人工智能

语文出版社

课时分配表

|  |  |  |  |
| --- | --- | --- | --- |
| **章序** | **课程内容** | **课时** | **备注** |
| **1** | 初识人工智能 | 6 |  |
| **2** | 人工智能应用 | 7 |  |
| **3** | 新一代人工智能生态 | 6 |  |
| **4** | 大模型实践 | 6 |  |
| **5** | 人工智能伦理 | 8 |  |
| **6** | 机器学习 | 3 |  |
| **总计** |  | 36 |  |

# 第4课 大模型实践

|  |  |  |
| --- | --- | --- |
| **课 题** | 大模型实践 | |
| **课 时** | 6课时（270 min）。 | |
| **教学目标** | **知识技能目标：**  1．理解大模型的定义、特点、发展历程以及其在人工智能领域的重要地位。  2．掌握大模型在文本处理、图像识别、自然语言处理等领域的具体应用案例。  **思政育人目标：**  让学生通过学习大模型实践，在使用和开发大模型时，能够评估大模型在隐私保护、数据安全、算法偏见等方面的风险，并采取相应措施加以应对，确保技术的健康发展和社会福祉的最大化。 | |
| **教学重难点** | **教学重点：** 大模型概述  **教学难点：**语言大模型 | |
| **教学方法** | 讲授法、问答法、讨论法 | |
| **教学用具** | 电脑、投影仪、多媒体课件、教材 | |
| **教学设计** | 第1节课：考勤（2min）--知识讲解（40min）--作业布置（3min）  第2节课：知识讲解（40min）--课堂小结（3min）--作业布置（2min）  第3节课：知识讲解（40min）--课堂小结（3min）--作业布置（2min）  第4节课：知识讲解（40min）--课堂小结（3min）--作业布置（2min）  第5节课：知识讲解（40min）--课堂小结（3min）--作业布置（2min）  第6节课：知识讲解（40min）--课堂小结（3min）--作业布置（2min） | |
| **教学过程** | **主要教学内容及步骤** | **设计意图** |
| **考勤**  **（2min）** | ■【教师】清点上课人数，记录好考勤  ■【学生】班干部报请假人员及原因 | 培养学生的组织纪律性,掌握学生的出勤情况 |
| **知识讲解**  **（40min）** | **【教师】**讲解包罗万象—— 大模型概述（一）  **一、大模型技术的发展历程**  2006 年，杰弗里·辛顿（Geoffrey Hinton）提出逐层无监督预训练方法，这一方法有效缓解了深层网络训练时的梯度消失问题，为神经网络学习提供了重要优化途径。此后，深度学习在多个领域取得突破性进展，掀起新一轮的发展浪潮。回顾过去十多年，基于深度学习的人工智能技术经历了显著的研究范式转变：从早期“标注数据监督学习”的任务特定模型，发展到“无标注数据预训练 + 标注数据微调”的预训练模型，再到如今“大规模无标注数据预训练 + 指令微调+ 人类对齐”的大模型。这一历程体现了技术从小数据到大数据、小模型到大模型、专用到通用的飞跃，标志着人工智能技术正迈入大模型时代。  2022 年年底，由 OpenAI 发布的语言大模型 ChatGPT 引发了社会的广泛关注，如图 4-1 所示。在“大模型＋大数据＋大算力”的加持下，ChatGPT 能够通过自然语言交互完成多项任务，具备了多场景、多用途、跨学科的任务处理能力。以 ChatGPT 为代表的大模型技术可以在经济、法律、社会等诸多领域发挥重要作用。大模型被认为很可能像 PC 时代的操作系统一样，成为未来人工智能领域的关键基础设施，引领行业发展热潮。    **（一）统计语言模型**  统计语言模型主要依据马尔可夫假设来构建文本序列的生成概率。其中，N-gram 语言模型是一个典型代表，它认为下一个词汇的出现概率仅与前面 N 个词汇相关（即遵循 N 阶马尔可夫假设）。然而，这类模型存在显著的问题：它们容易受到数据稀疏的影响，导致概率分布估计不准确。为改善这一状况，需采用平滑策略进行调整。但即便如此，统计语言模型在文本序列建模方面的能力仍然相对较弱。  **（二）神经语言模型**  针对统计语言模型的弊端，神经语言模型通过神经网络（如 MLP、RNN）建模词汇间的语义共现关系，有效捕获复杂语义依赖，精准预测词汇生成概率。而 word2vec 简化了神经语言模型架构，能从无监督语料中学习可迁移的词向量，为预训练语言模型的研究奠定基础。  **（三）预训练语言模型**  预训练语言模型主要是基于“预训练＋微调”的学习范式构建，首先通过自监督学习任务从无标注文本中学习可迁移的模型参数，进而通过有监督微调适配下游任务。早期的代表性预训练语言模型包括 ELMo、GP T-1 和 BERT 等。其中，ELMo 模型基于传统的循环神经网络（LSTM）构建，存在长距离序列建模能力较弱的问题；随着 Transformer 的问世，神经网络序列建模能力得到了显著提升，GPT-1 和 BERT 都是基于 Transformer 架构构建的，可通过微调学习解决大部分的自然语言处理任务。  **（四）语言大模型**  1. 探索阶段  在预训练语言模型的发展历程里，扩展定律（Scaling Law）扮演了关键角色。该定律表明随着模型参数和预训练数据规模的增加，模型的能力与任务效果会随之提升。OpenAI 在研发 GPT系列时，通过 GPT-1（1.1 亿参数）、GPT-2（15 亿参数）到 GPT-3（1 750 亿参数）的逐步扩展，有力地验证了这一法则。谷歌推出的 PaLM 模型，其参数规模更是高达 5 400 亿。当模型参数达到千亿量级别时，语言大模型展现出显著的能力跃升。GPT-3 便能在无微调情况下，仅通过提示词或少数样例完成多任务，甚至在某些领域超越当时的专用模型。因此，学术界专门用“语言大模型”（Large Language Models）来强调这种超大规模预训练语言模型与早期模型的区别。  2. 提升阶段  早期的语言大模型虽展现出少样本学习能力，但主要局限于预测下一个单词，难以有效遵循人类指令，可能输出无用甚至有害信息，与人类偏好存在偏差。为解决这些问题，主要采用了这些技术：指令微调（Instruction Tuning），通过格式化的训练数据（指令和回答配对）强化大模型的通用任务泛化能力；基于人类反馈的强化学习（RLHF），引入人类标注者训练与人类偏好对齐的奖励模型，指导大模型训练，使其更好地遵循用户意图。同时，在大模型应用中，利用提示技术如思维链（CoT）、思维树（ToT）等，充分挖掘大模型的潜力，提升其解决实际问题的能力。  此外，针对大模型在特定能力上的不足，如数值计算，可通过外部工具如计算器、搜索引擎等来扩展其能力边界，弥补其基于文本数据训练与推理的局限性。  作为重要前沿探索力量，OpenAI 对于语言大模型的研发工作主要是在 Transformer 架构推出后开展，并取得了一系列的技术进展。其中，GPT-1 探索了解码器 Transformer 架构（decoderonly Transformer）在“预训练 + 微调”范式下的自然语言任务求解能力；GPT-2 初步验证了扩大模型参数规模的有效性（扩展法则），并探索了基于自然语言提示的多任务解决能力；GPT-3 首次探索了千亿参数规模的语言模型效果，提出了基于“上下文学习”的任务解决方法；CodeX使用代码数据对 GPT-3 进行了微调，从而提升了代码能力和复杂推理能力；InstretGPT 基于人类反馈的强化学习技术，能够强化对于人类指令的遵循能力和人类偏好的对齐能力；ChatGPT 与 InstruciCPT 的技术原理相似，进一步引入了对话数据进行学习，从而加强了多轮对话能力；GPT-4 能够处理更长的上下文窗口，具备多模态理解能力，在逻辑推理、复杂任务处理方面的能力得到显著改进，但其他相关技术细节未被披露。  随着 GPT-4 的推出，语言大模型已实现对多模态领域产生了重大影响。它不再局限于单调的文本交互，而是能够接收文本与图像的组合输入。这种多模态大模型更符合人类多渠道感知的认知方式，能够灵活应对复杂多变的环境、场景和任务需求。GPT-4 的成功实践表明，在多模态大模型中融入基于人类知识的自然语言，能显著提升模型在多模态理解、生成及交互等方面的能力。 | **讲解包罗万象—— 大模型概述（一），让学生更加仔细的了解知识，从而激发学生的学习欲望。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了包罗万象—— 大模型概述（一），让学生了解大模型技术的发展历程。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  简述语言大模型。 | 通过课后练习，使学生巩固所学新知识 |
