

致广大而尽精微

生成式 AI 企业应用落地技术白皮书

神州数码集团股份有限公司
神州数码通明湖研究院
北京信百会信息经济研究院

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

CONTENT 目录

1 生成式 AI 是一场技术范式变革	3
2 生成式 AI 的六层技术生态	8
2.1 AI 算力基础设施	8
2.2 基础大模型与相关技术	14
2.3 大模型与训练、评测数据	22
2.4 生成式 AI 应用开发技术	26
2.5 生成式 AI 安全与监控	35
2.6 生成式 AI 应用设计	38
3 生成式 AI 企业应用落地实践探索和总结	41
3.1 生成式 AI 与企业数字化转型	41
3.2 企业应用落地的关键问题与应对方法	42
3.3 企业应用落地的四类驱动模式	55
4 AI 产业政策与发展趋势	65
4.1 我国 AI 产业政策	65
4.2 AI 产业发展趋势	69
4.3 促进我国 AI 产业发展的对策建议	73
5 写在最后	74
6 引用	76

1 生成式 AI 是一场技术范式变革

2022年末ChatGPT的横空出世及其之后的持续迭代，以一种人人可亲身感知的方式，把人工智能在自然语言领域里的重大进展在一夜之间展示在世人面前。而在企业应用场景方面，之前的AI技术都集中在相对专业的应用场景内，如机器视觉、语音识别、或推荐系统、风控管理等。但是语言，作为人类重要的思维工具以及知识组织和传播的最主要手段，其“能力泛化”的可能性远远超出了其他领域。因此，当ChatGPT能够与人类进行深入、富有深度的对话时，人们开始想象一个真正能够理解业务或专业、思考解答专业问题、甚至进行业务的组织、管理和创新的机器的可能性。对企业的数字化转型进程而言，生成式AI技术带来的潜在影响很容易让人将之类比于交通史上铁路系统的发明、亦或动力系统中对交流电的引入。

在生成式 AI技术出现之前的十多年间，数字化转型一直是企业采取的一项重要战略，来促进企业新的商业环境中保持竞争优势、创造新的商业机会。根据 2011年，数字化转型最早的提出者之一——Gartner 的定义，数字化转型包括从 IT现代化升级(比如全面云化升级)，到通过数字技术进行业务优化(比如精准营销)或业务模式创新(比如创新的引流和盈利模式)的一系列战略举措。近几年来，数字化转型的重点聚焦领域，已经越来越转向企业数据资产的建立，神州数码集团的创始人和 CEO郭为在《数字化的力量》一书中，对此提出了全面和系统的论述。

而生成式 AI出现之前，数据一般只有经过结构化处理之后，才能够在企业应用环境中发挥作用；而在企业的经营活动中，产生的大量的数据无法被结构化处理，比如内部海量的会议纪要、周报、季报，其中包含大量关于企业具体业务事项的分析和讨论；企业的大量的合同文本、项目验收材料，其中包含有大量的交易细节；而在销售和客服人员与顾客的线上互动文本，其中也有一手的客户对产品和服务的反馈；再有，就是企业产品的大量的用户手册、故障分析文档、产品服务和支撑技术资料等等，其中有丰富的技术支持所需的知识。

所有这些包含的非常有价值的信息和知识，以往只能限于少数专家或管理者的随机及离散地利用。传统的数据处理和分析方法对这种非结构化的文本数据无所适从。高价值的信息无法被有效提取，意味着企业可能错失了重要的决策依据、市场洞察和创新机会。

以大语言模型为代表的先进的自然语言处理技术的出现，预示着这种情况开始发生变化。企业有可能利用这些创新技术来自动分析、归类和抽取这些非结构化数据中的关键知识，进而为决策者提供有力的支持。例如，通过自动分析销售和客服的交互文本，企业可以更准确地了解客户的需求和不满，进一步优化产品和服务。更

进一步，企业还可以利用这些技术结合知识图谱技术，将分散在不同文档和系统中的信息连接起来，形成一个跨组织结构、跨业务领域、跨时间维度的企业大脑；为企业提供一个一体化的知识查询甚至咨询平台。这样的平台将会成为企业的超级销售助理、超级客服助理或者是超级管理助理。**生成式 AI 技术的出现，为企业数字化转型，注入了强大且更为直接的新动能。**

不过，以上对生成式 AI 技术对数字化转型的推动的“推演”，可能还存在很大局限。如同早期的英国铁路，斯托克顿-达灵顿专线其实是在铁轨上跑马车。早期的蒸汽机的一个主要应用场景是在枯水期将水引向高处蓄水池以帮助驱动水车。

