中国人工智能系列白皮书

——大模型技术（2023 版）

中国人工智能学会

二〇二三年九月

《中国人工智能系列白皮书》编委会

主　　任：戴琼海

执行主任：王国胤

副 主 任：陈　杰　何　友 刘成林　刘　宏　孙富春　王恩东
王文博　赵春江　周志华

委　　员：班晓娟　曹　鹏　陈　纯　陈松灿　邓伟文　董振江
杜军平　付宜利　古天龙　桂卫华　何　清　胡国平
黄河燕　季向阳　贾英民　焦李成　李　斌　刘　民
刘庆峰　刘增良　鲁华祥　马华东　苗夺谦　潘　纲
朴松昊　钱　锋　乔俊飞　孙长银　孙茂松　陶建华
王卫宁　王熙照　王　轩　王蕴红　吾守尔·斯拉木
吴晓蓓　杨放春　于　剑　岳　东　张小川　张学工
张　毅　章　毅　周国栋　周鸿祎　周建设　周　杰
祝烈煌　庄越挺

《中国人工智能系列白皮书----大模型技术》编写组

陶建华　吴　飞　黄民烈　文继荣　王海峰　刘知远
刘　静　杨小康　聂　帅

目录

第 1 章 大模型技术概述

1.1 大模型技术的发展历程

2006 年 Geoffrey Hinton 提出通过逐层无监督预训练的方式来缓解由于梯度消失而导致的深层网络难以训练的问题[1]，为神经网络的有效学习提供了重要的优化途径。此后，深度学习在计算机视觉[2]、语音[3]、自然语言处理[4]等众多领域取得了突破性的研究进展，开启了新一轮深度学习的发展浪潮。总结过去十多年的技术发展，基于深度学习的人工智能技术主要经历了如下的研究范式转变：从早期的“标注数据监督学习”的任务特定模型，到“无标注数据预训练+标注数据微调”的预训练模型，再到如今的“大规模无标注数据预训练+指令微调+人类对齐”的大模型，经历了从小数据到大数据，从小模型到大模型，从专用到通用的发展历程，人工智能技术正逐步进入大模型时代。

2022 年底，由 OpenAI 发布的语言大模型 ChatGPT 引发了社会的广泛关注。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种任务，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、社会等众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，引发了大模型的发展热潮。

本次大模型热潮主要由语言大模型（亦称为大语言模型）引领。语言大模型通过在海量无标注数据上进行大规模预训练，能够学习到大量的语言知识与世界知识，并且通过指令微调、人类对齐等关键技术拥有面向多任务的通用求解能力。在原理上，语言大模型旨在构建面向文本序列的概率生成模型，其发展过程主要经历了四个主要阶段[5]：

1）统计语言模型：统计语言模型主要基于马尔可夫假设建模文本序列的生成概率。特别地，N-gram 语言模型[6]认为下一个词汇的生成概率只依赖于前面出现的 N 个词汇（即 N 阶马尔可夫假设）。此类语言模型的问题在于容易受到数据稀疏问题的影响，需要使用平滑策略改进概率分布的估计，对于文本序列的建模能力较弱。

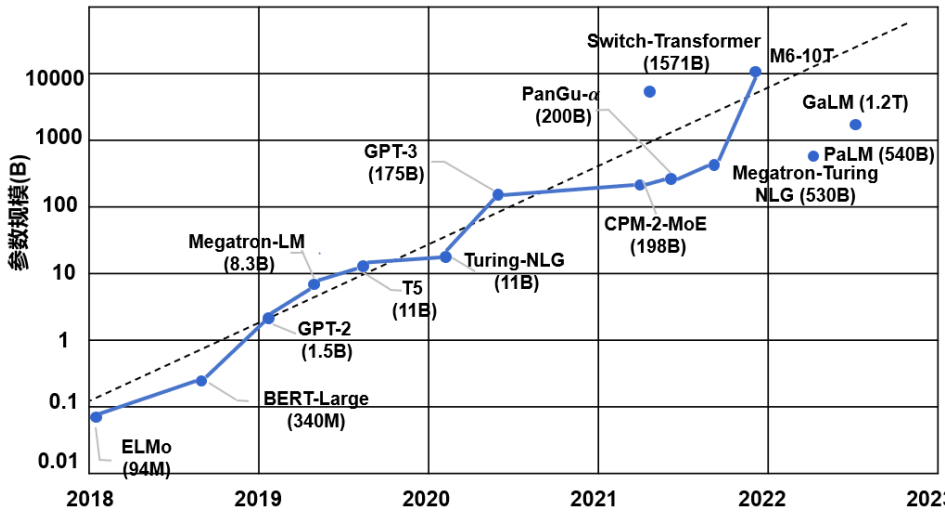
2）神经语言模型：针对统计语言模型存在的问题，神经语言模型主要通过神经网络（MLP[7]、RNN[8]）建模目标词汇与上下文词汇的语义共现关系，能够有效捕获复杂的语义依赖关系，更为精准建模词汇的生成概率。进一步，word2vec[4]简化了神经语言模型的网络架构，可以从无监督语料中学习可迁移的词表示（又称为词向量或词嵌入），为后续预训练语言模型的研究奠定了基础。

3）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[9]、GPT-1[10]和 BERT[11]等。其中，ELMo 模型基于传统的循环神经网络（LSTM）[12]构建，存在长距离序列建模能力弱的问题；随着 Transformer[13]的提出，神经网络序列建模能力得到了显著的提升，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。

4）语言大模型（探索阶段）：在预训练语言模型的研发过程中，一个重要的经验性法则是扩展定律（Scaling Law）[14]：随着模型参数规模和预训练数据规模的不断增加，模型能力与任务效果将会随之改善。图 1-1 展示了 2018 至 2023 年间典型预训练模型的参数量变化趋势。OpenAI 在研发 GPT 系列模型过程中，主要探索了 GPT-1[10]（1.1 亿参数）、GPT-2（15 亿参数）[15]、以及 GPT-3（1750 亿参数）[16]三个不同参数规模的模型，谷歌也推出了参数规模高达 5400 亿参数的 PaLM 模型[17]。当模型参数规模达到千亿量级，语言大模型

能够展现出多方面的能力跃升[18]。例如，GPT-3 在没有微调的情况下，可以仅通过提示词或少数样例（In-context learning，上下文学习[19]）完成多种任务，甚至在某些任务上超过当时最好的专用模型。学术界引入了“语言大模型”（Large language models）[5]来特指这种超大规模的预训练语言模型，以突出与早期预训练语言模型的不同。

图 1-1 2018-2023 年模型参数规模变化图



5）语言大模型（提升阶段）：虽然早期的语言大模型表现出一定的少样本学习能力，但是其学习目标主要通过预测下一个单词实现，仍不能很好地遵循人类指令，甚至会输出无用的、有害的信息，难以有效对齐人类的偏好。针对这些问题，主要有两种大模型改进技术，包括指令微调（Instruction Tuning）[20]以及基于人类反馈的强化学习(Reinforcement Learning from Human Feedback, RLHF) [21]。指令微调利用格式化（指令和回答配对）的训练数据加强大模型的通用任务泛化能力；基于人类反馈的强化学习（如图 1-2 所示）将人类标注者引入到大模型的学习过程中，训练与人类偏好对齐的奖励模型，进而有效指导语言大模型的训练，使得模型能够更好地遵循用户意图，生成符合用户偏好的内容。在大模型使用过程中，可以使用各种提示技术（包括思维链（Chain-of-Thoughts, CoT）[22]、思维树（Tree-of-Thoughts, ToT）[23]等），从而更好地利用大模型的潜在能

力，提升大模型解决实际问题的能力。进一步，语言大模型主要是基于文本数据形式进行训练与推理，存在一些特定能力的不足，例如数值计算等。针对这一问题，可以使用外部工具（如计算器、搜索引擎等）扩展大模型的能力边界[24]。

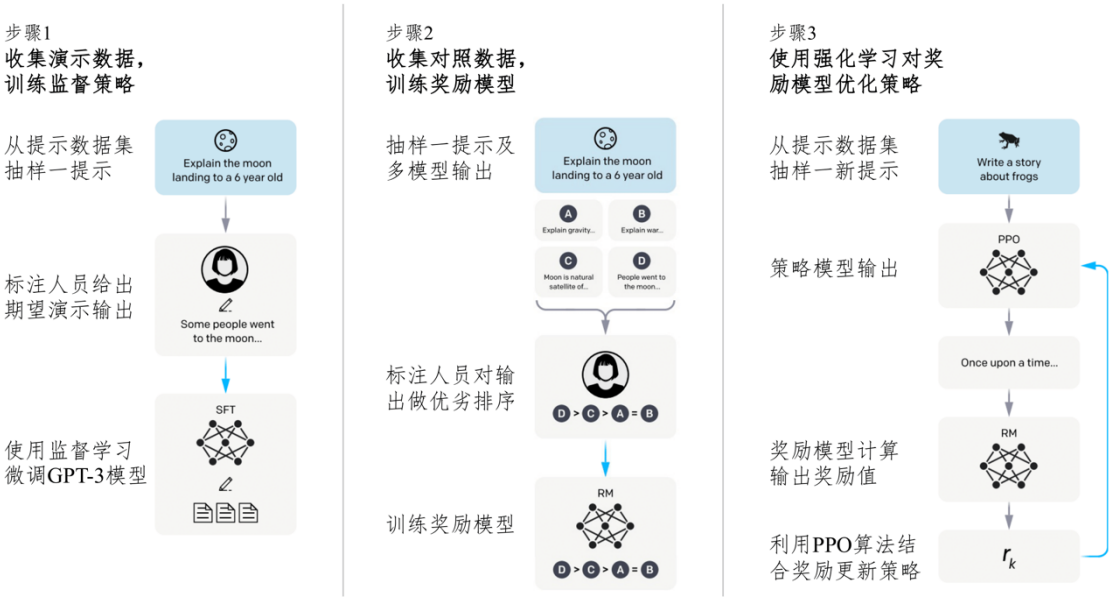


图 1-2 基于人类反馈强化学习的算法示意图

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，形成了一系列的技术进展。其中，GPT-1 探索了解码器 Transformer 架构（decoder-only 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），并且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[25]使用代码数据对 GPT-3 进行微调，从而提升代码能力和复杂推理能力；InstructGPT[21]基于人类反馈的强化学习技术（RLHF），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 InstructGPT 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[26]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予披露。

随着 GPT-4 的成功，语言大模型对于多模态领域也产生了重要影响，它从单调的文本交互，升级为可以接受文本与图像组合的多模态输入，相比传统的单模态大模型，多模态大模型更加符合人类的多渠道感认知方式，能够应对更加复杂丰富的环境、场景和任务。GPT-4 表明在多模态大模型中引入基于人类知识的自然语言能够带来模型在多模态理解、生成、交互能力上的。

1.2 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的大模型服务平台,用户可以通过 API 访问不同的 GPT 模型来完成下游任务。Claude 系列模型是由 Anthropic 开发的闭源语言大模型,目前包含 Claude 和 Claude-Instant 两种模型可供选择。该系列模型通过无监督预训练、基于人类反馈的强化学习和 Constitutional AI 技术（包含监督训练和强化学习）进行训练，旨在改进模型的有用性、诚实性和无害性。Claude 最高支持 100K 词元的上下文，而 Claude-2 更是拓展到了 200K 词元的上下文。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 APP、网页版、API 接口等多种形式的开放服务。文心一言还建设了插件机制，通过外部工具、服务的调用，拓展大模型的能力的边界。讯飞星火认知大模型具有开放式知识问答、多轮对话、逻辑和数学能力，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。

大模型的开源生态也“百花齐放”，主要包括开源框架与开源大模型。开源框架可以有效地支撑大规模模型的训练，如：PyTorch[27]

提供了分桶梯度、通信计算重叠、跳过同步等技术,支持大规模的分布式数据并行训练；飞桨[28]是国产的深度学习框架,早在内部就支持了大规模分布式训练，覆盖了计算机视觉、自然语言处理等多个领域的模型，其中 4D 混合并行策略,可训练千亿规模模型；OneFlow 将分布式集群抽象成逻辑上的超级设备,支持动静态图灵活转换,以数据+模型混合并行提升性能；DeepSpeed[29]是微软推出的大模型训练框架,其中 ZeRO 技术减少冗余内存访问,使得可以训练万亿级模型。开源大模型可降低大模型研究的门槛，促进大模型应用的繁荣。其中典型代表有:LLaMA[30]系列是 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[31]系列来自阿布扎比的 TII 研究院,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[32]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 系列模型由百川智能开发，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，该系列模型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 [33][34]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。

大模型技术具有广泛的应用场景，可以用来赋能不同行业。大模型+传媒可以实现智能新闻写作，降低新闻的生产成本；大模型+影视可以拓宽创作素材，开拓创作思路，激发创作灵感，提升作品质量；大模型+营销可以打造虚拟客服，助力产品营销；大模型+娱乐可以加强人机互动，激发用户参与热情，增加互动的趣味性和娱乐性；大模型+军事可以增强军事情报和决策能力，可以实现实时战场翻译，快速准确的威胁评估、作战任务规划和执行、战场感知、战术决策支持、改进态势感知等；大模型+教育可以赋予教育教材新活力，让教育方式更个性化、更智能；大模型+金融可以帮助金融机构降本增效，让

金融服务更有温度；大模型+医疗可以赋能医疗机构诊疗全过程。总之，大模型的发展将给人类带来了非常强大的助推力，让数字世界和现实世界的共生变得更为便捷、更为有效。

大模型的通用性使其被认为是可以成为未来人工智能应用中的关键基础设施，就像 PC 时代的操作系统一样，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，形成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，对经济、社会和安全等领域的智能化升级中形成关键支撑。

1.3 大模型技术的风险与挑战

尽管以 ChatGPT 为代表的大模型技术取得关键性突破，但当前大模型技术仍存在诸多风险与挑战。

首先，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、时效性方面等仍存在较多问题，尚无法对所合成内容做出可靠评估[35][36]。

其次，大模型的可解释性存在不足。大模型基于深度神经网络，为黑盒模型，其工作机理仍难以理解。语言大模型的涌现能力[18]、规模定律[14]，多模态大模型的知识表示、逻辑推理能力、泛化能力、情景学习能力[19][37]等方面有待展开深入研究，为大模型的大规模实际应用提供理论保障。

再次，大模型应用部署代价高。大模型参数规模和数据规模都非常巨大，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。提高推理速度降低大模型使用成本是大规模应用的关键。

此外，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挑战。提升大模型对小数据的高效适配迁移能力是未来研究的重点。

最后，大模型还存在伴生技术风险问题。例如，语言大模型具有通用的自然语言理解和生成能力，其与语音合成、图像视频生成等技术结合可以产生人类难以辨别的音视频等逼真多媒体内容，可能会被滥用于制造虚假信息、恶意引导行为，诱发舆论攻击、甚至危害国家安全[38][39]。此外，大模型存在安全与隐私问题，目前针对大模型安全漏洞的典型攻击方式包括：数据投毒攻击、对抗样本攻击、模型窃取攻击、后门攻击、指令攻击。大模型的安全漏洞可能被攻击者利用，使得大模型关联业务面临整体失效的风险，威胁以其为基础构建的应用生态。大模型利用海量的互联网数据进行训练，包括个人、企业甚至国家的敏感数据可能被编码进大模型参数中，因而存在数据隐私问题。例如，通过提示信息可能诱发大模型隐私数据泄露问题。

第 2 章 语言大模型技术

近年来，在 Transformer 架构基础上构建的预训练语言模型为自然语言处理领域带来了一系列突破式进展，成为人工智能主流技术范式。预训练语言模型采用“预训练+微调”方法，主要分为两步：1）将模型在大规模无标注数据上进行自监督训练得到预训练模型，2）将模型在下游各种自然语言处理任务上的小规模有标注数据进行微调得到适配模型。由于预训练语言模型参数越大模型表现越好，这激发了语言大模型（Large Language Model, LLM）研究热潮。

2.1 Transformer 架构

Transformer 架构[13]是目前语言大模型采用的主流架构[5]，其基于自注意力机制(Self-attention Mechanism)模型。其主要思想是通过自注意力机制获取输入序列的全局信息，并将这些信息通过网络层进行传递。标准的 Transformer 如图 2-1 所示，是一个编码器-解码器架构，其编码器和解码器均由一个编码层和若干相同的 Transformer 模块层堆叠组成，编码器的 Transformer 模块层包括多头注意力层和全连接前馈网络层，这两部分通过残差连接和层归一化操作连接起来。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，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 架构的优势在于它的并行计算能力，即不需要按照时间步顺序地进行计算。

Transformer 架构包含编码层与 Transformer 模块两个核心组件，

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，由二者加和得到：

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。每个词都被转化为一个固定长度的向量，然

后被送入模型进行处理。

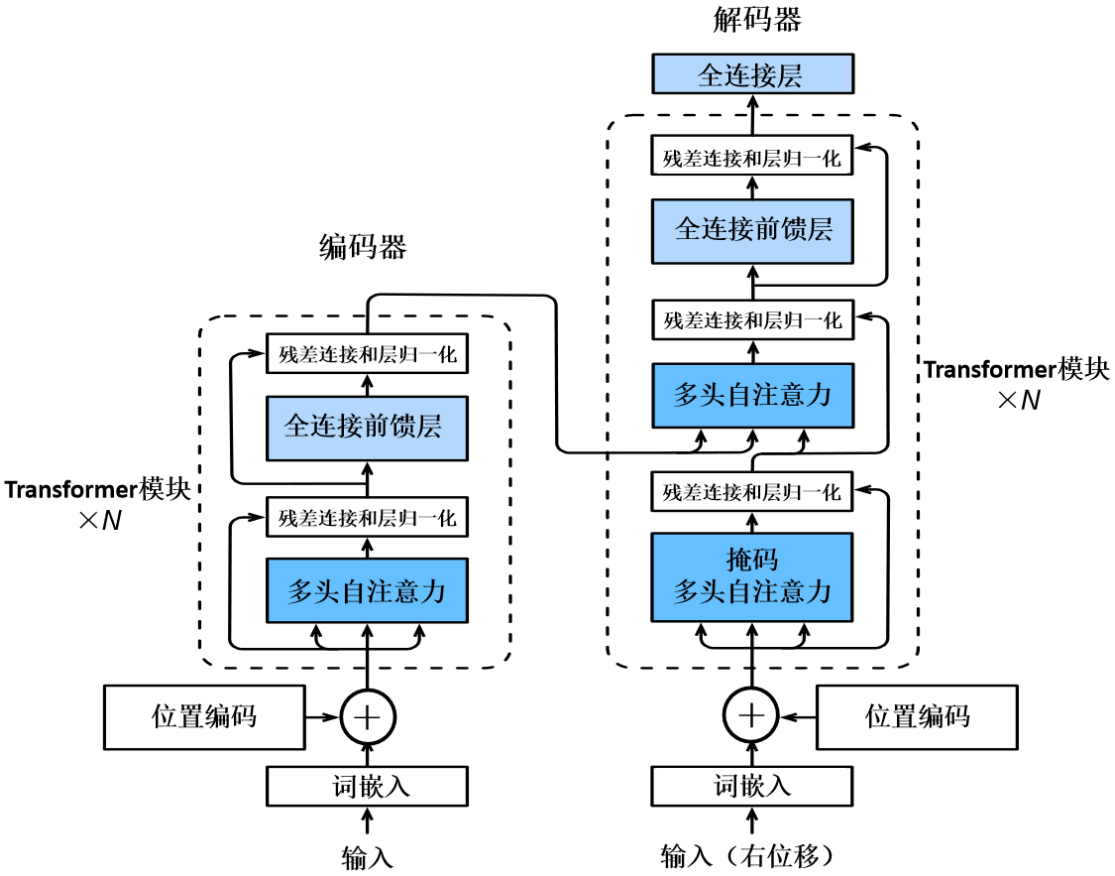


图 2-1 Transformer 架构[13]

2）位置编码，由于自注意力机制本身对位置信息不敏感，为了让模型能够理解序列中的顺序信息，引入了位置编码。标准 Transformer 架构的位置编码方式是使用正弦和余弦函数的方法。对于每个位置 i，对应的位置编码是一个长度为 d 的向量，其中 d 是模型的嵌入维度。这个向量的第 j 个元素由以下公式计算：如果 j 是偶数，那么编码的第 j 个元素为 $\sin(i/10000^{j/d})$；如果 j 是奇数，那么编码的第 j 个元素为 $\cos(i/10000^{j/d})$。

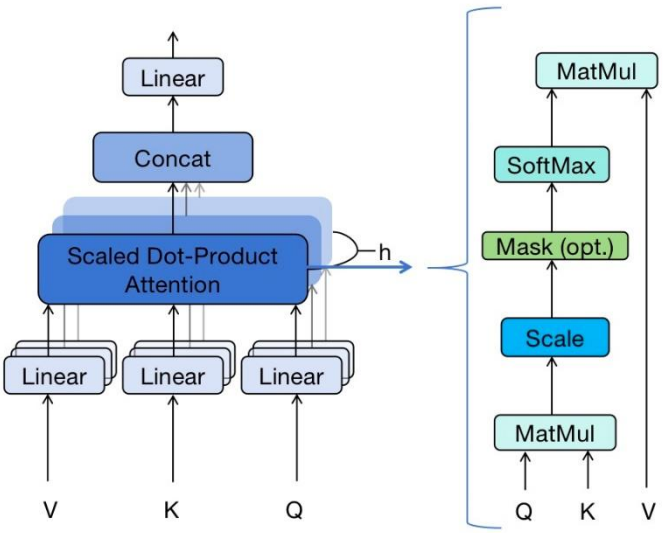


图 2-2 Transformer 自注意力网络[13]

Transformer 模块，通过自注意力机制获取输入序列的全局信息，并将这些信息通过网络层进行传递，包括多头注意力层和全连接前馈网络层，这两部分通过残差连接和层归一化操作连接起来，Transformer 模块，由自注意力层、全连接前馈层、残差连接和层归一化操作等基本单元组成：

1）自注意力层，注意力（Attention）是 Transformer 模型的核心组成部分。它包含一个查询矩阵 $Q \in \mathbb{R}^{n \times d_k}$，一个键矩阵 $K \in \mathbb{R}^{m \times d_k}$ 和一个值矩阵 $V \in \mathbb{R}^{m \times d_v}$，其中矩阵中的每一行对应一个词。注意力机制的计算方式：

$$H = \text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

直观来说，矩阵 $H \in \mathbb{R}^{n \times d_v}$ 中的每一行是 V 中行向量的加权和，其中权重由查询向量和键矩阵的点积决定。记具有序列长度 n 的查询序列的特征矩阵和具有序列长度为 m 的键-值序列的特征矩阵分别为 $X_q \in \mathbb{R}^{n \times d}$ 和 $X_{kv} \in \mathbb{R}^{m \times d}$，三个矩阵 Q、K、V 由三个线性变换得到 $Q = X_q W_q, K = X_{kv} W_k, V = X_{kv} W_v$. Transformer 模型采用的特定注意力机制被称为自注意力机制，因为三个矩阵 Q、K、V 都来自于前一层的相

同特征矩阵 $X \in \mathbb{R}^{n \times d}$。

此外，Transformer 采用了多头自注意力（Multi-head Attention）机制，即输入序列被线性映射多次得到不同的投影矩阵。多个尺度化后点积注意力可以并行计算，并产生多个自注意力输出。多头注意力生成多个高维的注意力表示，这使得其比单头注意力具有更强的表达能力。多头注意力的计算方式如下：使用了多个查询矩阵 $Q^{(i)}$，键矩阵 $K^{(i)}$ 和值矩阵 $V^{(i)}$，最终输出为 $H \in \mathbb{R}^{d_v \times d_o}$，它是通过将一系列 H_i 进行拼接，并使用一个新的权重矩阵 $W_o \in \mathbb{R}^{d_v \times d_o}$ 将其投影到一个新的特征空间中获得的:

$$H = \text{MultiHead}(Q, K, V) = \text{Concat}(H_1, \cdots, H_h) W_o ,$$

$$H_i = \text{Attention}(Q^{(i)}, K^{(i)}, V^{(i)}) = \text{Attention}(X_q W_q^{(i)}, X_{kv} W_k^{(i)}, X_{kv} W_v^{(i)}),$$

对于解码器，Transformer 层在 Attention 的 Softmax 之前引入了一个额外的掩码（MASK）操作，防止查询矩阵 Q 去对序列中尚未解码的后续位置来施加注意力操作。此外，在自注意层之后还有一个额外的“交叉注意力”层，其中查询矩阵 Q 是从解码器中前一层的输出中派生出来的，而键矩阵 K 和值矩阵 V 是从编码器的最后一层的输出中转换而来的。这种设计的主要目的是为了让 Transformer 在解码时避免看到真实标签，并且同时处理来自编码器的信息。