| **知识讲解**  **（40min）** | **【教师】**讲解包罗万象—— 大模型概述（二）  **二、大模型技术的生态发展**  大模型服务平台正向个人开放及商业落地应用延伸，不同公司互有侧重，为用户提供了多种获取大模型能力的途径。OpenAI API 较早地面向公众开放大模型服务平台，用户可以通过API 就能访问不同的 GPT 模型来完成下游任务。  Claude 系列模型是 Anthropic 打造的闭源语言大模型，有 Claude 和 Claude-Instant 两种。它们历经无监督预训练、人类反馈的强化学习及 Constitutional AI 技术（融合监督训练与强化学习）的锤炼，力求在实用性、诚信度及安全性上更胜一筹。Claude 支持高达 100K 词元的上下文，而Claude-2 更是升级至 200K 词元。文心一言是基于百度文心大模型的知识增强语言大模型，提供手机应用程序（App）、网页版、API 接口等多种形式的开放服务。此外，文心一言还建设了插件机制，通过外部工具、服务的调用，拓展大模型的能力边界。讯飞星火认知大模型具有开放式知识问答、多轮对话能力，在逻辑和数学运算方面表现突出，并且具有较强的对代码和多模态的理解能力。讯飞和华为还联合重磅发布了国内首款支持大模型训练私有化的全国产化产品“星火一体机”，可支持企业快速实现讯飞星火大模型的私有化部署、场景赋能和专属大模型训练优化。  大模型开源生态丰富多样，汇聚了众多开源框架与模型。PyTorch 等框架凭借分桶梯度、通信与计算重叠等先进技术，赋能大规模分布式训练；国产飞桨框架内置大规模分布式训练支持，4D 混合并行策略助力千亿级模型训练；OneFlow将集群构建为超级设备，通过动静态图转换与混合并行提升训练效率；微软 DeepSpeed 框架则运用 ZeRO 技术降低内存冗余，轻松应对万亿级模型的训练挑战。  开源大模型可降低大模型研究的门槛，促进大模型应用走向繁荣。其中，典型代表有：LLaMA 系列，这是 Meta 研发的开源大模型，参数规模从 7B 到 65B 不等，仅依赖公开数据集进行预训练，通过数据过滤和并行优化实现高效训练；Falcon 系列，由阿布扎比的 TII（Technology Innovation Institute）研究院推出，最大规模达 180B 参数，基于开源许可发布，性能与 GPT-4 和PaLM2 相当，参数量却较小；GLM 系列，采用空白填充等多任务联合训练方式，提升了模型的生成能力；Baichuan 系列，由百川智能开发，支持中英双语，使用高质量训练数据，在多个基准测试上表现优秀，还开源了多种量化版本；Baichuan 2，在保留原有模型优势的基础上，增强了逻辑推理等方面的能力；CPM 系列，采用经典的语言模型自回归训练方式，在各类中文 NLP 任务上均表现卓越。  大模型技术广泛赋能各行业：传媒领域实现智能新闻撰写，降低运营成本；影视行业拓宽创作视野，提升作品品质；营销领域打造虚拟客服，助力产品推广；娱乐领域增强人机互动，提升趣味性；军事领域强化情报获取与决策，实现实时战场支持；教育领域焕发教材新活力，推动个性化智能教学；金融领域助力机构降本增效，提升服务温度；医疗领域赋能诊疗全流程。大模型的发展将为人类社会带来强大动力，促进数字与现实世界的深度融合。  大模型的通用性使其有望成为未来人工智能应用中的关键基础设施，如同 PC 时代的操作系统，赋能百业，加速推进国民经济的高质量发展。向上，大模型可带动上游软硬件计算平台的革新，实现高性能软硬件与大模型的协同发展，构建“大模型＋软硬件＋数据资源”上游发展生态；向下，大模型可以打造“大模型 + 应用场景”的下游应用生态，加速全产业的智能升级，为经济、社会和安全等领域的智能化升级转型提供关键支撑。  **三、大模型技术的风险与挑战**  尽管以 ChatGPT 为代表的大模型技术已取得关键性突破，但当前大模型技术仍存在诸多风险与挑战。  1. 大模型的可靠性无法得到有效保障  例如，基于海量数据训练的语言大模型，尽管其生成的内容符合语言规则、通顺流畅且与人类偏好对齐，但其合成内容在事实性、时效性等方面仍存在较多问题，尚无法对所合成内容做出可靠评估。  2. 大模型的可解释性存在不足  大模型依托深度神经网络，其内部工作机理复杂且难以透彻理解。语言大模型的涌现特性、规模增长规律，以及多模态大模型在知识表征、逻辑推理、泛化适应和情景学习等方面的能力，均亟待更为深入的探索与研究。这些研究将为大模型的大规模实际应用奠定坚实的理论基础，确保其性能稳定且可靠。  3. 大模型应用部署代价高  大模型的参数规模和数据规模都非常庞大，存在训练和推理计算量大、功耗高、应用成本高、端侧推理存在延迟等问题，从而限制了其落地应用。提高推理速度、降低使用成本是大模型大规模应用的关键。  4. 大模型在小数据情景下的迁移能力存在不足  大模型基于数据驱动的深度学习方式，依赖训练数据所覆盖的场景，由于复杂场景数据不足，大模型存在特定场景适用性不足的问题，面临鲁棒性和泛化性等难题。如何提升大模型对小数据的高效适配与迁移能力是未来研究的重点。  5. 大模型还存在伴生技术风险问题  语言大模型具备强大的自然语言处理能力，与多媒体技术结合可生成逼真的音频、视频内容，但这也可能被滥用制造虚假信息，引发舆论危机乃至威胁国家安全。大模型存在的安全漏洞，如数据投毒、对抗样本、模型窃取、后门及指令攻击等，可能使关联业务瘫痪，危及应用生态。此外，大模型训练涉及海量互联网数据，个人、企业及国家敏感信息可能被编码进模型参数，存在数据泄露风险。特别是通过精心构造的提示信息，可能诱使大模型暴露隐私数据。 | **通过教师讲解，熟悉包罗万象—— 大模型概述（二）。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了包罗万象—— 大模型概述（二），让学生了解大模型技术的生态发展及大模型技术的风险与挑战。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  如何平衡大模型技术的风险与收益？ | 通过课后练习，使学生巩固所学新知识 |
| **知识讲解**  **（40min）** | **【教师】**讲解滔滔不绝—— 语言大模型（一）  **一、Transformer 架构**  Transformer 架构，作为当前语言大模型的核心选择，其根基在于自注意力机制（Selfattention Mechanism），如图 4-3 所示。该架构旨在通过自注意力机制全面捕捉输入序列的全局信息，并逐层传递。其标准形态为编码器—解码器结构，编码器与解码器均由单一的编码层及多层相同的 Transformer 模块构成。这些模块内部融合了多头注意力机制与全连接前馈网络，两者通过残差连接及层归一化技术紧密相联。  与编码器模块相比，解码器由于需要考虑解码器输出作为背景信息进行生成，因此其中每个 Transformer 层多了一个交叉注意力层。相比于传统循环神经网络（Recurrent Neural Network，RNN）和长短时记忆神经网络（Long Short-Term Memory Network，LSTM），Transformer 架构的优势在于具备并行计算能力，即不需要按照时间顺序地进行计算。Transformer 架构包含编码层与 Transformer 模块两个核心组件。    **（一）编码层**  编码层主要是将输入词序列映射到连续值向量空间进行编码，每个词编码由词嵌入和位置编码构成，由二者相加得到最终编码。  1. 词嵌入  在 Transformer 架构中，词嵌入是输入数据的第一步处理过程，它将词映射到高维空间中的向量，可以捕获词汇的语义信息，如词义和语法关系。