第 1 章

ChatGPT 与 GPT 的发展

随着科技的不断发展，过去科幻作品中强大的人工智能（Artificial Intelligence，AI）正在逐步变为现实，AI 大模型也进入了人们的视野，其中最值得注意的就是 ChatGPT。作为人工智能生成内容（Artificial Intelligence Generated Content，AIGC）在自然语言处理（Natural Language Processing，NLP）领域的代表，它在文本创作、人机交互和科学研究等各个方面都展示了强大的语言理解和生成能力，引起了工业界和学术界的广泛关注。

ChatGPT、GPT 和大模型是什么？它们有什么样的关系？又应该如何理解和应用呢？本章将从 ChatGPT 出发，介绍其基本概念、技术体系和典型应用，到 GPT 的模型架构和发展历程，再到大模型及其研究现状，以便读者对 GPT 和大语言模型的整体概念和关系有全面和深入的了解。

1.1 ChatGPT开启人工智能新时代

1.1.1 ChatGPT 基本概念

2022 年 11 月 30 日，OpenAI 公司正式发布人工智能聊天机器人 ChatGPT，上线仅两个月其活跃用户数就破亿，成为历史上增长最快的消费者应用程序[1]。它强大的功能改变了许多人的工作和生活方式，在全世界掀起了 AI 新浪潮。ChatGPT 不仅展示了技术的进步，更开启了新的交互模式，使机器人与人类之间的对话变得前所未有的流畅和实用。如图 1–1 所示，ChatGPT 能够像人类一样自然地回答问题并进行自我介绍。

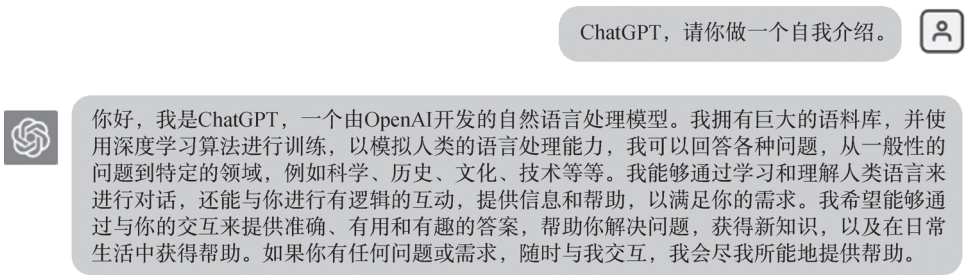


图1-1　ChatGPT的自我介绍

作为ChatGPT的开发者，OpenAI是一家颇具影响力的人工智能研究机构。OpenAI总部位于旧金山，由特斯拉的马斯克、Sam Altman 及其他投资者在 2015 年共同创立，目标是开发造福全人类的 AI 技术[2]。它在自然语言处理、机器学习、计算机视觉等领域均有突出贡献，GPT 系列模型是其核心研究成果。这一系列模型因其出色的内容生成能力被

部署于多种应用之中，给人们的生活带来了极大的改变。

ChatGPT 是一种基于 Transformer 架构的深度学习模型。Transformer 架构在 2017 年首次被提出，它采用了“自注意力机制”，该机制能有效处理序列数据，同时关注输入数据的所有位置，进行全局理解，通过并行化处理提高了训练效率和模型处理复杂文本的能力[3]。与传统依赖规则或简单机器学习技术的 NLP 算法相比，ChatGPT 能更好地理解复杂语言的结构和语境，在生成语言时表现出接近人类交流的水平。

GPT 模型的特点是“预训练”，即在执行特定任务前，已在大规模数据集上进行了训练，学习语言规律和知识，使模型具备强大的语言基础。这些数据集涵盖新闻、书籍、网站等。在此基础上利用机器学习，特别是深度学习分析大量数据，模型就可以自动生成文本、图像、音乐等内容。在开发 ChatGPT 时，OpenAI 着重关注对话质量和用户体验，如通过增强上下文理解能力，记住对话历史，生成连贯的响应。通过预训练和微调，ChatGPT 展示了在语言生成方面的强大能力，提供了高质量的交互体验，并在特定领域实现了不同的应用。例如编写详尽的用户手册、生成精确的产品描述，甚至帮助制定营销策略和执行计划。

随着 ChatGPT 的流行，各个行业、各个领域都开始思考如何利用 ChatGPT，并将其优势最大化。科技企业在思考如何掌握新一轮 AI 竞赛的主动权。教育工作者在思考当教材知识已经由语言模型掌握后，以应试为主的学校教育应该如何调整。而对于大部分转型中的非数字原生企业而言，则更倾向于思考如何真正使用AI技术提升管理效率和客户体验。

在客户服务领域，ChatGPT 被许多大型零售商和电信公司用于自动化处理常见查询并解决问题，显著提高了响应速度和客户满意度。在教育行业，ChatGPT 帮助教师提供个性化的学习材料，同时作为一个虚拟学习伴侣，与学生进行互动，帮助他们解答问题和学习新知识。在内容创作领域，ChatGPT 协助内容创作者快速生成文章初稿、创意文案和营销材料，极大地提升了生产效率和创新性。在软件开发方面，ChatGPT 通过提供代码建议和错误修正来帮助开发者提高编程效率。在医疗行业中，ChatGPT 通过提供医疗信息查询和病患教育服务，帮助医生和患者更有效地沟通。这些应用示例不仅证明了 ChatGPT 的技术成熟度，也彰显了其在各行各业的深远影响。

1.1.2　ChatGPT 技术体系

ChatGPT 基于 Transformer 架构，通过微调和人类反馈的强化学习进行训练，这种方法通过人类干预来增强机器学习以获得更好的效果。在训练过程中，人类训练者扮演着用户和人工智能助手的角色，并通过近端策略优化算法进行微调。由于 ChatGPT 强大的

性能和海量参数，它包含了更多不同种类的数据，能够处理更多的小众主题。ChatGPT现在可以进一步处理回答问题、撰写文章、文本摘要、语言翻译和生成计算机代码等任务。

ChatGPT的核心技术构建在三大支柱上：预训练、有监督微调（Supervised Fine Tuning，SFT）和基于人类反馈的强化学习（Reinforcement Learning from Human Feedback，RLHF）。这3个阶段共同构成了ChatGPT能够理解和生成语言并实现与人类自然“对话”的基础。

（1）预训练

ChatGPT最早基于OpenAI的GPT–3.5自回归语言模型，2023年已升级至最新的GPT–4架构，在理解和生成自然语言文本方面有了更为显著的改进，能够处理更复杂的任务和更广泛的主题。ChatGPT通过在大规模无标注语料数据上进行自监督预训练，使模型具备基本的语言理解和生成能力。预训练是构建ChatGPT的关键第一步，其核心任务是让模型通过大量文本学习语言的深层规律和结构。这一阶段，模型在没有特定任务的情况下进行训练，以捕捉语言的普遍性质和模式。通过从大量文本中提取语言规律，ChatGPT能够构建对语言的一般性理解。预训练的原理如图1–2所示。

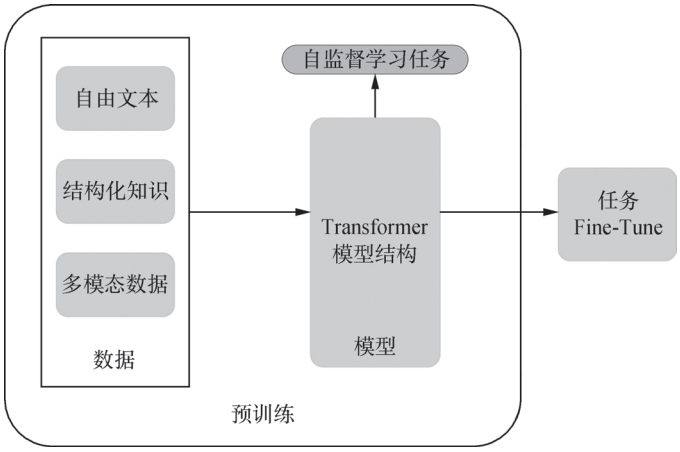


图1–2　预训练的原理

在预训练过程中，数据集被划分为训练集、验证集和测试集。数据集的质量和多样性直接影响模型的性能和适用范围。该阶段，GPT模型关注语言模式的两个方面：单词共现频率和语句结构。共现频率是指在特定语境中单词同时出现的频率，帮助模型理解词语关联性。语句结构学习涉及复杂的语法和句法分析，使模型学会组织单词、形成合理语句。

图1–3展示了目前主流自然语言预训练方法[4][5]，其中GPT系列模型采用了自回归语

言建模预训练方法，即根据前 $i-1$ 个单词预测第 i 个单词。自回归模型是生成式模型的一种特例，即基于前面若干个已生成值预测新目标值。自回归模型在时间序列分析、语音信号处理和自然语言处理等领域有广泛的应用。其中，在序列生成问题中，自回归模型特别重要，比如机器翻译、文本生成、语音合成等任务。

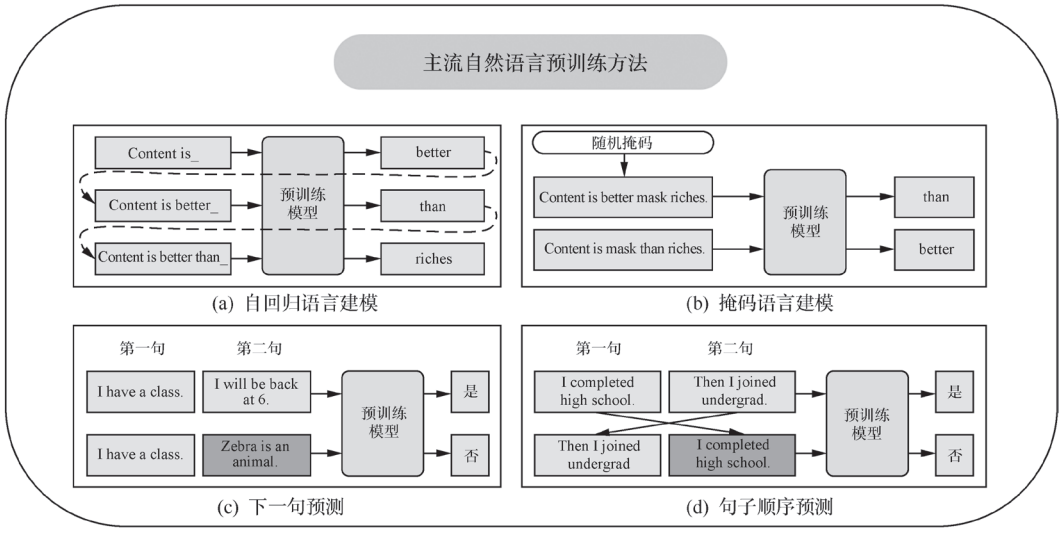


图1-3　主流自然语言预训练方法

具体来说，GPT 使用 Transformer 架构中的 Decoder 部分，在预测单词时，基于先前已生成的所有单词进行自回归计算，得到下一个单词的概率分布，并通过不断重复这一过程，生成连贯且符合上下文的文本。自回归任务在本质上与生成任务契合，使 GPT 系列模型在文本生成方面表现出色，能够处理从对话生成到文章写作等多种任务。