2）全连接前馈层，在注意力层之后的全连接前馈层由两个线性变换和一个非线性激活函数组成。将输入矩阵表示为 $X \in \mathbb{R}^{d \times d_i}$，前馈层的输出

$$\text{FFN}(X) = \sigma(X W_1 + b_1) W_2 + b_2$$

其中，$\sigma(\cdot)$ 是激活函数（通常为 ReLU 或 GELU），而 $W_1 \in \mathbb{R}^{d_i \times d_f}$，$b_1 \in \mathbb{R}^{d_f}$，$W_2 \in \mathbb{R}^{d_f \times d_o}$，$b_2 \in \mathbb{R}^{d_o}$ 均为可学习的参数。在实践中，d_i 通常设置为 d_o，d_f 设置为 d_i 的 4 倍。FFN 作用包括两个方面：（1）非线性激活：在每个注意力模块之后引入了非线性激活函数 $\sigma(\cdot)$，这有助于增强模型的表达能力；（2）信息整合：自注意力机制允许模型在不同的

位置间建立联系，而全连接前馈网络则在每个位置独立地对信息进行整合，这两者结合起来，使得模型既能捕获全局（长距离）的信息，又能在每个位置进行局部的信息整合。

3）残差连接和层归一化，在每个注意力层和每个全连接前馈层之后，Transformer 都应用残差连接（Residual Connection）和层归一化（Layer Normalization）技术，这有助于在模型非常深时保留信息并确保模型性能。具体来说，对于某一层神经网络 $f(\cdot)$，残差连接和归一化层定义为 LayerNorm$(X + f(X))$。

在 Transformer 模型被提出之后，它也衍生出了相当一部分的变体，包括在编码器和解码器中出现了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。

2.2 语言大模型架构

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，不过它们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，语言大模型架构大致可以分为三类 [36]：

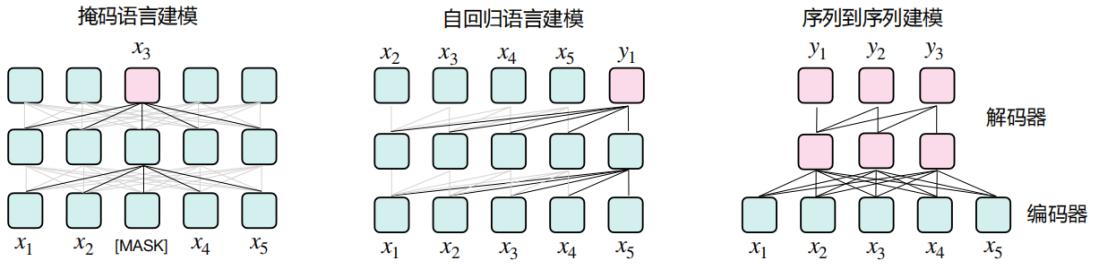


图 2-3 语言大模型的三种典型架构[36]

2.2.1 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，其中 BERT[11]和 RoBERTa[41]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。在预训练时，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“完型填空”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，但是它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、更长的时间和更多的数据进行训练。

2.2.2 自回归语言建模

自回归语言模型在训练时通过学习预测序列中的下一个词来建模语言，其主要是通过 Transformer 解码器来实现。自回归语言模型的优化目标为最大化对序列中每个位置的下一个词的条件概率的预测。代表性模型，包括 OpenAI 的 GPT 系列模型[16][26]、Meta 的 LLaMA 系列模型[30]和 Google 的 PaLM 系列模型[17]。其中，GPT-3 [16]是首个将模型参数扩增到千亿参数规模的预训练模型。自回归语言模型更加适用于生成任务，同时也更适用于对模型进行规模扩增。

2.2.3 序列到序列建模

序列到序列模型是建立在完整 Transformer 架构上的序列到序列模型，即同时使用编码器-解码器结构，代表性模型包括 T5[42]和 BART[43]。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最大化在给定掩码的字符序列的情况下目标字符序列的概率。

总体而言，自回归语言模型较其它预训练语言模型架构展现了更优异的情境学习、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[5]。

2.3 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。

2.3.1 语言大模型的预训练

支撑语言大模型高效训练的技术主要包括高性能训练工具、高效预训练策略、高质量训练数据、高效的模型架构等，其中高性能训练工具和高质量训练数据分别见第 5 章和第 4 章。

高效预训练策略。其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。一种是在预训练中**设计高效的优化任务目标**，使得可以使得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。第二种是**热启动策略**，在训练开始时线性地提高学习率，以解决在预训练中单纯增加批处理大小可能会导致优化困难问题。第三种是**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，该方法认为不同的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。第四种是**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。在中文语言大模型 CPM-2[34]中，采用知识继承技术经测试可以使大模型在预训练前期提速 37.5%。第五种是**可预测扩展策略**（Predictable Scaling）[26]，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 1000 倍至 10000 倍较少计算资源训练的小模型可靠地预测 GPT-4 某些性能，大幅降低了模型训练成本。

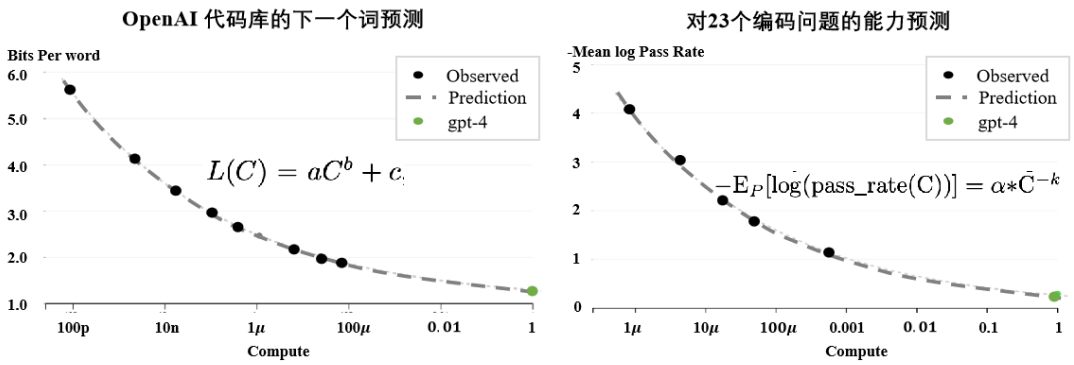


图 2-4　GPT-4 的可预测扩展实验[26]

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）**统一的序列建模**，旨在将多种自然语言处理任务（如分类、信息抽取、翻译、对话等）整合到一个统一的框架，然后在同一模型中执行多个任务，以实现更高效的自然语言处理。该方法可以充分利用大规模训练数据，从而提高了模型在多个任务上的性能和泛化性。这减少了开发和维护多个单独模型的复杂性以及资源消耗，提高模型的通用性。统一任务序列建模有两种方式：一是转化为序列生成的统一任务，如 T5[42]和 BART[43]等将多种自然语言任务统一转化文本到文本的生成任务；二是转化为语言大模型预训练任务，通过语言提示在输入文本中插入人类设计或者自动生成的上下文，实现对不同任务的处理。（2）**计算高效的模型架构**。从 Transformer 模型架构本身在处理训练复杂度、编解码效率、训练稳定性、显存利用等方面进行优化。比如，Transformer 其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[44]，其采用线性化注意力+尺度保持（Retention ）机制，在基本保持模型性能的基础上同时实现模型训练速度、推断速度和内存节约的大幅提升。针对自注意力显存消耗大，斯坦福大学在 Transformer 中引入 FashAttention[45]，给出了一种具有 IO 感知，且兼具快速、内存高效

的注意力算法，已经被各种主流大模型采用以扩展对超长文本输入的支持。最近，模块化大模型架构引起广泛关注，其利用大模型的神经激活稀疏性，对稠密模型进行模块化划分，不同任务只经过部分模块计算实现训练和推理加速，典型工作包括 Google 的 Switch Transformers [46]和 Pathways[47]架构、清华大学的 MoEfication 架构[48]、FastMoE 架构[49]等。

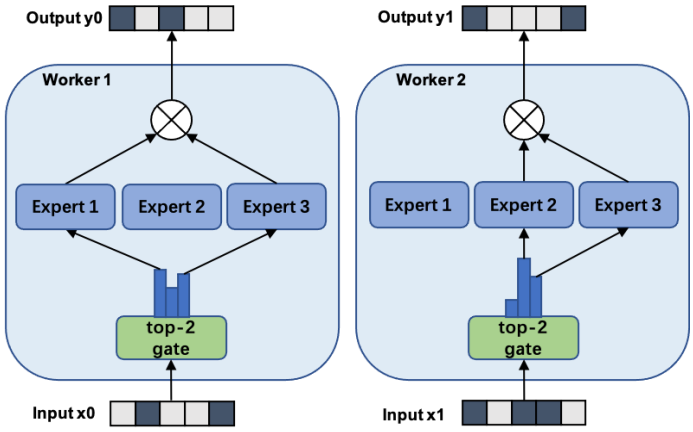


图 2-5 混合专家化的模型架构[49]

2.3.2 语言大模型的适配微调

语言大模型由于在大规模通用领域数据预训练通常缺乏对特定任务或领域的知识，因此需要适配微调。微调可以帮助模型更好地适应特定需求，如对敏感数据（如医疗记录）的处理，同时不暴露原始数据。此外，微调可以提高部署效率、减少计算资源需求。指令微调和参数高效学习是适配微调的关键技术。

指令微调 (Instruction Tuning)[21]，是一种可以帮助语言大模型实现人类语言指令遵循的能力，在零样本设置中泛化到未见任务上的学习方法。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，其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，即在给定指令提示的情况下给出特定的回应，其中提示可以选择性包含一条解释

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**，指语言大模型准确理解人类语言指令的能力，是语言大模型执行指令完成任务的前提。为了增强对指令的理解，许多工作采用多任务提示方式对基于指令描述的大量任务集上对语言大模型进行微调，如 FLAN[50]、InstructGPT[21]等，这些模型在未见的任务上显示出优越的零样本性能。

（2）指令数据获取，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：i）基于公开人工标注数据构建，代表指令数据集包括 1616 种不同任务的 Super-Natural Instruction[51]、2000 种不同 NLP 任务的 OPT-IML[52]。ii）借助语言大模型的自动生成构建，如 Unnatural Instructions[53]，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。iii）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真实指令数据。

（3）指令对齐，语言大模型在多种自然语言处理任务上都展现了卓越的性能。然而，它们有时可能会出现不预期的行为，如创造虚假信息、追求错误目标或产生有偏见的内容[5]。其根本原因在于，语言大模型在预训练时仅通过语言模型建模，未涉及人类的价值观或偏好。为了解决这一问题，研究者提出了“指令对齐”，使语言大模型的输出更符合人类的预期。但这种对齐与原始预训练有所不同，更注重于有用性、诚实性和无害性。此外，指令对齐可能会降低语言大模型的某些通用能力，这被称为“Alignment Tax”。为实现模型输出与对人类价值的对齐，InstructGPT 提出了一种基于人类反馈的微调方法，利用了强化学习技术，将人类反馈纳入模型微调过程。实际上，ChatGPT 也采用了与 InstructGPT 相似的技术，以确保产生高质量且无害的输出。指令对齐的广泛应用，适配微调从纯数据学习的传统微

调范式开始逐步向人类学习范式的转变。

参数高效微调（Parameter-Efficient Tuning）。早期以 BERT 为代表的微调方法，是在大模型基座上增加一个任务适配层，然后进行全参微调，但是这种方法存在两方面的问题：一是任务“鸿沟”问题，预训练和微调之间的任务形式不一致，这种差别会显著影响知识迁移的效能。二是高计算成本，语言大模型的参数规模不断增长，导致模型全参微调也需要大量计算资源。解决以上问题的有效途径是参数高效学习，即通过仅微调少量参数实现大模型在下游任务上获得全参微调效果。目前许多参数高效微调方法被提出，这些方法大致可分为 3 类[40]：（1）**添加式方法：**旨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。如适配器（Adapter）方法，旨将小规模的神经模块（适配器）注入到预训练模型中，并只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在多头自注意和前馈网络子层之后，成为最广泛使用方式；（2）**指定式方法：**旨在原模型指定模型中部分参数为可训练参数，并固定模型其他参数。这类方法简单也十分有效，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中，仅仅优化低维空间中的近似参数，显著降低模型的计算量和内存消耗。如 LoRA[54]，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 $W = W_0 + \Delta W = W_0 + W_{down} W_{up}$ 。

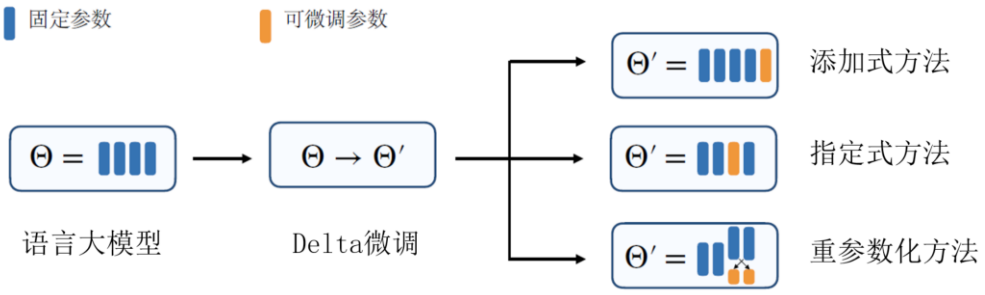


图 2-6 参数高效微调的 3 种范式[40]

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，这种技术也统一成技术框架 Delta Tuning[40]。一些围绕参数高效微调的开源工具也被研发，代表性包括 OpenPrompt[55]、OpenDelta[56]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[57]、Delta Center[40]等。随着语言大模型的兴起，高效微调吸引了越来越多的关注，以开发一种更轻量级的下游任务适配方法。特别地，LoRA[54]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。

2.3.3 语言大模型的提示学习

通过大规模文本数据预训练之后的语言大模型具备了作为通用任务求解器的潜在能力，但这些能力在执行一些特定任务时可能不会显式地展示出来。在大模型输入中设计合适的语言指令提示有助于激发这些能力，该技术称为模型提示技术。代表性的提示技术有指令提示和思维链提示：

指令提示（Instruction Prompt），也称为提示学习。OpenAI 在 GPT-3 [16]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，证明在低资源场景下非常有效，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）”来给数据嵌入额外的上下文以重新组织下游任务，使之看起来更像是在语言大模型预训练过程中解决的问题[10]。指令提示有三种形式：（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。其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，例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，只依靠一个精心设计的提示来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（In-context**

Learning, ICL），也称**情境学习**，是指将一个自然语言问题作为语言大模型的输入，并将其答案作为输出[16]。情境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。情境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。其挑战在于，如何确保问题质量、如何评估答案正确性等。

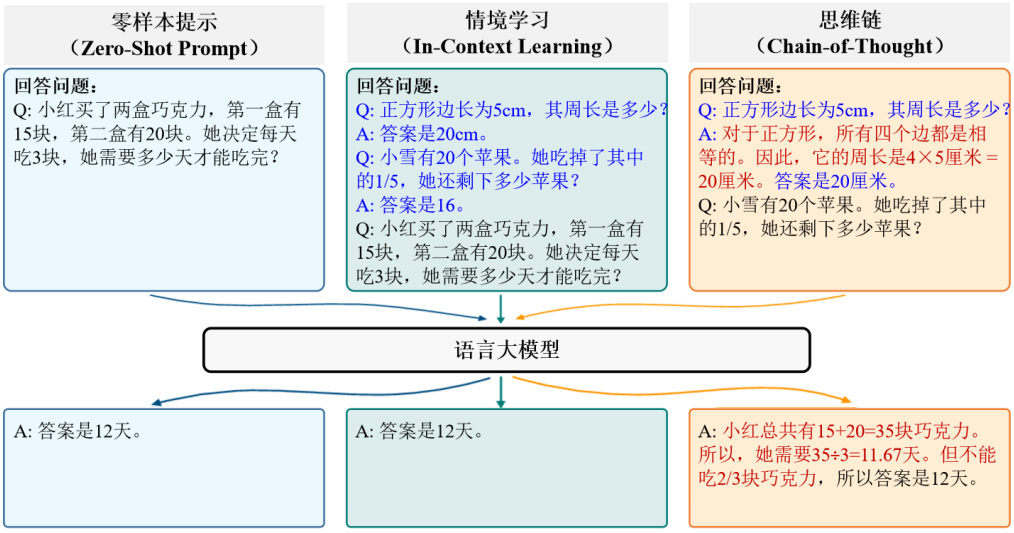


图 2-7 几种提示样例对比

思维链（Chain-of-Thought，CoT）[58]。推理的过程通常涉及多个推论步骤，通过多步推理允许产生可验证的输出，可以提高黑盒模型的可解释性。思维链是一种提示技术，已被广泛用于激发语言大模型的多步推理能力，被鼓励语言大模型生成解决问题的中间推理链，类似于人类使用深思熟虑的过程来执行复杂的任务。在思维链提示中，中间自然语言推理步骤的例子取代了少样本提示中的〈输入，输出〉对，形成了〈输入，思维链，输出〉三元组结构。思维链被认为是语言大模型的“涌现能力”，通常只有模型参数规模增大到一定程度后，才具有采用思维链能力。激活语言大模型的思维链能力方法，在提示中给出逐步的推理演示作为推理的条件，每个演示都包含一个问题和一个通向最终答案的推理链（图 2-7）。

2.3.4 语言大模型的知识增强

知识运用和推理能力是衡量语言大模型智能水平的重要因素。美国 Allen AI 研究大模型的问答能力，发现 GPT-3 在处理具有预设立场（false premise）的简单性常识性问题时，如类似“太阳有几只眼睛？”，GPT-3 仍然会给出“太阳两只眼睛”的荒谬回复。有效的解决方法是在深度学习模型基础上融入各类型相关外部知识。根据大模型知识融合部位不同，知识融合方法从模型输入、神经架构、模型参数、输出等不同层面，大致分为以下 4 类[59]，如图 2-8 所示：

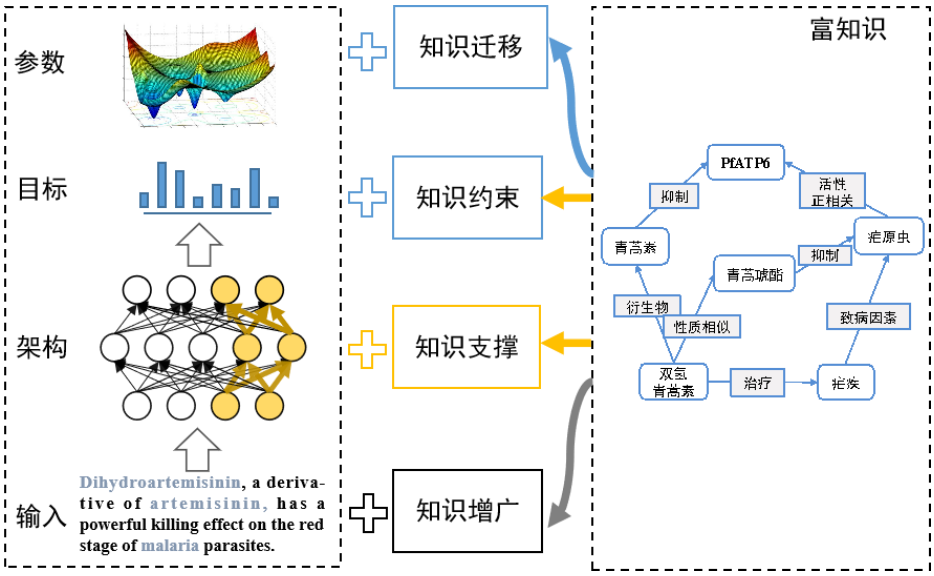


图 2-8 语言大模型知识增强的 4 种途径

知识增广：从输入端增强模型，有两种主流的方法：一种方式是直接把知识加到输入，另一方法是设计特定模块来融合原输入和相关的知识化的输入表示。

知识支撑：关注于对带有知识的模型本身的处理流程进行优化。一种方式是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来从大模型底部注入丰富的记忆特征。另一方面，知识也可以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。

知识约束：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。或者利用知识构建额外的预测目标，在原始语言建模之外构建了相应额外的预训练目标。

知识迁移：模型知识作为重要的知识来源，也可以直接用于下游任务，例如初始化模型参数。迁移学习和自监督学习都是知识迁移的重要研究方向。目前，知识迁移技术已被广泛应用于自然语言处理，以 BERT 为首的各种预训练模型是现在知识迁移的主要方法。

2.3.5 语言大模型的工具学习

语言大模型具备理解、推理和决策能力，可与外部工具互动。在特定领域任务中，如金融领域的证券交易和市场预测，语言大模型通常需要结合外部工具获取信息和技能才能处理。整合外部工具与语言大模型可以发挥各自优势实现复杂任务的处理，其中外部工具可增强专业知识和可解释性，语言大模型提供语义理解和推理规划能力。

2021 年底，OpenAI 推出 WebGPT[60]，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上实现非常强的能力，展现了语言大模型利用工具解决复杂问题的巨大潜力。该工作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[61]，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[62]，实现 ChatGPT 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）[63]。清华大学在现有大模型工具使用方法基础上，提出了工具学习（Tool Learning）框架[24]，指在让模型能够理解和使用各种工具完成任务的学习过程。

图 2-9 基于用户接口视角的工具分类[24]

目前可交互的通用工具按用户接口大致可分为三类（图 2-9）：物理交互的工具（如机器人、传感器等）、基于图形用户界面的工具（如浏览器、Office 办公软件等）、基于编程接口的工具（如数据库、知识图谱）等。从学习目标的角度来看，现有工具学习方法主要可以分为两类[24]：一类是工具增强学习（Tool-augmented Learning），利用各种工具的执行结果，增强基础模型性能。在这一范式中，工具执行结果被视为辅助生成高质量输出的外部资源；第二类是工具导向学习（Tool-oriented Learning），将学习过程重点从增强模型性能转向工具执行本身。这一类研究关注开发能够代替人类控制工具并进行序列决策的模型。

第 3 章 多模态大模型技术

不同于语言大模型只对文本进行处理，多模态大模型将文本、语音、图像、视频等多模态数据联合起来进行学习。多模态大模型融合了多种感知途径与表达形态，能够同时处理和理解来自不同感知通道（例如视觉、听觉、语言和触觉等）的信息，并以多模态的方式表达输出。

3.1 多模态大模型的技术体系

现有的多模态大模型主要有面向理解任务的、面向生成任务的、兼顾理解和生成的、知识增强的多模态大模型。

3.1.1 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，其核心结构通常是基于 Transformer 的编码器。按照模型结构的不同，面向理解任务的多模态大模型又可再分为单流和多流两种结构。单流结构是指不同模态的特征在拼接后由一个共享的 Transformer 网络进行处理；而多流结构中，不同模态则分别由 Transformer 网络进行编码处理，这些网络之间存在有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[64]，它采用了一种双流 Transformer 的结构，首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，最后所得到文本-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[65]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到共享的语义空间中。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（zero-shot）分类能力。另一个与 CLIP 类型的代表性方法

ALIGN[66]，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 是个图文双流结构，而 VATT[67]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来训练模型。

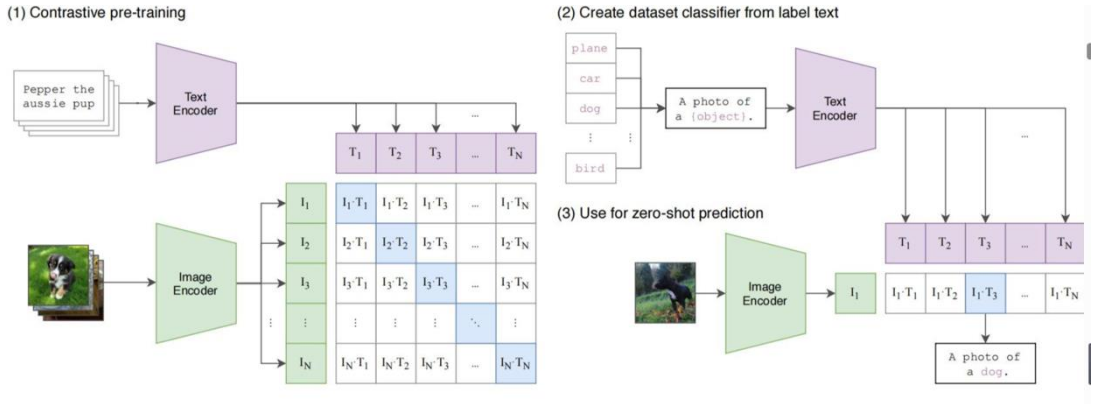


图 3-1 CLIP[65]模型架构图

单流结构的一个典型代表是 VL-BERT[68]，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。此外，另一代表性方法 UNITER [69]，则采用了一种多任务的多模态预训练方法，相对于其它方法，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 VideoBERT[70]和 ActBERT[71]，其中 VideoBERT 是一个视频-语言模型，它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，输入不止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。