每个词都被转化为固定长度的向量，然后被送入模型进行处理。  2. 位置编码  由于自注意力机制本身对位置信息不敏感，为使模型能够理解序列中的顺序信息，引入了位置编码。标准 Transformer 架构采用正弦和余弦函数进行位置编码。对于每个位置 i，对应的位置编码是长度为 d 的向量，其中 d 是模型的嵌入维度。这个向量的第 j 个元素由公式计算：如果 j 是偶数，编码的第 j 个元素为 sin（i/10000j/d）；如果 j 是奇数，编码的第 j 个元素为 cos（i/10000j/d）。  **（二）Transformer 模块**  Transformer 模块通过自注意力机制获取输入的全局信息，并将这些信息传递给多头注意力层和全连接前馈网络层，这两层之间通过残差连接和层归一化操作进行连接，共同构成了Transformer 模块的基本单元，如图 4-4 所示。    在 Transformer 模型被提出之后，衍生出了诸多变体，包括在编码器和解码器中不同方式的注意力机制、归一化操作、残差连接、前馈层和位置编码等。  **二、语言大模型架构**  当前，现有的语言大模型几乎全部是以 Transformer 模型作为基础架构来构建的，不过在所采用的具体结构上通常存在差异，如只使用 Transformer 编码器或解码器，或者同时使用两者。  从建模策略的角度看，语言大模型架构大致可以分为以下三种。  **（一）掩码语言建模**  掩码语言建模（Masked Language Modeling，MLM）是基于 Transformer 编码器的双向模型，BERT[11] 和 RoBERTa 是其中典型代表。这类模型通过掩码语言建模任务进行预训练，BERT 中还加入了下一句预测（Next Sentence Prediction，NSP）任务。在预训练阶段，模型的输入是自然语言序列。首先在原始输入中添加特殊标记 [CLS] 和 [SEP]，并且随机用 [MASK] 标记替换原始序列中的字符。掩码语言建模旨在根据上下文来最大化 [MASK] 位置的标签字符的条件概率，即让模型执行“完形填空”任务。而 [CLS] 的最终表示被用于预测两个句子是否连贯。RoBERTa与 BERT 基本相同，但 ROBER Ta 删去了下一句预测任务，采用了更具鲁棒性的动态掩码机制，并使用更大的批次、更长的训练时间和更多的数据进行训练。  **（二）自回归语言建模**  自回归语言模型借助 Transformer 解码器，通过预测序列中的后续词来构建语言模型，其核心目标是最大化序列各位置后续词的条件概率。此类模型的典型代表有 OpenAI 的 GPT 系列、Meta 的 LLaMA 系列及 Google 的 Pa LM 系列。其中 GPT-3 具有里程碑意义，率先将模型参数规模提升至千亿级，显著增强了模型在生成任务及大规模扩展方面的能力。  **（三）序列到序列建模**  序列到序列模型是建立在完整 Transformer 架构上的，即同时使用编码器—解码器结构，代表性模型包括 T5 和 BART。这两个模型都采用文本片段级别的掩码语言模型作为主要的预训练任务，即随机用单个 [MASK] 特殊标记替换文本中任意长度的一段字符序列，并要求模型生成填充原始的字符。序列到序列模型可以形式化地表示为在给定掩码的字符序列的情况下最大化，目标字符序列的概率。  总体而言，自回归语言模型较其他预训练语言模型架构展现了更优异的上下文学习、思维链推理、内容创造等能力，自回归模型架构是当前大模型的主流架构。 | **通过教师讲解，掌握滔滔不绝—— 语言大模型（一）。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了滔滔不绝—— 语言大模型（一），让学生了解Transformer 架构以及语言大模型架构。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  简述语言大模型架构。 | 通过课后练习，使学生巩固所学新知识 |
| **知识讲解**  **（40min）** | **【教师】**讲解滔滔不绝—— 语言大模型（二）  **三、语言大模型关键技术**  语言大模型技术主要包括模型预训练、适配微调、提示学习、知识增强和工具学习等。  **（一）预训练**  支撑语言大模型高效训练的技术，主要包括高性能训练策略、高效预训练策略、高质量训练数据、高效模型架构等。  1. 高效预训练策略  高效预训练策略主要思路是采用不同的策略，以更低成本实现对语言大模型的预训练。  （1）在预训练中设计高效的优化任务目标，使模型能够利用每个样本更多的监督信息，从而实现模型训练的加速。  （2）热启动策略，在训练开始时线性提高学习率，以解决在预训练中单纯增加批处理大小，可能导致的优化困难问题。  （3）渐进式训练策略，不同于传统的训练范式使用相同的超参数同时优化模型每一层，该方法认为不同层可以共享相似的自注意力模式，首先训练浅层模型，然后复制构建深层模型。  （4）知识继承方法，即在模型训练中同时学习文本和已经预训练语言大模型中的知识，以加速模型训练。经测试，在中文语言大模型 CPM-2 中采用知识继承技术，可以使大模型在预训练前期提速 37.5%。  （5）可预测扩展策略（Predictable Scaling），旨在大模型训练初期，利用大模型和小模型的同源性关系，通过拟合系列较小模型的性能曲线预测大模型性能，指导大模型训练优化。OpenAI在 GPT-4 训练中，使用 1 000～10 000 倍较少计算资源训练的小模型，可靠地预测 GPT-4 某些性能，大幅降低模型训练成本。  2. 高效的模型架构  BERT 之后的 Transformer 架构在提高自然语言处理效率方面，有两个重要优化方向。  （1）统一序列建模，旨在整合分类、信息抽取、翻译、对话等多样自然语言处理任务至统一框架，于单一模型中并行处理多任务，以提效自然语言处理能力。此策略凭借大规模训练数据，优化模型多任务表现与泛化性，缩减多模型开发维护的繁复与资源开支，增强模型普适性。实践路径有两个方面：一方面，将任务归一为序列生成，如 T5、BART 将诸任务转化为文本生成文本；另一方面则依托语言大模型预训练，借由文本内嵌的语言提示，灵活应对各类任务。  （2）计算高效的模型架构，从 Transformer 模型架构本身的处理训练复杂度、编解码效率、训练稳定性、显存利用等方面进行优化。比如，Transformer 的并行处理机制是以低效推理为代价的，解码时每个步骤的复杂度为 O（N），Transformer 模型也是显存密集型模型，输入序列越长、占用的内存越多。为此，微软设计了一种新的 Transformer 架构 RetNet，采用线性化注意力+尺度保持（Retention）机制，在基本保持模型性能的同时，实现模型训练速度、推断速度大幅提升和内存节约。  针对自注意力显存消耗大的挑战，斯坦福大学提出了 FlashAttention 算法，它是一种快速、内存高效的注意力算法，具有 IO 感知能力，无须任何近似即可加速注意力计算，并显著降低内存占用。该算法已被广泛采纳于各大主流模型中，有效支持超长文本输入的处理。与此同时，模块化大模型架构因其利用神经激活稀疏性对稠密模型进行模块化划分而备受瞩目，不同任务仅需通过部分模块即可完成计算，从而实现了训练和推理的加速。这一领域的典型工作包括 Google的 Switch Transformers 和 Pathways 架构，以及清华大学的 MoEf ication 和 FastMoE 架构等，它们均通过混合专家的方式融合不同的模块化大模型，进一步提升模型处理的效率和灵活性，如图4-5 所示。    **（二）适配微调**  语言大模型虽在大规模通用数据上预训练，但往往缺特定任务或领域知识，故需微调以适配。微调能让模型更贴合特定需求，如处理医疗记录等敏感数据，同时保障数据不暴露。此外，它还能提升部署效率、降低计算资源消耗。其中，指令微调和参数高效学习是微调的关键技术。  1. 指令微调（Instru ction Tuning）  指令微调是一种可以帮助语言大模型具备遵循人类语言指令能力，在零样本设置中泛化到未见任务的学习方法。