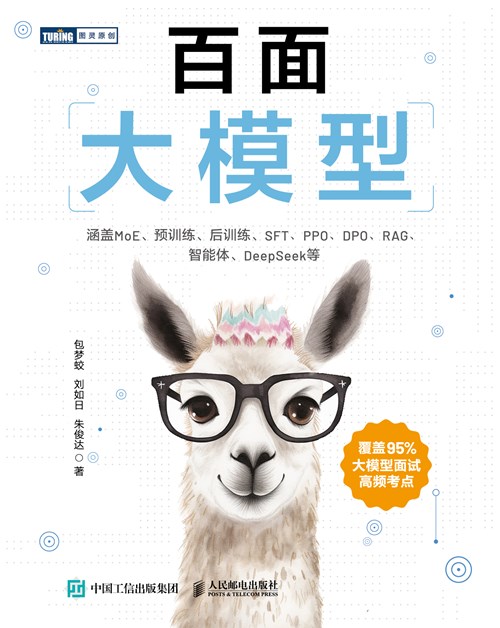
《百面大模型》带你拿Offer拿到手软

在当前信息爆炸、技术迅猛发展的时代背景下，人工智能技术的演进已深刻地影响了社会的各个层面。大模型作为人工智能领域的一个关键分支，正逐渐成为推动技术进步的核心驱动力。

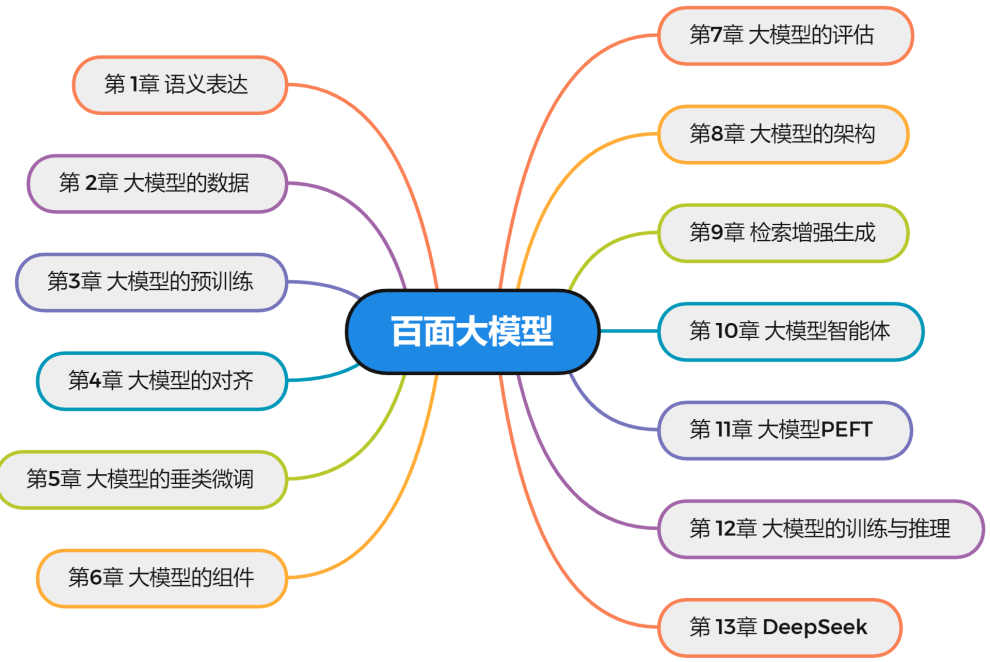
目前不管是独角兽还是一些国企单位，都在尽自己最大的能力在布局大模型，基于大模型的各种提升数字化劳动力的“人”变得越来越多。大模型的相关人才是缺乏的，但是招聘的需求却是越来越多。相关的面试准备是不是也开始让你无法好好备战，毕竟知识点太多了。好在《百面大模型》一书正式出版。



本书全彩印刷，其以独特的学术视角和深入浅出的叙述方式，为我们揭开了大模型的神秘面纱，引领读者深入探索一个充满无限可能的智能世界。

**这本书的内容如何**

作为一本用于面试的计算机技术图书，而且是关于LLM的前沿图书，大家可能对图书质量有所担忧。其实不用担心，本书内容涉及广泛，而且具有很强的针对性，我们来看一下这本的书的大纲：



我这里稍微啰嗦再次介绍一下每一章节的大致内容：

第1章 语义表达

本章首选介绍稀疏词向量、分布式语义假设和稠密词向量的概念，探讨如何处理词汇表外的词，分析了词、子词和字符分词的优缺点。之后讨论如何构建句子向量及衡量语义相似度，分析位置编码和BERT的不同嵌入类型的差异性，最后介绍典型的大模型语义建模方法。