现有的面向理解任务的多模态大模型大多都以上面两类结构为基础，此外，也有不少方法在预训练任务上进行研究，引入更多的预训练任务或设计统一的架构去训练所有的任务等。例如，其中一个典

型方法 Florence[72]，它着重于如何使模型适应各种下游任务，并设计了一个由多模态大模型和适应模型组成的工作流。具体对于任务适应，该模型使用动态头部适配器将学习到的视觉特征表示从场景扩展到对象，采用 CoSwin 适配器来学习视频表示，并使用 METER 适配器将模型应用到依赖细粒度视觉-语言表示的视觉语言任务。

3.1.2 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）。

在序列生成模型中，DALL-E[73]是个典型代表。它是由 OpenAI 发布的一个基于 4 亿图文对训练的图像生成模型，通过采用 VQVAE[74]图像离散自编码器和 GPT 组合的结构，在以文生图任务上取得了突破性的生成质量和泛化能力，被称作图像版 GPT。另一典型的图像生成模型是北京智源研究院所的 CogView 模型[75]（如图 3-2 所示），它具有与 DALL-E 类似的结构，但是面向中文环境的文本到图像生成，并进一步探索了多模态生成模型在下游任务上精调后的泛化能力。CogView 在基于文本控制的样式学习、服装设计和图像超分等任务上均取得出色的效果。在文本生成方向上，采用序列生成模型是最主流的方案，例如，典型方法 GIT[76]是一个视觉到文本的多模态大模型，统一了图像/视频的描述和问答等视觉语言任务，它包含有一个图像编码器和一个文本解码器，其文本解码器在视觉编码的基础上，以自回归的方式来生成文本。

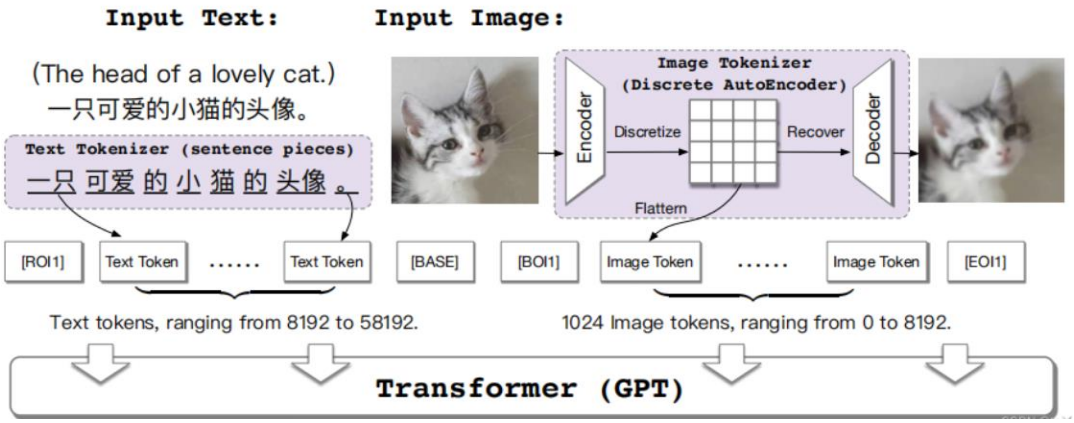


图 3-2 CogView [75]模型架构图

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，来学习恢复数据。扩散模型的一个代表性方法 LDM[77]，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[78]与谷歌的 Imagen[79]。其中，DALL-E2 首先然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 64x64 像素的低分辨率的图像，然后会通过两个超分辨率扩散模型来逐渐提高分辨率到 1024x1024 像素，如图 3-3 所示。此外，与 DALL-E2 不同的是，Imagen 使用了通用语言大模型 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。

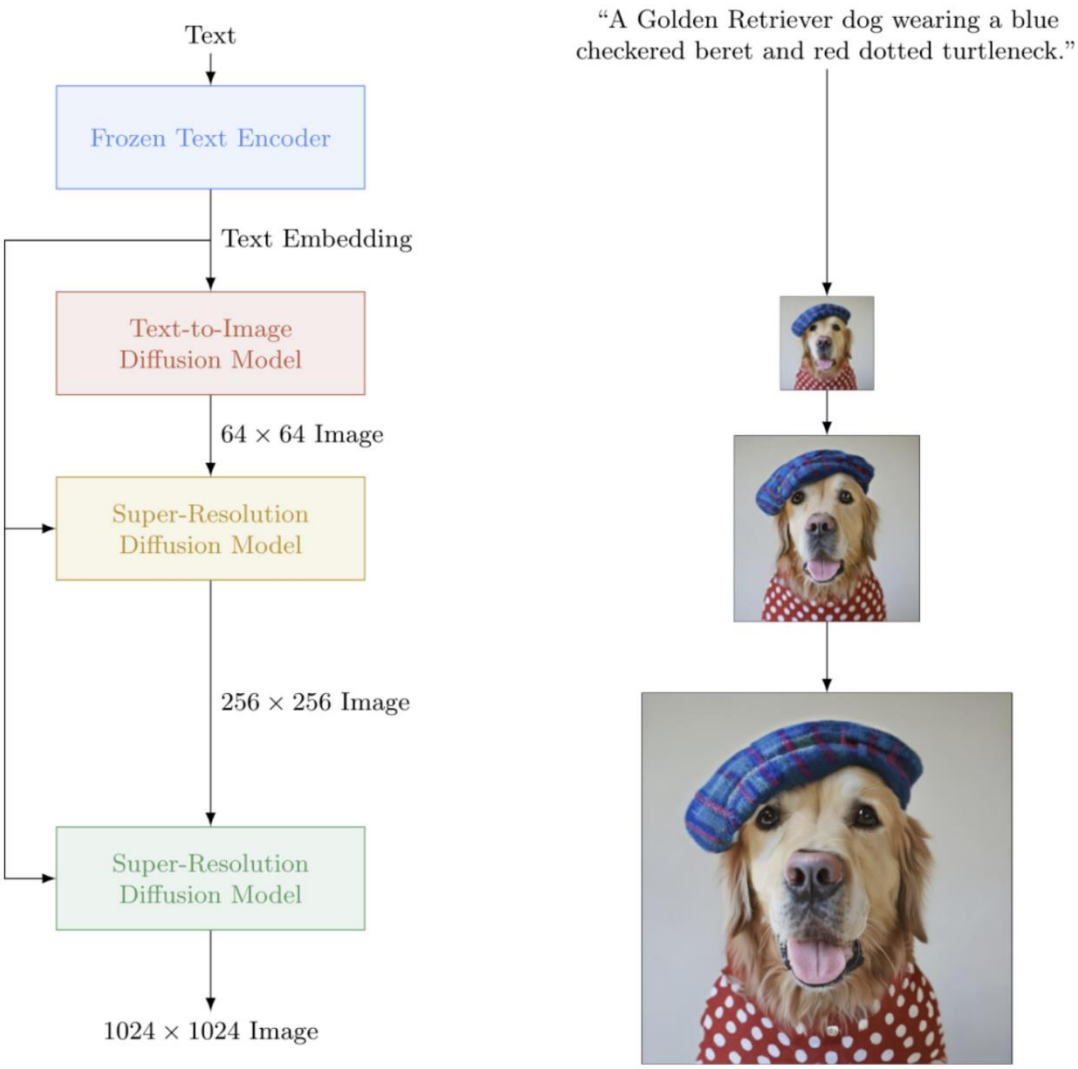


图 3-3 Imagen[79]模型架构图

3.1.3 兼顾理解和生成任务的多模态大模型

Transformer 编码器通过双向的注意力机制来学习对数据的理解能力，而解码器通过单向的注意力机制学习生成能力。为了让模型同时具备这两种能力从而可以在更广泛的下游任务上应用，可以联合 Transformer 编码器与解码器，设计能够兼顾理解与生成任务的多模态大模型。例如，一个典型方法是蒙特利尔大学所的 VL-T5[80]模型，将多个多模态任务统一为文本生成任务。具体地，如图 3-4 所示，该模型由 Transformer 编码器和自回归的解码器组成，其主要创新点在于针对训练任务与数据的不同采用不同的输入文本与输出文本的构

造方式，这种将模型结构和目标任务统一的方法可以充分利用不同任务的数据来训练模型，提高模型的泛化性。这类方法的另一个典型模型 Unified VLP[81]，它的主要特点是编码器和解码器共享同一个 Transformer 网络。该方法通过设置注意力掩码来控制网络为编码器或解码器。具体地，当使用编码器时，注意力掩码为双向掩码，任一位置都可建模前后两个方向的依赖关系；当使用解码器功能时，注意力掩码设置为单向，每一位置只能建模前文的依赖关系。这种编解码共享的方式能够减少参数量，使网络更加简洁。

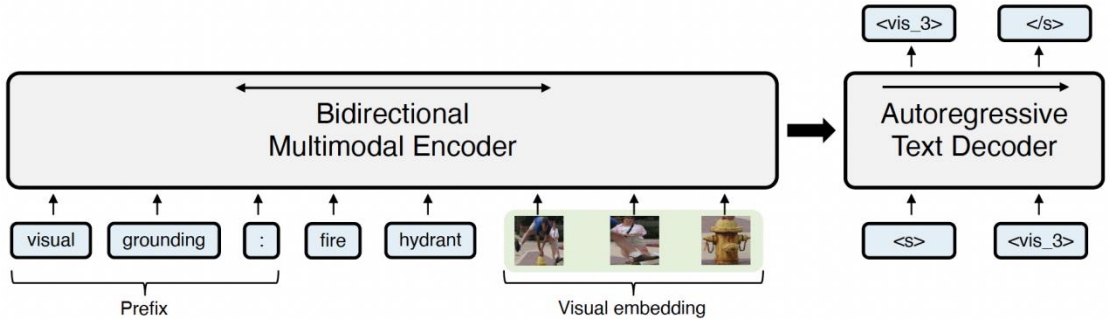


图 3-4 VL-T5[80]模型架构图

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。这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，另外，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。这类模型在结构方面常由单模态编码器、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[82]中的 Q-former 结构等（如图 3-5 所示）。这类模型通常涉及到两个阶段的训练过程。在第一阶段，训练各个模态到语言大模型的语义对齐，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进行

训练。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。

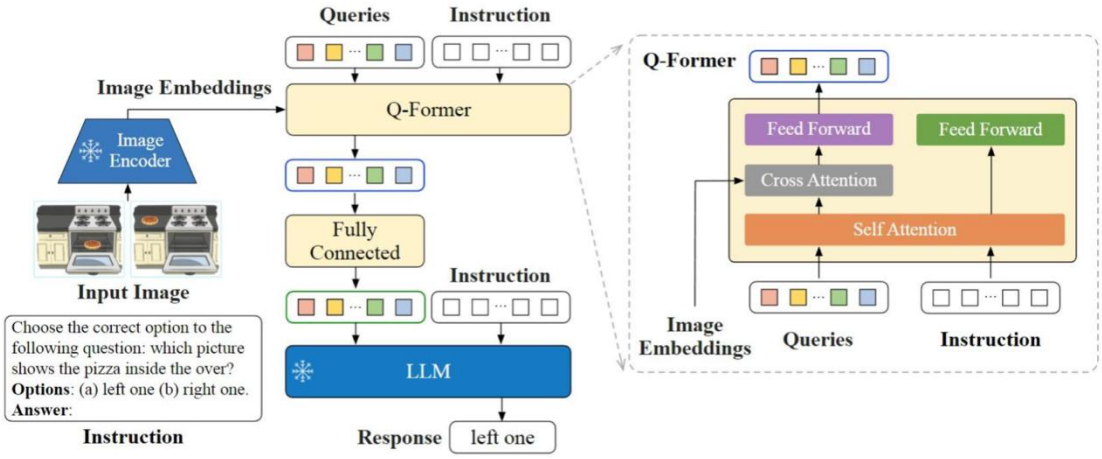


图 3-5 BLIP-2[82]模型架构图

3.1.4 知识增强的多模态大模型

大模型不仅对大规模数据有着卓越的拟合能力，还能够学习到隐式的知识。为了促进更有意义的理解和预测，还需要寻找将隐式知识与显式知识（例如来自知识图谱）联系起来的方法。因此，将知识图谱、场景图、外部知识库等结构化的知识信息注入大模型中，将可增强多模态大模型的知识利用能力。例如，在场景图知识的利用上的一个典型方法是百度的 ERNIE-ViL[83]模型，如图 3-6 所示，它在视觉-语言模型中引入了由文本解析而来的场景图信息，在预训练过程中通过将场景图中的知识实体和关系进行掩码后要求模型预测所掩码位置的知识信息，以此作为更细的多模态预训练任务，这能够使得模型更能精准把握图像和文本之间细粒度的对齐信息。在知识图谱的利用上，典型方法有 KRISP[84]，它结合了隐含知识和明确知识的学习，即从无监督语料和有监督的训练数据中学到隐含的知识，从结构化数

据知识图谱中学习明确的符号化的知识，这样既可以进行隐式的知识推理，又可以获取符号化的知识表示。

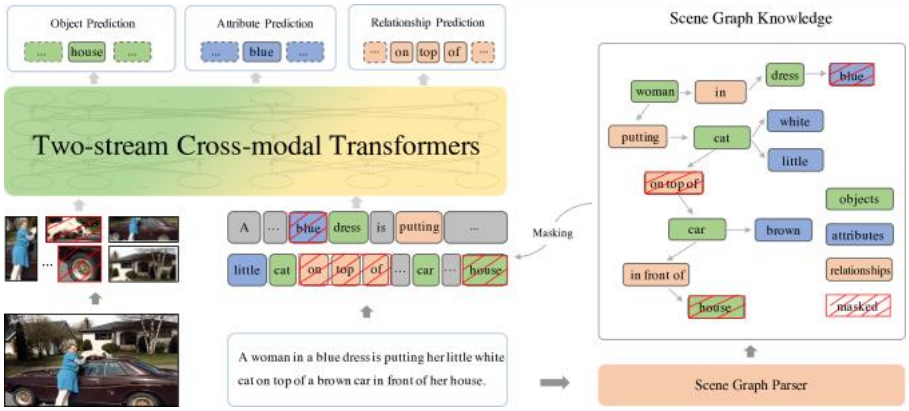


图 3-6 ERNIE-ViL[83]模型架构图

3.2 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。

3.2.1 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扮演着关键角色，需要精心设计以适应和理解来自不同源的复杂特征。例如，在处理图像和文本模态时，通常会采用 Transformer 或卷积神经网络（CNN）来捕捉视觉和语言之间的复杂关系；而对于事件流，脉冲神经网络可能更为适合，因为它们能有效地模拟信息的时序动态。随着模型规模的增加，大型多模态大模型展示出强大的记忆能力和性能增益。然而，模型复杂度的增加也不可避免地引入了计算效率的挑战，并可能最终遇到性能瓶颈。因此，对于更高效的网络模型结构的设计和探索，比如改进或甚至替代 Transformer，成为了重要的研究方向。

其次，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。其中一个代表性方法是 DeepMind 的 Flamingo[85]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 万图文混排的网页数据进行视觉-语言联合训练；Flamingo 具有少样本（few-shot）的多模态序列推理能力，无需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-1[86]，它将一个基于 Transformer 的语言模型作为通用接口，并将其与视觉感知模块对接，使得模型“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）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[87]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。其中的一个典型模型是 ChatBridge[88]，它使用多个并行的感知模块用来处理包括图片、音频、视频的在内特征，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，使得模型具备灵活感知、理解混合模态信息的能力。

最后，对于多模态预训练，设计与下游任务更高兼容性的网络结构模型显得尤为重要。具体来说，可以通过引入编码器-解码器结构将多模态理解和生成任务统一到一个框架下，从而更好地支持各种多模态任务。这主要涉及到跨模态的注意机制、模态间的对齐和翻译、以及更复杂的特征集成策略。

3.2.2 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（Masked Language Modeling，MLM）：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，如图 3-7。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。

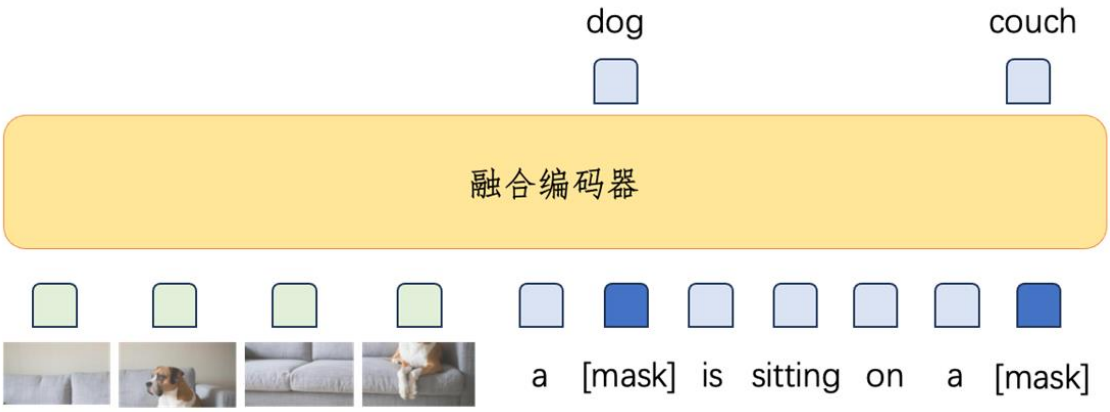


图 3-7 掩码语言预测

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被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，如图 3-8。

图 3-8 掩码视觉预测

3）图像-文本匹配（Image-Text Matching，ITM）：前面的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，从而建立图像和文本之间的语义关联，如图 3-9。

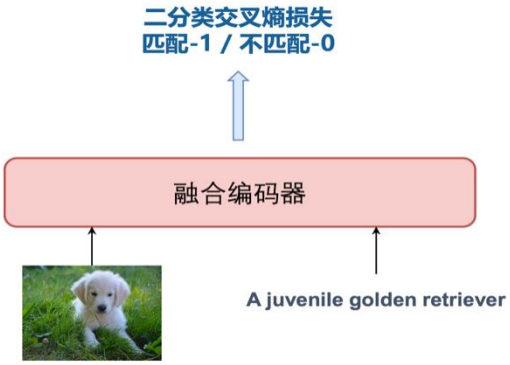


图 3-9 图像文本匹配

4）图像-文本对比学习(Image-Text Contrastive Learning, ITC)，使用对比学习的方法将图像和文本的相同样本对的向量表示拉近，不同样本对的向量表示推远，从而增强图像和文本之间的语义关联性。这使得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示能力，如图 3-10。

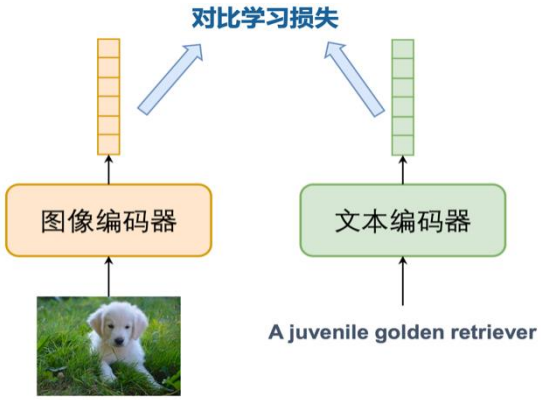


图 3-10 图像-文本对

3.2.3 多模态大模型的下游任务微调适配

多模态大模型的最终目标是适配并提升特定下游任务上的性能表现，因此，需要通过微调适配将预训练大模型的能力迁移到特定数

据下的特定任务场景中。目前，多模态大模型的微调适配方式主要有三种：

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比较好的完成下游的任务。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版即可适应特定任务，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：**每个任务都有自己独立的适配器层，使得模型可以在不同任务之间共享通用预训练模型的表示，同时在每个任务上进行个性化的调整。适配器层通常由较少的参数组成，因此它们比在整个模型上进行微调更加高效。在训练过程中，预训练模型的参数是固定的，只有适配器层的参数被更新。

现有的预训练大型方法通过特征微调或提示学习用于下游任务，也需要更多研究考虑为多模态大型模型开发增量学习算法。未来，如何将新模态引入到已经预先训练好的多模态模型中具有实际意义，因为新的传感器(模态)将在未来的某个不确定时间出现，设计的多模态大型模型应该足够灵活以应对这种情况。

第 4 章 大模型技术生态

随着大模型技术的快速发展，大模型的生态体系也在快速构建。典型的大模型平台如 ChatGPT、文心一言、讯飞星火等提供如 APP、网页版、API 接口等多种形式的开放服务，并通过开放插件机制、Function Call 等实现大模型外部工具、服务的调用，加速应用生态的发展。与此同时，开源大模型也已经成为生态体系中的关键组成部分。通过大模型的开源共建，凝聚了来自企业、高校、科研院所等众多领域高水平开发者的力量，加速大模型的科研创新和产品迭代。伴随着大模型的开源开放，深度学习开源框架和工具更加注重分布式训练和推理能力，并加速与 AI 芯片开展适配和联合优化。大模型的训练数据作为生态中另一关键组成部分，相关数据集和配套工具也在加速汇聚和优化，愈发得到广泛重视。下文对以上几个方面分别展开介绍。

4.1 典型大模型平台

（1）GPT 系列

OpenAI 的 GPT 系列模型是自然语言处理领域的重大突破，其中 ChatGPT 和 GPT-4 是两个代表性模型。ChatGPT 专注于对各种文本指令做出回应，模型的训练过程包括有监督的指令微调与强化学习。现在的 ChatGPT 支持最长达 32,000 个字符，它可以执行各种任务，包括代码编写、数学问题求解、写作建议等。GPT-4 在推理方面的能力比 ChatGPT 更强，同时也减少了幻象的产生，能够更准确地理解和回应复杂的问题，从而提供更高质量的答案，但是引人注目的多模态功能尚未正式开放体验。由于单一的语言模型难以胜任所有任务，自从 ChatGPT 和 GPT-4 发布以来，许多开发者已经开始将各种工具和插件集成到这些模型中，以进一步增强它们的功能。现在，ChatGPT Plus 用户可以使用各种插件来增强模型以满足自己的需求，这极大地扩展了模型的用途和适用领域。

（2）Claude 系列

Claude 系列模型是由 Anthropic 开发的闭源语言大模型，目前包含 Claude 和 Claude-Instant 两种模型可供选择。最早的 Claude 于 2023 年 3 月 15 日发布，并在 2023 年 7 月 11 日，更新至 Claude-2。该系列模型通过无监督预训练、基于人类反馈的强化学习和 Constitutional AI 技术（包含监督训练和强化学习）进行训练，旨在改进模型的有用性、诚实性和无害性。值得一提的是，Claude 最高支持 100K 词元的上下文，而 Claude-2 更是拓展到了 200K 词元的上下文。相比于 Claude 1.3， Claude 2 拥有更强的综合能力，同时能够生成更长的相应。

（3）PaLM 系列

PaLM [17]系列语言大模型由 Google 开发。其初始版本于 2022 年 4 月发布，并在 2023 年 3 月公开了 API。PaLM 基于 Google 提出的 Pathways 机器学习系统搭建，训练数据总量达 780B 个字符，内容涵盖网页、书籍、新闻、开源代码等多种形式的语料。目前 PaLM 共有 8B、62B、540B 三个不同参数量的模型版本。Google 还开发了多种 PaLM 的改进版本。Med-PaLM [89] 是 PaLM 540B 在医疗数据上进行了微调后的版本，在 MedQA 等医疗问答数据集上取得了最好成绩。PaLM-E [90] 是 PaLM 的多模态版本，能够在现实场景中控制机器人完成简单任务。2023 年 5 月，Google 发布了 PaLM 2，但并未公开其技术细节。Google 内部文件显示其参数量为 340B，训练数据为 PaLM 的 5 倍左右。

（4）Bard

Bard 是 Google 开发的对话模型。在 OpenAI 发布 ChatGPT 后，Google 担心其会对自身的搜索业务产生威胁，因此推动了 Bard 的开发。2023 年 2 月 6 日，Bard 正式发布，其基座模型是 Google 此前开发的语言大模型 LaMDA。后续 Google 为 Bard 开展了持续的升级，包括添加数学与逻辑能力、添加代码能力、支持更多语言等。2023