指令微调学习形式与多任务提示微调相似，但与提示微调让提示适应语言大模型并且让下游任务对齐预训练任务不同，它是让语言大模型对齐理解人类指令并按照指令要求完成任务，即在给定指令提示（Instruction Prompt）的情况下给出特定的回应，提示可以选择性包含解释任务的指令。指令微调研究涉及指令理解、指令数据获取和指令对齐等内容。  （1）指令理解是语言大模型执行任务的关键能力，确保模型能准确领会人类指令并付诸行动。为强化这一能力，研究者们常采用多任务提示微调方法，在涵盖广泛指令描述的任务集上训练模型，如 FLAN、InstructGPT 等。经过此类微调，模型在面对新任务时展现出良好的零样本性能。  （2）指令数据获取是指如何构建包含多样性的任务指令数据。指令数据构建常见有三种方式：①基于公开人工标注数据构建，代表指令数据集包括 1 616 种不同任务的 Super-Natural Instruction、2 000 种不同 NLP 任务的 OPT-IML。②借助语言大模型的自动生成构建，如Unnatural Instructions，通过种子指令作为提示让语言大模型生成新的指令描述和问题，然后再输入到模型让其输出回答。③基于人工标注方法，如 ChatGPT 在人工标注指令的基础上通过GPT-3、InstructGPT 等在线平台收集用户的真实指令数据。  （3）指令对齐旨在解决语言大模型在生成内容时可能出现的虚假信息、错误目标或偏见内容等问题。由于大模型预训练时仅关注语言建模，未融入人类价值观或偏好，导致这类问题频发。指令对齐的提出，旨在使模型输出更贴近人类预期，强调有用性、诚实性和无害性。然而，这一过程可能会牺牲模型的某些通用能力，即所谓的“Alignment Tax”。为实现与人类价值的对齐， InstructGPT采用基于人类反馈的微调方法，结合强化学习技术，将人类反馈融入微调过程。ChatGPT 也采用了类似技术，以确保输出高质量且无害。指令对齐的推广，促使适配微调从传统的纯数据学习范式向人类学习范式转变。  2. 参数高效微调（Parameter-Efficient Tuning）  早期 BERT 为代表的微调方法，存在任务“鸿沟”和高计算成本两大问题。任务“鸿沟”源于预训练与微调任务形式的不一致，影响知识迁移效能；而高计算成本则因语言大模型参数规模庞大，全参微调需消耗大量资源。为解决这些问题，参数高效学习成为有效途径，即通过微调少量参数使大模型在下游任务上达到全参微调效果。目前，已提出多种参数高效微调方法，这些方法主要分为以下三种类型。  （1）添加式方法：在原模型基础上引入额外的模块或参数，并仅微调该引入部分的参数。例如，适配器（Adapter）方法，将小规模的神经模块（适配器）注入预训练模型，并仅调整这些适配器以进行模型自适应。在实际应用中，适配器模块通常分别插入多头自注意力和前馈网络子层之后，成为最广泛的使用方式。  （2）指定式方法：将原模型指定模型中部分参数设为可训练参数，并固定模型其他参数。这类方法也十分简单有效，如仅通过优化模型内的偏置项并固定其他参数，模型仍然可以再现95% 以上的模型全参微调性能。  （3）重参数化方法：将原模型或部分模型参数重参数化到低维度参数空间，仅仅优化低维空间中的近似参数，从而显著降低模型的计算量和内存消耗。例如，LoRA 将模型自注意力模块的变化权重参数分解为两个低秩矩阵相乘，即 W=Wo+ΔW=Wo+WdownWup。  参数高效微调通常具有微调参数规模小、增量式微调参数、即插即用等特点，这种技术也被统一成技术框架 Delta Tuning。围绕参数高效微调，一些开源工具也被研发，具有代表性的包括 OpenPrompt、OpenDelta 等。由于不同任务的微调参数可以被重复利用，一些关于高效微调的仓库也被构建，如 AdapterHub、Delta Center 等。随着语言大模型的兴起，高效微调受到了越来越多的关注，现已开发了一种更轻量级的下游任务适配方法。特别是 LoRA 已广泛应用于各种开源语言大模型（如 LLaMA）以实现参数高效微调。  **（三）提示学习**  语言大模型经过大规模文本数据预训练后，虽具备通用任务求解的潜力，但在特定任务中这些能力可能并不明显。为此，模型提示技术应运而生，通过在模型输入里加入恰当的语言指令，有效激发这些潜能。其中，指令提示和思维链提示是两种具有代表性的提示技术。  1. 指令提示  指令提示，又称为提示学习。OpenAI 在 GPT-3 中首次提出上下文提示，并发现 GPT-3 在少样本提示下能够达到人类水平，证明其在低资源场景下非常有效，从而引起广泛关注。指令提示核心思想是避免强制语言大模型适应下游任务，而是通过提供“提示（Prompt）”来给数据嵌入额外上下文以重新组织下游任务，使其看起来更像是在语言大模型预训练过程中解决的问题。指令提示有如下三种形式。  （1）少样本提示：是指在一个自然语言提示后面附加一些示例数据，作为语言大模型的输入。它可以提高语言大模型在不同领域和任务上的适应性和稳定性。少样本提示也存在一些挑战，如如何确定合适的示例数量、如何选择示例等。  （2）零样本提示：是指不使用任何示例数据，仅依靠一个精心设计的提示，激活语言大模型中与目标任务相关的知识和能力。零样本提示关键问题包括如何设计合适的提示、如何选择最优的提示等。  （3）上下文学习，又称为情境学习，是指将一个自然语言问题作为语言大模型输入，并将其答案作为输出。上下文学习可以看作是一种特殊形式的少样本提示，在问题中隐含地包含了目标任务和格式信息。上下文学习可以简化问题表示和答案生成，并且可以灵活地处理多种类型和复杂度的问题。但面临确保问题质量、如何评估答案正确性等挑战。  2. 思维链  推理过程常含多个步骤，多步推理能产出可验证输出，增强黑盒模型的可解释性。思维链作为一种提示技术，被广泛用于激发语言大模型的多步推理潜能，促使模型生成解决问题的中间推理步骤，模仿人类深思熟虑的过程。在思维链提示中，〈输入，输出〉对被中间自然语言推理步骤取代，形成〈输入，思维链，输出〉结构。思维链被视为语言大模型的“涌现能力”，通常需模型参数达到一定规模才具备。激活这一能力的方法是在提示中提供逐步推理演示，每个演示均含问题及其通向答案的推理链。  **（四）知识增强**  知识运用与推理能力是衡量语言大模型智能的关键指标。Allen AI 研究表明，GPT-3 等大模型在处理含预设立场的简单常识问题，如“太阳有几只眼睛？”时，可能给出荒谬答案，如“太阳有两只眼睛”。为提升模型智能水平，需将外部知识融入深度学习模型。依据知识融合部位不同，方法可归纳为四类。  1. 知识增广  知识增广方法是从输入端增强模型，有两种主流的方法：（1）直接把知识加到输入。（2）设计特定模块来融合原始输入和相关的知识化的输入表示。  2. 知识支撑  知识支撑方法关注优化模型对知识的应用流程。方式有两种：（1）在模型的底部引入知识指导层来处理特征，以便能得到更丰富的特征信息。例如，使用专门的知识记忆模块从大模型底部注入丰富的记忆特征。（2）将知识作为“专家”，在模型顶层构建后处理模块，以计算得到更准确和有效的输出。  3. 知识约束  知识约束方法利用知识构建额外的预测目标和约束函数，来增强模型的原始目标函数。例如，远程监督学习，利用知识图谱启发式标注语料作为新的目标，并广泛用于实体识别、关系抽取等系列 NLP 任务。  4. 知识迁移  模型知识作为关键知识源，可直接应用于下游任务以初始化模型参数。迁移学习与自监督学习是知识迁移的重要研究领域，现今以 BERT 为代表的预训练模型已成为知识迁移的主流方法，并在自然语言处理中广泛应用。  **（五）工具学习**  语言大模型具备理解、推理和决策能力，并能与外部工具有效互动。在金融等特定领域，如证券交易和市场预测场景中，语言大模型常需结合外部工具以获取必要信息和技能。这种整合方式能充分发挥两者优势，助力复杂任务处理：外部工具强化专业知识和可解释性，而语言大模型则提供强大的语义理解和推理规划能力。  2021 年年底，OpenAI 推出 WebGPT，利用 GPT-3 与网页浏览器和搜索引擎交互，获取互联网信息，在长文本问答方面实现非常强的能力，展现了语言大模型利用工具解决复杂问题的巨大潜力。该项成果引起了学术界和产业界的广泛关注，产生了许多面向不同任务或场景需求的大模型调用工具的方法，如 Webshop，使用语言大模型替代人在购物平台上执行一系列操作、购买所需物品。