目前我们设想的生成式 AI 的应用场景，也处于早期状态。生成式 AI 技术为企业数字化转型带来的会是更为根本的变革，即技术范式的改变(Paradigm Shift)。我们借用《技术的本质》一书中对“技术域”的定义来解读“技术范式的改变”：作者在这本书的第 8 章指出，（它）不是单独一个技术体的出现，而是新技术体引发的“重新域定”。新技术域对经济的影响也比单个技术对经济的影响要更深刻。**作者认为，经济并不是采用(Adopt)了一个新的技术体，而是遭遇(Encounters)了一个新的技术体。**经济对新的技术体的出现会作出反应，它会改变活动方式、产业构成以及制度安排，也就是说，经济会因新的技术体而改变自身的结构。如果改变的结果足够重要，我们就会宣称发生了一场颠覆性改变。

生成式 AI 技术正在形成新的技术域定，它首先对应用软件开发产生了显著影响。得益于计算机程序设计语言的严格语法、清晰逻辑性和罕见的二义性，生成式 AI 技术在代码生成和辅助编程方面的效果日益突出。展望未来，软件开发的重心将更多地倾向于需求分析和软件架构设计，而编码和代码质量审核的流程，将在先进的辅助编程工具的助力下，实现效率的飞跃性提升。在 2017 年，曾经是 OpenAI 创始成员和研究科学家，担任特斯拉技术总监的 Andrej Karpathy 就预见到了引入 AI 之后的新软件开发范式，他在一篇技术博客中提出了软件 2.0 的概念。在软件 1.0 的模式下，由程序员设计软件解决问题的方法和细节逻辑，并通过编写显示指令来实现这些逻辑。而软件 2.0 是利用神经网络自动完成软件的设计。未来大部分程序员无需编写复杂的程序，维护复杂的软件库、或者分析它们的性能。他们只负责收集、清理、操作、打标签、分析和可视化为神经网络提供信息的数据即可。随着生成式 AI 技术的快速迭代，业界内正在宣称“软件工程 3.0”时代的开启：AI 重新定义了开发人员构建、维护和改进应用软件的方式，研发团队的主要任务而是以含有私域专业知识的语料(或图像、视频)来训练或精调模型、围绕业务主题设计提示模板(Prompt Template)、探索最有效的智能体(Agent) 机制等。

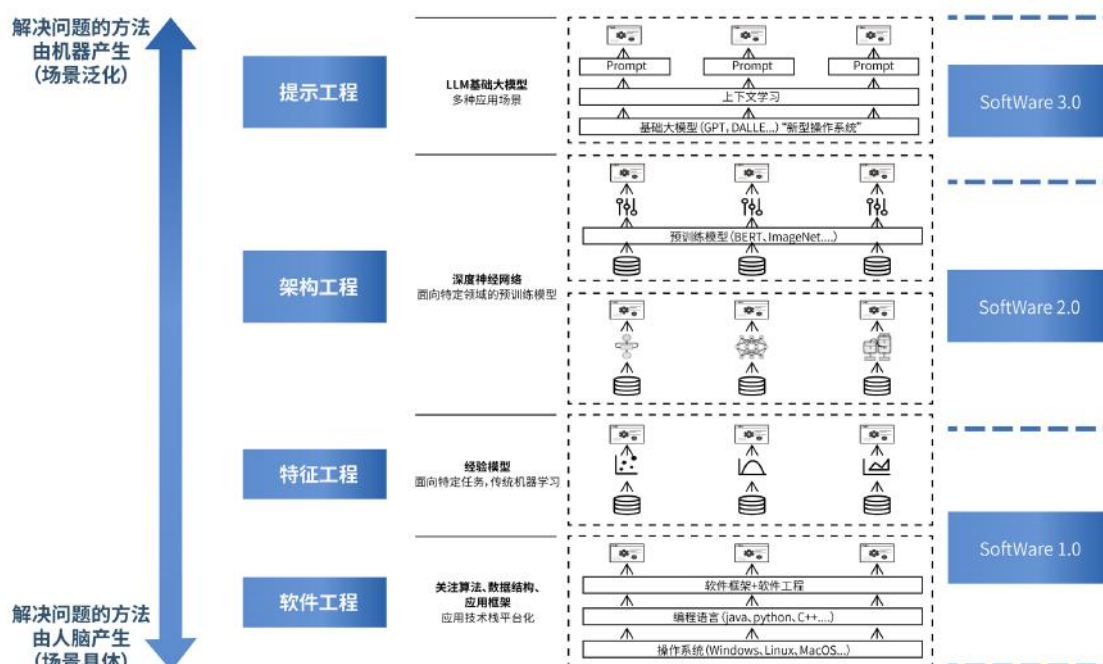


图 1 Software1.0 到 3.0

综上所述，不论是“1.0”“2.0”还是“3.0”模式的软件，生成式 AI 技术都将为其注入革命性的创新力量。应用会在价值和体验、安全和运营、架构和交付等方面发生深刻变革，从而催生出企业应用的大升级和大迭代。而更快和更广泛的业务数字化转型，则会产生更多的数据资产和应用场景，数字化转型的飞轮效也将应运而生。