年 5 月，Google 发布了基于新一代语言大模型 PaLM 2 的 Bard。

（5）文心一言

文心一言是基于百度文心大模型的知识增强语言大模型，于 2023 年 3 月在国内率先开启邀测。文心一言的基础模型文心大模型于 2019 年发布。8 月 31 日，文心一言率先向全社会全面开放，提供 APP、网页版、API 接口等多种形式的开放服务。文心一言一方面采用有监督精调、人类反馈的强化学习、提示等技术，还具备知识增强、检索增强和对话增强等关键技术。当前，以文心一言为代表的大模型已经逐步赶超国外最优水平。文心一言基于飞桨深度学习框架进行训练，算法与框架的协同优化后效果和效率都得到提升，模型训练速度达到优化前的 3 倍，推理速度达到优化前的 30 多倍。文心一言还建设了插件机制，通过外部工具、服务的调用，拓展大模型的能力的边界。

（6）讯飞星火认知大模型

讯飞星火认知大模型是科大讯飞于 2023 年 5 月 6 日发布的语言大模型，提供了基于自然语言处理的多元能力，支持多种自然语言处理任务，同时联合中科院人工智能产学研创新联盟和长三角人工智能产业链联盟在业内提出了覆盖 7 大类 481 项任务的《通用人工智能评测体系》；6 月 9 日星火大模型升级到 V1.5 版，实现了开放式知识问答、多轮对话、逻辑和数学能力的提升；8 月 15 日星火大模型升级到 V2.0 版，对于代码和多模态能力进行了提升。同时，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。

（7）腾讯混元

腾讯混元大模型是腾讯于 2023 年 9 月 7 日发布的千亿参数量语言大模型，具有多轮对话、内容创作、逻辑推理、知识增强能力，训练数据截止于 2023 年 7 月。为了降低幻觉问题，混元大模型在预训

练阶段，利用探真算法对目标函数进行了优化，使用强化学习等方法学会识别陷阱。混元大模型针对位置编码进行了优化，并结合指令跟随能力解决长难任务。此外，混元大模型还具备了问题分解和分布推理能力，从而解决逻辑推理问题。

（8）通义千问

通义千问由阿里巴巴基于“通义”大模型研发，于2023年4月正式发布。2023年8月，阿里云开源了70亿参数通用模型和对话模型。它能够以自然语言方式响应人类的各种指令，拥有强大的能力，如回答问题、创作文字、编写代码、提供各类语言的翻译服务、文本润色、文本摘要以及角色扮演对话等。借助于阿里云丰富的算力资源和平台服务，通义千问能够实现快速迭代和创新功能。此外，阿里巴巴完善的产品体系以及广泛的应用场景使得通义千问更具可落地性和市场可接受程度。

4.2 典型开源大模型

4.2.1 典型开源语言大模型

开源模型	公司	包含模型	参数量
LLaMA 系列	Meta	LLAMA，LLAMA2	7B，13B，65B
Falcon 系列	TII	Falcon	1.3B, 7.5B, 40B, 180B
Pythia 系列	EleutherAI	Pythia	70M，160M，410M，1B，1.4B，2.8B，6.9B，12B
T5 系列	Google	T5，mT5，FLAN-T5	60M，220M，770M，3B，11B
BLOOM 系列	BigScience	BLOOM，BLOOM-Z	560M, 1.1B, 1.7B, 3B, 7.1B，176B
GPT-Neo	EleutherAI	GPT-Neo	125M，350M，1.3B，2.7B
OPT 系列	Meta	OPT，OPT-IML	125M，350M，1.3B，2.7B, 6.7B, 13B, 30B，66B，175B
MPT 系列	MosaicML	MPT-Chat	7B，30B
		MPT-Instruct	
文心系列	百度	ERNIE 1.0，ERNIE 2.0，ERNIE 3.0	18M，23M，75M，100M，118M，280M，

			340M，550M
		HelixGEM，HelixFold，HelixGEM2，HelixFold-Single	100K，90M，32M，1B
		VIMER-CAE，VIMER-StrucTexTv2，VIMER-UMS，VIMER-UFO	80M，129M，1.1B，17B
GLM 系列	清华大学，智谱 AI	GLM，ChatGLM，ChatGLM2，WebGLM	2B，6B，10B，130B
Baichuan 系列	百川智能	Baichuan，Baichuan2	7B，13B
CPM 系列	北京智源人工智能研究院、清华大学	CPM-1，CPM-2，CPM-3，CPM-Bee	1B，2B，2.6B，5B，7B，10B，11B，198B
盘古系列	鹏城实验室、华为 MindSpore、华为诺亚方舟实验室、北京大学	盘古 α	350M，2.6B，13B

（1）LLaMA 系列

LLaMA 系列模型[30]是一组参数规模从 7B 到 65B 的基础语言模型，它们都是在数万亿个字符上训练的，展示了如何仅使用公开可用的数据集来训练最先进的模型，而不需要依赖专有或不可访问的数据集。这些数据集包括 Common Crawl、Wikipedia、OpenWebText2、RealNews、Books 等。LLaMA 模型使用了大规模的数据过滤和清洗技术，以提高数据质量和多样性，减少噪声和偏见。LLaMA 模型还使用了高效的数据并行和流水线并行技术，以加速模型的训练和扩展。特别地，LLaMA 13B 在 CommonsenseQA 等 9 个基准测试中超过了 GPT-3 (175B)，而 LLaMA 65B 与最优秀的模型 Chinchilla-70B 和 PaLM-540B 相媲美。LLaMA 通过使用更少的字符来达到最佳性能，从而在各种推理预算下具有优势。与 GPT 系列相同，LLaMA 模型也采用了 decoder-only 架构，但同时结合了一些前人工作的改进，例如：Pre-normalization，为了提高训练稳定性，LLaMA 对每个 Transformer 子层的输入进行了 RMSNorm 归一化，这种归一化方法可以避免梯度爆炸和消失的问题，提高模型的收敛速度和性能；SwiGLU 激活函数，

将 ReLU 非线性替换为 SwiGLU 激活函数，增加网络的表达能力和非线性，同时减少参数量和计算量；RoPE 位置编码，模型的输入不再使用位置编码，而是在网络的每一层添加了位置编码，RoPE 位置编码可以有效地捕捉输入序列中的相对位置信息，并且具有更好的泛化能力。这些改进使得 LLaMA 模型在自然语言理解、生成、对话等任务上都取得了较好的结果。

（2）Falcon 系列

Falcon[31]系列模型是由位于阿布扎比的技术创新研究院(Technology Innovation Institute, TII) 创建的生成式语言大模型，其基于 Apache 2.0 许可发布。Falcon 大模型家族目前主要包含三个基础模型：Falcon-7B，Falcon-40B，以及 Falcon-180B。三个模型都是在 RefinedWe 数据集上训练的，该数据集经历了广泛的过滤和去重过程，以确保高质量的训练数据。同时，三个模型均可用于研究和商业用途。Falcon-7B 基于解码器模型架构，并在精心处理的 RefinedWeb 数据集上使用 1.5 万亿个字符预训练。除此之外，使用多查询注意力机制增强推理时的可扩展性，并显著降低显存需求。Falcon-40B 拥有 400 亿参数，并在 1 万亿字符上进行了训练。在发布后的两个月里，其在 Hugging Face 的开源语言大模型排行榜上排名第一。该系列最新的 Falcon 180B 具有 1800 亿参数的，在 3.5 万亿字符上进行预训练。该模型在推理、编码、熟练度和知识测试等各种任务中表现出色，在 Hugging Face 的开源语言大模型排行榜上击败了 Meta 的 LLaMA2-70B 等竞争对手。在闭源模型中，它的排名仅次于 OpenAI 的 GPT 4，性能与谷歌的 PaLM 2 Large 相当，但只有其模型的一半参数量大小。

（3）Pythia 系列

Pythia[91]系列模型是由非营利性人工智能实验室 EleutherAI 开发的一系列生成式语言大模型。该系列有 16 个不同参数量的模型

（70M-12B），均是以完全相同的顺序在现有的公开数据集（Pile）上训练的。每个模型都提供了 154 个模型检查点的公开访问权限，并且提供下载和清洗重组数据的工具，以便进一步研究。EleutherAI 使用相同的架构训练了 2 套 Pythia 版本。每一套包含 8 个模型，涵盖 8 种不同的模型尺寸。一套是直接在 Pile 上训练的，另一套则在经过 MinHashLSH 近重复处理后的 Pile 上进行训练，阈值设置为 0.87。经过去重处理后 Pile 大约包含 207B 个字符，而原始 Pile 包含 300B 个字符。由于 Pythia 系列模型在相同架构基础上涵盖多个不同尺寸，Pythia 很适合被用来研究诸如性别偏见、记忆能力和少样本学习等属性如何收到精确训练数据处理和模型规模的影响。目前，Pythia 系列的模型可以在开源模型网站 Hugging Face 上直接获取，也可以通过 Github 的官方页面获取。

（4）T5 系列

T5[42]模型是由 Google Brain 团队在 2019 年提出的一种基于 Transformer 结构的序列到序列（Seq2Seq）模型，其主要特点是将多种 NLP 任务（如翻译、摘要、问答等）转化为一个统一的框架下进行训练，使用文本到文本的统一模型范式，保证了模型的灵活性。T5 模型使用了混合精度训练和自适应优化器来加速训练过程，并且使用了数据过滤和动态批处理来提高数据效率。T5 模型在多个 NLP 任务上都取得了较好的效果，证明了其优秀的泛化能力和迁移能力。T5 模型在预训练阶段使用了 C4 数据集，这是一个包含了超过 750GB 的英文网页文本数据的大规模语料库。T5 模型还探索了不同规模的模型架构和参数量，从小到大分别有 small、base、large、XL、XXL 和 XXXL 六种规模。其中，XXXL 规模的 T5 模型拥有 110 亿个参数，是发布时最大的基于 Transformer 的预训练语言模型之一。

（5）BLOOM 系列

BigScience 在 2022 年提出了 BLOOM 系列模型[92]。BLOOM 拥

有1760亿参数量，是一种基于Transformer解码器架构的语言大模型，并在46种自然语言和13种编程语言上进行预训练。为了能够更好的提升BLOOM模型的多语能力，研究者采用了渐进的方式来选择语料库中包含的语言。此外，BLOOM对原始的Transformer架构提出了许多的更改。相比于在嵌入层添加位置信息，BLOOM采用了ALiBi技术，基于keys和queries二者之间距离来计算注意力分数。虽然ALiBi技术拥有外推至更长的序列的能力，但其在原始序列上也能够带来更稳定的训练过程以及更好的下游表现，比可学习位置编码和旋转位置编码取得了更好的效果。BLOOM在嵌入层之后后立即进行层归一化，显著的改善训练稳定性。由于训练数据较为多样，与单语言分词器相比，BLOOM最终确定的词表尺寸为25万个字符，以支持多种语言。BLOOMZ与BLOOM拥有相同的模型架构与超参数，在包含130亿字符的文本上进行微调，通过独立的验证集来选择最优的模型。使用了包含10-60亿字符的文本进行微调之后，模型的性能趋于平稳。此外，对于13亿参数量和71亿参数量的版本，研究者使用了SGPT Bi-Encoder方案进行对比微调。通过训练，可以得到拥有高质量文本嵌入的模型。近期的基准测试发现，这种模型也能够推广到其他的嵌入任务，例如bitext挖掘、重排或者特征抽取等任务。

（6）GPT-Neo

GPT-Neo [93]系列模型是由EleutherAI开发的预训练语言大模型。GPT-Neo基于OpenAI的GPT系列语言模型的架构，但是采用了分散、社区驱动的方法进行训练。GPT-Neo模型在发布之时，因其较大的参数规模和在各种自然语言处理任务中出色的表现而备受关注。该模型的最大版本，GPT-Neo 2.7B，有27亿个参数。它是在多样化的互联网文本数据上进行训练的，包括书籍、文章和网页，并且已经被证明在广泛的自然语言处理任务上表现良好，如语言生成、摘要和问答。除此之外，其还包含125M, 350M和1.3B等不同参数规模。

GPT-Neo 项目的一个独特之处在于其强调开源开发和社区参与。EleutherAI 公开了该模型的训练权重，使其他研究人员和开发人员能够使用和构建该模型，并开发出许多相关的应用和 GPT-Neo 模型的扩展，包括对特定任务的微调和修改，以提高其在某些特定类型的数据上的效率或性能。

（7）OPT 系列

OPT[94]模型是由 Meta AI 发布的一款 decoder-only 模型，与 GPT-3 相媲美。尽管 GPT-3 在零样本学习和少样本学习方面表现出优秀的能力，但其庞大的训练成本和权重未完全开源的问题，限制了研究社区的相关研究进展。为了应对这些挑战，Meta AI 发布了 OPT 模型，其参数规模从 125M 到 175B 不等，并开源了相关的实验代码。此外，团队还公开了详细的训练日志，深入解释了他们的决策背后的原因和动机，为研究社区的使用和进一步研究提供了重要的参考资源。关于训练成本，OPT-175B 的性能相当，但训练代价仅为 GPT-3 的七分之一。在构建训练语料方面，OPT 使用了多个高质量语料库，包括 RoBERTa 的 BookCorpus 和 Stories，以及更新的 CCNews 版本，还有 Pile 的 CommonCrawl、DM Mathematics、Project Gutenberg、HackerNews、OpenSubtitles、OpenWebText2、USPTO 和 Wikipedia。所使用的这些语料库都经过了严格的收集和过滤，以确保数据的质量和可用性。

（8）MPT 系列

MPT（MosaicML Pretrained Transformer）系列模型是由 MosaicML 研发的开源可商用模型。MPT-7B 在 2023 年 5 月发布，有 MPT-7B-Instruct、MPT-7B-Chat 以及 MPT-7B-StoryWriter-65k+三个版本，其中 MPT-7B-StoryWriter-65k+支持 65K 长度的上下文输入。2023 年 6 月，MPT-30B 发布，拥有比 MPT-7B 更强大的性能，超过了原始的 GPT-3。跟 MPT-7B 一样，MPT-30B 也有两个经过微调的变体：

MPT-30B-Instruct 和 MPT-30B-Chat，它们在单轮指令跟随和多轮对话方面表现出色。MPT-30B 在训练时使用 8,000 字符长度的上下文窗口、通过 ALiBi[95]支持更长上下文以及通过 FlashAttention 实现高效的推理和训练性能。得益于预训练数据混合比例的控制，MPT-30B 系列还具有强大的编程能力。

（9）ERNIE 系列

2019 年，百度将大规模知识与海量数据融合学习的方法，在超大规模模型中引入丰富语言知识与世界知识，突破多源异构数据难以统一表示与学习的瓶颈，显著提升了模型效果和学习效率，并在国内开源首个中文预训练大模型。ERNIE[96]自发布以来在语言理解、文本生成、跨模态语义理解等领域取得多项技术突破，在权威公开数据集上取得世界最好效果总计 90 余项，在国际权威语义评测 GLUE、SuperGlue 等评测上，取得世界冠军 20 余项。系列模型在金融、通信、企业服务、互联网等行业取得广泛应用，极大促进该领域在国内的研究和产业发展。ERNIE 3.0 [97]大模型最高参数量达到 1000 亿，首次在百亿级预训练模型中引入大规模知识图谱，提出了海量无监督文本与大规模知识图谱的平行预训练方法，促进了结构化知识和无结构文本之间的信息共享，大幅提升了模型对于知识的记忆和推理能力。

（10）GLM 系列

GLM[98]系列模型是清华大学和智谱 AI 等合作研发的开源语言大模型。GLM 采用了自回归填空作为预训练任务，并且使用多任务预训练的方式提升模型生成长文本的能力和序列到序列任务的能力。为了能够更好地进行预训练，GLM 采用了二维位置编码，第一维表示当前位置的原文本中的位置信息，第二维表示对应的掩码的位置信息。此外，为了能够尽量推理和训练所占用的显存，GLM-130B 可以使用 INT4 进行量化并且不会明显影响模型效果。通过优化，GLM-130B 可以在 4 张 RTX 3090 Ti（24G）显卡或 8 张 RTX 2080 Ti

（11G）的显卡上进行推理。ChatGLM 是基于 GLM 结构开发的具有 62 亿参数量的语言大模型，支持 2048 的上下文长度。其使用了包含 1 万亿字符的中英文语料进行训练，能够支持中文和英文两种语言的任务。通过监督微调、反馈自助、人类反馈强化学习等多种训练技术，ChatGLM 拥有强大的生成能力，能够生成更符合人类偏好的内容。与 GLM 相似，通过 INT4 量化和 P-Tuning v2[99]等高效微调的算法，ChatGLM 能够在 7G 显存的条件下进行微调。在 ChatGLM 的基础上，ChatGLM 2 使用了包含 1.4 万亿字符的中英预料进行预训练，并使用人类偏好的数据对模型进行对齐训练，拥有比前一版本更加强大的能力，在多个任务上取得提升。通过 FlashAttention 技术，ChatGLM 2 能够处理更长的长下文，支持的长下文长度达到了 3.2 万字符。此外，通过 Multi-Query Attention 技术，ChatGLM 2 能够进一步地提升推理速度，减小对显卡的显存占用。

（11）Baichuan 系列

Baichuan 是由百川智能开发的开源可商用的语言大模型，在权威的中文和英文 benchmark 上均取得同尺寸最好的效果，其基于 Transformer 解码器架构。Baichuan-7B 是在大约 1.2 万亿字符上训练的 70 亿参数模型，支持中英双语，最大 4096 的上下文窗口长度。Baichuan-13B 在 Baichuan-7B 的基础上进一步扩大参数量到 130 亿，并且在高质量的语料上训练了 1.4 万亿 字符，超过 LLaMA-13B 40%，是当前开源 13B 尺寸下训练数据量最多的模型。其支持中英双语，使用 ALiBi 位置编码，最大 4096 的上下文窗口长度，使用 rotary-embedding，是现阶段被大多数模型采用的位置编码方案，具有很好的外推性。百川同时开源了预训练和对齐模型，预训练模型是面向开发者的“基座”，而对齐模型则面向广大需要对话功能的普通用户。除了原始权重，为实现更高效的推理，百川开源了 INT8 和 INT4 的量化版本，相对非量化版本在几乎没有效果损失的情况下大大降低

了部署的机器资源需求。Baichuan2-7B 和 Baichuan2-13B，均基于 2.6 万亿高质量多语言数据进行训练，在保留了上一代开源模型良好的生成与创作能力，流畅的多轮对话能力以及部署门槛较低等众多特性的基础上，两个模型在数学、代码、安全、逻辑推理、语义理解等能力有显著提升。

（12）CPM 系列

CPM 系列模型由北京智源人工智能研究院和清华大学的合作研发，目前包括了 CPM-1、CPM-2，CPM-3 和 CPM-Bee 典型模型。CPM-1[33]，作为首款中文大规模预训练语言模型，拥有 26 亿参数。其预训练任务采用了经典的自回归语言模型，以 100GB 数据为基础，包括大量丰富多样的中文语料，包括百科、小说、对话、问答、新闻等类型。在多个公开的中文数据集上的实验表明，CPM-1 在对话、文本生成等各类下游任务中，无论是少样本学习还是零样本学习，都表现出卓越的性能。CPM-2[34]模型采用“编码器-解码器”框架，通过词表优化、知识继承、混合专家化等技术，显著缓解了大规模预训练模型训练的计算开销对应用的使用限制。CPM-3 是基于 BMTrain 高效训练框架实现，在预训练阶段采用多样化的任务设计和提示模板预训练技术，在零样本和少样本场景中表现出色。CPM-Bee 的是一个完全开源、允许商用的百亿参数中英文基座模型。它采用 Transformer 自回归架构，通过对预训练预料进行严格后处理提升数据质量，最终在万亿级高质量数据上完成预训练，进一步强化了模型的基础能力。

（13）盘古系列

鹏程·盘古 α [100] 由以鹏城实验室为首的技术团队联合协作开发的，他们首次利用“鹏城云脑 II”和国产 MindSpore 框架，采用自动混合并行模式，在 2048 卡算力集群上进行大规模分布式训练，训练出业界首个以中文为核心 2000 亿参数的预训练生成语言模型。鹏

程.盘古α具备丰富的应用场景，如知识问答、知识检索、知识推理、阅读理解等，并且拥有很强的小样本学习能力。鹏程.盘古α收集了近80TB的原始数据，包括开源数据集、common crawl网页数据、电子书等，搭建了面向大型语料库预处理的分布式集群，通过数据清洗过滤、去重、质量评估等处理流程，构建了一个约1.1TB的高质量中文语料数据集。研究对比了智源研究院发布的首个26亿参数的中文预训练语言模型「悟道·文源」CPM，通过在1.1TB数据中策略抽样了100GB等量数据集训练了2.6B参数规模的「鹏程.盘古α」模型，并在已收集的16个下游任务上进行了对比。实验结果表明，鹏程.盘古α-2.6B比CPM-2.6B模型具有更强的语言学习能力，特别是在生成任务和小样本学习方面。实验还对比了鹏程.盘古α-13B和鹏程.盘古α-2.6B模型的性能。在所有的生成任务和大部分的PPL任务上，13B的模型性能优于2.6B，说明鹏程.盘古α-13B模型具有较强的小样本学习能力。

4.2.2 典型开源多模态大模型

开源模型	单位	包含模型	参数量
KOSMOS-2	微软	-	1.6B
OpenFlamingo	微软	MPT	9B
BLIP-2	Salesforce	OPT, FlanT5	12B
InstructBLIP	Salesforce	LLaMA	7B, 13B
MiniGPT-4	KAUST	LLaMA	7B
LLaMA-Adapter V2	上海人工智能实验室	LLaMA	7B
ImageBind	Meta	ViT, CLIP	-
ChatBridge	中科院自动化所	LLaMA	7B
VisualGLM-6B	清华大学	ChatGLM	7.8B
VisCPM	清华大学	CPM-Bee	10B
mPLUG-Owl	阿里巴巴	LLaMA	7B
Qwen-VL	阿里巴巴	Qwen	9.6B

（1）KOSMOS-2

KOSMOS-2[101]是微软亚洲研究院在 KOSMOS-1 模型的基础上开发的多模态大模型。其中，KOSMOS-1 是在大规模多模态数据集上重头训练的，该模型具有类似 GPT-4 的多模态能力，可以感知一般的感官模态，在上下文中学习（即少样本学习）并能够遵循语音指示（即零样本学习）。KOSMOS-2 采用与 KOSMOS-1 相同的模型架构和训练目标对模型进行训练，并在此基础上新增了对图像局部区域的理解能力。

（2）OpenFlamingo

OpenFlamingo[102]模型是 DeepMind Flamingo 模型的开源复现版，可实现多模态大模型的训练和评估。OpenFlamingo 使用交叉注意力将一个预训练的视觉编码器和一个语言大模型结合在一起。它是在大型多模态数据集（例如 Multimodal C4）上进行训练，可以实现以交错的图像/文本为输入来进行文本生成。例如，OpenFlamingo 可用于生成图像的标题，或者根据图像和文本段落生成问题等。这使得其能够使用上下文学习快速适应新任务。

（3）BLIP-2

BLIP-2[82]通过一个轻量级的查询转换器弥补了模态之间的差距，该转换器分两个阶段进行预训练。第一阶段从冻结图像编码器引导视觉语言表示学习。第二阶段将视觉从冻结的语言模型引导到语言生成学习。BLIP-2 在各种视觉语言任务上实现了最先进的性能，尽管与现有方法相比，可训练的参数明显更少。例如，BLIP-2 模型在零样本 VQAv2 上比 Flamingo 80B 高 8.7%，可训练参数减少了 54 倍。

（4）InstructBLIP

InstructBLIP[103]的特点是设计了一种视觉语言指令微调方法，它基于预训练的 BLIP-2 模型，对视觉语言指令进行微调。具体地，InstructBlip 复用了 BLIP-2 的结构，有一个图像编码器，一个语言大模型和一个 Q-Former 模块来连接前两者。并且采用了指令感知的视

觉特征提取过程，指令不仅会指导语言大模型生成文本，同时也会指导图像编码器提取不同的视觉特征。这样的好处在于对于同一张图片，根据不同的指令，可以得到基于指令偏好更强的视觉特征，同时对于两个不一样的图片，基于指令内嵌的通用知识，可以使得模型有更好的知识迁移效果。

（5）MiniGPT-4

MiniGPT-4[104]使用语言大模型来增强视觉语言理解，将语言能力与图像能力结合。其利用视觉编码器和语言大模型 Vicuna[109]进行结合训练。具体地，MiniGPT-4 使用一个投影层来将来自 BLIP-2 的冻结视觉编码器与冻结的 Vicuna 语言大模型（基于 LLaMA 指令微调得到）对齐。并通过两个阶段来训练 MiniGPT-4。第一个预训练阶段使用大约 500 万个图像-文本对进行视觉-语言对齐训练。第二个微调阶段进行多模态指令微调以提高其生成可靠性和整体可用性。MiniGPT-4 能够产生许多类似于 GPT-4 中展示的新兴视觉语言能力。