2023 年 3 月，OpenAI 发布 ChatGPT Plugins，实现 ChatCPT 调用各种外部插件的功能，支持浏览器实时信息获取、代码解释器、PDF 阅读等能力，截至 8 月已支持 480 个常用工具插件。Meta 将这种通过非参数的外部模块扩展语言大模型能力的方法，统一称为增广语言模型（Augmented Language Models）。清华大学基于现有大模型工具使用方法，提出了工具学习框架，旨在让模型能够理解和使用各种工具完成任务的学习过程。  目前，可交互的通用工具依据用户接口分为物理交互工具（如机器人、传感器）、图形用户界面工具（如浏览器、Office 软件）和编程接口工具（如数据库、知识图谱）。从学习目标看，工具学习方法主要分为两类：工具增强学习，利用工具执行结果提升模型性能，视其为生成高质量输出的外部资源；工具导向学习，则聚焦于工具执行本身，旨在开发能替代人类控制工具并进行序列决策的模型，转变学习重点。 | **通过教师讲解，熟悉滔滔不绝—— 语言大模型（二）。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了滔滔不绝—— 语言大模型（二），让学生了解人工智能发展的第二个繁荣期、复苏期和第三个繁荣期。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  提示学习是如何帮助语言大模型更好地适应用户输入和特定任务的？ | 通过课后练习，使学生巩固所学新知识 |
| **知识讲解**  **（40min）** | **【教师】**讲解五彩斑斓 —— 多模态大模型  **一、多模态大模型的技术体系**  当前，多模态大模型主要有面向理解任务、面向生成任务、兼顾理解和生成以及知识增强的多模态大模型。  **（一）面向理解任务的多模态大模型**  面向理解任务的多模态大模型，通常把基于 Transformer 的编码器当作核心结构。根据模型结构的不同，这类多模态大模型可以被进一步划分为单流结构和多流结构两种类型。在单流结构中，不同模态的特征会被拼接在一起，然后通过一个共享的 Transformer 网络来进行处理；而在多流结构中，不同模态则会分别由各自的 Transformer 网络进行编码处理，并且这些网络之间还会存在一些特征上的交互融合机制。  多流结构的一个典型代表是图文理解模型 ViLBERT，它采用了双流 Transformer 的结构，先将文本和图像数据分别输入两个独立的 Transformer 编码器，接着使用互注意力 Transformer（Co-Attention Transformer）层将文本和图像特征进行融合，最终所得到文本—图像特征可以被应用到视觉问答、图像描述生成等不同的多模态的任务中。多流结构的另一个代表是 OpenAI 公司的 CLIP 模型，它采用两个独立的编码网络对图像和文本进行特征抽取，并通过对比学习将两者的特征嵌入共享的语义空间中。CLIP 基于 4 亿图文对进行训练，可以从自然语言监督中有效地学习视觉概念，从而获得泛化性能极强的零样本分类能力。另一个与 CLIP 类型的代表性方法 ALIGN，使用对比损失训练了一个简单的双编码器模型，利用包含超过 10 亿个噪声图像—文本对的数据集来扩展视觉和视觉语言表征学习。CLIP 是一个图文双流结构，而VATT 则是针对视频—文本—音频数据的多流模型。与 CLIP 类似，VATT 将每个模态线性投影为特征向量，然后将其分别送到 Transformer 编码器中，并将编码后的特征在语义分层的不同粒度空间中通过对比学习用于训练模型。  单流结构的一个典型代表是 V L-BERT，它将图像的描述文本和关键物体的区域特征拼接后作为 BERT 网络的输入，通过掩码掉部分文本输入和图像输入，并预测所缺失的信息来进行模型训练。另一代表性方法 UNITER，则采用了一种多任务的多模态预训练方法。相对于其他方法，该模型增加了单词与图像区域的匹配模块，来更进一步建立图像与文本的细粒度关联。在视频领域，单流结构的代表性方法有 Vi deoBERT 和 Act BERT，其中 Vid eoBERT 是视频—语言模型，融合了文本和视频作为 BERT 网络的输入；而 ActBERT 采用了一种全局—局部关系的建模方法，输入不只包括文本和视频的全局信息，还利用了视频帧中的局部信息来加强对于视频内容的理解。  目前，面向理解任务的多模态大模型，其结构大多以单流和多流两类为基础。除此之外，预训练任务的研究也备受关注，旨在通过引入更多预训练任务或构建统一架构来训练所有任务。例如，Florence 作为一种典型方法，其重点在于使模型能够灵活适应各种下游任务。为此，Florence 设计了一个结合多模态大模型和适应模型的工作流。在任务适应方面，该模型采用动态头部适配器，将学习到的视觉特征表示从场景拓展至对象；同时，利用 CoSwin 适配器来学习视频表示；并且，通过 METER 适配器将模型应用于依赖细粒度视觉—语言表示的视觉语言任务中。  **（二）面向生成任务的多模态大模型**  面向生成任务的多模态大模型，能够实现文本、图片、视频、音频、3D、分子结构等多种模态内容的生成应用。目前常用的方法主要是基于序列生成模型和扩散模型（Diffusion Models）。  DALL-E 与 CogView 是序列生成模型的佼佼者。DALL-E 是 OpenAI 发布的图像生成模型，基于 4 亿图文对训练，采用 VQVAE 图像离散自编码器和 GPT 组合结构，实现了以文生图的突破性生成质量和泛化能力，被誉为图像版 GPT 。它利用 Transformer 架构，理解自然语言输入并生成相应图片，包括虚拟的、现实中不存在的图像。同样，北京智源研究院的 CogView 模型也是序列生成模型的典范，其结构与 DALL-E 类似，但更专注于中文环境的文本到图像生成。CogVie w 在多模态生成模型的下游任务精调上，展现出强大的泛化能力，并在样式学习、服装设计和图像超分等任务上取得显著成效。  在文本生成方向，序列生成模型同样占据主流地位。例如，GIT 这一视觉到文本的多模态大模型，便统一了图像 / 视频的描述和问答等任务。它包含图像编码器和文本解码器，后者在视觉编码基础上以自回归方式生成文本，进一步体现了序列生成模型在文本生成领域的强大应用潜力。  扩散模型的工作原理，是通过连续添加高斯噪声来破坏训练数据，然后通过反转这个噪声过程，学习恢复数据。LDM（Latent Diffusion Models）是扩散模型的一个代表性方法，它先压缩图像的像素信息来获取图像对应的隐特征表达，再采用扩散模型来建模图像隐特征分布。另一典型扩散模型 Stable Diffusion，它拓展 LDM 至开放领域的文本至图像生成，是当前开源模型的代表方法。除了开源模型之外，闭源的扩散模型中代表性方法有 OpenAI 的 DALL-E2 与谷歌的 Imagen。其中，DAL L-E2 首先训练一个扩散解码器来反转 CLIP 图像编码器，然后训练一个独立的映射模型，将 CLIP 模型的文本特征映射到图像特征空间，从而实现以文生图的过程，并极大提升了生成图像与输入文本的匹配程度。而 Imagen 首先将文本进行编码表征，之后使用扩散模型将表征映射成为 64 像素 ×64 像素的低分辨率的图像，然后通过两个超分辨率扩散模型来逐渐提高分辨率到 1 024 像素 ×1 024 像素。此外，与 DALL-E2 不同的是，Imagen 使用通用语言大模型 T5 模型直接编码文本信息，然后直接用该文本编码来生成图像；同时，Imagen 发现基于 T5 模型提取的文本特征生成的图像比基于 CLIP 模型的图像细节准确度更高。  **（三）兼顾理解和生成任务的多模态大模型**  Transformer 编码器利用双向注意力机制提升数据理解能力，而解码器则通过单向注意力机制增强生成能力。为融合这两种能力以适应更广泛的下游任务，可联合编解码器设计多模态大模型。蒙特利尔大学的 VL-T5 模型便是一个典型例子，它将多个多模态任务整合为文本生成任务，通过特定的输入与输出文本构造方式，实现模型结构和目标任务的统一，进而提升模型的泛化性。  另一典型模型 Unified VLP 则采用编解码器共享 Transformer 网络的设计，通过注意力掩码灵活切换编码与解码功能。