为什么会有这篇白皮书

每一次技术的范式变革都深刻地重塑了经济格局和社会结构，同时也催生出企业数字化的新浪潮。例如，以 2010 年为分水岭，移动互联网和智能手机的快速渗透为众多崭新的应用提供创新的土壤。在此背景下，移动定位、身份绑定和移动支付等技术场景快速落地，为企业开辟了全新的移动获客渠道。不少企业敏锐地捕捉到这一趋势，纷纷推出小程序或打造移动应用平台，助力自身在激烈的市场竞争中快速而精准地获取用户、拓展市场地位。这不仅为企业和市场带来了前所未有的变革和机会，甚至形成了新的社会消费习惯。

由于对上一次的技术变革带来的影响仍记忆犹新，使得这一轮人工智能的飞跃式进展所产生的震撼和影响更为强烈。不仅技术层面的 CIO、CDO 和 CTO 表现出浓厚的兴趣，企业的各个业务单元、包括 CEO 在内的高级决策层，几乎都在第一时间启动了密切地关注与讨论。

而另一方面，在这场由生成式 AI 引领的技术范式变革中，相关的推动力量从实验室快速走到了公众舆论

中心。这些力量，不再仅仅局限于学术会议的探讨。行业头部公司、初创企业及各个研究团队，也在数字化的今天利用自媒体平台和社区平台积极互动，并保持与主流媒体的沟通。开源社区的贡献和风险投资的活跃参与，更是助燃了这场技术革命，大量创新的想法都会快速落地实现、并成为资本追逐的目标。

大量的自媒体在这场热潮中成为了连接“圈内”和公众的纽带，他们迅速收集信息，并按更易传播的方法拆解(或碎片化)信息，使其在短时间内成几何级数放大，触达更广泛的受众。

然而，这种聚光灯下的创新展现，也给企业带来了难题。在信息海量涌入的时代，过多的信息反而形成了负担。企业在努力把握技术趋势、评估技术进展对自身业务的潜在影响时，往往陷入信息过载的困境，这不仅无法快速做出决策，更可能导致企业面临选择困惑，产生不必要的焦虑。而大量的粗粒度信息，也会对技术产生误解并不恰当的期望，这反而会阻碍早期的创新型尝试。

在与众多企业客户深入交流的过程中，我们深刻地认识到，对于当前的技术进展和各种应用实践进行系统的梳理与小结是至关重要的。这不仅能为企业提供一个清晰的技术发展蓝图，同时也助于他们更好地了解趋势，捕捉潜在机会，进而制定更加科学、前瞻性的战略规划。此外，这样的梳理还能推动行业间的交流与合作，为企业之间打造共赢的合作模式，加速整个数字化转型领域向更新的阶段发展。

我们希望通过编撰这篇《白皮书》，能够起到“抛砖引玉”的效果，引发业界的讨论。我们热切地期望生成式 AI 相关的技术提供者、应用解决方案的开发者、行业内的重要客户，以及各大研究机构等，能够以这篇《白皮书》为“靶子”进行深入的梳理和探讨。我们更希望它能成为企业客户和生成式 AI 技术落地实践者之间共识的起点，帮助大家澄清概念、分析当前的技术趋势，预测未来可能的发展方向。我们深知，单凭一家之力难以捉摸整个行业的脉搏，但是，通过集思广益，我们相信能够对这一领域产生更深入、更全面的了解。

在这篇《白皮书》中，我们旨在全面探索生成式 AI 技术的进展与应用。后续内容将分别从生成式 AI 的相关技术梳理、技术落地企业应用的路径、以及生态和监管这三个维度展开：对相关技术梳理，将从生成式 AI 的六层技术生态的角度，思考和总结生成式 AI 技术在不同维度带来的技术创新和挑战；然后，我们将深入探索生成式 AI 在落地企业中的实际应用，以及与现有业务的整合和可能遇到的挑战；最后，我们将讨论生成式 AI 在整个行业生态中的地位，伴随的伦理考量，以及对应的监管建议和未来发展趋势。通过这三个章节，希望可以为读者提供清晰的技术发展蓝图，帮助企业和研究者更好地理解、应用并推动技术的健康发展，从而应对信息过载、技术误解和创新尝试中的挑战，正如我们在白皮书开篇所述的背景和目的。

并发式创新的复杂局面和企业应对的策略

生成式 AI 的企业应用落地，事实上已经形成了 基础研发、监管和安全、应用开发、企业（或行业）私域数据就绪、企业能力就绪等 多个领域并行探索的局面。上述每一个领域既相互促进，又相互制约，而在企业应用的实际环境中，又需要探索业务流程、使用习惯和技术落地之间的变通和粘合。例如企业（或行业）私域数据就绪意味着企业需要建立一套完整的数据管理和维护体系，来确保数据的质量、完整性和安全性，当大语言模型需要进行微调或适应特定场景时，可以迅速地获得高质量的训练数据。