第2章 大模型的数据

本章首先列举大模型训练常用的开源数据集，之后分析了不同训练阶段所需的数据量的差异性问题，强调了数据质量和多样性的重要性，之后讨论了模型扩展策略及灾难性遗忘问题。

第3章 大模型的预训练

本章介绍了预训练与监督微调两者的区别与联系，探讨大模型的涌现能力及优化训练效率的方法。对于模型预训练、监督微调和强化学习三阶段进行了详细的描述，对于训练过程中的显存优化及并行计算的通信开销的相关技术也作了较为深入的探讨。

第4章 大模型的对齐

本章介绍如何构造对齐数据。介绍了PPO与DPO算法，讨论奖励模型训练及PPO稳定化的方法，以及其他偏好对齐方法的阐述。

第5章 大模型的垂类微调

本章介绍了垂类微调及词表扩充方法，探讨如何外推序列长度及常用的损失函数（如交叉熵、z-loss等），以及对一些检索增强生成（RAG）等技术的介绍。

第6章 大模型的组件

本章先介绍了Transformer架构，之后比较不同位置编码方法的特点，以及各种注意力机制的优化方法。之后讨论了归一化技术，分析了BatchNorm、LayerNorm、RMSNorm及Dropout的实现与影响，最后介绍了一些参数初始化Tips.

第7章 大模型的评估

本章先介绍了大模型的评测标准和原则，之后分析生成式模型的评测指标及修复方法，然后探讨自动化评估和对抗性测试的方法以及备案流程。

第8章 大模型的架构

本章介绍了因果解码器架构并分析其成为主流的原因，之后探讨模型融合方法及MoE架构的优势。

第9章 检索增强生成（RAG）

本章首先介绍RAG的核心组件及评估方法，之后分析RAG中的召回策略和重排技术，最后讨论了RAG在实际应用中的工程挑战。

第10章 大模型智能体

本章主要讨论智能体的组成，介绍了智能体的规划、记忆、工具调用等模块。在智能体框架（XAgent、AutoGen）上分析了不同智能体框架的特点和实践方法。

第11章 大模型PEFT（参数高效微调）

本章主要介绍了LoRA的设计思路和实现流程。

第12章 大模型的训练与推理

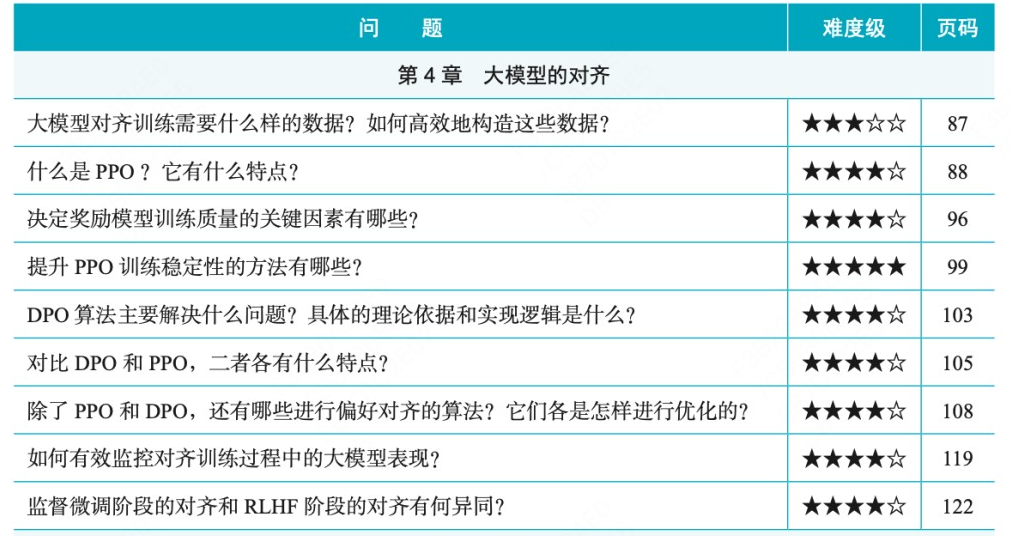
本章综述大模型的解码策略，分析训练和推理阶段的加速技术。在专家并行与量化中探讨了MoE训练的专家并行及模型量化方法。同时介绍数据并行、张量并行和流水线并行策略。