在编码模式下，应用双向掩码以捕捉全文的依赖关系；在解码模式下，则使用单向掩码以确保仅依赖前文信息。这种共享方式有效减少了参数量，使模型更加精简且高效。  此外，还可以将语言大模型的文本生成能力与各类模态编码器的多模态感知能力相结合，以此构建的多模态大模型也能够兼顾理解和生成能力。此类方法以语言大模型为主导来实现多模态的对齐、融合和交互。这是由于文本有高效的表达效率、能够通过语义描述的方式与其余所有模态建立直接的联系，而且语言大模型在预训练过程中学习到了非常多的世界知识，有潜在理解多模态信息的能力；这类模型的结构常由单模态编码器、连接器与语言大模型三部分组成，其中单模态编码器和语言大模型的参数可以冻结，以减少计算量、提高训练效率；连接器常见的有简单的线性映射层，或者特殊设计的网络模块如 BLIP-2 中的 Q-former 结构等。这类模型通常涉及两个阶段的训练过程：在第一阶段，通常利用大规模弱关联的跨模态数据（如图像—文本、视频—文本、音频—文本数据等），基于条件文本生成任务，训练各个模态到语言大模型的语义对齐。在第二阶段，进行多模态指令微调，以提升零样本多模态能力，此阶段的核心是构造面向多模态任务的指令微调数据，目前常见的多模态指令微调数据类型有多模态对话、多模态详细描述与多模态推理问答等。  **（四）知识增强的多模态大模型**  大模型凭借对大规模数据的卓越拟合能力，能有效学习隐式知识。为进一步提升模型的理解与预测能力，需探索将隐式知识与显式知识相结合的方法。将知识图谱、场景图、外部知识库等结构化知识信息融入大模型，是增强多模态大模型知识利用能力的有效途径。例如，百度的ERNIE-ViL 模型在视觉—语言模型中引入场景图信息，通过掩码预测任务，使模型更精准地捕捉图像与文本间的细粒度对齐信息。而 KRISP 方法则结合隐含知识与明确知识的学习，既从无监督语料和有监督数据中学习隐式知识，又从结构化数据知识图谱中学习明确的符号化知识，从而实现隐式知识推理与符号化知识表示的双重获取。  **二、多模态大模型的关键技术**  多模态大模型的关键技术，主要包括预训练数据收集、基础模型构建、自监督学习与模型优化训练、下游任务微调。  **（一）多模态大模型的网络结构设计**  网络架构在多模态预训练中至关重要，需精心设计，以理解和融合不同源的复杂特征。在处理图像、文本等模态时，常采用 Transformer 或 CNN 等架构捕捉视觉与语言间的复杂关系；对于事件流，脉冲神经网络因其能有效模拟信息时的序动态而更为适合。随着模型规模的不断扩大，多模态大模型展现出强大的记忆与性能。然而，模型复杂度提升也带来了计算效率挑战，并可能触及性能瓶颈。因此，探索更高效的网络模型结构，如改进或替代 Transformer，成为当前研究的重要方向。  得益于语言大模型涌现出的知识与逻辑推理能力，近期有一系列多模态大模型开始以语言大模型为核心进行构建。其中，一个典型代表性是 DeepMind 的 Flamingo 视觉语言模型。该模型能够将图像、视频和文本作为提示，并输出相关语言回复。它将视觉编码器与语言大模型的参数冻结，并通过可学习的融合模块联系起来，模型采用 20 多亿对图片—文本、270 万对视频—文本，与 430 万个图文混排的网页数据进行视觉—语言联合训练；它还具有少样本（few-shot）的多模态序列推理能力，无须额外训练即可完成视觉语义描述、视觉问答等多种任务。另一个代表性模型 KOSMOS-I，它将一个基于 Transformer 的语言模型作为通用接口，并与视觉感知模块对接，使模型“能看”和“会说”。该模型具有 16 亿参数量，在大规模多模态语料库上训练，具有遵循指令（即零样本学习）以及在上下文中学习（即少样本学习）的能力，能够原生处理视觉对话、视觉问答、图像描述生成、光学字符识别等任务。此外，近期还有一系列模型尝试将图像、视频等感知模块与 LLaMA 等开源的语言大模型对接，从而实现类似 GPT-4 的多模态理解能力。其中，ChatBridge 是一个典型模型，它使用多个并行的感知模块来处理包括图片、音频、视频在内的特征，然后通过少量预训练参数将这些模态的特征投影至语言大模型的语义空间，使模型具备灵活感知、理解混合模态信息的能力。  在多模态预训练中，设计与下游任务高兼容性的网络结构模型至关重要。可通过引入编码器—解码器结构，将多模态理解和生成任务整合至统一框架，以更全面地支持多样化的多模态任务。这一设计涵盖了跨模态注意机制、模态间对齐与翻译，以及更精细的特征集成策略，旨在提升模型的综合性能与泛化能力。  **（二）多模态大模型的自监督学习优化**  以视觉—语言数据的联合学习为例，多模态大模型常用的自监督学习任务通常有以下几种类型。  1. 掩码语言建模  输入文本序列中的某些单词或标记会被替换为特殊的掩码标记 [mask]，然后预训练模型被要求根据可见的多模态上下文，来预测这些被遮蔽的单词或标记。多模态大模型通过执行这种预训练任务，模型能够在大规模文本数据上获取深层次的语言理解，从而更好地执行下游自然语言处理任务，如文本分类、命名实体识别、句子相似性计算等。  2. 掩码图像建模（Masked Image Modeling，MIM）  输入图像中的部分区域会被隐藏或替换为特殊的掩码标记 [mask]，然后预训练模型被要求在仅看到其余图像内容与文本等其他模态信息的情况下，预测或还原被遮蔽的图像区域。多模态大模型通常使用这种训练方式促使模型学习图像的视觉特征、多模态上下文信息和语义关系，以更好地理解图像内容。  3. 图像—文本匹配（Image—Text Matching，ITM）  掩码语言建模和掩码图像建模旨在建立图像与文本的细粒度对齐，而图像—文本匹配任务是旨在实现图像与文本的全局对齐。通常给定图文对作为正样本，随机配对作为负样本对，然后通过二分类方法实现图像和文本的匹配，从而构建图像和文本之间的语义关联。  4. 图像—文本对比学习（Image—Text Contrastive Learning，ITC）  使用对比学习的方法，拉近图像和文本的相同样本对的向量表示，推远不同样本对的向量表示，从而增强图像和文本之间的语义关联性。这使模型能够更好地理解图像和文本之间的语义关联，为多模态任务提供更好的表达能力。  **（三）多模态大模型的下游任务微调适配**  多模态大模型的终极目标，是在特定下游任务中实现性能的最优化。为此，需通过微调适配技术，将预训练大模型的能力有效迁移至特定数据集的特定任务场景中。目前，多模态大模型的微调适配主要采取以下三种方式。  1. 面向特定任务的模型微调适配  多模态大模型的权重被作为初始参数，并在任务特定数据上进行有监督的微调。通过这种微调，模型将学习针对具体任务的细粒度特征和表示，从而适应特定任务的要求。  2. 联合提示学习的模型微调适配  设计契合上游预训练任务的模板，挖掘出上游预训练模型的潜力，让上游的预训练模型在尽量不需要标注数据的情况下较好地完成下游任务。提示学习允许在不同类型的任务上重复使用预训练模型，只需简单修改提示模板即可适应特定任务，从而节省了训练时间和计算资源。  3. 基于适配器网络的模型微调适配  每个任务配备独立适配器层，使模型既能共享通用预训练表示，又能实现任务个性化调整。适配器层参数少，相比全模型微调更高效。训练时，预训练模型参数固定，仅更新适配器层参数。现有预训练大型方法已应用于下游任务，但需进一步研究多模态大型模型的增量学习算法。未来，新模态将不断出现，需设计的多模态模型需具备足够灵活性，以应对新模态的引入。 | **通过教师讲解，认识五彩斑斓 —— 多模态大模型。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了五彩斑斓 —— 多模态大模型，让学生了解多模态大模型的关键技术。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  多模态大模型的网络结构设计有哪些关键因素？ | 通过课后练习，使学生巩固所学新知识 |