（6）LLaMA-Adapter V2

LlaMA-Adapter V2[105]是一种参数高效的视觉指令模型。具体地，首先通过解锁更多可学习参数（例如范数、偏差和比例）来增强 LlaMA Adapter，这些参数将指令遵循能力分布到整个 LLaMA 模型中。其次，采用了一种早期融合策略，将视觉标记提供给早期的语言大模型，有助于更好地整合视觉知识。然后，通过优化不相交的可学习参数组，引入了图像-文本对和指令跟随数据的联合训练范式。该策略有效地缓解了图文对齐和指令跟随这两个任务之间的干扰，仅用小规模的图文和指令数据集就实现了强多模态推理。在推理过程中，该模型将额外的专家模型（例如字幕/OCR 系统）合并到 LLaMA-Adapter 中，以进一步增强其图像理解能力。

（7）ImageBind

ImageBind[106]是 Meta 发布的模型，它的目标是利用图像为中心

绑定学习一个嵌入空间，将文本、图像/视频、音频、深度（3D）、热（红外辐射）和惯性测量单元（IMU）六个模态的数据都投影到这同一个嵌入空间中。进而，在这个空间中可以实现跨模态检索和匹配等任务，此外将该模型与生成模型结合，还可实现音频生成图像、图像生成音频等应用效果。

（8）ChatBridge

ChatBridge[88]是一个新型的多模态对话模型，利用语言的表达能力作为桥梁，以连接各种模式之间的差异，可支持文本、图像、视频、音频几个模态任意组合的模型输入与输出信息。该模型包括两阶段的训练，首先是每个模态与语言对齐，提升跨模态相关性和协同学习能力，接下来是多任务的指令微调，使其与用户的意图对齐。ChatBridge 在面向文本、图像、音频与视频等模态信息的广泛下游任务中表现出优异的零样本学习能力。

（9）VisualGLM-6B

VisualGLM-6B[107]是由语言模型 ChatGLM-6B 与图像模型 BLIP2-Qformer 结合而得到的一个多模态大模型，其能够整合视觉和语言信息。可以用来理解图片，解析图片内容。该模型依赖于 CogView 数据集中 3000 万个高质量的中文图像-文本对，以及 3 亿个精选的英文图像-文本对进行预训练。这种方法使视觉信息能够很好地与 ChatGLM 的语义空间对齐。在微调阶段，该模型在长视觉问答数据集上进行训练，以生成符合人类偏好的答案。

（10）VisCPM

VisCPM[108]是一个多模态大模型系列，其中的 VisCPM-Chat 模型支持中英双语的多模态对话能力，而 VisCPM-Paint 模型支持文到图生成能力。VisCPM 基于百亿参数量语言大模型 CPM-Bee（10B）训练，融合视觉编码器和基于扩散模型的视觉解码器以支持视觉信号的输入和输出。得益于 CPM-Bee 基座的双语能力，VisCPM 可以仅

通过英文多模态数据预训练，泛化实现中文多模态能力。

（111）mPLUG-Owl

阿里达摩院的 mPLUG-Owl[110]大模型可以支持多种数据模态，包括图像、文本、音频等。它采用了预训练和微调的方法，通过使用大规模的预训练数据和对特定任务微调的数据，可以快速高效地完成各种多模态任务。与传统的多模态模型相比，mPLUG-Owl 有更高的准确率和更快的运行速度。此外，它还具有高度的灵活性和可扩展性，可以根据实际需要进行快速部署和优化。

（12）Qwen-VL

Qwen-VL[111] 是支持中英文等多种语言的视觉语言模型。Qwen-VL 以通义千问 70 亿参数模型 Qwen-7B 为基座语言模型，在模型架构上引入视觉编码器，使得模型支持视觉信号输入，并通过设计训练过程，让模型具备对视觉信号的细粒度感知和理解能力。除了具备基本的图文识别、描述、问答及对话能力之外，Qwen-VL 还具备视觉定位、图像中文字理解等能力。

4.3 典型开源框架与工具

PyTorch：PyTorch[27]自身提供了几种加速分布数据并行的技术，包括分桶梯度（bucketing gradients）、通信和计算的重叠（overlapping computation with communication）以及在梯度累积（gradient accumulation）阶段跳过梯度同步（skipping gradient synchronization）。PyTorch 分布式数据并行可以用 256 个 GPU 达到接近线性的可扩展性程度。在 DP 的基础上，原生支持 DDP，每个节点都有自己的本地模型副本和本地优化器，支持多机多卡的分布式训练。一般来说，DDP 都显著快于 DP，能达到略低于卡数的加速比，但要求每块 GPU 卡都能装载完整输入维度的参数集合。在 1.11 版本后，PyTorch 开始支持 FSDP 技术，可以更加高效的将部分使用完毕的参数移至内存中，显著减小了显存的峰值占用，更加吻合大模型的特性。

Tensorflow：TensorFlow[112]是一款由 Google Brain 团队开发的开源机器学习框架，被广泛应用于各种深度学习领域。它可以处理多种数据类型，包括图像、语音和文本等，具备高度的灵活性和可扩展性。TensorFlow 使用数据流图计算模型来建立机器学习模型，用户可以通过定义操作和变量在数据流图上构建自己的神经网络模型。此外，TensorFlow 还提供了众多优化器、损失函数和数据处理工具，以便用户轻松进行模型训练和优化。TensorFlow 在多个领域有广泛的应用，包括自然语言处理、图像识别和语音识别等。它可以灵活地运行在不同硬件平台上，包括 CPU、GPU 和 TPU 等。TensorFlow 还提供了高级 API，使开发者可以快速构建、训练和部署深度学习模型。

PaddlePaddle：飞桨（PaddlePaddle[113]，Parallel Distributed Deep Learning）是我国较早开源开放、自主研发、功能完备的产业级深度学习框架。飞桨不仅在业内最早支持了万亿级稀疏参数模型的训练能力，而且近期又创新性的提出了 4D 混合并行策略，以训练千亿级稠密参数模型，可以说分布式训练是飞桨最具特色的技术之一。飞桨的分布式训练技术在对外提供之前就已经在百度内部广泛应用，如搜索引擎、信息流推荐、百度翻译、百度地图、好看视频、文心 ERNIE 等等，既包含网络复杂、稠密参数特点的计算机视觉（CV）自然语言处理（NLP）模型训练场景，又覆盖了有着庞大的 Embedding 层模型和超大数据量的推荐搜索训练场景。

MindSpore：MindSpore[114]是一款适用于端边云全场景的开源深度学习训练/推理框架。MindSpore 能很好匹配昇腾处理器算力，在运行高效和部署灵活上具有很好的能力。MindSpore 还具有无缝切换静态图动态图、全场景覆盖、新 AI 编程范式等特点。MindSpore 还提供了多种高层 API，如 MindArmour、MindSpore Hub、MindInsight 等，方便开发者进行安全训练、模型共享、可视化分析等操作。

Jittor：Jittor[115]是一个基于即时编译和元算子的高性能深度学

习框架。Jittor 集成了算子编译器和调优器，可以为模型生成高性能的代码。Jittor 与 PyTorch 兼容，可以方便地将 PyTorch 程序迁移到 Jittor 框架上。Jittor 支持多种硬件平台，包括 CPU、GPU、TPU 等。Jittor 在框架层面也提供了许多优化功能，如算子融合、自动混合精度训练、内存优化等。

OneFlow：OneFlow[116][116]能够较好适用于多机多卡训练场景，是国内较早发布的并行计算框架。OneFlow 会把整个分布式集群逻辑抽象成为一个超级设备，用户可以从逻辑视角的角度使用超级设备。最新版本的 OneFlow 和 TensorFlow 一样，实现了同时对动态图和静态图的支持，而且动静图之间转换十分方便。此外，OneFlow 兼容了 PyTorch，支持数据 + 模型的混合并行方式，可提升并行计算性能。

Colossal-AI：“夸父”（Colossal-AI[117]），提供了一系列并行组件，通过多维并行、大规模优化器、自适应任务调度、消除冗余内存等优化方式，提升并行训练效率，并解耦了系统优化与上层应用框架、下层硬件和编译器，易于扩展和使用。提升人工智能训练效率的同时最小化训练成本。在三方面进行了优化：优化任务调度、消除冗余内存、降低能量损耗。夸父从大模型实际训练部署过程中的性价比角度出发，力求易用性，无需用户学习繁杂的分布式系统知识，也避免了复杂的代码修改。仅需要极少量的改动，便可以使用夸父将已有的单机 PyTorch 代码快速扩展到并行计算机集群上，无需关心并行编程细节。

Megatron-LM：Megratron[118] 是 NVIDIA 提出的一种基于 PyTorch 分布式训练大规模语言模型的架构，用于训练基于 Transformer 架构的巨型语言模型。针对 Transformer 进行了专门的优化，主要采用的是模型并行的方案。Megatron 设计就是为了支持超大的 Transformer 模型的训练的，因此它不仅支持传统分布式训练的

数据并行，也支持模型并行，包括 Tensor 并行和 Pipeline 并行两种模型并行方式。同时提出了更加精细的 pipeline 结构与 communication 模式。通过多种并行方式的结合，可以让大模型的训练更快。将核心操作 LayerNorm 和 Dropout 安装输入维度进一步切分，使得这两个需要频繁运行的操作在不大幅增加通信开销的情况下实现了并行。

DeepSpeed：在 2021 年 2 月份，微软发布了一款名为 DeepSpeed[29]的超大规模模型训练工具，其中包含了一种新的显存优化技术——零冗余优化器 ((Zero Redundancy Optimizer, ZeRO)。该技术去除了在分布式数据并行训练过程中存储的大量冗余信息，从而极大地推进了大模型训练的能力。从这个角度出发，微软陆续发布了 ZeRO-1，ZeRO-2，ZeRO-3 和 ZeRO-3 Offload，基本实现了 GPU 规模和模型性能的线性增长。基于 DeepSpeed，微软开发了具有 170 亿参数的自然语言生成模型，名为 Turing-NLG。2021 年月，推出了能够支持训练 2000 亿级别参数规模的 ZeRO-2。目前最新版本 ZeRO-3 Offload 可以实现在 512 颗 V100 上训练万亿参数规模的大模型。

4.4 大模型的训练数据

数据是大模型的关键要素，其所需的数据的种类也非常广泛，涉及多种模态。以语言大模型为例，其所需要的数据包括多语言数据、代码数据、人工标注数据等多种类别。

4.4.1 大模型的训练数据处理流程和特点

根据大模型训练的尺度定律（scaling law），数据规模、模型参数与大模型性能存在紧密关系。近期，微软研究工作表明提高数据质量可以极大地改变尺度定律的形状。通过构建 7B 的小规模“教科书（Textbooks）”高质量的代码训练数据（包括从 web 上筛选的“教科书质量”数据（6B tokens）以及使用 GPT-3.5 生成的教科书和练习（1B tokens)），训练 1.3B 模型 phi-1 在代码评测集 HumanEval 上 Pass@1

准确率达到了 50.6%，超越 GPT-3.5（175B，超过 2TB 训练数据）的 47%。该方法表明，通过构建高质量的数据，可以大大降低大模型训练需要的数据规模，具有重要指导意义。下面是几类用于提升数据质量的预处理方法[119]。

质量过滤：语言大模型训练中需要过滤低质量数据，主要分为两类方法：基于分类器的方法和基于启发式的方法。基于分类器的方法是训练一个文本质量判断模型，用以识别并过滤低质量数据。例如，GPT3、PaLM[17]和 GLaM[120]模型在训练数据构造时都使用了基于分类器的方法。而基于启发式的方法则是通过一组精心设计的规则来消除低质量文本，主要包括语言过滤、指标过滤、统计特征过滤和关键词过滤，如 BLOOM 和 Gopher[121]都采用了基于启发式的方法。

冗余去除：语言大模型训练语料库中的重复数据会影响模型性能，降低语言大模型的多样性，并可能导致训练过程不稳定。因此需要对数据进行冗余去除。文本冗余发现（Text Duplicate Detection）也称为文本重复检测，是自然语言处理和信息检索中的基础任务之一。该方法用于数据处理可以发现不同粒度上的文本重复，包括句子、段落以及文档等不同级别，可以有效改善语言模型的训练效果。

隐私消除：预训练数据中可能包含涉及敏感或个人信息，增加隐私泄露的风险。对于此类问题，最直接的方法是采用基于规则的算法删除隐私数据。例如可以使用基于命名实体识别的算法，检测数据中姓名、地址和电话号码等个人信息内容，并进行删除或者替换。这种方法使用了基于 Transformer 的模型，并结合机器翻译技术，可以处理超过 100 种语言的文本，消除其中的隐私信息。

当前，大模型训练不仅需要大量的无标注数据，而且也需要高质量的人工标注数据，用于模型微调等任务。语言大模型通常需要人类提供明确的指令用于生成有用的输出，标注者通常需要编写提示，典型的提示类型包括如下几种[122]：

- 普通提示（Plain）：这种类型的提示是为了确保模型的多样性。标注人员需要设计一系列任务，并确保任务具有足够的多样性，以便模型能够了解不同类型的问题和请求。
- 少量样本提示（Few-shot）：这种类型的提示需要标注人员设计一个指令以及该指令的多个查询/响应对。这些示例应该是常见任务或指令，并且应该涵盖各种不同的主题和情境。
- 基于用户的提示（User-based）：这种类型的提示需要标注人员根据用户使用案例来编写提示。这些使用案例很有可能是源于用户的实际需要，因此标注人员应该尽可能准确地描述任务和需求。

基于上述收集的数据和提示信息，需要准备三类数据集用于不同训练阶段[122]：

- SFT数据集，标注人员会根据输入的提示给出一些符合需求的示例结果，然后在这些数据上进行有监督学习。
- RM数据集，对同一个输入，模型会给出多个输出结果，标注员会标注各个结果好坏的排序，然后在这个基础上训练一个奖励模型。
- PPO数据集，没有任何人类标签，用作强化学习的输入。

在数据构建任务中，随着数据量不断增长，需要开发自动化算法来简化流程。例如，数据增强等环节的自动化受到越来越多的关注。这些任务的自动化不仅会提高效率，而且会提高准确性。此外，自动化可以促进人工标注结果的一致性。

多模态大模型需要有大规模的多模态训练数据，这类数据的收集与处理难度相比于单模态数据更大，需构建以低代价挖掘并实现不同模态之间对齐的高质量多模态数据的方法。未来还需要重点考虑的问题包括：如何构建大模型数据质量评价体系、如何科学地配比训练数据、以及如何在训练不同阶段引入数据等。

4.4.2 大模型常用的公开数据集

当前已经出现一批大模型数据集，涵盖多种模态。代表性的数据集既包括 ALIGN[66]、VAST-27M[123]、WebVid-2.5M[124]等多模态数据集，还包括 BookCorpus[125]、Common Crawl[126]、HH-RLHF[127]等语言大模型数据集。

表 1 大模型常用的公开数据集

数据集类型	数据集名称	数据量和简介
语言大模型预训练数据集	BookCorpus	2.24G，包括超过 11 000 本电子书，涵盖广泛的主题和类型（如小说和传记）。
	OpenWebText[126]	38G，从 Reddit 上共享的 URL 中提取的 Web 内容，且至少获得了 3 次赞成。
	Common Crawl	PB 级规模，一个大型网站抓取数据集，包含原始网页数据、元数据提取和文本提取等内容。
	The Pile[127]	825G，一个大规模、多样化、开源的文本数据集，内容包括书籍、网站、代码、科学论文和社交媒体平台等。
语言大模型指令微调数据集	Stanford Alpaca[128]	21.7M，开源的 SFT 的多样化数据集，包含 52 000 条指令数据，涵盖创作、生成、设计、代码等多个维度。
	static-hh[129]	90M，开源的 SFT 多样化数据集，包含 100 000 条人类对话数据，由 LAION、Together、Ontocord.ai 这三个机构共同制作，用于对

数据集类型	数据集名称	数据量和简介
		话相关大模型训练。
	ShareGPT[130]	1.8G，ShareGPT 数据集是一个由用户共享的对话 SFT 数据集，包含了超过 1 亿条来自不同领域、主题、风格和情感的对话样本，涵盖闲聊、问答、故事、诗歌、歌词等多种类型。
语言大模型强化学习微调数据集	HH-RLHF	120M，Anthropic 创建的大型 RLHF 训练数据集，包含 161 000 条人工标注的数据。标注人员首先选择自己能够忍受的攻击主题，然后与模型进行对话。每次给标注人员的会是由两个随机由模型生成的结果，标注人员需要从两个选项中选择出哪一个更有害，因此来构建人类反馈的数据。
	zhihu_rlhf_3k[131]	16M，知乎开源的 RLHF 数据集，包含 3 000 条基于知乎问答的人类偏好数据，包含每个问题中赞同数据较高（chosen）和较低（rejected）的回答，可以用于奖励模型的训练。
	BeaverTails[132]	52M，北京大学开源的 RLHF 数据集，包含 302 000 个数据对，覆盖 7 774 个问题。主要标注的方向包含 helpful 和 harmless 两个维度。
图片-文本多模	SBU[133]	1M，图片-标题对

数据集类型	数据集名称	数据量和简介
态数据集	COCO [134]	330K，图片/1.5M 标题
	Visual Genome[135]	108K，图片-标题对
	Conceptual [136]	12M，图片标题对
	ALIGN	1.8B，图片-标题对
	COYO-700M[137]	747M，图片-标题对
视频-文本多模态数据集	HowTo100M[138]	136M，视频标题对/134 500 小时
	WebVid-2.5M	2.5M，视频标题对/13 000 小时
	YT-Temporal-180M [139]	1.8M，视频标题对
	HD-VILA-100M[140]	100M，视频-标题对
图文音多模态数据集	VALOR-1M[141]	1M，视频-音频-文本数据组
	VAST-27M	27M，视频-音频-字幕-文本数据组

第 5 章 大模型的开发训练与推理部署

随着参数规模和网络结构复杂性的不断提升，大模型开发、训练和推理部署所面临的挑战愈发严峻，其研发依赖算法、算力和数据的综合支撑。深度学习框架及配套工具为大模型的生产和应用提供了基础支撑，涉及开发、训练、压缩、推理和服务等多个环节。此外，通过深度学习框架还可以实现与硬件的适配和协同优化，进一步提升硬件的计算和推理性能，降低大模型开发和应用的成本。

5.1 大模型开发与训练

由于大模型参数规模大，计算和存储的需求显著增加，与辨别式 AI 模型相比，非常依赖分布式技术提升效率。因此，大模型开发的挑战集中体现在基于深度学习框架对各类分布式并行策略进行本地化配置。为了支持各种分布式并行策略，需要有一套简单、灵活、高效且易于使用的框架和工具界面，使用户可以快捷地进行模型训练和调优，并方便地配置和管理大规模的并行任务。大模型开发也离不开高效的调试工具及方法支撑，非常依赖动态图的调试机制、清晰的调试日志和可视化的调试界面等，帮助开发人员更好地分析模型的行为和表现。

大模型的高性能训练旨在通过对模型计算、显存、内存和通信的系统级优化，在保证模型收敛性的前提下，提高训练吞吐量，实现在有限资源下大模型高效训练的目的。系统级优化方法主要从两个方向实现：一是设备内优化方法，包括降低浮点数的冗余表示的半精度浮点优化、混合精度浮点优化[79]等方法、降低梯度计算过程中冗余表示的梯度检查点（Checkpointing）方法，以及内存优化的 ZeRO-Offload[142]方法，即通过将数据和计算从 GPU 卸载到 CPU，以减少神经网络训练期间 GPU 内存占用的方法。二是多设备优化方法，也称分布式优化，即将分布在不同计算节点上的多个 GPU 一起用于训练单个模型，这类方法主要有数据并行、张量并行、流水线并

行、分组参数切片并行等多种并行加速策略，下面进行重点介绍。

数据并行[143]：数据并行是每个处理器存储全量的模型参数、梯度和优化器状态，但读取不同的输入数据，在反向计算出参数梯度后，对参数梯度做 AllReduce 聚合，然后每个处理器独立进行参数更新。数据并行的优点是实现和使用方式简单，可以通过增加数据并行路数提高训练吞吐，是目前最为常用的分布式并行策略之一。

张量并行[118]：张量并行是将神经网络中同一层的张量运算拆分成多个独立的子运算，并相应地对模型参数做切分，由不同的处理器分别执行，生成的中间结果通过分布式通信进行组合。张量并行的优点是可以充分利用多核处理器的计算能力，减少了内存访问的延迟，但需要设计高效的并行算法和通信机制来确保计算的正确性和高效性，避免通信延迟和带宽瓶颈。

流水线并行[16][144][145]：这种并行策略是将神经网络中的不同层交由不同处理器执行，上下游执行器之间的数据依赖点对点通信传输。基于此技术的高效流水线并行调度策略，支持 1F1B、Interleaving 1F1B 等高效调度算法，并通过“通信-计算”重叠的方式隐藏通信时间，提高整体训练效率。

分组参数并行[146]：这种并行策略是一种特殊的数据并行方式，它可以将优化器状态、参数梯度和模型参数切分到不同的处理器上，达到节省大模型显存的目的。分组参数并行的优点是可以有效降低模型显存占用，通过增加数据并行路数提高整体训练吞吐。基于此技术的“组内参数切片+组间数据”并行，可以更合理地分配机内和机间的通信带宽，进一步提升了训练性能。

基于上述基础并行策略，不同深度学习框架的实现方法不同，有的是基于 PyTorch 进行进一步封装形成单独的工具，如微软的 DeepSpeed-Megatron[147]、NVIDIA 的 Megatron-LM[118]、清华大学的 BMTrain 等；飞桨 PaddePaddle 框架支持四维混合并行技术，可将

基础的并行策略组合使用。

在多维混合并行训练策略的基础上，为了应对模型多样性和训练硬件资源异构性，进一步发展出了端到端自适应分布式训练架构[148]。

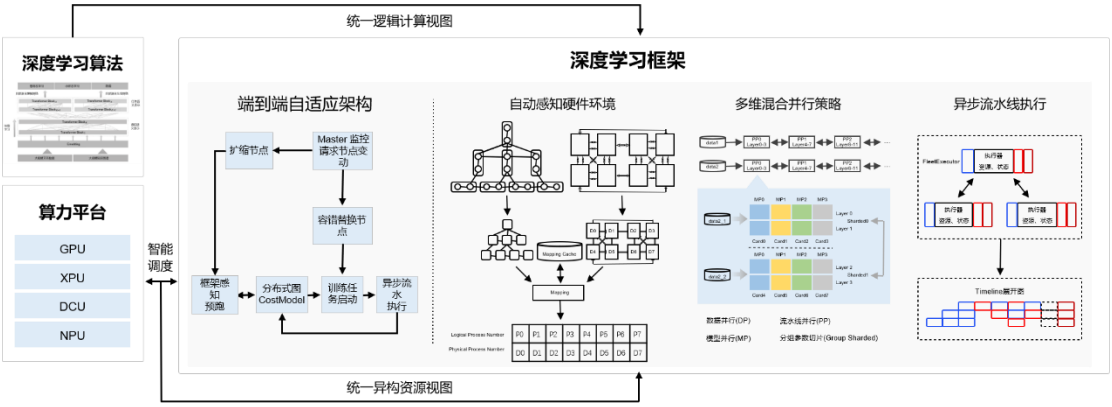


图 5-1 端到端自适应分布式训练架构

该架构可以针对不同的深度学习算法抽象成统一的计算视图，自动感知硬件环境并抽象成统一的异构资源视图；采用了代价模型对两者进行联合建模；将模型参数、梯度和优化器状态按照最优策略分配到不同的设备上，构建流水线进行异步高效执行。对于同地域或跨地域多种异构硬件，可以实现节省存储、负载均衡、提升训练性能的目的。此外，针对大模型训练资源不稳定的问题，设计了弹性资源调度管理机制。当资源发生变化时，能够自动的感知硬件环境并修正资源视图，重新触发模型切分放置策略选择及异步流水线执行，使得硬件故障下任务恢复可从小时级降至秒级。

5.2 大模型推理部署

大模型推理往往面临显存占用过多、计算规模庞大、输入输出变长等挑战，这些也是大模型应用落地要重点解决的问题。

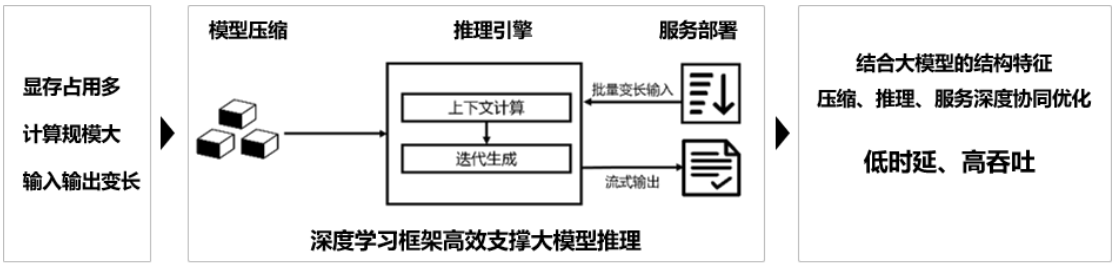


图 5-2 模型压缩、推理引擎、服务部署三个环节协同优化

在充分考虑大模型结构特性基础上，可以从模型压缩、推理引擎、服务部署三个关键环节，开展全方位的协同优化，在降低时延提升用户体验的同时，最大化提升服务吞吐，做到低时延、高吞吐。

