，如何在保证模型获得特定领域知识的同时，最大程度保留其通用能力？

Q135：请比较以下三种方案在垂直领域文本分类任务上的优缺点：(a) 直接使用通用 BERT 模型微调；(b) 在医疗文本上继续预训练 BERT 后再微调；(c) 从头开始用医疗文本预训练模型再微调。

Q136：在基于掩码语言建模的继续预训练中，应该如何设计掩码出现的位置和概率？

Q137：在微调过程中，为什么模型对学习率等超参数通常比预训练阶段更敏感？

Q138：在命名实体识别任务中，当 BERT 将单词拆分成多个词元时，如何解决标签对齐问题？

Q139：如何用领域数据训练一个在嵌入式设备上使用的小模型，同时处理文本分类、命名实体识别和语义搜索三个任务？

Q140：假设一个嵌入模型的训练语料主要由英文构成，其中文表现不佳，如何用较低的继续预训练成本，提升其中文能力？

Q141：对于一个关键场景的分类任务，例如将“严重不良反应”误分类为“轻微不良反应”比反向错误更危险，如何选择评估指标，解决数据集类别不平衡的问题，并修改损失函数？

第 12 章　微调生成模型

Q142：在 Llama-3 70B 开源模型基础上，如何微调模型以使其输出风格更简洁、更像微信聊天，并保证输出的内容符合中国的大模型安全要求？你认为需要准备多少数据，用多少 GPU 训练多长时间？

Q143：有人声称一篇文章是用 DeepSeek-R1 生成的，并给了你生成所用的完整提示词，你应该如何证实或证伪这个说法？如何量化计算这个提示词生成这篇文章的概率？（提示：利用困惑度）

Q144：计算一个拥有 96 个 Transformer 块，且每个块有 12 288 × 12 288 权重矩阵的模型，使用秩为 8 的 LoRA 后，需要微调的参数量是多少？微调过程中的每一步需要多少计算量？相比全量微调减少了多少？

Q145：QLoRA 中的分块量化如何解决了普通量化导致的信息损失问题？

Q146：现有一个若干篇文章组成的企业知识库，希望通过 SFT 方法让模型记住，如何将其转换成适合 SFT 的数据集？如何确定 SFT 所需数据集的大小？

Q147：如果微调数据模板中缺少了结束标记 `</s>` 会产生什么影响？

Q148：微调模型时，学习率、LoRA alpha、LoRA rank 等超参数通常应该如何设置？应该如何决定模型何时停止训练，是不是验证集损失函数越低效果就越好？

Q149：在微调过程中，损失函数应该仅计算输出部分，还是同时计算输入和输出部分？两种方案各有什么优缺点？

Q150：微调后的模型上线后发现一些反复出错的用例，应当怎样修改 SFT 数据集？

Q151：模型对话轮次较多后，出现模型重复用户的提问或者之前轮次的回答等“复读机”问题，应该怎样通过微调方法解决？

Q152：目前最流行的几个模型分别在什么领域表现较好？为什么有些模型在排行榜中表现突出，但在实际使用中表现不佳？

Q153：Chatbot Arena 的模型评估方法相比固定测试集有什么优缺点？

Q154：PPO 和 DPO 在计算效率上、实现复杂度上、训练稳定程度上有什么区别？

Q155：如果现有人类偏好数据集质量高但数量有限，应该用 PPO 还是 DPO ？

Q156：PPO 中的 Proximal（近端）是什么意思？如何防止模型在微调数据集以外的问题上泛化能力下降？如何防止模型收敛到单一类型高奖励回答？

Q157：PPO 中演员模型、评论家模型、奖励模型、参考模型的作用分别是什么？

Q158：PPO 是如何解决 RL 中经典的稀疏奖励和奖励黑客（reward hacking）问题的？

Q159：PPO 中的归一化优势函数、值函数剪裁、熵正则化等关键技巧有什么作用？

Q160：DPO 中 beta 参数是什么意思，增大或减小它会有什么影响？

Q161：设想一个网站上都是 AI 生成的内容，统计了每篇内容的平均用户停留时长，如何将其转化为 DPO 所需的偏好数据？对于小红书和知乎两种类型的网站，处理方式有什么区别？

Q162：对一个 ChatGPT 类型的网站，如何把用户行为转化为 DPO 数据？例如点赞点踩、重新生成、复制、分享、后续追问等。

Q163：什么是大模型的对齐问题？如何避免大模型输出训练语料中的个人隐私信息？

Q164：如何通过模型微调，尽量解决提示词注入的问题？

Q165：现有 100 条回答用户问题的规则，完全放在提示词中指令遵循效果不佳，如何构建微调数据集和利用 RL 训练，让模型微调后能够遵从这 100 条规则？

图解推理大模型

Q166：根据缩放定律，如何估算训练一个特定规模的大模型所需的预训练数据集大小和所需算力？

Q167：从大模型原理的角度说明，为什么 Llama-3 70B 模型不可能在不输出思维链的前提下，可靠地解决 24 点问题。（即输入 24 点的问题描述和 4 个 100 以内的整数，要求立即输出一个单词 Yes 或 No）

Q168：通过“let’s think step by step”提示词触发的思维链模式，与推理模型的原理有什么不同？同样是测试时计算，为什么推理模型的上限更高？