通过大规模文本数据的预训练，GPT 模型能够深层理解自然语言，包括词汇和句子的语义。这使得模型能生成语法结构准确、语义连贯的文本，并为后续特定任务学习打下基础。

（2）有监督微调

有监督微调是指在源数据集上预训练一个神经网络模型，即源模型，然后创建一个新的神经网络模型，即目标模型。目标模型复制了源模型上除了输出层外的所有模型设计及其参数。这些模型参数包含了源数据集上学习到的知识，并且这些知识同样适用于目标数据集。源模型的输出层与源数据集的标签紧密相关，因此在目标模型中不予采用[6]。微调时，为目标模型添加一个输出大小为目标数据集类别个数的输出层，并随机初始化

该层的模型参数。在目标数据集上训练目标模型时，将从头训练到输出层，其余层的参数都基于源模型的参数微调得到。

SFT 同样是 ChatGPT 训练中的关键阶段，能够确保模型在实际对话中提供恰当的回应。预训练结束后，当前的 ChatGPT 模型具有较强的语言生成能力，但难以理解人类输入不同指令的意图。因此，需要通过有监督微调引导模型按照人类意图进行答案的生成。具体而言，首先选取部分输入 Prompt，并人工为其构造符合人类意图的高质量答案。然后以上述数据中的 <Prompt，Answer> 模板对预训练模型进行精调，使模型初步具有理解人类意图、生成高质量答案的能力。

微调阶段将模型的通用性知识转化为专业性知识，学习在具体对话场景中根据用户意图提供反馈。与预训练数据集相比，微调数据集规模较小但专业且高质量，涵盖客户服务、技术支持等对话实例。这些数据由人工生成或抽取，并由专业人员标注，以确保模型学习不同请求的正确回应。微调过程中，需要调整模型权重以匹配特定场景的语言特征，使其识别用户意图，理解问题并生成相关回答。

图 1–4 表示的是将预训练模型的前 L–1 层的参数复制到微调模型，而微调模型的输出层参数随机初始化。在训练过程中，通过设置很小的学习率，从而达到微调的目的。

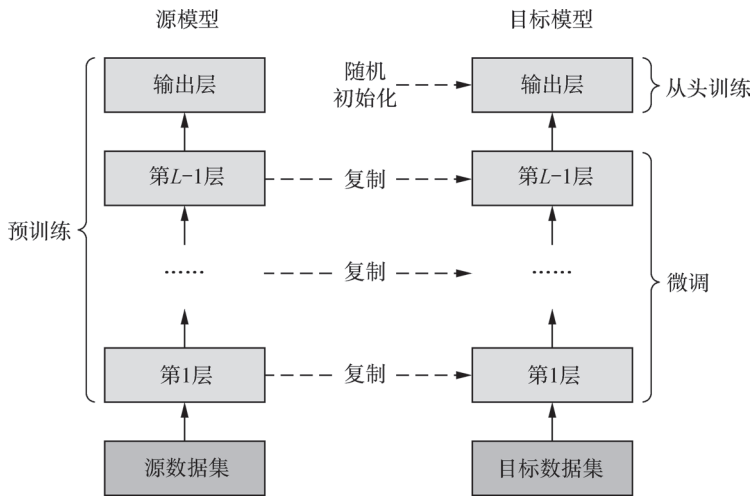


图1–4　有监督微调过程示例

微调技术的实施涉及选择损失函数、调整学习率及正则化技术。其中，损失函数衡量模型预测和实际标注之间的差距；学习率决定了模型权重调整的速度；正则化技术则能帮助模型保持泛化能力，避免在特定数据集上过度训练。通过有监督的微调，ChatGPT 提升

了特定领域的对话质量和信息准确性。微调后，模型需要通过自动评估（如准确率、召回率、F1 分数）和人工评估（用户体验和满意度）来确保表现符合标准。

（3）基于人类反馈的强化学习

基于人类反馈的强化学习（Reinforcement Learning from Human Feedback，RLHF）是基于人类对输出结果的反馈、对语言模型进行的强化学习。这一过程能够提升模型的对话质量，尤其注重使模型的回答更加符合人类的习惯[7]。如图 1–5 所示，评估式强化人工训练代理（Training an Agent Manually via Evaluative Reinforcement，TAMER）框架将人类引入到 Agent 的学习循环中，通过人类向 Agent 提供奖励反馈，指导 Agent 进行训练，从而快速达到训练任务的目标。RLHF 基于 TAMER 理念，结合强化学习算法和人类反馈，调整 Agent 的策略和奖励函数，实现更全面和长期的优化。RLHF 处理即时，反馈稀疏，适用于大规模复杂的任务，显著提升了 Agent 在复杂环境中的表现。

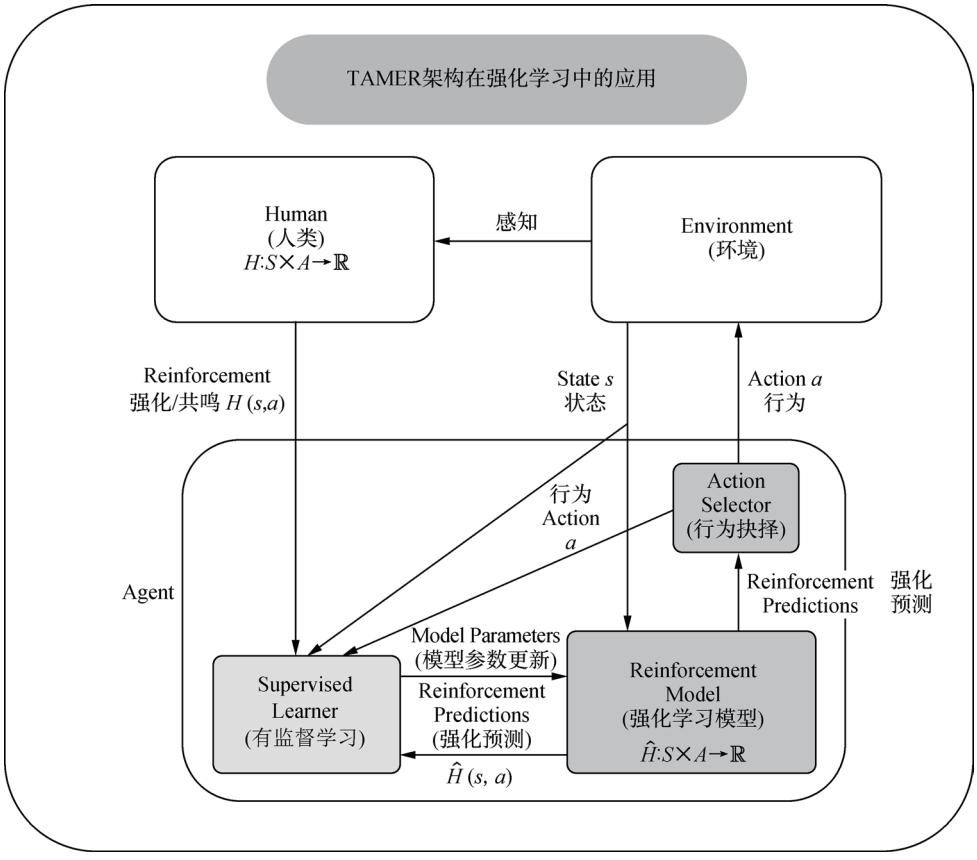


图1-5　TAMER架构在强化学习中的应用

在 RLHF 过程中，ChatGPT 在实际对话中尝试不同的回答方式，人类评估员根据回答的准确性、相关性和自然性提供反馈。这些反馈指导 ChatGPT 调整对话策略，提升对话质量。通过不断的试错和调整，ChatGPT 逐步优化其策略，实现更自然、更准确的对话交流。

总的来说，ChatGPT 的训练过程具体可以分为 3 个核心阶段，如图 1–6 所示。

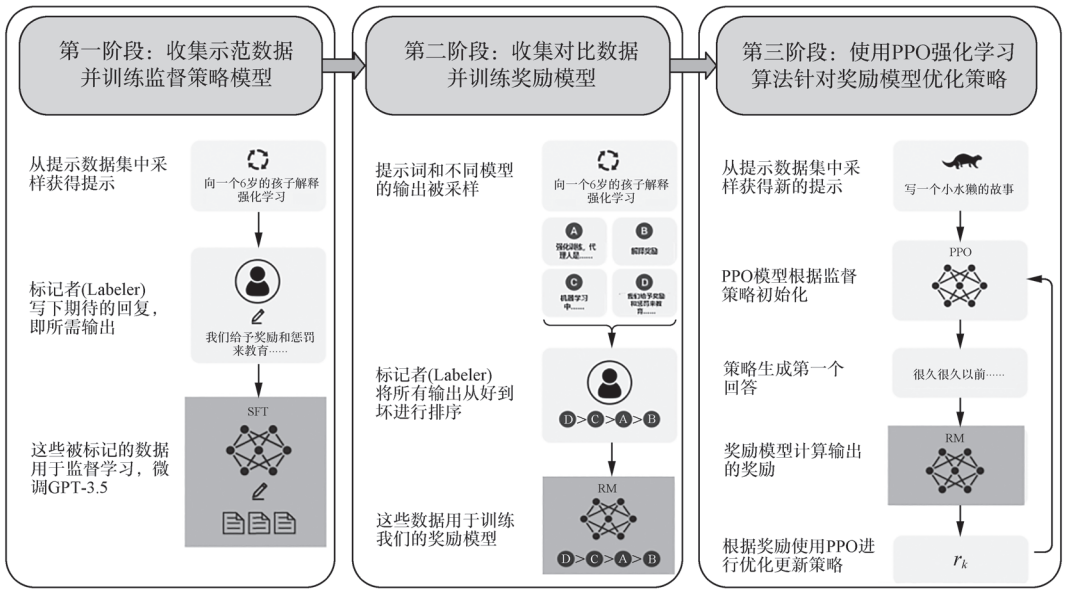


图1–6　ChatGPT训练过程

第一阶段：收集示范数据并训练监督策略模型。

为了让 GPT–3.5 初步具备理解指令的意图，研究人员首先会在数据集中随机抽取问题，由人类标注人员给出高质量的答案，然后用这些人工标注好的数据来微调 GPT–3.5 模型（获得 SFT 模型）。此时的 SFT 模型在遵循指令 / 对话方面已经优于 GPT–3，但不一定符合人类的偏好。

第二阶段：收集对比数据并训练奖励模型。

该阶段通过人工标注训练数据（约 33K 个）来训练回报模型。从数据集中随机抽取问题，使用生成模型生成多个回答，人类标注者对这些回答进行排名。然后，使用排序结果训练奖励模型。将多个排序结果两两组合，形成训练数据对，奖励模型（Reward Model，RM）接收一个输入并评价回答质量的分数[8]。这样，对于一对训练数据，可调节参数使得高质量回答的打分比低质量的打分要高。

第三阶段：使用 PPO 强化学习算法针对奖励模型优化策略。

近端策略优化（Proximal Policy Optimization，PPO）的核心思路在于将策略梯度中同策略的训练过程转化为异策略，即将在线学习转化为离线学习，这个转化过程称为重要性采样。PPO使用信任域优化方法来训练策略，这意味着它将策略的变化限制在与前一策略的一定范围内，以确保稳定性[9]。这一阶段利用第二阶段训练好的奖励模型，靠奖励打分来更新预训练模型的参数。在数据集中随机抽取问题，使用 PPO 模型生成回答，并用上一阶段训练好的 RM 模型给出质量分数。把回报分数依次传递，由此产生策略梯度，通过强化学习的方式更新 PPO 模型参数。