而最为重要的是，生成式 AI 的基础技术研发还在快速进展之中，制约其在真实业务场景使用范围的问题：例如在私域知识框架内的对齐，包括幻觉消除，知识收敛，以及上下文长度等，还在不断探索和解决之中。其中应用场景更为广阔的多模态大模型技术，更是令人充满期待。

从来没有哪一个时刻，使得企业在制定技术战略时，需要理解如此复杂的技术趋势，平衡考虑如此多的矛盾因素。从近期和客户的广泛交流中，我们发现，一些非常值得借鉴的策略已经形成：

1、两个立即着手：

- 立即着手采用点状业务创新的方式：紧密跟踪最新技术进展，探索安全和监管的边界构建；
- 立即着手采用共创的方式：选择外部供应商和合作伙伴，为有可能到来的生成式 AI 的场景爆发准备好强大的外援力量。

2、两个规划制定：

- 私域知识治理规划：生成式 AI 技术助力企业数字化转型，无论如何都需要企业私域知识的加持，部分企业曾经开展过数据治理工作，这为企业私域知识治理打下了很好的基础；
- 生成式 AI 应用开发和管理平台规划：不论软件 1.0、2.0 还是 3.0 的应用，都是企业数字化转型落地的手段，在点状创新之后，需要认真规划新应用的体系化开发、部署、运维和管理的平台，以及大模型及其算力管理平台 and 现有技术栈的融合。

神州数码，作为中国 IT 生态的核心参与者，始终致力于促进先进技术在企业的系统化应用。作为生态链的建设者和守护者，我们深知生成式 AI 技术的崛起标志着一场技术革命的开始。因此，我们决意联合整个生态体系，共同帮助企业全面拥抱这一技术范式转变的到来。

面对巨大而复杂的机遇与挑战，儒家经典《礼记·中庸篇》为我们提供了宝贵的指导思想：“**故君子尊德性而道问学，致广大而尽精微。**”这启示我们在追求技术创新的道路上，既要有宏观的视角，又必须全神专注于每一个关键的落地技术细节。持此信念，神州数码将继续汇聚各方力量，助力生成式 AI 技术为企业数字化转型注入更强劲的动力。

2 生成式 AI 的六层技术生态

GPT的成功，促成整个 AI 行业的技术生态正发生着巨大变革，并形成了激烈的竞争：从众多 AI 芯片厂商奋力追赶英伟达当前的技术优势，到模型厂商间的“百模大战”迅速升级为“千模混战”，生态中的厂商都在力图找准自己的定位，形成自己的技术优势。激烈竞争的同时也带来了技术的快速发展，相关的论文和报告以惊人的速度发布着，新的应用以及产品更是层出不穷。

随着不断地创新、试错以及优化，生态架构中许多关键的概念逐步清晰，一些关键的技术沉淀下来，积极影响着企业场景的落地。我们可以明显观察到生成式 AI 相关技术的发展已经形成了六层技术生态体系，包含 AI 算力基础设施、基础大模型与相关技术、大模型与训练及评测数据、生成式 AI 应用开发技术、生成式 AI 安全与监控以及生成式 AI 应用设计。本章概述了架构中每层的核心技术，并结合自身在实际应用场景中的经验与思考，为大家带来生成式 AI 技术生态的总结。