Q169：推理模型的 RL 与非推理模型的 RLHF 有什么区别？

Q170：根据 AlphaZero 玩桌游的研究，训练时计算和测试时计算的算力最优配比是多少？

Q171：如果需要针对垂直领域微调推理模型，过程奖励模型（PRM）和结果奖励模型（ORM）分别适合什么场景？

Q172：在 MCTS 方法中，如何平衡探索和利用？探索和利用分别使用什么方式来评估？

Q173：STaR 方法是如何让模型通过自我生成的推理数据来改进自身的？它有什么优缺点？

Q174：推理模型在后训练过程中，思维链会越来越长，这样结果的准确率提升了，但响应延迟也增加了。如何处理推理深度与响应延迟的权衡？

Q175：如何让推理模型根据问题复杂度、用户需求和系统负载自动调整推理深度？

Q176：为什么推理模型每个输出词元的成本一般高于架构和参数量相同的非推理模型？

Q177：在实时语音对话应用中，如何利用推理模型，又不让用户忍受过高的响应延迟？

Q178：如何用 RL 方法提升一个大模型的工具调用能力？如何训练模型，使其能够智能地决定何时依靠内部推理能力以及何时调用外部工具，例如写一段代码来解决复杂的推理问题，而不是在输出的推理过程中穷举所有可能？

Q179：提示工程、RAG、SFT、RL、RLHF 方法应该分别在什么场景下应用？例如：快速迭代基本能力（提示工程）、用户个性化记忆（提示工程）、案例库和事实知识（RAG）、输出格式和语言风格（SFT）、领域基础能力（SFT）、领域深度思考能力（RL）、领域工具调用能力（RL）、根据用户反馈持续优化（RLHF）。

DeepSeek-R1

Q180：DeepSeek-R1 与 DeepSeek-R1-Zero 的训练过程有什么区别，各自有什么优缺点？既然 R1-Zero 生成的推理过程可读性差，在非推理任务上的表现也不如 R1，R1-Zero 存在的价值是什么？ R1 训练过程是如何解决 R1-Zero 的上述问题的？

Q181：为什么说 DeepSeek-R1-Zero 可能开启了一条让模型智力水平超越人类的路径？

Q182：为什么 DeepSeek-R1 在创意写作任务中，只需较短的思考过程，就能写出比 DeepSeek-V3 基座模型有趣很多的内容？

Q183：DeepSeek-R1 为什么没有使用 PRM、MCTS、集束搜索等方法？

Q184：DeepSeek-R1 使用的 GRPO 与 PPO 有什么区别？优势值归一化是如何解决传统 PPO 算法中的值函数估计问题的？

Q185：GRPO 中的 KL 惩罚项有什么作用？为什么过大或过小的 KL 惩罚项会影响训练效果？

Q186：DeepSeek-R1 在 SFT 阶段，为什么要加入 20 万条与推理无关的训练样本？

Q187：DeepSeek 是如何把 R1 的推理能力蒸馏到较小的模型中的？如果我们要自己蒸馏一个较小的垂直领域模型，如何尽可能保留 R1 在特定领域的能力？

Q188：DeepSeek MLA 相比 MQA 占用的 KV 缓存事实上更多，那么 MLA 为什么比 MQA 更好？ MLA 是对哪个维度做了低秩压缩？

Q189：DeepSeek MLA 是如何解决 RoPE 位置编码与低秩 KV 不兼容的问题的？如果采用其他基于注意力偏置的位置编码，会有什么问题？

Q190：DeepSeek MoE 模型为什么前 3 层采用稠密连接而后续采用 MoE ？如果所有层都使用 MoE，会有什么影响？

Q191：DeepSeek MoE 和 Mixtral MoE 有什么区别？ DeepSeek MoE 的细粒度专家分割和共享专家隔离有什么优点？

Q192：DeepSeek MoE 中的专家负载均衡是如何解决路由崩溃问题的？

Q193：从大模型对语言中概念建模的角度分析，为什么 R1-Zero 的思维链会出现多语言混杂现象？

Q194：R1-Zero 的方法主要适用于有明确验证机制的任务（如数学、编程），如何将这一方法扩展到更主观的领域（如创意写作或战略分析）？

Q195：如果要在一个非推理模型基础上通过 RL 后训练出一个 1000 以内整数四则运算错误率低于 1% 的模型，基座模型预计最少需要多大，RL 过程预计需要多少 GPU 训练多长时间？（提示：TinyZero）

Q196：在 QwQ-32B 推理模型基础上，通过 RL 在类似 OpenAI Deep Research 的场景中强化垂直领域能力，如何构建训练数据集，如何设计奖励函数？

Q197：DeepSeek-R1 不支持多模态，如果要在 R1 基础上支持图片推理，例如学会走迷宫、根据照片推断地理位置，如何构建训练数据集，如何设计奖励函数？

Q198：DeepSeek-V3 的多词元预测方法在样本利用效率和推理效率方面相比一次预测一个词元，有什么优势？

Q199：DeepSeek-V3 的混合精度训练在哪些矩阵计算中使用了 FP8 量化？为了减少对模型精度的影响，DeepSeek-V3 是如何对激活值和权重做分组量化的？

Q200：DeepSeek 的 DualPipe 并行训练算法相比传统流水线并行有什么优势？它如何与专家并行协同工作，以解决 MoE 模型的负载均衡问题？