通过不断地重复第二和第三阶段进行迭代，就能够训练出更高质量的 ChatGPT 模型。

1.1.3　ChatGPT 典型应用

ChatGPT 的功能覆盖各个板块，应用场景也非常广泛，并有望持续拓展，具有很大的市场潜力。除了一些常见的应用，例如问答聊天、编写代码、撰写论文、图片处理等，ChatGPT 还被广泛应用于商业智能分析、教育辅助工具、游戏交互设计、在线购物体验优化及医疗诊断和健康管理等领域。这些具体应用不仅展示了 ChatGPT 在不同领域的强大功能，还凸显了其在推动各行业数字化转型中的重要作用。图 1-7 列举了“ChatGPT 能做的 49 件事”。

图1-7　ChatGPT能做的49件事

以下将分别介绍 ChatGPT 在商业办公、教育学习、游戏开发、电商购物和辅助医疗场景下的典型应用。

（1）商业办公

① 场景描述。在商业办公场景中，处理复杂的文字和统计图表是一个重要的步骤，尤其是财务行业。财务人员经常需要整理数据量大的报表，财务工作中常常会遇到一些复杂的财务问题，如税务政策解读、会计准则解释等。规划也是财务工作中的重要环节，它涉及资金的配置、投资决策等方面。最后还有报告的整理，这是商业工作中的重要成果之一，它需要准确地反映公司的财务状况和经营情况。

② 具体应用。ChatGPT 在财务工作中具有广泛的应用场景。它可以用于财务数据的处理和分析、解答财务问题、进行财务规划和预测，以及撰写财务报告等方面。使用 ChatGPT 可以提高财务工作的效率和准确性，使财务人员能够更好地发挥其专业能力和创造力。此外，商业办公还可以依靠 ChatGPT 智能问答功能的智能客服及利用 ChatGPT 编写销售文案等，例如依靠 ChatGPT 驱动的营销文案写作初创公司 Copy.ai，能在几秒钟内生成高质量的广告和营销文案。

结合大模型、Microsoft Graph 数据及 Microsoft 365 应用，微软于 2023 年 3 月 16 日发布全新 Microsoft 365 Copilot，将 GPT 的生成式 AI 能力与 Microsoft 365 应用中的数据相结合，全面应用于 Word、Excel、PowerPoint、Outlook、Teams 等办公套件，打破了传统的办公方式，能自动生成文档、电子邮件和 PPT 等，大大提升了商务办公人员的工作效率。Microsoft 365 Copilot 界面如图 1-8 所示。

图1-8　Microsoft 365 Copilot界面

（2）教育学习

① 场景描述。在中国的教育体系中，复习是其显著特征，课外补习班、高考的一轮二轮复习、考研培训班、火爆的刷题软件均是其体现，学生的负担一直比较重。然而，随着技术发展，由大模型通过预训练获得的知识已经覆盖了通用知识领域，使人们能够在任何时候都可以低成本地调用任何领域的知识。因此，教师和教培机构也需要调整产品设计，以适应这种变化。

② 具体应用。ChatGPT 不仅可以为学生服务，还可以用于教师减负和公司运营效率的提升，比如可汗学院和Coursera帮助教师使用提示生成课程材料；在数字化转型缓慢的日本，倍乐生早在2023年4月就宣布，在公司内部使用Azure OpenAI，以提升运营效率，并且其数字化转型部门已经开始观察和讨论 ChatGPT 的使用。GPT-4 深化语言学习软件 Duolingo 的对话功能，在 SuperDuolingo 的基础上，通过“角色扮演”和“解释答案”两大全新功能，打造了 Duolingo Max 协助语言教育产品，协助语言教育。Duolingo Max 界面如图 1-9 所示。

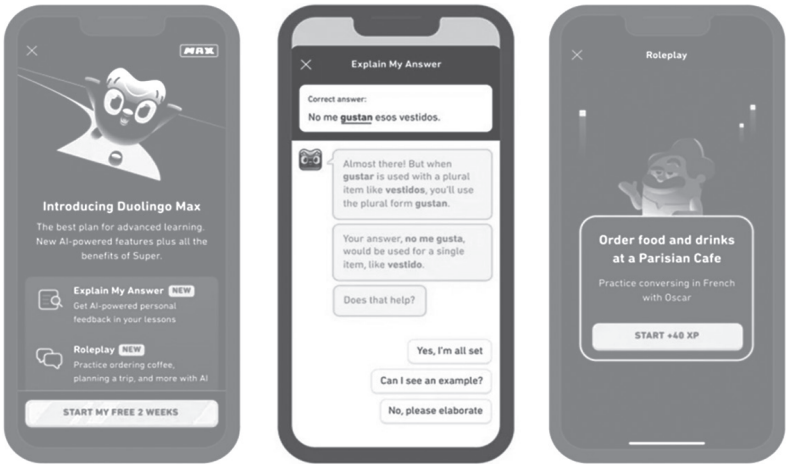


图1-9　Duolingo Max界面

（3）游戏开发

① 场景描述。游戏的开发制作是一个极其复杂的过程，其中每一个步骤都需要投入大量的人力和时间。随着人们更多地投入休闲娱乐，各类新型游戏和 VR 等技术的出现对游戏的要求也不断提高，使得游戏开发日益困难。首先，当代游戏拥有庞大的代码库，含有牵涉多种编程语言的数百万行代码，开发周期较长，成本过高，延误和超支的风险也成倍增加。这给开发人员带来了越来越大的压力，并导致了行业危机和职场倦怠。

② 具体应用。ChatGPT 可以帮助开发人员快速生成符合要求的代码，甚至可以编写小游戏。利用 ChatGPT 编写代码比人工编写速度更快、注释更清晰，理解更方便，并且没有冗余代码。不过现阶段还无法完全替代程序员，只能降低程序员大量重复的代码编写工作。例如 GitHub Copilot，是微软旗下代码托管平台 GitHub 推出的 AI 编程工具，主要定位是提供代码补全与建议。GitHub Copilot 界面如图 1–10 所示。

图1–10　GitHub Copilot界面

研究人员还开发出了一种名为 GameGPT 的模型，该模型可以整合多个 AI 智能体，自动完成游戏开发中的部分流程。这一突破性的技术为游戏开发带来了全新的可能。GameGPT 可以简化传统游戏开发流程中一些重复和死板的内容，比如代码测试。这样一来，大量开发人员就可以从繁杂的检验工作中解放出来，专注于 AI 所不能替代的、更有挑战性的设计环节，但该模型目前还处在概念形成阶段。

（4）电商购物

① 场景描述。随着信息技术的发展，电商行业获得了机会，更多消费者转向了线上购物，购物过程也发生了变化。过去人们需要反复挑选和对比才能订购，需要耗费大量的时间。但随着 ChatGPT 的出现，未来只需要告诉 ChatGPT 一声“请帮我买一套 ×× 元左右的适合明天户外骑行的装备”，它就能根据你的地址、预算、身高体重、购买记录、当地天气、送达时间等信息分析你的喜好和需求，简化购物流程，实现快速购物。

② 具体应用。Shopify 从 2022 年末开始引入大模型，并发布了 AI 工具包 Shopify Magic。Shopify Magic 包含商店搭建、营销、客服和后台管理等方面的 AI 能力，其中比较知名的

Sidekick，是一个基于 AI 的商业助手，能够通过自然语言回答商家提出的问题，帮助商家完成启动创意、分析业务发展等任务。另一重要能力“AI 导购”，也是基于 ChatGPT，为消费者提供消费决策信息。Shopify Magic 界面如图 1–11 所示。

图1–11　Shopify Magic界面

国内互联网公司中，阿里巴巴也推出了淘宝问问、通义万相、瓴羊 one 等基于大模型的产品，其中淘宝问问在 2023 年“双 11”预售日更新了带购物链接的版本。京东推出言犀电商大模型后，也推出了一系列云上 AIGC 产品，而百度也推出文心一言加持的百度优选，通过 AI 重拾电商梦想。

（5）辅助医疗

① 场景描述。在医疗行业中，医疗数据是涉及个人隐私的，因此医疗机构对于数据信息处理的要求也一直很严格。例如电子健康记录，它包含了患者所有医疗信息，如病史、诊断、治疗方案等。然而，电子健康记录通常包含大量的文本数据，医生需要花费大量的时间和精力来阅读和理解。一些医学影像的解读也需要专业的知识和丰富的经验，而且解读工作通常耗时较长。此外，在许多突发情况下或医疗资源紧张的地区，救援人员无法及时到达，需要进行远程医疗服务，因此对患者信息传达的准确性和实时性也有较高要求。

② 具体应用。国内外都在抢先落地医疗 AI 大模型的应用。2023 年上半年该领域就出现了两则重磅消息：一是，医联 MedGPT 与真人医生进行了多维度评测对比，且与三甲医院主治医生在比分结果上的一致性达到 96%；二是，谷歌 Med–PaLM 与临床医生进行医学问题回答测试，92.6% 的长篇答案符合科学共识，与临床医生生成的答案（92.9%）相当。

Be My Eyes 推出帮助盲人和视障人士的免费 App——Be My Eyes Virtual Volunteer。这是 GPT-4 助力的首款数字视觉助手，目标是为全球盲人和视障人士提供前所未有的可访问性和强大功能。用户可通过 App 将图像发送给 AI 驱动的虚拟志愿者，虚拟志愿者可回答有关该图像的任何问题，并提供即时视觉帮助，使得盲人和视障人士更好地驾驭物理环境、满足日常需求并获得更多的独立性。Be My Eyes Virtual Volunteer 界面如图 1-12 所示。

图1-12　Be My Eyes Virtual Volunteer界面

未来，ChatGPT 在金融、市场营销、教育等领域将会有更加广泛的应用。例如，金融机构可以通过 ChatGPT 实现金融资讯、金融产品的介绍，分析金融市场数据，帮助金融机构评估风险和预测市场趋势，提供决策支持；广告营销行业可以通过 ChatGPT 提供个性化推荐，分析用户的偏好、兴趣和历史数据，为用户定制个性化的推荐广告，提高广告点击率和转化率；银行、保险等公司的智能客服系统可以通过 ChatGPT 获取快速准确的客户支持和解答常见问题，同时为员工提供个性化的福利咨询和建议，提高员工满意度和福利使用效果。

总而言之，ChatGPT 的推出标志着 OpenAI 在 NLP 领域的重大突破，作为一项创新技术，其未来发展潜力巨大。ChatGPT 这样的 AI 大模型的出现，在向我们宣告一个事实：“AI 新时代已来。”其强大的模型容量、多样化的训练数据、上下文感知能力、自我学习能力及多语言支持能力都为用户提供了更优质的体验，这些创新特性也将推动人工智能技术的进一步发展。我们有理由相信，ChatGPT 及未来的 GPT 系列大模型将在各个领域带来更多的创新和应用，为人类的生活和工作带来更多的便利，创造更大的价值。

1.2　GPT引领人工智能发展热潮

1.2.1　GPT：生成式预训练转换器

GPT 的全称是 Generative Pre-trained Transformer，即生成式预训练转换器，其背景源于深度学习和 NLP 领域。在过去的几年里，随着计算能力的提升和大数据的出现，NLP 领域取得了突破性的进展。GPT 作为一系列 NLP 技术的集大成者，正是在这样的背景下应运而生的。