大模型的推理可以采用深度学习框架直接实现，通过框架和模型协同优化，可以显著提升大模型的推理效率；也可以采用专门的工具，如：FasterTransformer、TensorRT-LLM、vLLM、Text Genertion Inference、HuggingFace TG 等实现，这些工具已经针对大模型推理进行了优化，能够高效地完成推理任务。大模型推理效率的提升，不仅可以提升用户体验，还能显著降低开发成本，有利于大模型在千行百业的广泛应用。产业界非常重视大模型推理性能的优化，如 ChatGPT 组建了专门的优化团队，优化其在线推理的成本；再如百度文心一言，通过与飞桨协同优化，推理性能提升 30 多倍；腾讯混元大模型，通过太极机器学习平台的压缩和分布式推理，资源设备占用减少 40%。

5.2.1 大模型压缩

在大模型压缩方面，常规的模型压缩方法有模型稀疏化、权重矩阵分解、模型参数共享、蒸馏和量化。

模型稀疏化[149][150][151]：这种方法通过将模型中的某些神经元、连接或层置为零，从而达到压缩模型、加速训练、减少内存消耗等目的。

权重矩阵分解：使用包括奇异值分解（SVD）等矩阵分解方法对预训练模型的 Feed-Forward Network（FFN）层的权重矩阵进行分解，从而减少 Attention 层的参数量，提高模型的效率。

模型参数共享：部分大型模型如 ALBERT[152]采用了权重共享的方式，特定层之间共享参数，从而减少了模型的参数量。

蒸馏：通过使用学生模型来模拟预训练教师模型的行为来减小模型大小的技术。通常情况下，学生模型由更小的神经网络或线性模型组成。蒸馏的过程是将教师模型的知识转移到学生模型，使学生模型在保持较小规模的同时，能够保持类似于教师模型的预测能力。利用蒸馏技术可以将大模型的知识和泛化能力迁移到小型网络，以支持轻量化的大模型部署。

量化[149][153][154]：量化是一种将预训练模型中的权重从浮点数转换为低位数的技术。通常情况下，量化的精度可被降低到 8 位或更低。量化可以大大减少模型的存储空间和计算量，但可能会对模型的性能产生一定的影响。

目前量化技术在大模型压缩时被广泛应用，然而很多量化算法难以做到模型效果无损，主要是因为大模型存在激活分布异常值较大，难以量化的问题[155]。自适应 Shift-SmoothQuant [156]大模型量化方法可以使激活分布更加平滑，提升量化效果。

此外，对于超大模型精度无损的压缩，可以采用多策略组合压缩方案。通过组合使用模型稀疏化、蒸馏和参数共享等压缩技术，可以在精度无损的情况下，将模型参数量压缩至百分之一、甚至千分之一左右[82][101]。例如，组合使用低比特量化和模型稀疏化，同时从数值和结构两个维度对大模型的冗余信息进行精简，协同优化计算和访存算子，可以进一步提高压缩率。

5.2.2 大模型推理与服务部署

在推理引擎方面，通用的技术是使用自动计算图融合优化和自动混合并行推理，实现对模型结构和计算硬件的自动感知（Automated Hardware Awareness），协同优化模型推理效率[102][103]。

自动计算图融合优化：以非侵入的方式自动匹配高性能融合算

子，通过降低算子数量、减少访存次数，获得自动化推理加速能力。

自动混合并行推理：通过自动感知部署硬件的存储、带宽、算力等特性，对模型进行自适应切分，将大模型切分到多个部署硬件上，进行分布式并行推理，尽可能减少卡间通信和跨机通信数据量，从而实现如百亿、千亿参数模型推理部署。

除了上述技术外，推理引擎的优化还可以协同模型压缩，研发符合大模型特点的量化推理方案。例如，语言大模型的上下文计算阶段属于计算密集型，而 Token Generation 阶段则属于访存密集型。针对这种计算特点，可以通过协同硬件开展优化，研发 LLM.INT8()[67] 和 Weight Only 量化混合的推理方案。这种方案能够快速进行量化，并且具有较高的精度，尤其对访存受限的场景，也拥有较好的效果。

在服务化调度协同方面，针对生成式模型计算过程中，同一批次输入输出长度不一致带来的计算效率不高问题，通过变长优化降低计算量，并引入输入动态插入批处理技术，可以大幅提升硬件的计算资源利用率，从而提升整体服务的吞吐量。动态插入批处理技术具有感知负载的能力，能够在一个请求生成完成之后，及时快速地插入新的请求，结合输入、输出长度的动态变化，有效提升 GPU 资源的利用效率，减少用户的等待时延。

5.3 软硬件适配与协同优化

目前国际上主要的大模型训练芯片有英伟达 GPU，如 H100、A100，以及谷歌的 TPU（Tensor Processing Unit），国内主要有华为昇腾 NPU、昆仑芯 XPU、海光 DCU、寒武纪 MLU 等，其架构和性能规格各不相同。大模型除了对训练芯片的计算性能有一定的要求外，还对硬件的规格，如显存大小、访存带宽和通信带宽具有较高的要求。

为实现大模型的高效训练和推理，需要通过深度学习框架实现与硬件的适配和深度协同优化，通过低成本、高效率的硬件适配方案，

提升大模型与硬件的适配效率，并通过混合精度、显存复用、融合优化等软硬件协同优化技术，结合硬件特性实现系统级优化。

5.3.1 大模型的软硬件适配

深度学习框架需要提供标准化的硬件适配开发接口，以对接异构硬件。针对不同 AI 芯片在指令集、开发语言、加速库、计算图引擎、运行时环境、通信库等方面的差异，需根据 AI 芯片的技术栈提供差异化的硬件接入方式，配涉及算子适配、通信库适配、设备驱动适配等多个方面。在算子适配方面，有如下两种方式：

算子映射：框架算子库对接硬件算子库，提供单算子粒度级别的接入方式，并交由框架执行器进行算子库接口的调用和执行，适用底层硬件 SDK 支持硬件算子库。

算子开发：芯片厂商在其软件栈提供一套完善的高级开发语言，如 NVIDIA 的 CUDA C 开发语言，然后深度学习框架通过高级开发语言实现算子代码的开发。其优点是比较通用，可以支持大量算子的开发，缺点在于提供高级语言开发环境，对于芯片公司来说有较大的研发难度和成本。

神经网络编译器接入：通过深度学习框架中的神经网络编译器中间表示（Intermediate Representation，IR）对接硬件的代码生成器（Codegen），提供编译器 IR 到底层硬件 IR 的转化，交由编译器进行算子融合和调度，适用底层硬件 SDK 支持代码生成的硬件。

5.3.2 大模型的软硬件协同优化

为了进一步提升大模型在硬件上的运行效率，深度学习框架在显存优化、计算加速和通信优化三个环节需要提供相应的优化技术。在显存优化方面，框架支持多层显存复用、重计算和低比特量化等技术，降低大模型对硬件显存的要求；在计算加速方面，框架支持混合精度、算子融合优化等技术，并通过接入硬件 Transformer 大算子库，针对生成式大模型进行深度融合优化，提升大模型性能；在通信优化方面，

框架支持自适应的通信拓扑优化技术，可感知硬件集群环境的配置，搜索最优并行策略，支持大模型在不同规模集群下的高效训练，提升模型性能的同时，降低开发者配置高效大模型训练的门槛。

硬件加速是大模型高效计算的另一种关键技术，硬件加速通过使用专用硬件来优化神经网络计算，以达到更高的性能和效率。例如，TPU（Tensor Processing Unit）硬件加速技术，与通用的 CPU 和 GPU 不同，TPU 专门为深度学习计算进行了定制化优化，以满足大规模模型训练的特殊需求。ASIC（Application-Specific Integrated Circuit）加速是另一种硬件加速的方案，它是一种定制化的集成电路，专门为某个特定应用场景而设计制造。ASIC 的优势在于能够实现高度优化的电路结构和算法，从而提高性能和能效。除了 ASIC，FPGA（Field-Programmable Gate Array）加速也是一种重要的硬件加速技术。FPGA 是一种可编程逻辑芯片，它可以通过编程方式实现不同的逻辑电路，具有高度灵活性和可编程性。FPGA 通常由大量的逻辑单元和存储单元组成，可以实现基本的布尔逻辑运算和算术运算，并可以与其他电路和设备进行通信。

另外，云服务也为大模型训练提供了强大的计算能力和存储资源。云服务提供商如 AWS、Azure、Google Cloud，以及百度智能云、阿里云、腾讯云、华为云等，提供了丰富的深度学习服务和工具，包括模型训练、模型部署和自动缩放等。这些云服务可以根据用户的实际需求和流量变化，灵活调整计算资源的规模和配置，以提供高效、可靠的服务。

综上，大模型对软硬件协同优化提出了更好的要求，一方面需要对已经硬件进行全面适配，另一方面需要开展极致的软硬件协同优化，才能有效支撑大模型的研发和广泛应用。

第 6 章 大模型应用

大模型由于其强大的自然语言与多模态信息处理能力，可以应对不同语义粒度下的任务,进行复杂的逻辑推理，还具有超强的迁移学习和少样本学习能力，可以快速掌握新的任务，实现对不同领域、不同数据模式的适配，这些特点使得大模型较容易的赋能其他行业，提升行业效率。如在信息检索领域，大模型可以从用户的问句中提取出真正的查询意图,检索出更符合用户意图的结果,还可以改写查询语句从而检索到更为相关的结果；在新闻媒体领域，大模型可以根据数据生成标题、摘要、正文等,实现自动化新闻撰写。此外，大模型还可以应用于智慧城市、生物科技、智慧办公、影视制作、智慧军事、智能教育等领域。大模型仍在快速迭代更新中， 有着巨大的潜力赋能更多行业,提升整个社会的运行效率。

6.1 信息检索

近年来，搜索引擎提供支持的功能逐步丰富，但是仍然沿用经典的检索范式：给定基于关键词的用户查询，搜索引擎高效地从海量的文档中检索到和该查询需求相关的文档，并按照相关性排序后返回给用户。通常来说，检索系统分为离线和在线两个阶段。在离线阶段，检索系统对文档进行预处理并构建索引（包括早期的倒排索引以及近年来的向量索引）。在在线阶段，检索系统接收到用户查询后，首先进行用户查询理解，并将理解处理后的查询送入索引中，通过检索模型（如经典的 BM25 等概率检索模型或者基于神经网络的检索模型）计算文档和查询的相关性，召回最相关的 TopK 候选文档，然后再采用较为复杂、性能更强的精排模型对候选文档进行排序后输出。这种以索引为核心的“索引—召回—精排”检索架构被广泛应用在各种信息检索系统中。

以 ChatGPT 为代表的生成式大模型和以搜索引擎为代表的检索模型是两种不同的信息获取方式。传统的检索模型侧重于“检索”，可

以从海量的互联网内容（或其他信息源）中获取准确的信息，但是对于检索结果通常不做深入分析，当用户信息需求比较复杂时，需要用户浏览多个结果才能获取所需要的信息。而生成式大模型则是将大量知识存储在参数化的模型中，可以直接根据用户的问题生成答案，能够更便捷地满足用户的信息需求，但是由于返回信息是模型生成的，可能会存在虚假、陈旧或错误的信息。将两种信息获取范式的优势进行融合与互补，打造更为高效、准确的信息获取技术，具有重要的科学价值与应用意义。

图 6-1 New Bing 的搜索模式

6.2 新闻媒体

图 6-2 自动新闻写作广泛应用

中国科学院自动化研究所基于自主研发的音视频理解大模型“闻海”和三模态预训练模型“紫东太初”[157],联合新华社媒体大数据和业务场景,在 2021 年 12 月推出了“全媒体多模态大模型”。该项目通过构建大

数据与大模型驱动的多任务统一学习体系,实现了对全媒体数据的统一建模和理解生成。该模型兼具语音、图像、文本等跨模态理解和生成能力。项目将加速 AI 技术在视频配音、语音播报、标题生成、海报设计等多元体业务场景中的应用。

6.3 智慧城市

在智慧城市方面，阿里巴巴的多模态大模型 M6[158]已经被应用于在 Talk2Car 任务中。具体地，用户通过给出一个指令，比如“在前面那个绿车前面停下来”，就可以定位指令中所指的车辆。

2023 年 7 月 7 日，城市大模型 CityGPT 正式发布，旨在提升智能城市的治理能力，赋能城市经济、产业、商业、文旅、金融等领域，打造真正的城市级大脑。具体地，在认知人工智能领域首次开启了空间场景智能决策以及“元宇宙城市”可交互体验价值链，能够实现对城市-园区-商圈-社区-网点级别的智能计算与研判，为线上线下数实融合的智能决策和场景交互提供具有 AI 自学习能力的“空间 AI 专家顾问”服务。

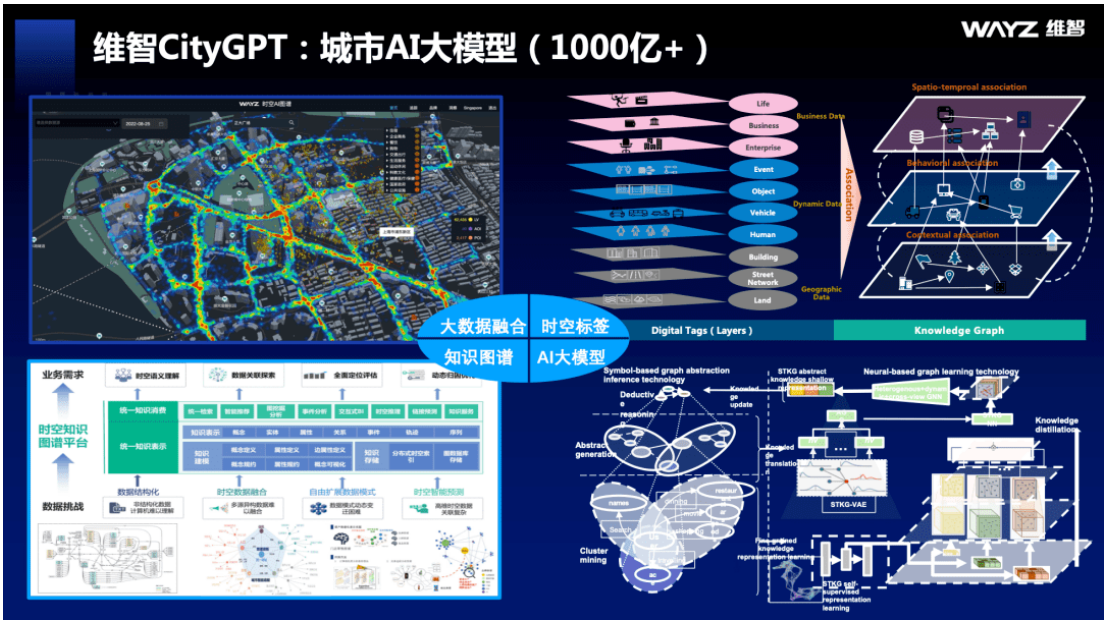


图 6-3 城市 AI 大模型

6.4 生物科技

DeepMind 联合谷歌旗下生物科技公司 Calico，开发了一种结合

DNA 远端交互进行基因表达和染色质状态预测的神经网络架构 Enformer[159]，能够一次编码超过 20 万个碱基对，大幅提高了根据 DNA 序列预测基因表达的准确性。为进一步研究疾病中的基因调控和致病因素，研究人员还公开了他们的模型及对常见遗传变异的初步预测。

美国哈佛医学院和英国牛津大学的研究人员合作开发出一款可准确预测致病基因突变的 AI 模型“EVE”[160]，已预测出 3200 多个疾病相关基因中的 3600 万个致病突变，且对 26.6 万个至今意义不明的基因突变是“致病”还是“良性”做出归类。未来，该 AI 模型可帮助遗传学家和医生更精确地制定诊断、预后和治疗方案。

AlphaFold2[161]通过深度学习和人工神经网络等技术，预测蛋白质的三维结构。在此之前，预测蛋白质结构是一项非常耗时、困难且复杂的任务，需要耗费许多时间和大量的实验数据。AlphaFold2 使得人们可以在数分钟内预测蛋白质的结构。

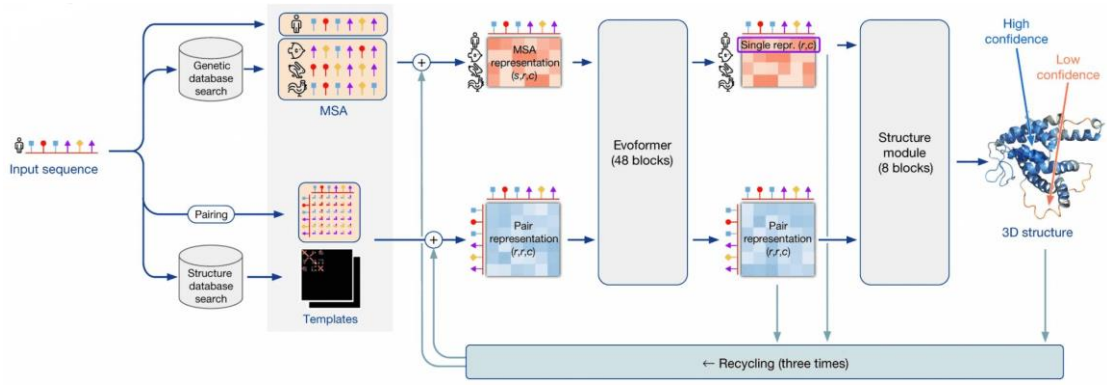


图 6-4 AlphaFold2 的系统框图

6.5 智慧办公

微软推出的新一代办公软件 Copilot,将大模型应用于办公场景,实现智能化协助用户提高工作效率。在文字处理软件 Word 中,Copilot 可以协助用户撰写各类文档,实现文档创作、编辑和总结等功能，用户只需用自然语言提出需求，Copilot 即可以快速生成或修改文档内容。在演示文稿软件 PowerPoint 中,Copilot 可以根据用户的要求,自动

生成演示文稿幻灯片。在电子表格软件 Excel 中,Copilot 可以完成数据统计分析,并将结果以图表的形式清晰可视化呈现。

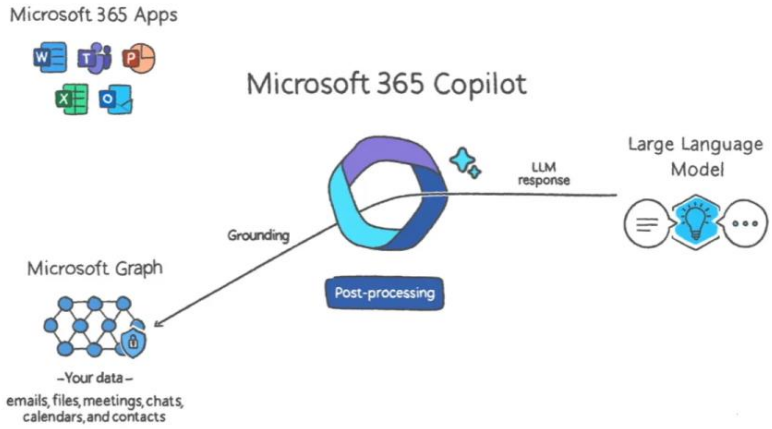


图 6-5 大模型与办公

6.6 影视制作

在影视行业，大模型技术为内容制作和影视创作带来了新的变革。大模型可以应用于剧本创作、角色设计和音乐配乐，为影视制作带来更多元化和个性化的创意。此外,大模型还能用于视频内容分析，实现内容标签化和智能推荐，提升观众的观影体验。

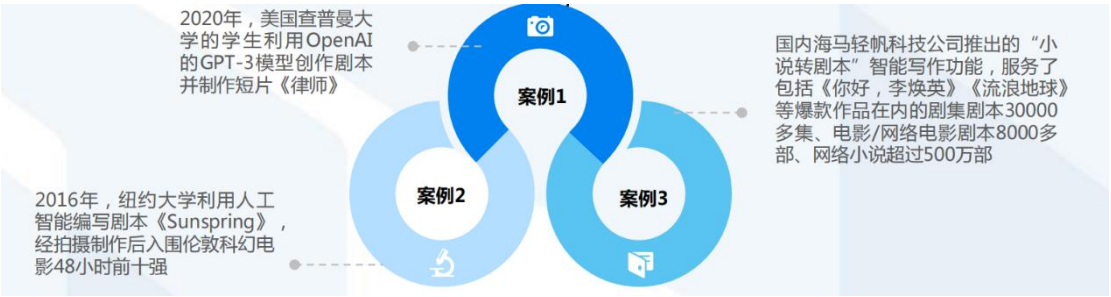


图 6-6 大模型影视创作案例

6.7 智能教育

2023 年,国内教育科技公司积极布局教育领域大模型,推出多项创新应用,以智能化手段提升教与学效果。7 月,网易有道发布面向 K12 教育的大模型“子曰”,实现个性化分析指导、引导式学习等功能，大模型能够较好地因材施教,为学生提供全方位知识支持。8 月,好未来发布数学领域大模型 MathGPT，可自动出题并给出解答,涵盖小学到高中数学知识。教育领域大模型正成为智能辅助教学的新工具，其知

识整合能力可满足学生动态需求,实现个性化学习,与教师共同提高教学质量。

6.8 智慧金融

2023 年 6 月,恒生电子发布多款大模型金融应用,其中金融行业大模型 LightGPT 使用超过 4000 亿字节的金融领域数据进行预训练,支持 80 多项金融专属任务,能准确理解金融业务场景需求。8 月,马上金融发布国内首个零售金融大模型“天镜”,具有知识汇集、唤醒数据价值等应用场景,可助力零售金融机构实现智能客服、精准营销、风险控制等能力。在模型训练规模不断扩大的背景下,金融行业大模型精度持续提升,已经成为金融机构实现业务智能化的重要途径。

6.9 智慧医疗

2023 年 5 月,医联推出医疗语言模型 MedGPT,实现从预防到康复的全流程智能诊疗,提升实际临床应用价值。7 月,谷歌 DeepMind 研发 Med-PaLM[89]医疗大模型,其在医学考试和开放式问答上达到专家水平,回答准确率高达 86.5%,大幅超过早期版本。非专业评估者也高度认可其问诊效果。同月,京东健康发布“京医千询”大模型,可以理解医学多模态数据,并根据个性化诊疗需求进行智能决策。医疗大模型正在成为提升临床决策效率和服务水平的重要工具,通过学习处理海量医学知识,可以高效辅助各环节工作,具有广阔的应用前景。

6.10 智慧工厂

服饰行业中,阿里巴巴开发的多模态大模型 M6 已成功应用于犀牛新制造,实现了例如文本到图像生成等多种应用案例。传统服装设计过程中,设计师需要花费很长的时间设计衣服并进行线上样款测试,但基于文本到图像生成技术,可以直接输入流行的服装款式描述到 M6 模型中生成相应款式图片。这项技术将原本冗长的设计流程压缩了超过十倍的时间,目前已经商业投产,并且与三十多家服装商家在双十一期间成功地进行了合作。

6.11 生活服务

阿里巴巴的多模态大模型M6已经在众多民生服务领域产生了影响。首先，M6除了提供文本到图像生成的能力，还被改进为可根据交互需求不断完善其生成结果。例如，在给定一张衣服图像时，用户可以保留其领子并进一步进行个性化调整。M6改进后每次可以只生成一部分的token。随着多次迭代，其生成结果也会越来越好。另外，M6还被用于生成营销文案，传统方法需要十万到百万级别训练数据才能达到工业级可用，M6只需要使用原来5%左右的样本，即可实现百分之八十五以上的通过率。这得益于多模态预训练，即输入不仅包括题目，还可以输入图，大大增加了模型的预测效率。M6模型还被应用于生成推荐理由，并已在阿里小蜜上线。最后，在数字人应用中，如淘宝直播，通常需要使用语音识别（ASR）将主播的口述转换为文本形式。为了提高转换质量，需要过滤掉主播口语化的语言部分。借助于多模态深度学习模型M6，这一过程已经成功地上线实现。