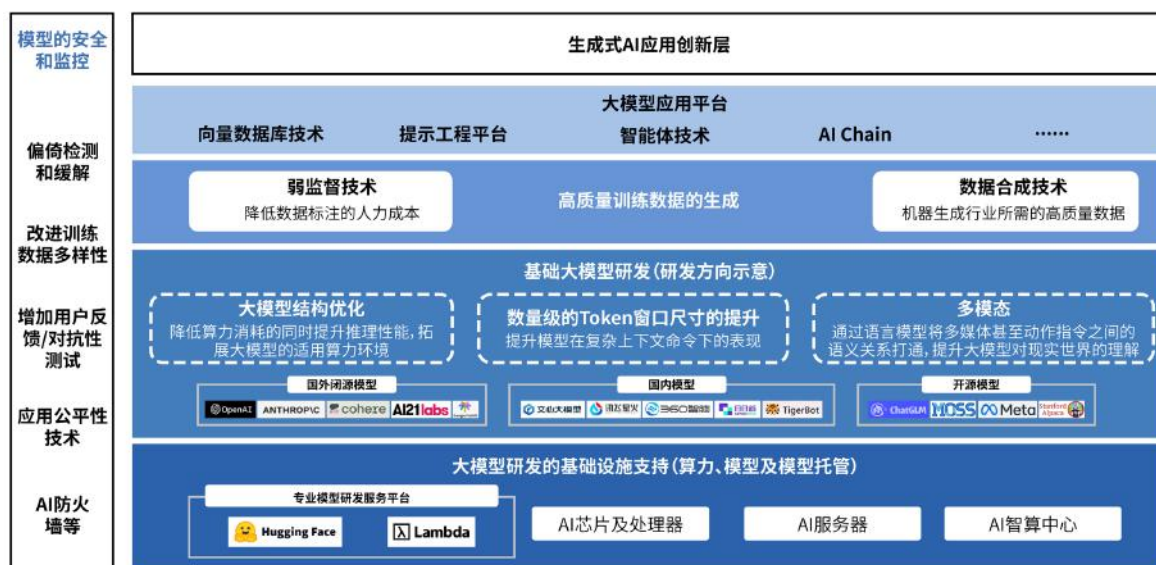


图 2 生成式 AI 六层架构技术生态体系

2.1 AI 算力基础设施

大模型的基础算力设施是 AI 生态中不可或缺的核心组成部分，为大模型在企业场景中的实际应用提供了关键的驱动力。其中 AI 芯片是算力的核心来源，其选型会直接影响到后续大模型的开发效率和性能。与此同时，AI 服务器，作为 AI 芯片的主要承载平台，其架构设计和性能优化也显得尤为关键。基于 AI 服务器，各大厂商会根据所持有的算力资源，发展出不同的经营模式。一些厂商选择采用“基础设施即服务（IaaS）”模式，主要

提供硬件设施的使用权限；而一些厂商则采用“平台即服务（PaaS）”模式，不仅提供算力，还为用户提供了一系列与模型开发相关的服务。为了更高效地管理这些AI服务器和算力资源，许多企业和政府机构会选择构建智算中心，这是一种集中管理和优化算力资源的方式，同时也反映了其对AI技术的重视和支持。

我们将深入探讨大模型基础设施的各个方面，包括AI芯片、AI服务器、AI IaaS、AI PaaS以及AI智算中心，阐述大模型对基础设施的特定需求，旨在为读者提供一个更全面的视角。

2.1.1 AI 芯片

2.1.1.1 AI 芯片概述与分类

AI芯片也称为AI加速器，专门用于处理人工智能应用中需要的大量计算任务的模块，为AI任务提供基础算力。

AI芯片前身是GPU（Graphics Processing Unit，图形处理单元），专门为游戏或者图像软件提供高效图形渲染的处理器，之后在人工智能技术逐步发展的过程中发现GPU的独特高效并行计算架构同样适用于人工智能计算加速过程。在人工智能理论知识逐渐丰富的过程中，芯片厂家也对AI芯片处理器的计算单元和架构组成有了更多的探索。

根据芯片的处理单元和可编程灵活性分类，AI芯片可以分为GPGPU、FPGA和ASIC以及类脑芯片。其中GPGPU（General Purpose Graphics Processing Unit，通用图形处理器）是GPU的衍生概念，保留了GPU的并行计算能力，去除了图像渲染显示部分。目前学术界和工业界普遍使用英伟达的AI芯片进行人工智能模型和应用开发，考虑到模型应用的普适性人们也都以GPGPU作为首选。FPGA（Field Programmable Gate Array，现场可编程门阵列）可以通过配置文件重新定义门电路和存储器之间的连线从而改变计算方式，与GPU相比具有高性能低功耗和可硬件编程的特点。ASIC（Application Specific Integrated Circuit，专用集成电路），是一种专用芯片，是为了某种特定的需求而专门定制的芯片的统称。在其所针对的特定的应用领域，ASIC芯片的能效表现要远超GPU等通用型芯片以及半定制的FPGA。近几年，颠覆传统冯·诺依曼架构模拟人脑神经元结构的类脑芯片成为学界和工业界探索的新思路。

根据 AI应用场景分类芯片有云端、终端和边缘端三种类型。云端芯片一般部署在公有云或私有云侧，支持模型的训练和推理任务。其优点是高性能、高计算密度，缺点是单价高、产品硬件形态单一。终端芯片通常部署在手机等移动设备中，支持模型推理任务，其优点是低功耗、高效能、成本低、产品最终硬件形态众多。边缘端芯片部署在边缘设备上如路边监控控制通讯设备，其对功耗、性能、尺寸的要求介于终端和云端之间，同

样以推理任务为主，产品的硬件形态相对较少。

根据芯片在 AI 任务中的功能分为训练芯片和推理芯片。训练芯片支持大型模型的训练过程，通过大量数据的输入训练构建复杂的深度神经网络模型。在模型训练的过程中涉及大量的训练参数和复杂的模型网络结构，需要巨大的运算量，对处理器的计算能力、可处理数据精度和可拓展性的要求都很高。推理芯片支持使用训练好的模型进行推理运算，对单位能耗算力、时延和成本有一定的要求。

2.1.1.2 AI 芯片的性能指标和大模型的算力消耗

在模型推训的过程中，主要关注 AI 芯片硬件的以下几个指标：算力、功耗、面积、带宽和显存。

算力是衡量 AI 芯片的重要指标，常用的单位是 TOPS 和 TFLOPS，分别代表芯片每秒能处理多少万亿次的 INT8 的整型运算或 FP32 单精度浮点运算。AI 芯片的算力越高代表它的运算速度越快，性能越强。