GPT 的含义其实就在它的 3 个首字母中，如图 1-13 所示。

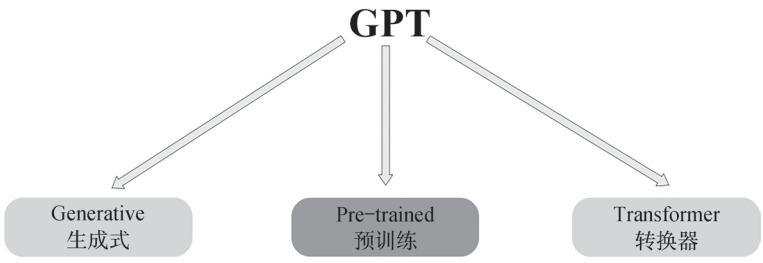


图1-13　GPT的含义

G：Generative，生成式。说明了 GPT 的能力是自发生成内容。GPT 不仅能够理解和解析给定的文本输入，而且能够基于这种理解产生全新的文本输出。这种能力使其能够执行多样的语言任务。

P：Pre-trained，预训练。说明了 GPT 模型在面对具体任务之前，已经在大规模文本数据集上学习到丰富的语言表示。这一阶段涉及的训练语料库广泛且多元，使模型能够捕捉到语言的统计规律和丰富多样的使用场景。预训练也让 GPT 模型在没有直接指导的情况下，通过自我学习，建立起对语言深层次的理解能力，这对于后续的任务特定微调至关重要。

T：Transformer，转换器。说明了 GPT 是基于 Transformer架构的语言模型。这一架构的创新之处在于自注意力机制，它支持模型在生成文本时评估输入序列中所有词元的相关性，不受传统序列模型限制。该机制适合处理依赖于长距离词元依赖关系的语言现象，使得模型在面对复杂的语言结构和细微的上下文变化时，能够灵活而准确地进行语言生成。

2017年，Google团队首次提出基于自注意力机制的Transformer模型，并将其应用于NLP。OpenAI应用了这项技术，于2018年发布了最早的一代大模型GPT–1，此后每一代GPT模型的参数量都呈爆炸式增长，2019年2月发布的GPT–2参数量为15亿，而2020年5月的GPT–3，参数量直接达到了1 750亿。

因此，ChatGPT的“一夜爆火”并不是偶然的，它是经过了很多人的努力，以及很长一段时间的技术演进得来的。自GPT–1发布以来，OpenAI系列模型每一代的性能都在不断实现突破和提高。

1.2.2 Transformer架构

Transformer架构是GPT的重要基础。Transformer是一种基于自注意力机制（Self–Attention Mechanism，SAM）的神经网络架构，广泛应用于自然语言处理领域的大模型中。谷歌和OpenAI在自然语言处理技术上的优化，都是基于这个模型。Transformer的核心部分是编码器和解码器，即Encoder和Decoder。Encoder把输入文本编码成一系列向量，Decoder则将这些向量逐一解码成输出文本。

在Transformer提出之前，自然语言处理领域的主流模型是循环神经网络（Recurrent Neural Network，RNN），使用递归和卷积神经网络进行语言序列转换。2017年6月，谷歌大脑团队在AI领域的顶级会议——神经信息处理系统大会（Conference and Workshop on Neural Information Processing Systems，NeurIPS）发表了一篇名为*Attention is All You Need*的论文，首次提出了一种新的网络架构，即Transformer，它完全基于自注意力机制，摒弃了循环递归和卷积[3]。

递归模型通常沿输入和输出序列的符号位置进行计算，来预测后面的值。但这种固有的顺序性质阻碍了训练样例内的并行化，这是因为内存约束限制了样例之间的批处理。而注意力机制允许对依赖项进行建模，无须考虑它们在输入或输出序列中的距离。Transformer避开了递归网络的模型体系结构，并且完全依赖于注意力机制来绘制输入和输出之间的全局依存关系。在8个P100 GPU上进行了仅12个小时的训练之后，Transformer就可以在翻译质量方面达到非常高的水平[3]，体现了很好的并行能力，成为当时最先进的语言模型。

Transformer网络结构如图1–14所示。Transformer是由一系列编码器（Encoder）和解码器（Decoder）组成，每一个编码器和解码器的内部结构均由多头注意力层和全连接前馈网络组成。虽然编码器和解码器都由多层组成，但其层结构不同，以适应各自功能，最终形成完整的网络结构。

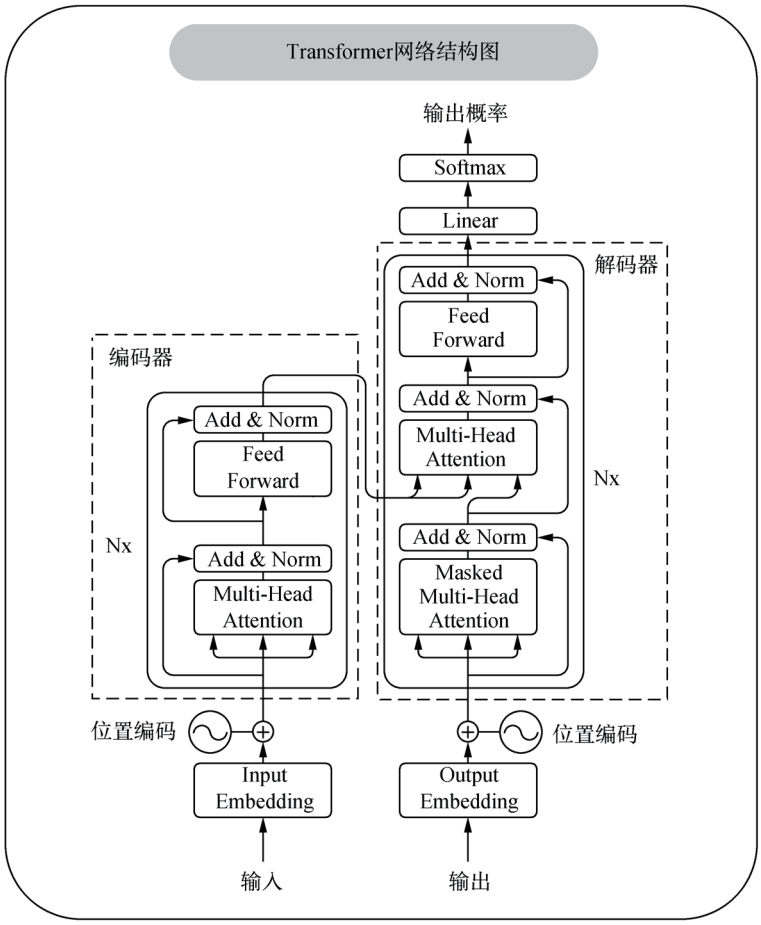


图1-14　Transformer网络结构图

编码器由 N 个编码器块串联组成，每个编码器块包含：

- 一个多头注意力层；
- 一个前馈全连接神经网络。

解码器也由 N 个解码器块串联组成，每个解码器块包含：

- 一个带掩码的多头注意力层；
- 一个接收编码器输出的多头注意力层；
- 一个前馈全连接神经网络。

除了GPT外，由谷歌在2018年开发的BERT也是最早一批出现的大模型[10]，同样基于Transformer架构。虽然BERT和GPT都是基于Transformer架构的大模型，两者也有所

区别。GPT 类似于 Transformer 的 Decoder 部分，而 BERT 的网络结构类似于 Transformer 的 Encoder 部分。

此外，BERT 是双向模型，而 GPT 是一个自回归模型。因此 BERT 模型会同时利用其前面和后面的文本信息，以更全面地理解该词语的语境，GPT 模型则利用以前生成的文本，来预测下一个词。如图 1–15 所示，图中的 Trm 代表 Transformer 层，E 代表 Token Embedding，即每一个输入的单词映射成的向量，T 代表模型输出的每个 Token（1 Token 约为 4 个英文字符）的特征向量[10]。

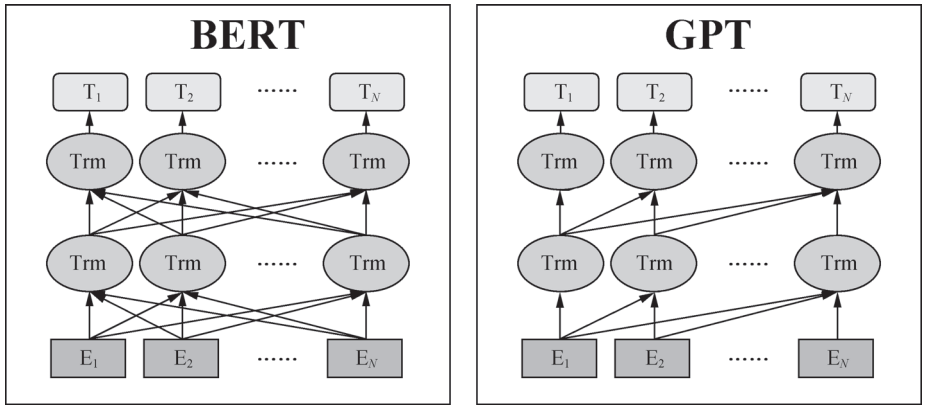


图1–15　BERT与GPT技术架构的区别

除了最基本的编码器和解码器网络结构，Transformer 还有许多重要的组成部分，通过这些内部的结构细节，Transformer 才能够完成一些复杂的任务，例如实现中英文翻译等。

（1）词嵌入向量和位置编码

输入序列（如文本或语音数据）进入 Transformer 模型时，首先通过嵌入层，将每个输入元素转换为密集的向量表示，捕捉其语义和句法特性，提供理解每个元素含义的必要信息。嵌入向量经过嵌入层处理后还需要加入位置信息，为此引入了位置编码来保留每个元素的位置信息。

我们使用词嵌入算法将每个词转换为一个词向量。输入句子的每一个单词的表示向量 $\boldsymbol{X}$，$\boldsymbol{X}$ 由单词的嵌入和单词位置的嵌入相加得到。如图 1–16 所示，“我有一只狗”这个句子中的单词分别被不同的词向量表示。

（2）自注意力机制

经过嵌入层和位置编码的输入数据将进入编码器层，可以得到句子所有单词的编码

信息矩阵。完成输入数据的向量化之后，模型结构的下一步就是自注意力层。

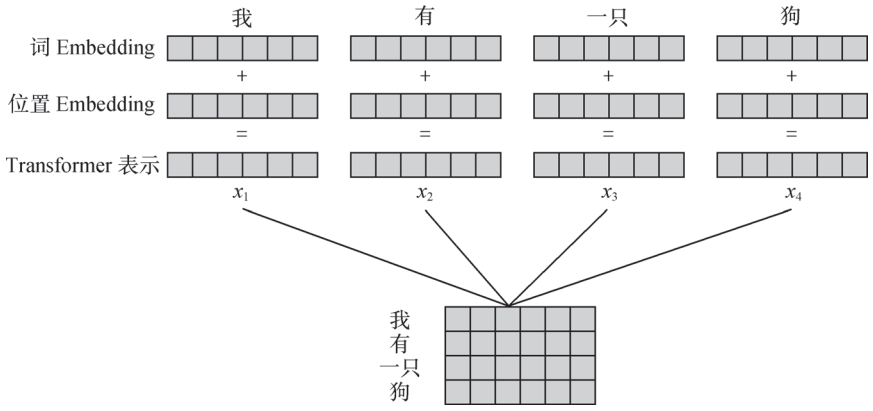


图1-16　输入数据的向量化表示

在自注意力机制中，模型生成查询（Query，$\boldsymbol{Q}$）、键（Key，$\boldsymbol{K}$）和值（Value，$\boldsymbol{V}$）3个向量。查询向量表示模型的注意力目标；键向量标识输入序列中的所有位置；值向量包含每个位置的实际信息。$\boldsymbol{Q}$，$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 均来自同一输入，通过公式将 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 之间两两计算相似度，依据相似度对各个 $\boldsymbol{V}$ 进行加权，就能得到注意力的计算结果。整个计算过程可以表示为