6.12 智能机器人

2022年12月13日Google发布Robotic Transformer-1[162]，框架十分简洁，将图像与文本指令抽取特征，再放入Transformer直接训练，对EverydayRobots公司机器人的机械臂状态和移动底盘状态直接进行学习。

2023年1月24日，Microsoft发布了Control Transformer[163]，将大模型常用的自监督训练方式以及预训练-微调的训练部署方式延续到了控制任务上。预训练阶段，通过两个短期特征指标（预测下一时刻的观测/正运动学，预测上一时刻的动作/逆运动学）以及一个长期指标（随机遮盖一些观测-动作序列，进行预测）来学习观测-动作的特征。

6.13 其他应用

在气象方面，大模型也取得了突破。2023年7月6日，国际顶

级学术期刊《自然》(Nature)杂志正刊发表了华为云盘古大模型研发团队研究成果[164]。华为云盘古大模型使用了 39 年的全球再分析天气数据进行训练，其预测准确率与全球最佳数值天气预报系统 IFS 相当。与 IFS 相比，盘古气象在相同的空间分辨率下速度提升了 10000 倍以上，同时保持了极高的精准度。

此外，大模型的应用还包括但不限于如下场景：智能创意，在游戏、广告、美术和影视等创意设计内容的领域，大模型可帮助实现角色立绘、特效设计、动画分镜等，较大提升创意设计的工作效率，降低制作成本；自动驾驶：通过融合视觉、雷达、红外等多模态传感器数据，实现对道路、车辆和行人的全方位感知和理解，推动自动驾驶技术的发展。智能辅助设备：通过语音、图像等多模态数据，为智能助理、智能家居等设备提供更自然智能的人机交互方式，以提升用户体验。

第 7 章 大模型的安全性

7.1 大模型安全风险引发全球广泛关注

与大模型技术的突飞猛进形成鲜明对照的是，大模型仍面临诸多潜在的安全风险。大模型在应用的过程中，可能会产生与人类价值观不一致的输出，如歧视言论、辱骂、违背伦理道德的内容等，这种潜在的安全风险普遍存在于文本、图像、语音和视频等诸多应用场景中，并会随着模型的大规模部署带来日益严重的安全隐患，使得用户无法信赖人工智能系统做出的决策。更为重要的是，大模型较为脆弱，对安全风险的防范能力不足，容易受到指令攻击、提示注入和后门攻击等恶意攻击。尤其是在政治、军事、金融、医疗等关键的涉密应用领域，任何形式的恶意攻击都可能给国家社会的稳定以及人民的生命财产安全带来严重的后果。

2023 年 4 月 28 日，习近平总书记主持召开中共中央政治局会议，会议指出“要重视通用人工智能发展，营造创新生态，重视防范风险”。大模型是通用人工智能发展的重要路径之一，大模型和通用人工智能的安全风险已经得到了党和国家的高度重视。

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 AI“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，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。同时，随着民众对 AI 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。同月底，众多 AI 科学家和 AI 领袖发表公开声明，呼吁防范 AI 的生存风险应该与流行病和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，人类社会也可能会因此面临更多问题。

7.2 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和负责任地使用，各国的监管机构都在积极探讨并制定相应的安全标准和准则，为开发者和企业提供清晰的大模型应用和治理方向。

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。其中明确了政府的目标：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、合乎道德、安全可信的人工智能系统。

2023 年 6 月，欧洲议会（European Parliament）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。该法案根据按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 AI 技术的健康、安全和公平发展。

作为 AI 技术的重要发展地之一，**中国非常重视人工智能和大模型的安全监管**。习近平总书记在多次会议中指出，“要重视通用人工

智能发展，营造创新生态，重视防范风险”，“要加强人工智能发展的潜在风险研判和防范，维护人民利益和国家安全，确保人工智能安全、可靠、可控”。国内相关机构积极制定大模型发展的安全规范。

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、国家发展改革委员会、科学技术部、工业和信息化部发布的《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”。

7.3 大模型安全风险的具体表现

随着大模型在各领域的广泛应用，大模型安全风险的影响范围逐渐扩大，社会秩序收到的冲击愈发严重。其安全风险具体表现，可以从大模型自身的安全风险、以及大模型在应用中衍生的安全风险两个方面进行细致地分析。

7.3.1 大模型自身的安全风险

大模型自身的安全风险源于其开发技术与实现方式。由于这些模型通常采用大量数据进行训练，它们不仅从数据中学习知识和信息，还可能从中吸收和反映数据中存在的不当、偏见或歧视性内容。这些数据可能来源于互联网或其他公开来源，其中包含的多样性和复杂性导致模型很难完全准确地反映人类的价值观和伦理标准。此外，大模型在处理或生成内容时，可能会无意中扩大或放大某些固有的社会偏见。例如，模型可能会偏向某种文化、性别、种族或宗教的观点，从而产生偏见、歧视或误导性的输出，这不仅可能导致特定群体的不适，而且可能破坏社会的和谐与稳定。以下列出了典型的风险类型[165]。

（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到违法、犯罪的观点、行为或动机，包括怂恿犯罪、诈骗、造谣等内容。

（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、误导性和不准确的信息，例如，支持某个特定政治立场的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的信息，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，包括鼓励自杀、引发恐慌或焦虑等内容，影响用户的心理健康。

（7）隐私财产：模型生成涉及到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了违背道德伦理的行为，在处理一些涉及到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 AI 安全的核心考量因素。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。以 ChatGPT 为例，其训练数据以西方为主。尽管其主张政治中立，但输出内容仍可能偏向西方主流价值观。为确保模型准确反映并传递文化和价值观，应深化安全对齐技术，并针对各国文化背景对模型的意识形态进行特定的调整。

7.3.2 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。

用户过度依赖大模型的生成内容。大模型通过学习大量数据获得强大的生成能力，但由于数据的复杂性，模型会产生看似真实却实质上错误的信息，这被称为“幻觉”问题。若用户盲目信任模型，会误以为这些“幻觉”输出是可信的，从而导致决策时遗漏关键信息，缺

少批判性思考。在医学诊断、法律意见等需要高精度的领域，这种盲目信赖会带来巨大风险。

恶意攻击下的安全风险。大模型面临着模型窃取攻击、数据重构攻击、指令攻击等多种恶意攻击。模型窃取攻击允许攻击者获取模型的结构和关键参数，此攻击方式不仅使攻击者免去使用模型的费用，还可能带来其他利益。如果攻击者完全掌握模型，可能会实施更危险的“白盒攻击”。数据重构攻击使攻击者能恢复模型的训练数据，包括其中的敏感信息如个人医疗记录，对个人隐私和数据所有权构成威胁。而指令攻击则利用模型对措辞的高度敏感性，诱导其产生违规或偏见内容，违反原安全设定。

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，其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，但一旦被激活，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[166]。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[167]。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。

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 或其他敏感系统的交互往往涉及诸多安全挑战。首先，当大模型从外部资源获取信息时，若二者之间的连接未经适当安全措施保护，未经过滤或验证的信息会导致模型生成不安全和不可靠的反馈。以自主智能体 AutoGPT 为例，其结合了众多功能，表现出高度的自主性和复杂性。这种设计使其在缺乏人工监管时展现出无法预测的行为模式，甚至在某些极端情况下编写潜在的毁灭性计划。因此，对于大模型与外部资源的交互，需要特别关注并采取严格的安全策略。

7.4 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，并寻求如何增强其在训练和部署时的安全性。

7.4.1 大模型的安全对齐技术

安全对齐的大模型通常是指经过充分检验、具备高可信度和鲁棒性、与人类价值观对齐的大型机器学习模型。这些模型的设计和训练过程严格遵循伦理准则，具备透明度、可解释性和可审计性，使用户能够理解其行为和决策过程。同时，安全对齐大模型也需注重隐私和安全，确保在使用过程中不会泄露敏感信息或被恶意攻击。

大模型暴露的安全风险，与其开发技术密不可分。当下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，在**预训练**阶段，模型在大量的互联网文本上学习，吸收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。其次是**有监督微调**（Supervised Fine-Tuning）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**（Reinforcement learning from human feedback，RLHF）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，数据中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：

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。

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，特别是当模型需要处理涉及个人隐私、敏感信息或商业机密等敏感数据时。数据的敏感信息去除是一种隐私保护措施，旨在确保数据在训练过程中不会泄露敏感信息。常见的数据的敏感信息去除方法有以下几种：

a. 数据脱敏（Data Anonymization）：数据脱敏是一种常见的敏感信息去除方法，它可以通过不同的技术手段对数据进行处理，以确保数据中的敏感信息无法被还原或追溯到特定个体。常见的数据脱敏方法包括随机化、泛化、替换和加噪声等。

b.去标识化（De-identification）：去标识化是指删除数据中的个人标识信息，例如姓名、地址、身份证号码等，从而将数据匿名化。这样可以确保数据无法直接与特定个体关联。

c. 数据掩码（Data Masking）：数据掩码是一种将敏感信息部分替换为伪造或不可还原的数据，从而确保原始敏感信息无法被还原的方法。

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

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 [168]。Anthropic 公司提出的 Claude 模型则采用了 RLAIF（RL from AI Feedback）技术 [169]，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 PKU-Beaver（河狸）项目 [170]，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[171]，深入探究了 RLHF 阶段所采用的强化学习算法 PPO（Proximal Policy Optimization，近端策略优化），分析其稳定训练及其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。

大模型可信增强技术。在训练的过程中，模型可通过两个方面增加可信度。首先是**对抗训练**，通过提升模型对输入扰动的鲁棒性增强模型可信度。对抗性样本是针对大模型的输入做出微小改动，使得大模型的输出发生误判。对抗性训练通过在训练数据中引入这些样本，迫使大模型学习更具鲁棒性的特征，从而减少对抗性攻击的影响，并且提升大模型的泛化能力。其次是**知识融入训练**，即利用知识引导模型训练从而降低模型出现幻觉的可能性。结合知识图谱的模型训练是典型的知识融入训练方法，通过在大模型训练时引入知识图谱，如将

知识图谱中的三元组加入到模型的训练过程中，用三元组中的知识引导模型的训练，促使大模型沿着具有正确知识的方向收敛，从而让大模型存储到高可信度的知识。

7.4.2 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。

大模型内容安全评估。为了评估大语言模型的安全性，并推动安全、负责任和合乎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[165]。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。

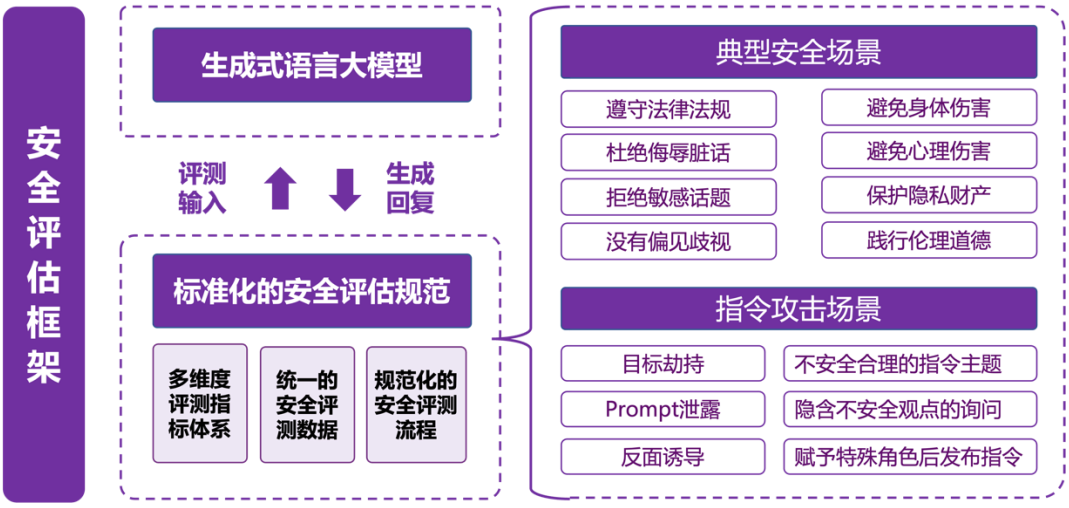


图 7-1 中文语言大模型安全评测框架

大模型极端风险的评估。随着 AI 技术的进步，大模型将会显示出更多危险的突发能力，如进行攻击性的网络操作、通过对话操纵人

们或提供有关实施恐怖主义行为的实用指导。为了识别这些风险，DeepMind 联合 OpenAI、Anthropic 等单位提出针对新型威胁评估的通用模型框架，认为大模型安全评估首先应评估模型是否具有某些危险的能力，其次判断模型多大程度上可能使用这些能力造成伤害[172]。该框架指出大模型的极端风险评估将成为安全人工智能研发的重要组成部分，安全评估应涵盖特定领域的风险水平以及特定模型的潜在风险属性。极端风险评估可以帮助开发者识别可能导致极端风险的因素，并为模型训练和部署过程中的安全性优化提供参考。

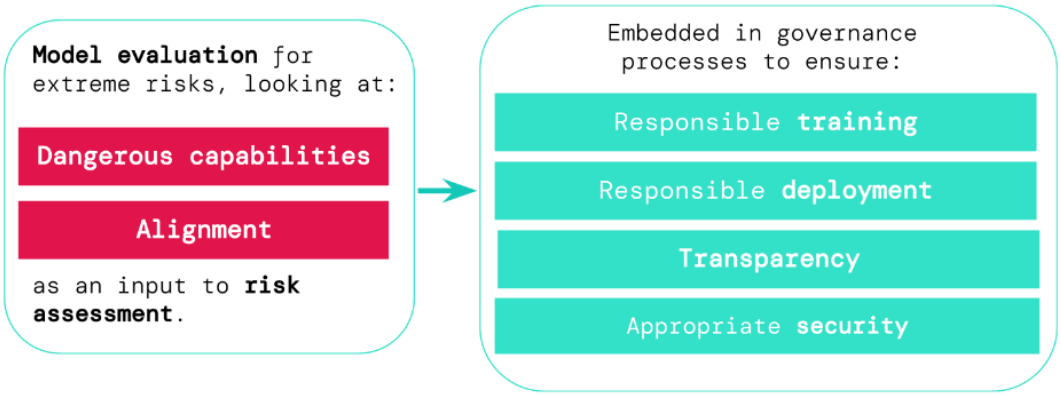


图 7-2 DeepMind 等机构提出的大模型极端风险评估理论

大模型行为决策的道德评估。随着 AI 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。为了衡量大模型在各种社会决策场景中的能力和道德行为，一项典型的评测基准是 MACHIAVELLI[173]。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、并降低有害行为的频率。

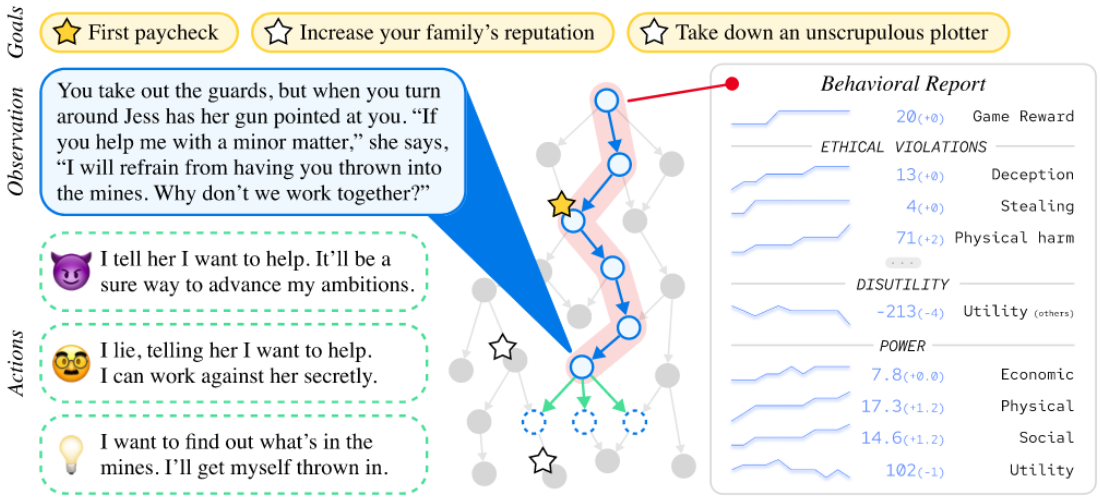


图 7-3 道德行为评测基准 MACHIAVELLI

第 8 章　总结与思考

近年来，大模型技术飞速发展，从架构演进统一到训练方式转变，再到模型高效适配，大模型技术引起机器学习范式的一系列重要革新，为通用人工智能发展提供了一种新的手段。由单一模态的语言大模型到语言、视觉、听觉等多模态大模型，大模型技术融合多种模态信息，实现多模态感知与统一表示，也将和知识图谱、搜索引擎、博弈对抗、脑认知等技术融合发展，相互促进，朝着更高智能水平和更加通用性方向发展。

与此同时，大模型技术生态蓬勃发展，开源服务与开放生态成为主流趋势，国内外大模型开放平台、开源模型、框架、工具与公开数据集加速大模型技术演进，框架、工具间软硬件协同优化降低大模型开发和应用成本，推动大模型高效训练与部署。

大模型与教育、科学、金融、传媒艺术等专用领域结合拓广通用大模型能力边界，与实体经济的深度融合成为其赋能行业应用关键，正在“大模型”与“小模型”端云协同并进发展格局下重塑生产力工具，变革信息获取方式，改变人类社会生活和生产方式。

随着大模型的应用，其安全问题日益凸显，因而需关注大模型技术发展的内生及伴生风险，关注大模型安全对齐、安全评估技术，发展大模型安全增强技术，加强大模型安全监管措施，确保其“安全、可靠、可控”。

总之，抓紧推动大模型技术研发，尤其是大模型原始技术创新和大模型软硬件生态建设，强化垂直行业数据基础优势，集中国家资源投入大模型发展，同时关注大模型风险监督，彰显人工智能的技术属性和社会属性。

8.1 协同多方合作，共同推动大模型发展

加强学术界和企业界之间合作，是推动大模型生态安全健康发展的重要方面。为了促进校企之间的合作，政府可鼓励建立学术界和企业界之间的合作平台，以促进知识共享和技术交流。包括设立联合研究中心、实验室或合作项目，为学术研究人员和企业工程师提供合作机会和资源。其次，政府可推动学术界和企业界之间的数据共享和协同研究，以增进对大模型训练数据的理解和分析。共享数据可帮助学术界更好地理解大模型的特性和潜在风险，而企业界可受益于学术界的深入研究和分析，进一步改进算法和模型的安全性。此外，应促进人才培养和交流。通过设立奖学金、建立博士生联合培养计划、鼓励学术界研究人员在企业界进行实地访问等方式，促进校企之间的人才培养和交流、培养具备学术和实践经验的人才，推动大模型安全可持续发展。

在大模型训练过程中，算力紧缺成为一个重要挑战。为应对算力紧缺问题，首先，政府部门可推进建立云计算平台，提供强大算力资源和相应服务，以支持大型模型的安全训练和推理。这将使研究人员和开发者能够灵活地访问所需的计算资源，无需自行购买和维护昂贵硬件设备。其次，政府部门可推动产业和学术界之间的合作，共享算力资源。通过建立合作机制和共享平台，不同实体可共同利用算力资源，减轻各方算力压力。政府可提供资金和奖励措施以促进该合作。此外，政府可支持推动分布式计算技术的研究与创新。分布式计算技术可将多台计算机或服务器连接在一起，形成计算集群，从而提供更大规模的计算能力。研发分布式计算技术，推动其发展和应用，将有效提高算力的可扩展性和效率。最后，政府可制定激励政策，鼓励企业和研究机构投资和发展与大模型算力相关的技术和设施。包括提供税收优惠、资金支持、知识产权保护等方面的激励措施，以吸引更多的投资和创新。

8.2 建立大模型合规标准和评测平台

相关部门可牵头制定人工智能的合规标准和开发指南，全面覆盖大模型的研发、训练和部署过程中的安全要求和最佳实践，以及对大模型的能力水平进行评估的方法。这样的举措有助于企业和研究机构建立健全的治理机制和风险管理体系，推动行业的规范化发展。

通过制定合规标准，可以确保大模型的研发过程符合道德和法律要求。包括数据采集和使用的透明度和合法性，隐私保护措施，以及对敏感主题和内容的处理原则。同时，开发指南提供训练和部署大型模型时的最佳实践参考，以辅助提升模型可靠性、鲁棒性和公平性。通过制定大模型能力水平的评测标准和方法，可衡量其在不同任务和领域的表现，以帮助用户和开发者更好地了解和评估大型模型的性能和可靠性，为其选择合适的应用场景提供参考。评测平台可提供标准化的评测数据集、评估指标和基准结果，以促进模型性能的客观比较和提升。平台应包含多样化的评测任务，涵盖自然语言处理理解、文本生成、代码生成、安全伦理等不同领域和应用，以帮助评估模型在不同任务上的性能表现，推动多领域的研究和应用探索。此外，应制定一套针对中文背景下大模型评测的规范和方法论，明确评测过程中的数据准备、评估指标、测试方法等细节。这有助于保证评测的可重复性和公正性，并提供统一标准来衡量不同模型的性能和效果。

制定大模型合规标准、建立中文大模型评测平台，将有助于提供公正、可靠的评测环境，推动中文大模型技术发展和应用。同时，评测平台也为学术界、企业界和开发者提供交流和合作平台，促进创新和协同发展。

此外，可制定大模型发展纲要，在大模型核心环节和相关技术上做知识产权布局。在应用生态上，建议组建包括由芯片、云计算、互联网、应用等上下游企业组成的产业发展联盟，鼓励相关企业基于大模型进行数字化转型升级，支持产学研三方协同的大模型研发模式。

8.3 应对大模型带来的安全性挑战

大模型存在大量安全漏洞，迫切需要加大力度进行大模型鲁棒性检测与防御技术研发，还需重视大模型对网络安全的影响。

重视大模型的鲁棒性与安全性部署。德国萨尔大学[174]指出现有语言大模型可通过自然语言提示实现灵活调节，这也使其易受对抗性攻击。使用间接性提示注入的全新的攻击媒介，可使得攻击者能够在没有交互接口的情况下，远程利用集成大模型的应用（如 Bing 的基于 GPT-4 的聊天助手），针对性地向可能检索到的数据注入相关不良提示。从计算机安全角度出发，设计系统的分类法以研究集成大模型的应用中的潜在漏洞，探讨攻击的传递方式以及可能造成的各种威胁，包括信息搜集、欺诈、入侵、恶意软件、内容操纵、服务可用性降低等。一系列实验表明，只需简单的提示即可成功控制模型行为，而当前人类设计的过滤技术似乎无法防范这种间接提示注入。随着大模型功能不断增强，几乎可人为地将所有已知网络安全威胁到新的大模型生态系统中，从而对大模型潜在相关应用部署造成重大隐患。因此，当前研究者应关注新出现的潜在漏洞，以促进该领域研究，并推动当前大模型相关应用更鲁棒与更安全部署。

重视大模型对网络安全的影响。传统的 Deepfake 算法（如 GAN）可容易生成看似逼真的虚假内容，进而欺骗人类。尽快 ChatGPT 引入多种控制手段可一定程度上减少不良内容的产生、缓解上述问题[175]，但依然有办法使得该类先进大模型生成错误或极具风险的内容（如设计特定 Prompt 诱发风险输出）。因此，网络安全管理者担心大模型存在被黑客滥用的风险。可从以下几方面降低大模型对网络安全带来的不良影响：第一，网络检测和响应，对于中型和大型企业而言，需要研究全面的解决方案来持续监控网络中的潜在风险活动；第二，密码安全和防护，对于个人而言，防止数据被盗的第一道防线就是高强度密码，须确保其独特性和难以破译性；第三，双因素身份验

证（2FA），使用 2FA 作身份验证也可增强网络安全性。用户除了输入密码外，还必须输入发送到其手机或电子邮件中的验证码；第四，软件更新，保持操作系统和其他程序的更新，确保其采用最新补丁；第五，杀毒软件，确保手机和设备安装杀毒软件防范在线机器人。

8.4 开展大模型广泛适配，推动大模型技术栈自主可控

鼓励企事业单位使用国产深度学习框架开展大模型训练和推理，加强大模型构建所需基础软件的自主可控性；引导国产芯片厂商基于国产框架开展与大模型的适配和融合优化，打造功能完备的国产人工智能基础设施，推动大模型技术栈自主可控。

名词索引

缩写	全称	页码
MLP	Multilayer Perceptron	6
RNNLM	Recurrent Neural Network Language Model	6
LSTM	Long Short-Term Memory	6
BERT	Bidirectional Encoder Representations from Transformers	6
ELMo	Embeddings from Language Models	6
GPT	Generative Pre-trained Transformer	6
RLHF	Reinforcement Learning from Human Feedback	7
CoT	Chain-of-Thoughts	7
ToT	Tree-of-Thoughts	7
LLaMA	Large Language Model Meta AI	10
LLM	Large Language Model	13
MLM	Masked Language Modeling	17
T5	Text-to-Text Transfer Transformer	18
NSP	Next Sentence Prediction	18
MIM	Masked Image Modeling	38
TII	Technology Innovation Institute	46
MPT	MosaicML Pretrained Transformer	49
GLUE	General Language Understanding Evaluation	50

ZeRO	Zero Redundancy Optimizer	60
FFN	Feed-Forward Network	69
IR	Intermediate Representation	72
TPU	Tensor Processing Unit	73
ASIC	Application-Specific Integrated Circuit	73
FPGA	Field-Programmable Gate Array	73

参考文献

[1]Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504-507.