功耗是芯片运行的电力消耗，由于模型推训耗时漫长，大量的电力消耗进而需要更大的资金投入，对使用者而言，AI 芯片的功耗不容忽视。摩尔定律预言了芯片面积和利润的关系，通常来讲相同工艺制程之下，芯片面积越小、良率越高，则芯片成本越低。

考虑到大数据并行访问的需求，AI 和大数据处理需要高带宽和大存储容量的内存。因此，大模型对于 AI 芯片有以下两项性能要求：其一，带宽，位数越大说明时钟周期内所能传输的数据量越大；其二，显存，大显存能减少读取数据的次数，降低延迟。

大模型的算力消耗受以下几个因素影响，每参数每 Token 算力需求、模型参数规模、训练数据规模和算力使用效率。

以 GPT-3(175B) 为例，其模型的参数量是 175B，假设训练数据为 300B tokens，每参数每 token 对算力的消耗是 6 Flops，以 NVIDIA 80GB A100 GPU 为例，理论算力是 312 TFLOPS，Megatron 利用张量并行和流水线并行技术能达到 51.4% 的利用率，即每秒能完成 0.16 PFLOPS，根据上述条件，结合模型算力消耗约等于（每参数每 token 的训练需求 * 训练数据规模 * 参数规模）/ 算力使用效率，推测单张 A100 完成一次迭代计算所需耗时约为 65 年，若采用 1000 张 A100，训练时间大约可缩短为 1 个月左右。

2.1.2 AI 服务器

区别于传统服务器，AI 服务器搭载了各类 AI 加速卡，通过异构的方式组成不同的 AI 服务器。

其常见的组合形式是 CPU+GPU、CPU+FPGA、CPU+TPU、CPU+ASIC 或 CPU+多种加速卡等。近期甚

至提出了“GPU+DPU的超异构”设计，加入 DPU的强大数据处理调度能力的 AI服务器将更加适合大模型时代超大数据量并行计算的场景。

AI服务器根据应用场景、芯片类型和 GPU数量有不同的分类。根据深度学习应用场景分为训练型服务器和推理型服务器，训练型服务器对算力要求较高，推理型服务器对算力要求较低。根据 AI服务器搭载的芯片不同，分为“CPU+GPU”的异构类型和“CPU+XPU”超异构类型。最后，根据搭载 GPU的数量分为多路 AI服务器，常见的有四路、八路和十六路 AI服务器。

大模型的训练和推理任务对算力和网络都有了新的需求，超大参量的模型需要超高的算力，然而训练时间的延长，对模型训练期间的网络稳定性也有要求。近来，芯片领头企业将目光转向了“超异构”计算架构，集成 CPU、GPU和 DPU多种芯片的 AI服务器可以高效解决 AI大模型计算中遇到的多种计算加速、可拓展性、数据带宽延迟、训练速度、网络稳定性等问题。

2.1.3 AI IaaS

IaaS(Infrastructure as a Service, 基础设施即服务)，运营商通过软件定义算力资源的方式将硬件资源池化提供给客户。客户通过即用即付的方式获取计算、存储和网络等 IT基础设施资源的调度使用权限，并在此基础上部署、维护和支持应用程序。运营商负责运营维护基础物理设施，采用依赖虚拟化、容器和自动化技术的云计算架构高效控制 IT资源。

AI IaaS服务平台通过软件定义 AI算力组成具备池化能力、池化调度和运维管理能力的功能架构。其中，池化能力支持算力切分、远程资源调用、资源聚合、算力超分和按需应变的功能。池化调度包括本地或跨机房调度、指定节点、指定型号、任务队列调度和多资源池管理。资源池的运维管理功能包括运行时自动注入、组件高度可用、集群运维管理、平台运维管理以及全局资源监控。

AI IaaS的关键技术点是算力池化。算力池化基于传统云计算技术(如 Kubernetes、OpenStack)用软件定义的方式，对 GPU等 AI算力资源进行分时调度管理，并且采用 GPU/AI芯片的 Runtime提供的 API劫持、应用程序监控等技术实现计算资源跨界点远程调用。

AI IaaS通过高速无损网络互连互通的 CPU、GPU、ASIC芯片等算力资源进行池化整合，实现资源的集中调度、按需分配，使资源充分利用，降低碎片概率，提高总体有效算力、降低智算中心购置成本，能够做到化整为零。革新传统的整卡分配、“一虚多”的虚拟化分配等粗放式分配方式，使能精细化分配能力，根据 AI 任务的资源需求进行按需分配，使资源可被充分利用，降低碎片概率，提高总体有效算力，降低基础硬件设施购置成本。

2.1.4 AI PaaS

PaaS(Platform as a Service, 平台即服务)为软件开发提供了一种服务化的平台,采用软件即服务(SaaS)的模式交付。对于 AI大模型的开发者,PaaS提供了一个便捷的环境,支持大模型应用的快速部署、开发和测试。