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}$$

例如，我们要将“The animal didn’t cross the street because it was too tired”翻译成中文，但句中的“it”所指代的是“animal”或“street”并不明确。对人来说，这是一个简单的问题，但是对算法来说却不那么简单。而自注意力机制使得模型不仅能够关注当前位置的词，而且能够关注句子中其他位置的词，从而可以更好地编码这个词。如图 1-17 所示，当对单词“it”进行编码时，有一部分注意力会集中在“The animal”上，并将它们的部分信息融入“it”的编码中。

（3）多头自注意力机制

多头自注意力机制是整个 Transformer 架构的核心，进一步完善了自注意力层的功能。每一组注意力用于将输入映射到不同的子表示空间，这使得模型可以在不同子表示空间中关注不同的位置。首先，通过 h 个不同的线性变换对 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 进行映射。然后，将不同的 Attention 拼接起来。最后，再进行一次线性变换。其基本结构如图 1-18 所示。

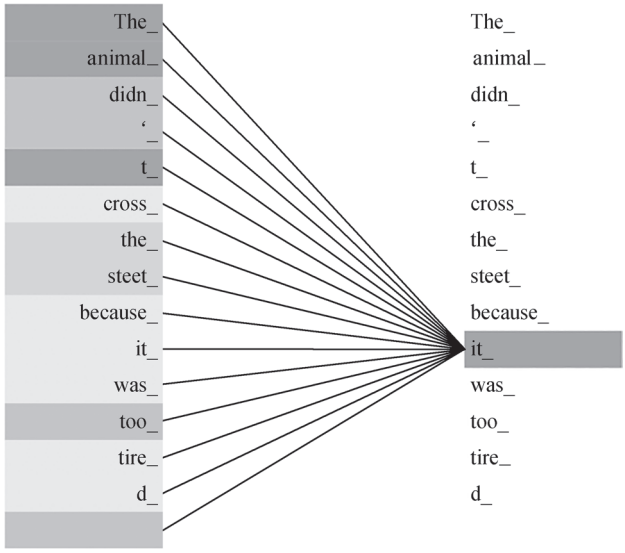


图1-17　自注意力机制示例

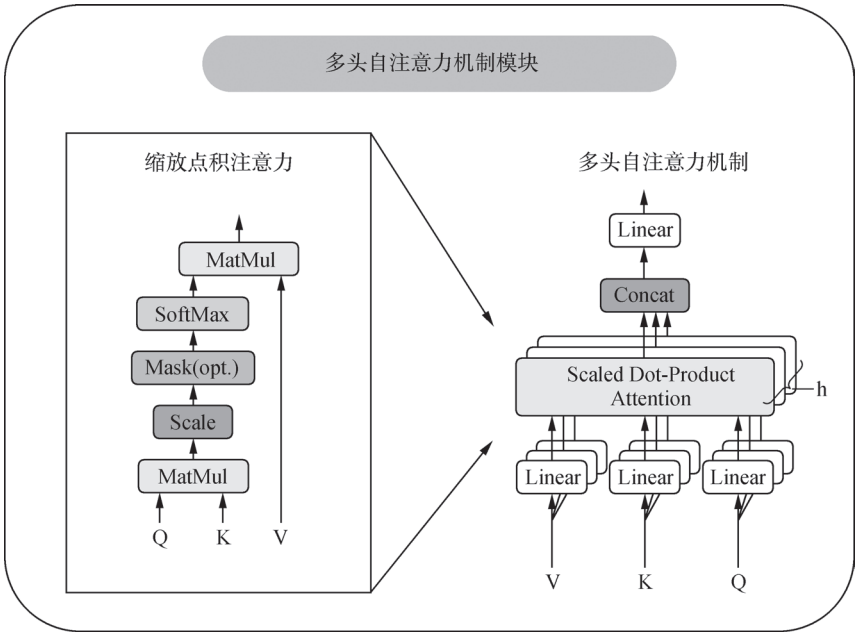


图1-18　多头自注意力机制模块

它不是只计算一次注意力，而是将输入分成更小的块，然后并行计算每个子空间上的缩放点积注意力。这种结构设计能让每个注意力机制去优化每个词汇的不同特征部分，

从而均衡同一种注意力机制可能产生的偏差，让模型能捕捉到不同层次的语义信息，增强模型的表达能力，提升模型效果。

1.2.3　GPT 发展历程

如图 1–19 所示，GPT 的发展历程主要可以分为两个阶段，在 ChatGPT 之前侧重于不断增大大模型的基础规模，并增强新能力。而 ChatGPT 和最新的 GPT–4 则更侧重于增加人类反馈强化学习，理解人类的意图，以提供更好的服务。

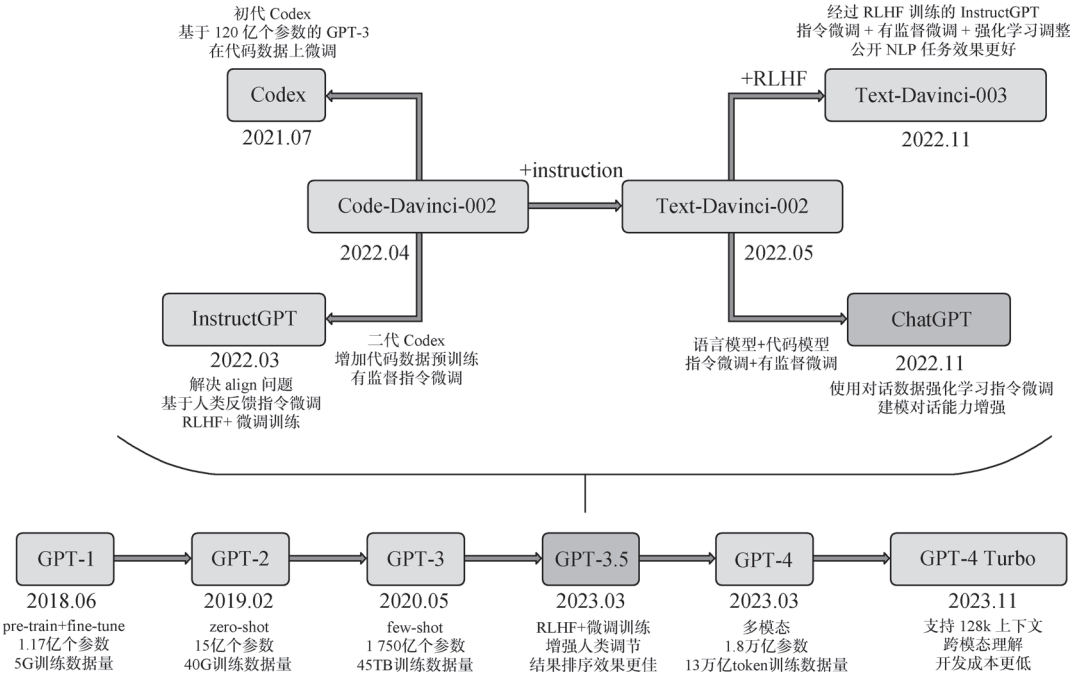


图1–19　GPT发展历程

① 2018 年 6 月，OpenAI 公司发表论文《通过生成式预训练提高语言理解能力》(*Improving Language Understanding by Generative Pre-training*)，正式发布了 GPT–1[11]。

- 基本思路：生成式预训练（无监督）+ 下游任务微调（有监督）。
- 基于 Transformer 的单向语言模型，Decoder 结构，共 12 层。
- 参数为 1.17 亿，训练数据量 5GB，模型规模和能力相对有限。
- 上下文窗口为 512 token。

② 2019 年，OpenAI 发表了最新进展，一篇《语言模型是无监督的多任务学习者》

（*Language Models are Unsupervised Multitask Learners*）的论文，提出语言模型是无监督的多任务学习，GPT-2 也随之诞生[12]。

· 基本思路：去掉有监督，只保留无监督学习。

· 48 层 Transformer 结构。

· 共 15 亿个参数，数据训练量提升至 40GB。

· 上下文窗口为 1 024 token。

③ 2020 年 5 月，OpenAI 公司发表论文《语言模型是少样本学习者》（*Language Models are Few-Shot Learners*），构建了 GPT-3 模型[13]。

· 基本思路：无监督学习 +in-context learning。

· 采用了 96 层的多头 Transformer。

· 参数增大到 1 750 亿，基于 45TB 的文本数据训练。

· 上下文窗口为 2 048 token。

④ 2022 年 3 月，OpenAI 再次发表论文《训练语言模型以遵循人工反馈的指令》（*Training language models to follow instructions with human feedback*），介绍了利用人类反馈强化学习（Reinforcement Learning from Human Feedback，RLHF），并推出了 InstructGPT 模型[14]。

· 基本思路：RLHF+ 微调训练。

· 增强了人类对模型输出结果的调节。

· 对结果进行了更具理解性的排序。

ChatGPT 是 InstructGPT 的衍生，两者的模型结构和训练方式都一致，只是采集数据的方式有所差异，ChatGPT 更加注重以对话的形式进行交互。

⑤ 2023 年 3 月，OpenAI 又发布了多模态预训练大模型 GPT-4，再次进行了重大升级。

· 基本思路：多模态。

· 上下文窗口为 8 195 token。

· 1.8 万亿个参数，13 万亿个训练数据。

· 强大的识图能力。

虽然目前 GPT-4 在现实场景中的能力可能不如人类，但在各种专业和学术考试上都表现出明显超越人类水平的能力，甚至 SAT 成绩（可以理解为美国高考成绩）已经超过了 90% 的考生，达到了考进哈佛、斯坦福等名校的水平。2024 年上半年，GPT-4 在 HumanEval 测试中表现优异，pass@1 得分 85.73%，pass@10 得分 98.17%，显著优于早期版本[15]，这代表着代码生成能力的提升。ChatQA 模型通过两阶段指令调优，提升了 GPT-4

在对话问答任务中的表现，包括监督微调和上下文增强指令调优，减少错误回答[16]。此外，中国的顶级国产人工智能大模型已达到 GPT–3.5 的水平，与 GPT–4 的技术差距正在缩小。中国拥有超过 130 个大语言模型，占全球总数的 40%[17]。

2024 年 5 月，OpenAI 对现有模型进行了一次全新升级，推出了桌面版 ChatGPT 及网页端用户界面（User Interface，UI）更新，并发布了 GPT–4o。其中，“o”代表“omni”，意为全能的。根据 OpenAI 官网给出的介绍，GPT–4o 可以处理文本、音频和图像任意组合的输入，并生成对应的任意组合输出。特别是音频，它可以在短至 232ms 的时间内响应用户的语音输入，平均 320ms 的用时已经接近人类在日常对话中的反应时间。与现有模型相比，GPT–4o 在视觉和音频理解方面尤其出色。GPT–4o 在英语文本和代码上的性能也与 GPT–4 Turbo 处于同一水平线上，在非英语文本上的性能有着显著提高，同时应用程序编程接口（Application Programming Interface，API）速度快，数据处理量大幅提升，成本则降低了 50%。在数据层面，根据传统的基准测试，GPT–4o 的性能与 GPT–4 Turbo 相比有优势，对比其他模型更是大幅领先[18]。