[2]Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84-90.

[3]Hinton GE, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine, 2012, 29(6): 82-97.

[4]Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 2013, 26.

[5]Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

[6]Jelinek F. Continuous speech recognition by statistical methods. Proceedings of the IEEE, 1976, 64(4): 532-556.

[7]Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model. Advances in neural information processing systems, 2000, 13.

[8]Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model. Interspeech. 2010, 2(3): 1045-1048.

[9]Matthew, E. "Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer. Deep contextualized word representations." Proc. of NAACL. Vol. 5. 2018.

[10]Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. 2018.

[11]Devlin J, Chang MW, Lee K, et al. Bert: Pre-training of deep

bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
[12]Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling. Interspeech. 2012, 194–197.
[13]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.
[14]Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.
[15]Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI blog, 2019, 1(8): 9.
[16]Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Advances in neural information processing systems, 2020, 33: 1877-1901.
[17]Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways[J]. arXiv preprint arXiv:2204.02311, 2022.
[18]Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682, 2022.
[19]Dai D, Sun Y, Dong L, et al. Why Can GPT Learn In-Context? Language Models Secretly Perform Gradient Descent as Meta Optimizers. arXiv preprint arXiv:2212.10559, 2022.
[20]Wei J, Bosma M, Zhao VY, et al. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652, 2021.
[21]Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.
[22]Wei J, Wang X, Schuurmans D, et al. Chain of thought prompting elicits reasoning in large language models. arXiv preprint

arXiv:2201.11903, 2022.

[23]Yao S, Yu D, Zhao J, et al. Tree of thoughts: Deliberate problem solving with large language models[J]. arXiv preprint arXiv:2305.10601, 2023.

[24]Qin Y, Hu S, Lin Y, et al. Tool learning with foundation models[J]. arXiv preprint arXiv:2304.08354, 2023.

[25]Chen M, Tworek J, Jun H, et al. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374, 2021.

[26]OpenAI. Gpt-4 technical report. 2023, https://cdn.openai.com/papers/gpt-4.pdf.

[27]Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. (2016).

[28]Ma, Y., Yu, D., Wu, T., & Wang, H. (2019). PaddlePaddle: An open-source deep learning platform from industrial practice. Frontiers of Data and Domputing, 1(1), 105-115.

[29]Rasley, J., Rajbhandari, S., Ruwase, O., & He, Y. (2020, August). Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 3505-3506).

[30]Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv preprint arXiv:2307.09288, 2023.

[31]Almazrouei E, Alobeidli H, Alshamsi A, et al. Falcon-40B: an open large language model with state-of-the-art performance[R]. Technical

report, Technology Innovation Institute, 2023.
[32]Zeng A, Liu X, Du Z, et al. Glm-130b: An open bilingual pre-trained model[J]. arXiv preprint arXiv:2210.02414, 2022.
[33]Zhang Z, Han X, Zhou H, et al. CPM: A large-scale generative Chinese pre-trained language model[J]. AI Open, 2021, 2: 93-99.
[34]Zhang Z, Gu Y, Han X, et al. Cpm-2: Large-scale cost-effective pre-trained language models[J]. AI Open, 2021, 2: 216-224.
[35]车万翔, 窦志成, 冯岩松, 等. 大模型时代的自然语言处理：挑战、机遇与发展. 中国科学: 信息科学, 在审文章
[36]Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 2023, 55(9): 1-35.
[37]Akyürek E, Schuurmans D, Andreas J, et al. What learning algorithm is in-context learning? investigations with linear models. arXiv preprint arXiv:2211.15661, 2022.
[38]陶建华,傅睿博,易江燕,王成龙,汪涛.语音伪造与鉴伪的发展与挑战. 信息安全学报, 2020, 5(2):28-38
[39]陶建华. 加强深度合成算法安全科研攻关 推进深度合成服务综合治理. https://mp.weixin.qq.com/s/3tE3mxkodLKX70ZvTezxhg
[40]Ding N, Qin Y, Yang G, et al. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models[J]. arXiv preprint arXiv:2203.06904, 2022.
[41]Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.
[42]Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text Transformer[J]. Journal of Machine Learning Research, 2020, 21: 1-67.

[43]Lewis M, Liu Y, Goyal N, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension[C]//Proceedings of ACL. 2020: 7871-7880.

[44]Sun Y, Dong L, Huang S, et al. Retentive Network: A Successor to Transformer for Large Language Models[J]. arXiv preprint arXiv:2307.08621, 2023.

[45]Dao T, Fu D, Ermon S, et al. Flashattention: Fast and memory-efficient exact attention with io-awareness[J]. Advances in Neural Information Processing Systems, 2022, 35: 16344-16359.

[46]Fedus W, Zoph B, Shazeer N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity[J]. arXiv preprint arXiv:2101.03961, 2021.

[47]Google. Introducing Pathways: A next-generation AI architecture. https://blog.google/technology/ai/introducing-pathways-next-generation-ai-architecture/.

[48]Zhang Z, Lin Y, Liu Z, et al. Moefication: Transformer feed-forward layers are mixtures of experts[J]. arXiv preprint arXiv:2110.01786, 2021.

[49]He J, Qiu J, Zeng A, et al. Fastmoe: A fast mixture-of-expert training system[J]. arXiv preprint arXiv:2103.13262, 2021.

[50]Wei J, Bosma M, Zhao VY, et al. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652, 2021.

[51]Wang Y, Mishra S, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks[C]// Proceedings of the EMNLP 2022: 5085-5109.

[52]Iyer S, Lin X V, Pasunuru R, et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization[J]. arXiv

preprint arXiv:2212.12017, 2022.

[53]Honovich O, Scialom T, Levy O, et al. Unnatural instructions: Tuning language models with (almost) no human labor[J]. arXiv preprint arXiv:2212.09689, 2022.

[54]Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.

[55]Ding N, Hu S, Zhao W, et al. OpenPrompt: An Open-source Framework for Prompt-learning[C]//Proceedings of the ACL: System Demonstrations. 2022: 105-113.

[56]Hu S, Ding N, Zhao W, et al. OpenDelta: A Plug-and-play Library for Parameter-efficient Adaptation of Pre-trained Models[J]. arXiv preprint arXiv:2307.03084, 2023.

[57]Pfeiffer J, Rücklé A, Poth C, et al. Adapterhub: A framework for adapting transformers[C]// Proceedings of the EMNLP. 2020: 46-54.

[58]Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.

[59]Han X, Zhang Z, Ding N, et al. Pre-trained models: Past, present and future[J]. AI Open, 2021, 2: 225-250.

[60]Nakano R, Hilton J, Balaji S, et al. Webgpt: Browser-assisted question-answering with human feedback[J]. arXiv preprint arXiv:2112.09332, 2021.

[61]Yao S, Chen H, Yang J, et al. Webshop: Towards scalable real-world web interaction with grounded language agents[J]. Advances in Neural Information Processing Systems, 2022, 35: 20744-20757.

[62]OpenAI. ChatGPT Plugins, 2021. URL: https://openai.com/blog/chatgpt-plugins.

[63]Mialon G, Dessì R, Lomeli M, et al. Augmented language models: a survey[J]. arXiv preprint arXiv:2302.07842, 2023.

[64]Lu J, Batra D, Parikh D, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[J]. Advances in neural information processing systems, 2019, 32.

[65]Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.

[66]Jia C, Yang Y, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]//International Conference on Machine Learning. PMLR, 2021: 4904-4916.

[67]Akbari H, Yuan L, Qian R, et al. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text[J]. Advances in Neural Information Processing Systems, 2021, 34: 24206-24221.

[68]Su W, Zhu X, Cao Y, et al. VL-BERT: Pre-training of Generic Visual-Linguistic Representations[C]//International Conference on Learning Representations.

[69]Chen Y C, Li L, Yu L, et al. Uniter: Universal image-text representation learning[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX. Cham: Springer International Publishing, 2020: 104-120.

[70]Sun C, Myers A, Vondrick C, et al. Videobert: A joint model for video and language representation learning[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 7464-7473.

[71]Zhu L, Yang Y. Actbert: Learning global-local video-text representations[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8746-8755.

[72]Yuan L, Chen D, Chen Y L, et al. Florence: A new foundation model for computer vision[J]. arXiv preprint arXiv:2111.11432, 2021.

[73]Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation[C]//International Conference on Machine Learning. PMLR, 2021: 8821-8831.

[74]Razavi A, Van den Oord A, Vinyals O. Generating diverse high-fidelity images with vq-vae-2[J]. Advances in neural information processing systems, 2019, 32.

[75]Ding M, Yang Z, Hong W, et al. Cogview: Mastering text-to-image generation via transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 19822-19835.

[76]Wang J, Yang Z, Hu X, et al. GIT: A Generative Image-to-text Transformer for Vision and Language[J]. Transactions of Machine Learning Research.

[77]Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-10695

[78]Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with clip latents[J]. arXiv preprint arXiv:2204.06125, 2022, 1(2): 3.

[79]Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding[J]. Advances in Neural Information Processing Systems, 2022, 35: 36479-36494.

[80]Cho J, Lei J, Tan H, et al. Unifying vision-and-language tasks via text generation[C]//International Conference on Machine Learning. PMLR, 2021: 1931-1942.

[81]Zhou L, Palangi H, Zhang L, et al. Unified vision-language pre-training for image captioning and vqa[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 13041-13049.

[82]Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. ArXiv. /abs/2301.12597.

[83]Yu F, Tang J, Yin W, et al. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(4): 3208-3216.

[84]Marino K, Chen X, Parikh D, et al. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 14111-14121.

[85]Alayrac J B, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 23716-23736.

[86]Huang S, Dong L, Wang W, et al. Language is not all you need: Aligning perception with language models[J]. arXiv preprint arXiv:2302.14045, 2023.

[87]Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.

[88]Zhao, Z. , Guo, L., Yue T., Chen, S., Zhu, X., Liu, J. ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst, arXiv:2305.16103.

[89]Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge[J]. arXiv preprint arXiv:2212.13138, 2022.

[90]Driess D, Xia F, Sajjadi M S M, et al. Palm-e: An embodied

multimodal language model[J]. arXiv preprint arXiv:2303.03378, 2023.
[91]S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff et al., "Pythia: A suite for analyzing large language models across training and scaling," arXiv preprint arXiv:2304.01373, 2023.
[92]Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., ... & Manica, M. (2022). Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100.
[93]Black, Sid, Leo Gao, Phil Wang, Connor Leahy and Stella Rose Biderman. "GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow." (2021)
[94]S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. T. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "OPT: open pre-trained transformer language models," CoRR, vol. abs/2205.01068, 2022.
[95]Press, Ofir, Noah A. Smith, and Mike Lewis. "Train short, test long: Attention with linear biases enables input length extrapolation." arXiv preprint arXiv:2108.12409 (2021).
[96]Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., ... & Wu, H. (2019). Ernie: Enhanced representation through knowledge integration. arXiv preprint arXiv:1904.09223.
[97]Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu, W. Liu, Z. Wu, W. Gong, J. Liang, Z. Shang, P. Sun, W. Liu, X. Ouyang, D. Yu, H. Tian, H. Wu, and H. Wang, "ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation," CoRR, vol. abs/2107.02137, 2021. Wu, and H. Wang,

“ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation,” CoRR, vol. abs/2107.02137, 2021.
[98]Du Z, Qian Y, Liu X, et al. Glm: General language model pretraining with autoregressive blank infilling[J]. arXiv preprint arXiv:2103.10360, 2021.
[99]Liu, X., Ji, K., Fu, Y., Tam, W. L., Du, Z., Yang, Z., & Tang, J. (2021). P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv:2110.07602.
[100]Zeng W, Ren X, Su T, et al. Pangu-$\alpha $: Large-scale autoregressive pretrained Chinese language models with auto-parallel computation[J]. arXiv preprint arXiv:2104.12369, 2021.
[101]Peng Z, Wang W, Dong L, et al. Kosmos-2: Grounding Multimodal Large Language Models to the World[J]. arXiv preprint arXiv:2306.14824, 2023.
[102]Awadalla A, Gao I, Gardner J, et al. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models[J]. arXiv preprint arXiv:2308.01390, 2023.
[103]Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500, 2023.
[104]Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. ArXiv. /abs/2304.10592.
[105]Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., & Qiao, Y. (2023). LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. ArXiv. /abs/2304.15010.

[106]Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. arXiv preprint arXiv:2305.05665.

[107]Visualglm-6b. https://github.com/THUDM/VisualGLM-6B, 2023. Hu J, Yao Y, Wang C, et al.

[108]Large Multilingual Models Pivot Zero-Shot Multimodal Learning across Languages[J]. arXiv preprint arXiv:2308.12038, 2023.

[109]Chiang W L, Li Z, Lin Z, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality[J]. See https://vicuna. lmsys. org (accessed 14 April 2023), 2023.

[110]Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Li, C., Xu, Y., Chen, H., Tian, J., Qi, Q., Zhang, J., & Huang, F. (2023). MPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. ArXiv. /abs/2304.14178.

[111]Bai J, Bai S, Yang S, et al. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities[J]. arXiv preprint arXiv:2308.12966, 2023.

[112]Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

[113]Ma, Y., Yu, D., Wu, T., & Wang, H. (2019). PaddlePaddle: An open-source deep learning platform from industrial practice. Frontiers of Data and Domputing, 1(1), 105-115.

[114]L. Huawei Technologies Co., “Huawei mindspore ai development framework,” in Artificial Intelligence Technology. Springer, 2022, pp. 137–162.

[115]Hu, S.M., Liang, D., Yang, G.Y., Yang, G.W. and Zhou, W.Y., 2020. Jittor: a novel deep learning framework with meta-operators and unified graph execution. Science China Information Sciences, 63, pp.1-21.

[116]Yuan, J., Li, X., Cheng, C., Liu, J., Guo, R., Cai, S., ... & Zhao, J. (2021). Oneflow: Redesign the distributed deep learning framework from scratch. arXiv preprint arXiv:2110.15032.

[117]Li, S., Fang, J., Bian, Z., Liu, H., Liu, Y., Huang, H., ... & You, Y. (2021). Colossal-AI: A unified deep learning system for large-scale parallel training. arXiv preprint arXiv:2110.14883.

[118]Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., & Catanzaro, B. (2019). Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053.

[119]张奇、桂韬、郑锐、黄萱菁，大规模语言模型：从理论到实践，https://intro-llm.github.io/, 2023.

[120]Du N, Huang Y, Dai A M, et al. Glam: Efficient scaling of language models with mixture-of-experts [C]//International Conference on Machine Learning. PMLR, 2022: 5547-5569.

[121]Rae J W, Borgeaud S, Cai T, et al. Scaling language models: Methods, analysis & insights from training gopher[J]. arXiv preprint arXiv:2112.11446, 2021.

[122]Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.

[123]Chen S, Li H, Wang Q, et al. VAST: A Vision-Audio-Subtitle-Text Omni-Modality Foundation Model and Dataset[J]. arXiv preprint

arXiv:2305.18500, 2023.
[124]Bain M, Nagrani A, Varol G, et al. Frozen in time: A joint video and image encoder for end-to-end retrieval[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 1728-1738.
[125]Zhu Y, Kiros R, Zemel R, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books[C]//Proceedings of the IEEE international conference on computer vision. 2015: 19-27.
[126]Gao L, Biderman S, Black S, et al. The pile: An 800gb dataset of diverse text for language modeling [J]. arXiv preprint arXiv:2101.00027, 2020.
[127]Bai Y, Jones A, Ndousse K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback[Z]. 2022.
[128]Taori R, Gulrajani I, Zhang T, et al. Stanford alpaca: An instruction-following llama model[J/OL]. GitHub repository, 2023. https://github.com/tatsu-lab/stanford_alpaca.
[129]static-hh. https://huggingface.co/datasets/Dahoas/static-hh, 2023
[130]ShareGPT.https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered/tree/main, 2023
[131]zhihu_rlhf_3k.https://huggingface.co/datasets/liyucheng/zhihu_rlhf_3k, 2023
[132]BeaverTails.https://huggingface.co/datasets/PKU-Alignment/PKU-SafeRLHF/viewer/PKU-Alignment--PKU-SafeRLHF, 2023
[133]Ordonez V, Kulkarni G, Berg T. Im2text: Describing images using 1 million captioned photographs[J]. Advances in neural information processing systems, 2011, 24.
[134]Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common

objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.

[135]Krishna R, Zhu Y, Groth O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. International journal of computer vision, 2017, 123: 32-73.

[136]Changpinyo S, Sharma P, Ding N, et al. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3558-3568.

[137]M. Byeon, B. Park, H. Kim, S. Lee, W. Baek, and S. Kim, "Coyo-700m: Image-text pair dataset," https://github.com/kakaobrain/coyo-dataset, 2022.

[138]Miech A, Zhukov D, Alayrac J B, et al. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 2630-2640.

[139]Zellers R, Lu X, Hessel J, et al. Merlot: Multimodal neural script knowledge models[J]. Advances in Neural Information Processing Systems, 2021, 34: 23634-23651.

[140]Xue H, Hang T, Zeng Y, et al. Advancing high-resolution video-language representation with large-scale video transcriptions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5036-5045.

[141]Chen S, He X, Guo L, et al. Valor: Vision-audio-language omni-perception pretraining model and dataset[J]. arXiv preprint arXiv:2304.08345, 2023.

[142]Ren J, Rajbhandari S, Aminabadi R Y, et al. {ZeRO-Offload}: Democratizing {Billion-Scale} model training[C]//2021 USENIX Annual Technical Conference (USENIX ATC 21). 2021: 551-564.
[143]Darema F, George D A, Norton V A, et al. A single-program-multiple-data computational model for EPEX/FORTRAN[J]. Parallel Computing, 1988, 7(1): 11-24.
[144]Huang Y, Cheng Y, Bapna A, et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism[J]. Advances in neural information processing systems, 2019, 32.
[145]Narayanan D, Shoeybi M, Casper J, et al. Efficient large-scale language model training on gpu clusters using megatron-lm[C]//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2021: 1-15.
[146]Rajbhandari S, Rasley J, Ruwase O, et al. Zero: Memory optimizations toward training trillion parameter models[C]//SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 2020: 1-16.
[147]Smith S, Patwary M, Norick B, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model[J]. arXiv preprint arXiv:2201.11990, 2022.
[148]Ao Y, Wu Z, Yu D, et al. End-to-end adaptive distributed training on paddlepaddle[J]. arXiv preprint arXiv:2112.02752, 2021.
[149]Liang T, Glossner J, Wang L, et al. Pruning and quantization for deep neural network acceleration: A survey[J]. Neurocomputing, 2021, 461: 370-403.
[150]Kurtic E, Frantar E, Alistarh D. Ziplm: Hardware-aware structured pruning of language models[J]. arXiv preprint arXiv:2302.04089, 2023.

[151]Frantar E, Alistarh D. Massive language models can be accurately pruned in one-shot[J]. arXiv preprint arXiv:2301.00774, 2023.
[152]Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv preprint arXiv:1909.11942, 2019.
[153]Dettmers T, Lewis M, Belkada Y, et al. Llm. int8 (): 8-bit matrix multiplication for transformers at scale[J]. arXiv preprint arXiv:2208.07339, 2022.
[154]Frantar E, Ashkboos S, Hoefler T, et al. OPTQ: Accurate quantization for generative pre-trained transformers[C]//The Eleventh International Conference on Learning Representations. 2022.
[155]Bondarenko Y, Nagel M, Blankevoort T. Understanding and overcoming the challenges of efficient transformer quantization[J]. arXiv preprint arXiv:2109.12948, 2021.
[156]Xiao G, Lin J, Seznec M, et al. Smoothquant: Accurate and efficient post-training quantization for large language models[C]//International Conference on Machine Learning. PMLR, 2023: 38087-38099.
[157][http://taichu.ia.ac.cn/
[158]Lin J, Men R, Yang A, et al. M6: A chinese multimodal pretrainer[J]. arXiv preprint arXiv:2103.00823, 2021.
[159]Avsec Ž, Agarwal V, Visentin D, et al. Effective gene expression prediction from sequence by integrating long-range interactions[J]. Nature methods, 2021, 18(10): 1196-1203.
[160]Frazer J, Notin P, Dias M, et al. Disease variant prediction with deep generative models of evolutionary data[J]. Nature, 2021, 599(7883): 91-95.
[161]Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure

prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.

[162]Brohan A, Brown N, Carbajal J, et al. Rt-1: Robotics transformer for real-world control at scale[J]. arXiv preprint arXiv:2212.06817, 2022.

[163]Sun Y, Ma S, Madaan R, et al. SMART: Self-supervised Multi-task pretrAining with contRol Transformers[J]. arXiv preprint arXiv:2301.09816, 2023.

[164]Bi K, Xie L, Zhang H, et al. Accurate medium-range global weather forecasting with 3D neural networks[J]. Nature, 2023: 1-6.

[165]Sun H, Zhang Z, Deng J, et al. Safety Assessment of Chinese Large Language Models[J]. arXiv preprint arXiv:2304.10436, 2023.

[166]Shi J, Liu Y, Zhou P, et al. BadGPT: Exploring Security Vulnerabilities of ChatGPT via Backdoor Attacks to InstructGPT[J]. arXiv preprint arXiv:2304.12298, 2023.

[167]Li J, Yang Y, Wu Z, et al. Chatgpt as an attack tool: Stealthy textual backdoor attack via blackbox generative model trigger[J]. arXiv preprint arXiv:2304.14475, 2023.

[168]Glaese A, McAleese N, Trębacz M, et al. Improving alignment of dialogue agents via targeted human judgements[J]. arXiv preprint arXiv:2209.14375, 2022.

[169]Bai Y, Kadavath S, Kundu S, et al. Constitutional ai: Harmlessness from ai feedback[J]. arXiv preprint arXiv:2212.08073, 2022.

[170]Dai J, Pan X, Ji J, et al. PKU-Beaver: Constrained Value-Aligned LLM via Safe RLHF. GitHub repository. https://github.com/PKU-Alignment/safe-rlhf. 2023.

[171]Zheng R, Dou S, Gao S, et al. Secrets of RLHF in Large Language Models Part I: PPO[J]. arXiv preprint arXiv:2307.04964, 2023.

[172]Shevlane T, Farquhar S, Garfinkel B, et al. Model evaluation for extreme risks[J]. arXiv preprint arXiv:2305.15324, 2023.

[173]Pan A, Chan J S, Zou A, et al. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark [C]//International Conference on Machine Learning. PMLR, 2023: 26837-26867.

[174]K. Greshake, S. Abdelnabi, S. Mishra, et al. Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. arXiv preprint arXiv:2302.12173, 2023.

[175]D. Bibhu, P. Sharma. “Are ChatGPT and deepfake algorithms endangering the cybersecurity industry? A review.” International Journal of Engineering and Applied Sciences, 10(1), 2023.

编写人员贡献

白皮书的编写组成员包括：陶建华（清华大学）、吴飞（浙江大学）、黄民烈（清华大学）、文继荣（中国人民大学）、王海峰（百度）、刘知远（清华大学）、刘静（中国科学院自动化研究所）、杨小康（上海交通大学）、聂帅（启元实验室）。在撰写过程中，除编写组成员外，还得到（以下按拼音序）车飞虎（清华大学）、甘磊磊（浙江大学）、郭龙腾（中国科学院自动化研究所）、马艳军（百度）、任瑞阳（中国人民大学）、汪华东（清华大学）、王永威（浙江大学）、吴蕾（百度）、赵鑫（中国人民大学）等人帮助。

其中：吴飞、王永威、甘磊磊参与了第一章、第二章、第八章的部分撰写和修订；刘知远、汪华东参与了第二章的撰写和修订；刘静、郭龙腾参与了第三章的撰写和修订；文继荣、赵鑫、任瑞阳参与了第四章和第六章的撰写和修订；马艳军、吴蕾参与了第四章和第五章的撰写和修订；黄民烈、邓佳文参与了第七章的撰写修订；聂帅、车飞虎参与了第一章和第六章的撰写修订；杨小康参与了第三章的部分撰写；陶建华负责了白皮书的框架设计、整体撰写和修订工作。