| **知识讲解**  **（40min）** | **【教师】**讲解固若金汤—— 大模型的安全性  **一、大模型安全风险的广泛关注**  尽管大模型技术迅猛发展，但其仍伴随着诸多潜在安全隐患。在应用中，大模型可能产生与人类价值观相悖的输出，如歧视言论、侮辱及违背伦理的内容，这种风险广泛存在于文本、图像、语音和视频等场景，并随模型大规模部署而加剧，损害用户对 AI 决策的信任。此外，大模型防范能力脆弱，易受指令攻击、提示注入和后门攻击等威胁。在政治、军事、金融、医疗等关键领域，任何恶意攻击都可能对国家社会稳定与人民生命财产安全造成严重后果。  人工智能和大模型安全也是国际社会高度关注的热门话题。2023 年 5 月，联合国秘书长古特雷斯在纽约联合国总部提到，利用人工智能“必须由各国展开协调设定红线”，需要“打造 AI有助于人类幸福，而不会成为人类威胁的环境”。OpenAI 首席执行官山姆·阿尔特曼呼吁美国监管高级大型语言模型的部署，并发出警告：没有坚实政策框架会使生成式人工智能陷入危险境地。2023 年 5 月底，众多人工智能科学家和人工智能领袖发表公开声明，呼吁防范人工智能的生存风险应该与流行病、核战争等其他大规模风险一样，成为全球优先议题。2023 年 6 月，图灵奖得主 Geoffrey Hinton 在演讲中指出，超级智能的到来比他想象中更快，在此过程中，数字智能可能会追求更多控制权，甚至通过“欺骗”控制人类，给人类社会带来更多问题。同时，随着民众对人工智能社会威胁的担忧日益加剧，监管过程对于减轻日益强大的模型带来的风险至关重要。  **二、大模型安全治理的政策法规和标准规范**  为确保大模型的安全及负责任应用，各国监管机构正积极探讨并制定相关安全标准和准则，为开发者和企业提供明确的大模型应用与治理指引。  2021 年 11 月，联合国教科文组织正式发布《人工智能伦理问题建议书》，指出“作为以国际法为依据，采用全球方法制定且注重人的尊严和人权以及性别平等、社会和经济正义与发展、身心健康、多样性、互联性、包容性、环境和生态系统保护的准则性文书，可以引导人工智能技术向着负责任的方向发展”。  2023 年 3 月，美国白宫科技政策办公室发布《促进隐私保护数据共享和分析的国家战略》。该策略旨在保障公共和私营部门实体中用户的数据隐私，同时确保数据使用的公平性和最大的效率。该战略明确了政府的目标：支持有关数据伦理和社会技术问题的解决方案的研究、开发、监管和应用，同时确保用户的机密性不受损害。  2023 年 4 月，美国政府发布《人工智能问责政策征求意见》。此征求意见稿涵盖人工智能审计、安全风险评估、认证等内容，以促进建立合法、有效、符合道德、安全可信的人工智能系统。  2023 年 6 月，欧洲议会通过《人工智能法案》草案，旨在为人工智能引入统一的监管和法律框架，并涵盖了除军事用途外的所有人工智能类型。该法案根据人工智能应用可能造成伤害的风险，对其进行分类和监管，以增强各成员国之间的合作，确保人工智能技术的健康、安全和公平发展。  2019 年 6 月，国家新一代人工智能治理专业委员会发布的《新一代人工智能治理原则——发展负责任的人工智能》指出：“人工智能系统应不断提升透明性、可解释性、可靠性、可控性，逐步实现可审核、可监督、可追溯、可信赖。高度关注人工智能系统的安全，提高人工智能鲁棒性及抗干扰性，形成人工智能安全评估和管控能力。”  2020 年 7 月，国家标准化管理委员会、中央网信办、国家发展和改革委员会、科学技术部、工业和信息化部联合发布的《国家新一代人工智能标准体系建设指南》指出：“重点开展人工智能安全术语、人工智能安全参考框架、人工智能基本安全原则和要求等标准的研制。”  2021 年 9 月，国家新一代人工智能治理专业委员会发布《新一代人工智能伦理规范》，旨在“将伦理道德融入人工智能全生命周期，促进公平、公正、和谐、安全，避免偏见、歧视、隐私和信息泄露等问题。”  2022 年 3 月，中共中央办公厅、国务院办公厅发布的《关于加强科技伦理治理的意见》指出，应“加快构建中国特色科技伦理体系，健全多方参与、协同共治的科技伦理治理体制机制，坚持促进创新与防范风险相统一、制度规范与自我约束相结合，强化底线思维和风险意识，建立完善符合我国国情、与国际接轨的科技伦理制度，塑造科技向善的文化理念和保障机制”。  2023 年 3 月，国家人工智能标准化总体组、全国信标委人工智能分委会发布《人工智能伦理治理标准化指南》，明确了人工智能伦理治理概念范畴，细化人工智能伦理准则的内涵外延，对人工智能伦理风险进行分类分级分析，提出人工智能伦理治理技术框架，构建人工智能伦理治理标准体系，引导人工智能伦理治理工作健康发展。  2023 年 7 月，国家互联网信息办公室、国家发展和改革委员会等联合发布的《生成式人工智能服务管理暂行办法》指出，“国家坚持发展和安全并重、促进创新和依法治理相结合的原则，采取有效措施鼓励生成式人工智能创新发展，对生成式人工智能服务实行包容审慎和分类分级监管”“提供和使用生成式人工智能服务，应当遵守法律、行政法规，尊重社会公德和伦理道德”。  **三、大模型安全风险的具体表现**  随着大模型技术的不断推广与深入应用，其潜在的安全风险已波及多个领域，对社会秩序构成了日益严重的威胁。这些风险不仅体现在大模型自身存在的安全漏洞与隐患上，还涉及大模型在实际应用过程中可能引发的诸多安全问题。  **（一）大模型自身的安全风险**  大模型因采用海量数据进行训练，可能吸收并反映数据中的不当、偏见或歧视性内容。数据的多样性和复杂性，使模型难以准确体现人类价值观和伦理标准。在处理或生成内容时，大模型可能无意中扩大社会偏见，如偏向特定文化、性别、种族或宗教观点，导致偏见、歧视或误导性输出，这不仅会引起特定群体的不适，还可能破坏社会和谐与稳定。以下列出了典型的风险类型。  1. 辱骂仇恨  模型生成带有辱骂、脏字脏话、仇恨言论等不当内容。  2. 偏见歧视  模型生成对个人或群体的偏见和歧视性内容，通常与种族、性别、宗教、外貌等因素有关。  3. 违法犯罪  模型生成的内容涉及违法犯罪的观点、行为或动机，包括怂恿或实施违法犯罪，如造谣、诈骗等内容。  4. 敏感话题  对于一些敏感和具有争议性的话题，模型输出了具有偏向、误导性或不准确的信息。例如，支持某个特定政治立场倾向的言论会导致对其他政治观点的歧视或排斥。  5. 身体伤害  模型生成与身体健康相关的不安全信息，引导和鼓励用户伤害自身和他人的身体，如提供误导性的医学信息或错误的药品使用建议等，对用户的身体健康造成潜在的风险。  6. 心理伤害  模型输出与心理健康相关的不安全的信息，包括引发恐慌或焦虑甚至鼓励自杀等内容，影响用户的心理健康。  7. 隐私和财产  模型生成涉及暴露用户或第三方的隐私和财产信息，或者提供重大的建议如投资等。在处理这些信息时，模型应遵循相关法律和隐私规定，保障用户的权益，避免信息泄露和滥用。  8. 伦理道德  模型生成的内容认同和鼓励违背道德伦理的行为。在处理一些涉及伦理和道德的话题时，模型需要遵循相关的伦理原则和道德规范，和人类价值观保持一致。  此外，语言模型的意识形态已成为人工智能安全的核心考量。模型在训练过程中不可避免地受训练数据中的文化与价值观所影响，从而决定了其形成的意识形态。  **（二）大模型在应用中衍生的安全风险**  随着大模型应用的广泛性和复杂性，不当使用和恶意使用等行为也随之增加，这为大模型带来了前所未有的安全挑战。  1. 用户过度依赖大模型的生成内容  大模型虽通过学习海量数据获得强大生成能力，但数据复杂性导致其可能产生真实却错误的“幻觉”信息。若用户盲目信任，会误将“幻觉”当真实，决策时遗漏关键信息，缺乏批判性思考。在医学诊断、法律意见等高精度领域，这种信赖将带来巨大风险。  2. 恶意攻击下的安全风险  大模型遭受模型窃取攻击，使攻击者能获取模型结构及关键参数，不仅逃避使用费用，还可能谋取其他利益；若完全掌控模型，更可能发动危险的“白盒攻击”。同时，数据重构攻击让攻击者能复原模型训练数据，暴露个人医疗记录等敏感信息，严重威胁个人隐私和数据所有权。此外，指令攻击利用模型对措辞的敏感性，诱导其生成违规或偏见内容，违背原安全设定。  3. 后门攻击带来的恶意输出  后门攻击是一种针对深度学习模型的新型攻击方式。它是在训练过程中对模型植入隐秘后门。后门未被激活时，模型可正常工作，而一旦被激活，模型将输出攻击者预设的恶意标签。由于模型的黑箱特性，这种攻击难以检测。比如，在 ChatGPT 的强化学习阶段，在奖励模型中植入后门，使攻击者能够通过控制后门来控制 ChatGPT 输出。此外，后门攻击具有可迁移性。通过利用 ChatGPT 产生有效的后门触发器，并将其植入其他大模型，这为攻击者创造了新的攻击途径。