此外，OpenAI 最新的声明表明，GPT–5 将在不久后发布，并将实现从高中生智力水平到博士生智力水平的飞跃。与 GPT–4 相比，预计 GPT–5 将在语义理解、上下文关联和语言生成方面表现出显著进步。未来的应用将包括更加精准的机器翻译、更自然的对话系统及更高效的文本摘要生成。随着 GPT 技术的不断突破，未来的 GPT–5 不仅局限于文本处理，还将融合视觉、音频等多模态数据，为用户提供更全面和深刻的分析与解决方案，形成更综合的智能系统。

1.3　大模型

1.3.1　大模型概述

一般来说，在 ChatGPT 之前，被公众关注的 AI 模型主要是用于单一任务的。比如引燃了整个人工智能市场并促使其爆发式发展的“阿尔法狗”（AlphaGo），它基于全球围棋棋谱的计算，在 2016 年轰动一时的“人机大战”中击败围棋世界冠军李世石。但是从本质上来说，这种专注于某个具体任务而建立的 AI 数据模型，和 ChatGPT 相比，只能叫“小模型”。

大模型是指具有庞大的参数规模和复杂程度的机器学习模型，我们所提到的大模型通常是大语言模型（Large Language Model，LLM）的简称。语言模型是一种人工智能模

型，它被训练后可以理解和生成人类语言，而“大”的意思是指模型的参数量非常大，是相对于“小模型”而言的。LLM 可以处理多种自然语言任务，如文本分类、问答、对话等，是通向人工智能的重要途径，其主要特点如下。

· 规模大：通常是具有数百万到数十亿参数的神经网络模型，模型大小可以达到数百 GB 甚至更大，需要大量的计算资源和存储空间来训练和存储。

· 能力强：能够学习更细微的模式和规律，从而更好地预测未来数据并处理不同类型的数据，表现出更强的泛化能力和表达能力。

· 精度高：可以处理更复杂的数据集，从而减少了数据预处理和特征提取的时间和成本，能够更好地捕捉数据的复杂关系，提高了模型精度和准确性。

图 1–20 从任务处理能力不断增强的角度分析了 LLM 的发展过程。可以看出，语言模型并不是一个专门针对 LLM 的新的技术概念，而是随着 AI 技术的发展而逐渐发展起来的。早期的语言模型主要针对文本数据进行建模和生成，而最新的语言模型（例如，GPT–4）则专注于复杂的任务求解。从语言建模到任务解决，是科学思维的重要飞跃，也是理解语言模型在研究史中发展脉络的关键[19]。

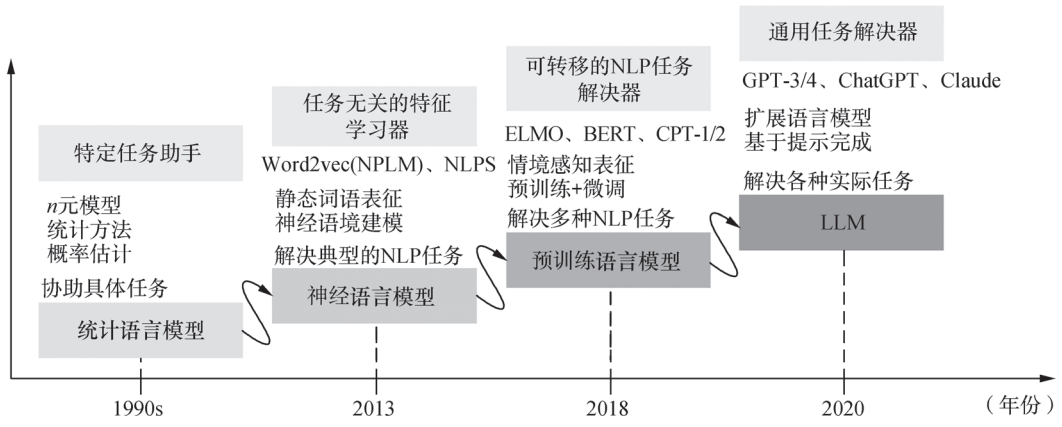


图1–20　LLM的发展过程

首先，统计语言模型主要在一些特定的任务（例如，提取任务或言语任务）中进行辅助，其中预测或估计的概率可以增强特定任务方法的性能。随后，神经语言模型专注于学习任务无关的表征，旨在减少人类特征工程的工作量。此外，预训练的语言模型学习了上下文感知的表示，可以根据下游任务进行优化。对于最新一代的语言模型，LLM 是通过探索模型容量的缩放效应来增强的，可以被认为是通用的任务求解器。综上所述，在发展过程中，语言模型所能解决的任务范围得到了极大的扩展，各方面性能也有了显

著的提升。

如图 1–21 所示，大模型进化树图追溯了近些年大模型的发展历程，其中重点凸显了某些最知名的模型，同一分支上的模型关系更近[20]。实心块表示开源模型，空心块则是闭源模型。基于 Transformer 的模型中，仅编码器模型是最左边的分支，仅解码器模型是最右边的分支，编码器 – 解码器模型则是中间的分支。

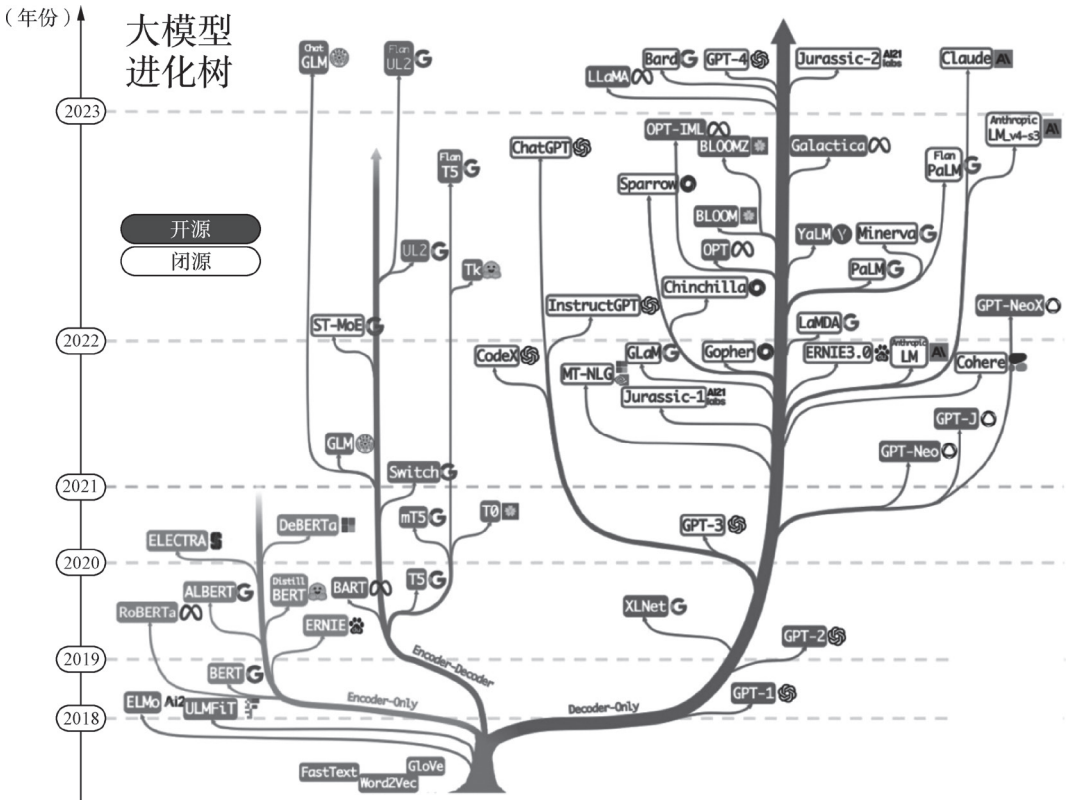


图1–21　大模型进化树

1.3.2　大模型研究现状

2023 年 10 月 12 日，分析公司 stateof.ai 发布了《2023 年人工智能现状报告》（*State of AI Report 2023*）。该报告指出，OpenAI 发布的 GPT–4 仍然是全球最强大的 LLM，该模型展示了专有技术与次优开源替代方案之间的能力鸿沟，同时也验证了基于人类反馈进行自学习的模型能力（RLHF）。ChatGPT 是世界上用户增速最快的产品，在其带领下，

生成式 AI 工具在图像、视频、编程、语音等领域取得了突破性的进展，带动了约 180 亿美元的风险投资和企业投资，拯救了风险投资界 [21]。生成式 AI 推动了生命科学的进步，大模型正不断地实现技术突破，特别是在生命科学领域，在分子生物学和药物发现方面取得了有意义的进展。

2023 年 12 月 14 日，《自然》（*Nature*）公布了 10 位 2023 年度人物，值得注意的是，聊天机器人 ChatGPT 因为占领了 2023 年的各种新闻头条，深刻影响了科学界乃至整个社会，被破例作为第 11 个“非人类成员”纳入榜单，以表彰生成式人工智能给科学发展和进步带来的巨大改变。目前，国内外对 GPT 大模型的研究不断深入，纷纷开始研发自己的大模型，并且应用的场景也越来越丰富。以 ChatGPT 为代表的大模型，正式开启了 AI 2.0 时代。

（1）国外

在美国，OpenAI、Anthropic 等初创企业和微软、Google 等科技巨头带领着美国在大模型的道路上飞速前进，同时各大公司也在不断提升自身的竞争力。谷歌给 Anthropic 投资了 3 亿美元以应对 ChatGPT 的威胁。2024 年 2 月，OpenAI 还发布了文生视频大模型“Sora”，展现了强大的内容理解和视频生成能力。

在欧洲，成立了一个旨在占领全球人工智能高地的研究机构 ELLIS。芬兰的 Flowrite，是一个基于 AI 的写作工具，可以通过输入关键词生成邮件、消息等内容。荷兰的全渠道通信平台 MessageBird 推出了自己的 AI 平台 MessageBird AI，可以理解客户信息的含义并做出相应的响应。2024 年 2 月，欧洲生成式 AI 独角兽 Mistral AI 发布了最新大模型 Mistral Large。

2023 年 4 月，英国政府宣布向负责构建英国版人工智能基础模型的团队提供 1 亿英镑的起始资金，以帮助英国加速发展人工智能技术。英国政府表示，该投资将用于资助由政府和行业共建的新团队，以确保英国的人工智能“主权能力”。此外，最大律师事务所之一麦克法兰也宣布与法律领域生成式 AI 企业 Harvey 达成技术合作，全面应用生成式 AI。

在日韩，日本的 7–11 连锁便利店将大模型用于产品创意和规划，提升产品研发效率。而本田也将大模型用于汽车设计之中。韩国也是最早加入大模型研发的国家之一。目前，韩国在大模型领域的代表有 NAVER、Kakao、KT、SKT 及 LG。韩国的半导体企业正在积极结盟，以应对大模型发展带来的算力挑战。

（2）国内

近年来，国内在大模型领域取得了显著进展。从科研机构到企业，都加大了对大模型的投入力度，在算法、算力、数据等方面取得了重要突破。国内已经出现了一批具有国际

竞争力的大模型，并在多个领域得到了广泛应用。

许多人可能会认为，中国的 AI 大模型是从“文心一言”开始的。但“文心一言”其实只是一个类 ChatGPT 的产品，背后驱动它的 AI 大模型，无论是百度、阿里还是腾讯、华为，都早有布局。