第13章 DeepSeek

最后一章是架构创新相关的介绍，分析了DeepSeek的混合专家设计、MLA（多头局部注意力）和多词元预测技术。最后介绍DeepSeek-R1的训练流程及优化方法。

从各个章节的内容可以看出，《百面大模型》内容全面覆盖了大模型的理论基础、技术实现、应用案例以及未来展望等多个维度。通过丰富的实例和深入的技术剖析，展示了大模型如何在各行各业中发挥其巨大的潜力和价值。

另外本书对面试的问题进行了一些难度的分类，这点还是非常好评的，减少很多的弯路：



**这本书解决了哪些面试痛点**

既然是一本协助我们进行面试的图书，《百面大模型》解决了在AI大模型相关的求职面试中，候选人常遇到概念模糊、缺乏系统性理解、实战经验不足等尖锐的问题。《百面大模型》针对这些痛点，提供了清晰、深入且实用的解决方案，比如说：

**1. 解决“大模型知识零散，难以体系化”的问题**

痛点：面试官常考察大模型的全栈知识，但许多候选人对Transformer、预训练、微调、对齐等概念的理解是割裂的。

本书思路：本书系统梳理大模型技术栈，从词向量→预训练→对齐→推理优化进行层层递进，帮助读者建立完整的知识框架。

**2. 解决“底层原理说不清，被问就卡壳”的问题**

痛点：面试官深挖LLM相关的技术细节（如：RoPE如何实现长度外推？），但很多人只会用API，不懂原理。

本书思路：本书系统深入解析核心算法，如：RoPE（第6章）：数学推导+外推能力实验，MHA、MQA、GQA的区别与代码实现。通过数学公式、代码片段、实验对比，让读者真正理解技术本质。

**3. 解决“训练优化问题答不上来”的问题**

痛点：面试常问“如何降低显存占用？”“怎么优化训练速度？”，但候选人往往缺乏实践经验。

本书思路：详解大模型训练中的显存计算、并行策略、通信开销。例如：本书深度阐述了数据并行 vs. 张量并行 vs. 流水线并行的通信开销计算，减少显存占用的ZeRo等模型训练优化技术。

**4. 解决“前沿技术不了解，被问新论文就懵”的问题**

痛点：面试官可能考察最新技术（如MoE、RAG、Agent），但很多人只熟悉传统Transformer。

本书思路：本书涵盖了2023-2024年大模型前沿技术，例如：MoE架构，检索增强生成（RAG）召回、重排、工程优化，大模型智能体，DeepSeek的创新等新技术，帮助读者紧跟技术趋势在面试中展现行业洞察力。

总结一下就是：

《百面大模型》直击大模型面试中的核心痛点：

☑ 知识零散 → 提供系统化的技术框架

☑ 原理不清 → 深入解析算法和数学推导

☑ 训练优化不会 → 详解显存计算、并行策略

☑ 对齐技术不懂 → 完整剖析PPO、DPO、奖励模型

☑ 评估方法缺失 → 提供评测指标和badcase修复策略

☑ 前沿技术不熟 → 覆盖MoE、RAG、Agent等最新进展

无论是校招生准备大模型岗位，还是在职工程师跳槽面试，本书都能帮助快速构建知识体系，掌握高频考点，提升面试通过率。

**这本书适合谁读**

《百面大模型》适合广泛的读者群体。对于人工智能领域的研究者和工程师，这本书是了解大模型最新进展和技术细节的宝贵资料。通过阅读这本书，他们可以掌握前沿的技术方法，提升自己的研究水平和工程能力。推荐的读者为：

☑ 高校教师和学生：大模型学习和研究的参考书，丰富的实践案例。

☑ AI从业者：掌握LLM核心知识，构建系统知识图谱

☑ AI感兴趣读者：足够的背景知识和案例，理解大模型在现代社会中的应用和影响。

**总结：从“应试”到“实战”的AI能力跃迁**

### 最后总结一下这本书额度特点就是：从“应试”到“实战”的AI能力跃迁。《百面大模型》的本质是用面试题作矛，以工程实践项目为盾，帮助我们的广大AI开发者： （1）突破“纸上调参”陷阱和窘境，掌握生产环境下的大模型的建设能力。 （2）从“单点技术”进化到系统工程思维，不局限于简单的数据闭环。 （3）在AI技术快速迭代中建立可持续学习框架，而非追逐碎片化知识点。

### 正如在人工智能领域所体现的“先发优势”竞争逻辑——那些提前半年掌握系统化能力的个体，将享受到职业价值的指数级增长。所以，加油吧！

另外作者得荐大家关注一下：