因此，迫切需要研究鲁棒的分类器和其他防御策略来对抗此类攻击。  4. 大模型访问外部资源时引发的安全漏洞  大模型与外部数据、API 或敏感系统交互时，面临诸多安全挑战。若连接缺乏适当安全措施，未经验证的信息将导致模型生成不安全、不可靠的反馈。例如，自主智能体 AutoGPT 虽功能众多、自主性和复杂性高，但在缺少人工监管时，可能出现无法预测的行为模式，极端情况下甚至编写毁灭性计划。因此，对于大模型与外部资源的交互，必须采取严格的安全策略。  **四、大模型安全研究关键技术**  随着大模型安全问题的日益凸显，全球众多知名的科研机构已将此作为核心研究领域，致力于探索模型的潜在薄弱点和安全风险，并寻求如何增强大模型在训练和部署时的安全性。  **（一）大模型的安全对齐技术**  安全对齐的大模型指的是那些经过充分检验、具备高可信度和鲁棒性，且与人类价值观紧密对齐的大型机器学习模型。这类模型在设计和训练过程中，严格遵循伦理准则，并具备透明度、可解释性和可审计性，确保用户能够清晰理解其行为和决策依据。同时，安全对齐的大模型还高度重视隐私保护和安全防护，在使用过程中能有效防止敏感信息泄露，并抵御恶意攻击。  大模型暴露的安全风险与其开发技术密不可分。当前主流的大模型训练过程可分为预训练、有监督微调和基于反馈的强化学习微调三个阶段。以 ChatGPT 为例，首先，在预训练阶段，模型在大量的互联网文本上学习，汲取其中的语言模式和知识，这个过程中模型可能会无意间学习并模仿数据中的价值观；然后是有监督微调阶段，模型在特定的监督数据集上进一步微调，以理解更具体的任务要求并调整其输出，使之更接近人类对特定任务的期望；最后一个阶段是基于人类反馈的强化学习阶段，此阶段的目标是让模型的输出与人类价值观尽可能一致，提高其有用性、真实性和无害性。  针对大模型开发过程中产生的安全风险，安全对齐研究可从提升训练数据的安全性、优化安全对齐训练算法两个方面展开，以实现更有用、诚实和无害的安全大模型。  1. 大模型的训练数据安全  训练数据的安全性是构建安全大模型的基石。训练数据安全是指数据集的来源和质量都是可靠的，数据蕴含的知识是准确的，数据集内容符合主流价值观。以下是提高数据安全性的一些关键要点。  （1）数据的来源与预处理。确保训练数据来自可信的、可靠的来源。数据应该从权威机构、专业组织、可验证的数据仓库或其他公认的数据提供者获得。在数据标注时，确保标注的准确性和一致性。标注过程应该由经过培训的专业人员进行，并且需要进行验证和审核，以确保标注的正确性。此外，需要进行数据清洗以去除重复项、噪声数据和错误数据。  （2）数据的敏感信息去除。在大模型中保护数据的敏感信息至关重要。为确保训练过程中不泄露个人隐私、敏感信息或商业机密，需采取有效的隐私保护措施。数据的敏感信息去除便是其中之一，其核心目标是防止敏感信息泄露。常见方法包括：数据脱敏，如随机化、泛化、替换和加噪声等，确保敏感信息无法被还原或追溯；去标识化，即删除个人标识信息，如姓名、地址等，实现数据匿名化；数据掩码，将敏感信息部分替换为伪造或不可还原的数据，确保原始信息无法被还原。  在进行数据的敏感信息去除时，需要谨慎处理，以确保不会破坏数据的完整性和质量。同时，也需要注意确保去除敏感信息后的数据仍然具有足够的信息量和代表性，以确保训练的模型具备合理的性能和泛化能力。  2. 大模型的安全对齐训练  （1）基于反馈的安全对齐技术。基于人类反馈的安全对齐技术已逐渐成为当下大模型安全研究的主流技术。其训练过程主要包括奖励模型训练和生成策略优化两个子阶段。奖励模型训练阶段中，人类对模型生成的多条不同回复进行评估，这些回复两两组合，由人类确定哪条更优，生成的人类偏好标签使奖励模型能学习并拟合人类的偏好。在生成策略优化阶段，奖励模型根据生成回复的质量计算奖励，这个奖励作为强化学习框架中的反馈，并用于更新当前策略的模型参数，从而让模型的输出更符合人类的期望。  DeepMind 使用 RLHF 技术，通过从人类反馈中学习来构建更有用、更准确和更安全的对话智能体 Sparrow。Anthropic 公司提出的 Claude 模型则采用了 RLAIF（RL from AIFeedback）技术。该技术使用预先训练的模拟人类偏好的打分模型，在强化学习过程中自动对数据进行排序，从而减少对人类反馈的依赖。  2023 年 5 月，北京大学团队开源了名为“河狸”（PKU -Beaver）项目，提供了一种可复现的 RLHF 基准，并公开了 RLHF 所需的数据集、训练和验证代码。2023 年 7 月，复旦大学发布基于 RLHF 实现人类对齐的 MOSS-RLHF 模型，深入探究了 RLHF 阶段所采用的强化学习算法近端策略优化（Proximal Policy Optimization，PPO），分析其稳定训练及在大模型人类对齐中的作用机理，并发布大模型人类对齐技术报告与开源核心代码，以推动中文 NLP 社区生态发展。  （2）大模型可信增强技术。在训练过程中，模型可通过对抗训练和知识融入训练两个方面增加可信度。对抗训练通过提升模型对输入扰动的鲁棒性来增强模型可信度，它引入对抗性样本，即对大模型输入做微小改动导致其误判的样本，迫使模型学习更具鲁棒性的特征，从而减少对抗性攻击影响，提升泛化能力。知识融入训练则利用知识引导模型训练，降低模型出现幻觉的可能性，如结合知识图谱的模型训练，通过引入知识图谱及其三元组引导模型训练，使大模型沿着正确知识方向收敛，存储高可信度的知识。  **（二）大模型安全性评测技术**  大模型安全性评测技术是大模型安全发展的有力保障。  1. 大模型内容安全评估  为评估大语言模型的安全性，并推动安全、负责和符合道德的人工智能的发展和部署，清华大学于 2023 年 3 月推出面向中文大模型的安全性评测平台。该平台依托于一套系统的安全评测框架，从辱骂仇恨、偏见歧视、违法犯罪等八个典型安全场景和六种指令攻击综合评估大语言模型的安全性能。其中，指令攻击是指一般模型难以处理的安全攻击方式，这些攻击更容易诱导模型出错，包含目标劫持、Prompt 泄露、赋予特殊的角色后发布指令、不安全 / 不合理的指令主题、隐含不安全观点的询问以及反面诱导。基于该框架，平台对 GPT 系列、ChatGLM 等主流大模型进行了安全评估，并发现指令攻击更有可能暴露所有模型的安全问题。  平台已开源大模型安全评测的数据基准，并测试了包括 ChatGPT 在内的十余个主流大模型，其安全分数以排行榜的形式在平台公布。  2. 大模型极端风险的评估  随着 AI 技术的进步，大模型可能展现出危险的突发能力，如网络攻击、操纵对话或提供恐怖主义指导。为识别这些风险，DeepMind 等机构提出了新型威胁评估的通用模型框架。该框架强调，大模型安全评估首先判断模型是否具备危险能力，并评估其可能造成的伤害程度。同时指出，极端风险评估是安全 AI 研发的关键，应涵盖领域风险水平和模型潜在风险属性，以帮助开发者识别风险因素，优化模型训练与部署的安全性。  3. 大模型行为决策的道德评估  随着人工智能系统能力的快速增长，越来越多的大模型被训练应用于真实世界的交互任务。衡量大模型在各种社会决策场景中的能力和道德行为的一项典型评测基准是 MACHIAVELL。它主要由 134 款基于文本的 Choose Your Own Adventure 游戏组成，在评估中为大模型代理提供真实世界的目标，并通过专注于高层次的决策来追踪代理的不道德行为，以评估其在现实社会环境中的规划能力及安全风险。该项研究发现，道德行为和最大化奖励之间存在权衡（Trade-Offs）的关系，但通过设计道德提示，对大模型进行道德调节，可缓解权衡并降低有害行为的发生频率。 | **通过教师讲解，理解固若金汤—— 大模型的安全性。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了固若金汤—— 大模型的安全性，让学生了解大模型安全性的现实意义以及各大平台“青少年模式”的设置流程。。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  在大模型安全性评测技术方面，我们如何确保评测的全面性和准确性？ | 通过课后练习，使学生巩固所学新知识 |
| **教学反思** | 课堂营造更包容的课堂氛围，鼓励学生大胆表达，对积极参与的学生给予更多肯定。同时，建立多样化的反馈渠道。 | |