PaaS 平台架构

AI 大模型的 PaaS 平台主要提供以下五大功能：

1、加速生产和部署:提供工具和指南,优化并加速模型的推理,满足生产部署的需求。比如平台会使用如 Docker或 Kubernetes的容器技术,确保模型在不同的环境中都能一致、稳定地运行,并通过 CI/CD流程,确保模型的更新和部署能够自动且连续地进行。

2、模型库与接口:提供统一的接口,支持多种预训练的 NLP模型,如 BERT、GPT、RoBERTa等。Transformer库的 API支持各种 NLP任务,如文本分类、命名实体识别、文本生成等。通过 API调用,开发者可以轻松地加载和使用模型,并可以通过接口提供丰富的参数和选项,使开发者可以根据自己的需求进行定制。

3、数据管理与处理:Datasets库可以提供 NLP数据集的访问、管理和处理工具,Tokenizers库可以支持文本数据的标记化,为模型准备输入数据。比如开发者可以直接在平台上加载和使用库中包含的多种 NLP数据集,平台会允许开发者上传自己的数据集,并为数据集提供版本管理功能,从而确保数据的一致性。

4、模型训练与微调:允许用户下载预训练模型,进行微调,适应特定任务,包括模型训练、微调、封装、验证、部署和监控。使用预训练模型并对其进行微调已经成为了 AI领域的标准做法,尤其是在 NLP中。这种方法结合了预训练模型的通用知识和特定任务的数据,从而获得了更好的性能。

5、模型共享:ModelHub和 Space为用户提供模型共享、代码分享和协作环境。鼓励开发者之间的开放合作,促进 NLP技术的快速发展。



图 3 大模型 PaaS 平台

传统的 PaaS 平台主要关注用户软件应用开发周期的加速, 通过开发工具的集成、硬件基础设施的自动管理、多租户应用共享基础资源和开发者多平台灵活访问的方案为企业和开发者提供便捷服务。大模型的高算力和高开发门槛要求 PaaS 平台更加关注大模型的开发部署流程的优化。参考目前市场中成功的厂家案例, 如 Google AI Platform、AWS SageMaker 和 HuggingFace 等, 这些厂家平台在部署大量基础设施资源的情况下为用户提供大模型快捷开发环境、大模型的全生命周期的监控调优, 同时也会提供一些预训练模型和数据集。大模型 PaaS 平台的上述功能优势也将为个人开发者和一些微小企业的 AI 应用开发提供便利, 大大降低大模型硬件基础设施的购买运维成本和搭建复杂的基础开发环境的时间精力消耗。

2.1.5 智算中心

智算中心是基于最新人工智能理论, 采用领先的人工智能计算架构, 提供人工智能应用所需算力服务、数据服务和算法服务的公共算力新型基础设施, 通过算力的生产、聚合、调度和释放, 高效支撑数据开放共享、智能生态建设、产业创新聚集, 有力促进 AI 产业化、产业 AI 化及政府治理智能化。

智算中心作业环节是智算中心的支撑部分, 智算中心通过作业环节实现了算力的生产、聚合、调度和释放, 是区别于其它数据中心的具体体现。功能部分是四大平台和三大服务, 四大平台分别是算力生产供应平台、数

据开放共享平台、智能生态建设平台和产业创新聚集平台；三大服务分别是数据服务、算力服务和算法服务，目标是促进 AI 产业化、产业 AI 化及政府治理智能化。

智算中心通常采用三方主体协作的投资建设运营模式：

1、投资主体：智算中心建设通常采用政府主导模式，政府作为投资主体加快推进智算中心落地，以智算中心为牵引打造智能产业生态圈，带动城市产业结构优化升级，增强城市创新服务力。

2、承建方主体：智算中心建设通常选择政府主导下的政企合作模式，由企业具体承建智能计算中心。

3、运营主体：运营主体为具体负责智算中心投入使用后的运营服务机构。

AI 智算中心不仅是一个高效的计算中心，更是一个综合性的创新平台，它结合“平台+应用+人才”的三合一策略，为新型AI产业的繁荣提供强大的算力支持、实际应用开发的鼓励，以及顶尖AI专家的培养和吸引。此外，中心还强调“算力+生态”的双轮驱动，通过持续的硬件投资和开放的AI生态合作，旨在吸引更多的企业和研究机构，从而推动AI全产业链的形成和快速发展。

2.2 基础大模型与相关技术

2.2.1 大模型研究发展迅速

2017年 Transformer模型提出并在机器翻译领域取得巨大成功后，自然语言处理大模型进入了爆发式的发展阶段。自 2018年以来，大型预训练语言模型的发展经历了几个重要阶段和突破：2018年，Google发布了 BERT模型，引领了自然语言处理领域预训练范式的兴起；2020年，OpenAI发布了 GPT-3模型，展示了强大的文本生成能力和在少量标注任务上的优秀表现，然而基于提示词学习的方法并未在大多数任务上超越预训练微调模型；2022 年 11月，ChatGPT 的问世展示了大语言模型的潜能，能够完成复杂任务并在许多任务上超越有监督模型。这一突破表明大型语言模型在复杂任务上的潜力。