2023 年 3 月 16 日，基于文心大模型，百度发布了“文心一言”，成为中国第一个类 ChatGPT 产品。科大讯飞于 2023 年 5 月 6 号发布中国版 ChatGPT“讯飞星火认知大模型”，具有文本生成、语言理解、知识问答、逻辑推理、数学能力、代码编程和多模态能力七大核心能力。2024 年 5 月 15 日，字节跳动豆包大模型在火山引擎原动力大会上正式发布。截至 2024 年 7 月，豆包大模型日均 Token 使用量已突破 5 000 亿，平均每家企业客户日均 Token 使用量较 5 月 15 日模型发布时增长 22 倍。

（3）标准组织

如今国际标准化组织（International Organization for Standardization，ISO）、国际电工委员会（International Electrotechnical Commission，IEC）等组织都已围绕关键术语等开展标准研究。2023 年 3 月，欧洲电信标准化组织（European Telecommunication Standards Institute，ETSI）提出了有关人工智能透明度和可解释性的标准规范，旨在生成更多可解释的模型，同时保持高水平的模型性能。

第三代合作伙伴计划（3rd Generation Partnership Project，3GPP）规范包括了 AI 在网络架构中的部署和使用，涵盖了 AI 算法和架构的规范，还涉及了 AI 数据的处理和管理标准。目前，3GPP 有 4 个工作组在进行 AI/ 机器学习（Machine Learning，ML）标准化方面的研究工作，分别包括 AI/ML for Air Interface、AI/ML for RAN、AI/ML for 5GS 及 AI/ML for OAM。

2023 年 11 月，在由上海人工智能实验室与商汤科技联合主办的电气电子工程师学会（Institute of Electrical and Electronics Engineers，IEEE）“人工智能大模型”标准大会上，中国电子技术标准化研究院、上海人工智能实验室和华为云等 11 家单位共同发起成立了 IEEE 大模型标准工作组。该工作组将协同国内外大模型产业力量，制定大模型在技术规范、测评方法、安全可信、可靠决策等领域的国际先进标准，为全球大模型产业技术创新和发展提供更好的支撑。

（4）学术研究

此外，多模态技术的发展，使得 GPT 大模型的多模态处理能力也在不断提升，模型功能也越来越多样化。以往的单模态模型通常只能处理一种类型的数据，例如文本、图像或声音，缺乏对复杂环境的全面理解。而具有多模态能力的大模型能够同时处理多种类

型的数据，例如将视觉和语言信息相结合，以实现更深层次的理解和交互，并在更广泛的场景中得到应用。

例如，香港中文大学多媒体实验室联合上海人工智能实验室的研究团队提出一个统一多模态学习框架——Meta-Transformer，采用全新的设计思路，通过统一学习无配对数据，可以理解 12 种模态信息，并在 12 个不同的模态上完成不同的感知任务，如图 1-22 所示。该工作不仅为当前多模态学习提供了强大的工具，也给多模态领域带来新的设计思路[22]。

图1-22　多模态学习

在情感识别领域，新型多模态情感识别（Multi-Modal Emotion Recognition，MMER）系统通过一种联合多模态变压器和关键的交叉注意力融合的 MMER 方法，提高了预测的准确性。首先使用独立的主干网络捕获每个模态的时空依赖关系，然后通过融合架构整合各模态嵌入，有效捕获模态间和模态内的关系[23]。

针对自回归视觉语言模型在对齐任务中表现不够理想的问题，新加坡国立大学 Show Lab 和 Microsoft Azure AI 合作提出了 COntrastive-Streamlined MultimOdal 框架（CosMo）[24]，将对比损失引入到文本生成模型中，把语言模型战略性地分成专用的单模态文本处理和熟练的多模态数据处理组件，在增强了模型在涉及文本和视觉数据任务中的性能的同时，

显著减少了可学习参数，提高了多模态任务的分类和生成能力。

在高分辨率遥感影像领域，多模态技术还结合了如数字表面模型、RGB 和近红外等多模态数据，用于地表覆盖分类中的语义分割任务。通过一种轻量级多模态数据融合网络（Lightweight Multimodal Fusion Network，LMFNet），能够达成同时处理各种数据类型并提取特征的目的。具体而言，LMFNet 在 US3D 数据集上实现了 85.09% 的平均交并比（Mean Intersection over Union，mIoU），显著优于现有方法。与单模态方法相比，LMFNet 在 mIoU 上提高了 10%，且仅增加了 0.5M 的参数量[25]。

1.3.3　典型的大模型

国外大模型的主要发布机构有 OpenAI、Anthropic、Google 及 Meta 等，这些模型参数规模以千亿级为主，也出现了一些万亿级的超大模型。

发展至今，国外的头部大模型主要有 4 个，如图 1–23 所示。

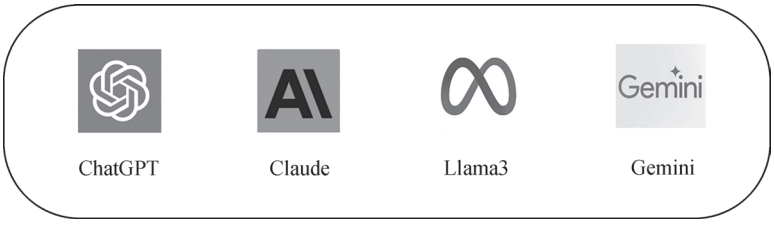


图1–23　国外头部大模型

（1）ChatGPT

ChatGPT 是由 OpenAI 研发的 AI 人工智能技术驱动的自然语言处理工具。它使用了 Transformer 神经网络架构，拥有语言理解和文本生成能力，可以通过连接大量的语料库来训练模型，这些语料库包含了真实世界中的对话，使得 ChatGPT 能够“上知天文下知地理”，还能根据聊天的上下文进行互动，生成高质量的回答，实现与真正人类几乎无异的对话交流。除了与人类进行交流外，ChatGPT 还具备完成多种复杂任务的能力。比如，它可以撰写邮件、视频脚本、文案、翻译、代码及学术论文等。

ChatGPT 是如今市面上火爆的对话 AI，也是最早“出圈”的 AI 大模型。这不仅是因为其强大的功能，更因为其在不同应用场景中的卓越表现。无论是在企业办公中提高效率，还是在创作领域提供灵感，ChatGPT 都展现出非凡的实用价值和广泛的适应性。此外，随着技术的不断进步和应用场景的扩展，ChatGPT 还在不断学习和优化，为用户提供越来越精准和个性化的服务，这使得它在全球范围内的受欢迎程度持续上升。

2024 年 5 月 14 日，OpenAI 还推出了 ChatGPT 桌面版应用程序。同年 6 月，苹果公司也在 2024 年度 WWDC 全球开发者大会上，正式宣布和 OpenAI 公司合作，未来将会在 Siri 中加入 ChatGPT。

（2）Claude

Claude 是由 Anthropic 开发的一种高级语言模型，在自然语言理解和生成方面取得了显著进展。凭借其广泛的训练数据和创新的算法，Claude 在各种与语言相关的任务中脱颖而出，在某些方面与 ChatGPT 相媲美甚至超越了 ChatGPT。

2023 年 11 月，Anthropic 宣布将加强与谷歌的合作伙伴关系，这一合作将进一步提升 Claude 的技术实力和市场影响力。同时，Anthropic 还得到了亚马逊的支持，这意味着 Claude 将受益于亚马逊的云计算基础设施和资源，为其模型训练和部署提供更强大的支持。随着技术的不断发展，Claude 不仅在语言生成和理解方面取得了突破，还在多个应用领域中展示了其广泛的适应性。例如，Claude 在医疗、金融、法律等专业领域中的文本分析和处理能力，使其成为专业人士的重要辅助工具。此外，在客户服务、内容创作和教育等领域，Claude 也展现出卓越的表现，为用户提供更智能和高效的解决方案。

Anthropic 在研发 Claude 时，还注重了模型的安全性和伦理性，确保其在提供强大功能的同时，不会对用户造成潜在风险。随着 Claude 的不断进化和改进，它在全球范围内的应用和认可度也在持续上升。Anthropic 通过与科技巨头的合作，致力于将 Claude 打造成下一代语言模型的标杆，推动人工智能技术的发展和应用，造福更多的行业和用户。2024 年 3 月 4 日，Anthropic 推出了最新的 Claude 3 大模型，包含性能由弱到强的 Haiku、Sonnet 和 Opus 3 种型号。Claude 3 也是多模态大模型，具有强大的“视觉能力”，用户可以上传照片、图表、文档和其他类型的数据，并进行分析和提问。

（3）Llama 3

Llama 3 是 Meta AI（Facebook 母公司）推出的开源大语言模型。在架构层面，Llama 3 选择了标准的仅解码器式 Transformer 架构，采用包含 128K token 词汇表的分词器。Llama 3 在 Meta 自制的两个 24K GPU 集群上进行预训练，使用了超过 15T 的公开数据，其中 5% 为非英文数据，涵盖 30 多种语言，训练数据量是前代 Llama 2 的 7 倍，包含的代码数量是 Llama 2 的 4 倍。

Llama 3 真正的特点是它专门在公开可获得的数据集上进行训练，使其对研究人员和开发人员更加可用。Llama 3 经过了广泛的测试和微调，以确保它的反应符合人类的偏好，并且不会泄露敏感信息。Llama 3 已经在多种行业基准测试上展现了先进的性能，提供了包括改进的推理能力在内的新功能，是目前市场上效果很好的开源大模型。

2024 年 7 月，Meta 还宣布将要发布 Llama 3 的更新版本。新的 Llama 3 模型可以进行 8 种语言的对话，编写更高质量的计算机代码，并可以解决复杂的数学问题。该模型拥有 4 050 亿个参数，远远超过了 2023 年发布的先前版本，但仍小于竞争对手的领先模型。据报道，OpenAI 的 GPT–4 模型有 1 万亿个参数，亚马逊正在投资一个有 2 万亿个参数的模型。此外，Meta 还将继续发布拥有 80 亿个和 700 亿个参数的轻量级 Llama 3 新版本。

（4）Gemini

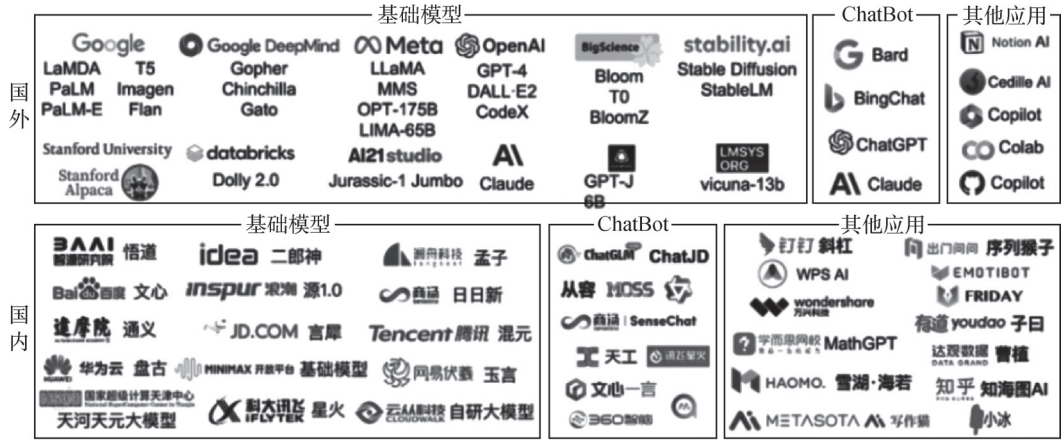
Gemini 是由谷歌在 2023 年 12 月 6 日发布的原生多模态大模型。Gemini 的原身是同样由谷歌发布的大模型 Bard。2023 年 3 月，谷歌推出 Bard，在全球大部分地区提供 40 多种语言版本。Bard 基于谷歌的 LaMDA 模型，后来升级为更强大的 PaLM 模型，可以帮助编写代码、调试和解释代码。最初在有限地区推出之后，Bard 于 2023 年 5 月扩展到了更多的国家。2024 年 2 月 8 日，谷歌宣布将聊天机器人 Bard 正式改名为 Gemini。

目前，大多数模型都通过训练单独的模块，然后将它们拼接在一起来近似多模态，无法在多模态空间进行深层次、复杂的推理。而 Gemini 最大亮点之一就是原生支持多模态，具有处理不同形式数据（语言 + 听力 + 视觉）的能力。因此，Gemini 可以泛化并无缝理解、操作和组合不同类型的信息，并且它的能力在几乎每个领域都很强。Gemini 可同时识别文本、图像、音频、视频和代码 5 种类型信息，还可以理解并生成主流编程语言（如 Python、Java、C++）的高质量代码，并拥有全面的安全性评估。它的首个版本为 Gemini 1.0，包括 3 个不同体量的模型，分别是用于处理“高度复杂任务”的 Gemini Ultra、用于处理多个任务的 Gemini Nano 和用于处理“终端上设备的特定任务”的 Gemini Pro。

同时，Google 也很注重安全性上的创新。通过制定积极主动的政策，并根据多模态能力的独特考虑因素进行调整，在 Gemini 中加入了对偏见和有害内容的全面评估，确保其多模态能力不会导致安全风险 [26]。

除了这几个比较熟悉的模型之外，新的 GPT 大模型仍在不断出现。在这场全球参与的竞争中，我国紧跟步伐，也开发了许多大模型。包括科大讯飞的“星火”、腾讯的“混元”、阿里的“通义”、华为的“盘古”、百度的“文心”及中国移动的“九天”系列等。

2023 年 8 月 31 日，首批国产大模型产品陆续通过《生成式人工智能服务管理暂行办法》的备案，正式上线面向公众提供服务。与此同时，更多企业的大模型也在迅速布局和推出。数据显示，截至 2023 年 10 月，国内 10 亿个参数规模以上的大模型厂商及高校院所共计 254 家，意味着“百模大战”正从上一阶段的“生下来”走向“用起来”的新阶段。图 1–24 展示了目前国内外厂商开发的一些主要大模型。



数据来源：InfoQ研究中心

图1-24　国内外各类大模型

此外，OpenAI 于 2024 年 2 月 18 日凌晨发布的新的文生视频大模型“Sora”，再次引起了轰动。Sora 是在 GPT 模型及图像生成模型 DALL-E 基础上开发而成的视频生成大模型，能够根据文本指令生成逼真或富有想象力的场景视频，展示了模拟物理世界的潜力。随着 Sora 的问世，视觉领域生成式人工智能达到了新的高度。如图 1-25 所示，通过相应的提示词，Sora 模型生成了一段戴墨镜女子走在东京街头的场景。

提示词：一位时尚女性走在东京街头，街道两旁是温暖发光的霓虹灯和动画城市标志。她穿着黑色皮夹克、红色长裙和黑色靴子，提着一个黑色钱包。她戴着太阳镜，涂着红色唇膏。她自信而随意地走着。街道潮湿而反光，形成彩色灯光的镜面效果。许多行人走来走去。

图1-25　Sora模型生成的视频场景

Sora 模型采用了扩散模型技术，基于 DALL · E 3 的再标记技术，从类似静态噪声的帧开始，逐步通过多个步骤精细化视频内容。该技术能够生成高度描述性的视觉训练数据，从而准确地将文本指令转换为视频中的动作[27]。Sora 模型架构如图 1–26 所示[28]。

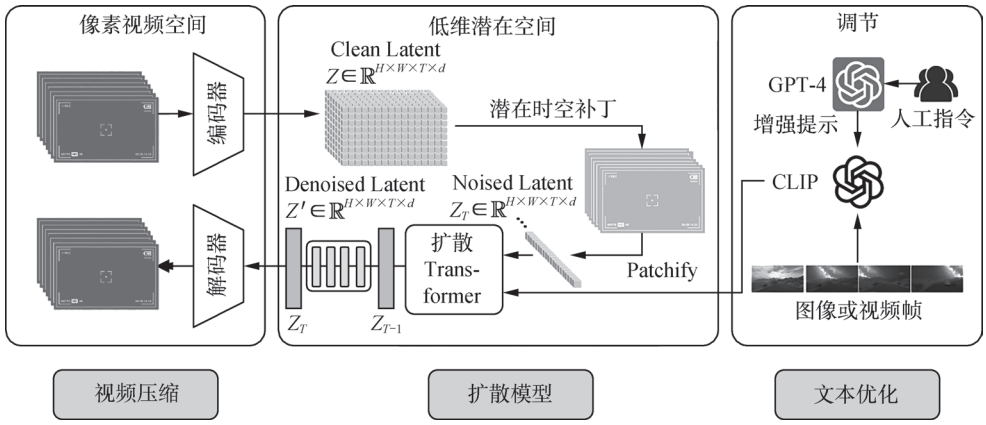


图1–26　Sora模型架构[28]

Sora 模型还集成了视觉和文本信息，利用视觉 – 语言模型（如 CLIP 和 Vision Transformer）实现文本到视频的生成。这种多模态整合使模型能够理解并生成复杂的多模态内容，提高了模型的应用广泛性和生成效果[29]。利用扩散模型，逐步将噪声转化为图像，Sora 能够生成具有较高细节和质量的视频。在这个过程中，Sora 使用 U–Net 架构，通过预测和减轻每一步的噪声来生成图像。

同时，Sora 模型还集成了幻觉检测机制，以确保生成视频的真实性和一致性。该机制能够检测并校正文本到视频生成中的各种幻觉，包括提示一致性幻觉、静态幻觉和动态幻觉，增强了模型在生成高质量视频时的可靠性[29]。

Sora 模型不仅能根据文本提示生成视频，还能对静态图像进行动画处理，或无缝填充视频中缺失的帧，增强视觉内容的流畅性和完整性。此能力表明 Sora 模型在模拟真实世界方面迈出了重要的一步，是通向通用人工智能（Artificial General Intelligence，AGI）的重要里程碑。然而，尽管 Sora 在视频生成方面取得了重要进展，它在模拟复杂场景的物理效果方面仍有改进的空间。例如，它可能无法准确模拟物体交互的物理变化，如食物的消耗或物体的空间细节[30]。

1.4 本章小结

总的来说，GPT的出现，极大地推动了自然语言处理技术的进步和应用。通过先进的Transformer架构和大规模预训练方法，这些模型能够处理复杂的语言任务，并且在众多领域实现了突破性的进展。随着技术的不断演进和优化，GPT及其衍生技术将继续影响和改变我们与机器互动的方式，推动人工智能技术在各行各业中的广泛应用和发展。

本章从大家最熟悉的“ChatGPT”出发，首先介绍了其基本概念和技术体系，以及在不同领域的典型应用。其次，本章介绍了生成式预训练转换器GPT的含义和最重要的Transformer架构，以及大语言模型的发展。此外，本章还回顾了GPT的发展历程，探讨了当前GPT的研究现状。最后，介绍了几个典型的GPT大模型，这些模型在不同的自然语言处理任务中展现了强大的性能和广泛的应用。本章通过对这些内容的介绍，帮助读者全面了解GPT和ChatGPT的发展背景、技术体系及它们对人工智能领域的重要影响。

参 考 文 献

[1] 刘禹良，李鸿亮，白翔，等. 浅析 ChatGPT：历史沿革，应用现状及前景展望[J]. 中国图象图形学报，2023，28(4): 893–902.

[2] Roumeliotis K I, Tselikas N D. ChatGPT and Open–AI models: A preliminary review[J]. Future Internet, 2023, 15(6): 192.

[3] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv: 1706.03762, 2017.

[4] 陈德光，马金林，马自萍，等. 自然语言处理预训练技术综述[J]. Journal of Frontiers of Computer Science & Technology, 2021, 15(8).

[5] 李耕，王梓烁，何相腾，等. 从ChatGPT到多模态大模型：现状与未来[J]. 中国科学基金，2023，37(5): 724–734.

[6] Prottasha N J, Sami A A, Kowsher M, et al. Transfer learning for sentiment analysis using BERT based supervised fine–tuning[J]. Sensors, 2022, 22(11): 4157.

[7] Bai Y, Jones A, Ndousse K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback[J]. arXiv preprint arXiv:2204.05862, 2022.

[8] Gao L, Schulman J, Hilton J. Scaling laws for reward model overoptimization[C]// International Conference on Machine Learning. PMLR, 2023: 10835–10866.

[9] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv:1707.06347, 2017.

[10]Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[11]Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. 2018, https://cdn.openai.com.

[12]Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI blog, 2019, https://cdn.openai.com/better-language-models.

[13]Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. arXiv preprint arXiv:2005.14165, 2020.

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

[15]Li D, Murr L. HumanEval on Latest GPT Models—2024[J]. arXiv preprint arXiv:2402.14852, 2024.

[16]Liu Z, Ping W, Roy R, et al. ChatQA: Building GPT-4 level conversational QA models[J]. arXiv preprint arXiv:2401.10225, 2024.

[17]GT Staff Reporters. Technical gap between China's AI large models with GPT-4 is narrowing: expert. Global Times Annual Conference, December 23, 2023. [Online]. Available: https://www.globaltimes.cn/page/202312/1304189.shtml.

[18]Sonoda Y, Kurokawa R, Nakamura Y, et al. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" cases[J]. Japanese Journal of Radiology, 2024: 1-5.

[19]Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.

[20]Yang J, Jin H, Tang R, et al. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond[J]. arXiv preprint arXiv:2304.13712, 2023.

[21]State of AI Report 2023. 2023, https://www.stateof.ai.

[22]Zhang Y, Gong K, Zhang K, et al. Meta-transformer: A unified framework for multimodal learning[J]. arXiv preprint arXiv:2307.10802, 2023.

[23]Waligora P, Zeeshan O, Aslam H, et al. Joint multimodal transformer for dimensional emotional recognition in the wild[J]. arXiv preprint arXiv:2403.10488, 2024.

[24]Wang A J, Li L, Lin K Q, et al. COSMO: Contrastive streamlined multimodal model with interleaved pre-training[J]. arXiv preprint arXiv:2401.00849, 2024.

[25]Wang T, Chen G, Zhang X, et al. LMFNet: An Efficient Multimodal Fusion Approach for Semantic Segmentation in High-Resolution Remote Sensing[J]. arXiv preprint arXiv:2404.13659, 2024.

[26]AZHAR A. Google Launches Gemini, Its Largest and Most Capable AI Model[EB/OL] // AIwire. (2023-12-07)[2024-09-15]. https://www.aiwire.net/2023/12/07/googlelaunchesgemini/.

[27]Video generation models as world simulators | OpenAI[EB/OL]. [2024-09-15]. https://openai.com/index/video-generation-models-as-world-simulators/.

[28]Liu Y, Zhang K, Li Y, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models[J]. arXiv preprint arXiv:2402.17177, 2024.

[29]Chu Z, Zhang L, Sun Y, et al. Sora Detector: A Unified Hallucination Detection for Large Text-to-Video Models[J]. arXiv preprint arXiv:2405.04180, 2024.

[30]Synced:AI technology & industry review[EB/OL]. [2024-09-15]. https://syncedreview.com/.