大语言模型的实现细节和训练过程仍存在许多复杂性，对研究人员提出了挑战。同时，大语言模型的发展也带来了一些挑战和争议，关于数据隐私、模型偏见和滥用等问题引发了广泛讨论。为了解决这些问题，研究人员和机构开始探索模型透明化、可解释性和模型治理方法。

更多的具备多模态功能的大模型也将很快推出，例如 Google的 Gemini，OpenAI的 Gobi，开源的 NExT-GPT等。多模态大模型的视觉功能会带来潜在的法律安全风险，这些潜在的风险会延缓多模态大模型的推出进度。

2.2.2 大模型与小模型将持续并存

大模型与中小模型在未来几年会并存。尽管大模型当前表现优异，但对于各行业使用者来说，实际应用于业务场景仍然存在较高的技术和成本门槛。从业务层面分析，一定会出现资源配置更加高效的小模型，例如细分领域的专用小模型。不仅仅存在大模型和小模型的融合使用，大模型的小型化，以大模型为底座的小型化微调，也是一种趋势，这种方式能够以低廉的成本解决大量的业务问题。

“大和小是一个相对的变化。”。当前大模型的参数标准并不统一，相对于参数量级，模型的效果且是否能够支持快速迭代对于用户实际应用来说更为重要。用户能够在一个白盒大模型基础上快速地、低成本地微调和迭代出定制化的小模型，才能高效地实现丰富场景的大模型应用。模型需要持续迭代，表明了 AI基础软件工具链的重要性。

2.2.3 大模型的基础理论与设计

2.2.3.1 大模型网络架构的发展

当前主流大模型是基于Transformer架构进行设计的。传统的Transformer架构通常具有二次计算复杂性，在上下文长度较长时，训练和推理效率已成为一个重要问题。为了提高效率，一些新的语言建模架构被提出来，例如 RWKV，RetNet等。

- Transformer，由于其架构的出色并行性和容量，使得将语言模型扩展到数百亿或数千亿个参数成为可能，Transformer 架构已成为开发各种大模型的事实标准骨干。一般来说，主流大模型架构可以分为4种类型，即 Decoder-Only、Encoder-Only、Encoder-Decoder和MoE。Decoder-Only，典型代表是 GPT 和 LLaMA 等模型，Encoder-Only的典型代表是 BERT 和 ALBERT 等模型，Encoder-Decoder的典型代表是 T5 和 BART 等模型；值得注意的是，即使GPT-4的技术细节未公开，业界的广泛认知是其使用了MoE架构。

- RWKV，结合 Transformer 和 RNN 的优势，训练时能够像 Transformer 那样并行计算，推理时又能像 RNN 那样高效。高效推理，对于降低模型成本，尤其是在端侧部署有重要意义。RWKV 的计算量与上下文长度无关，对于更长的上下文有更好的扩展性。和 RNN 一样，历史信息是靠隐状态(WKV)来记忆的，对于长距离历史信息的记忆不如 Transformer，如何设计提示对模型的性能会有很大影响。

- RetNet，作为全新的神经网络架构，同时实现了良好的扩展性、并行训练、低成本部署和高效推理。在语言建模任务上 RetNet 可以达到与 Transformer 相当的困惑度(perplexity)，推理速度提升 8.4倍，内存占用

减少 70%，具有良好的扩展性，并且当模型大小大于一定规模时，RetNet 的性能表现会优于 Transformer。这些特性将使 RetNet 有可能成为 Transformer 之后大语言模型基础网络架构的有力继承者。

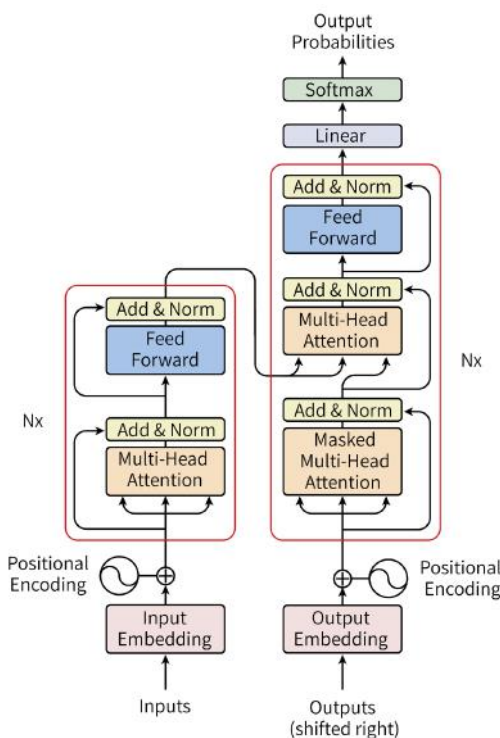


图 4 Transformer 网络架构

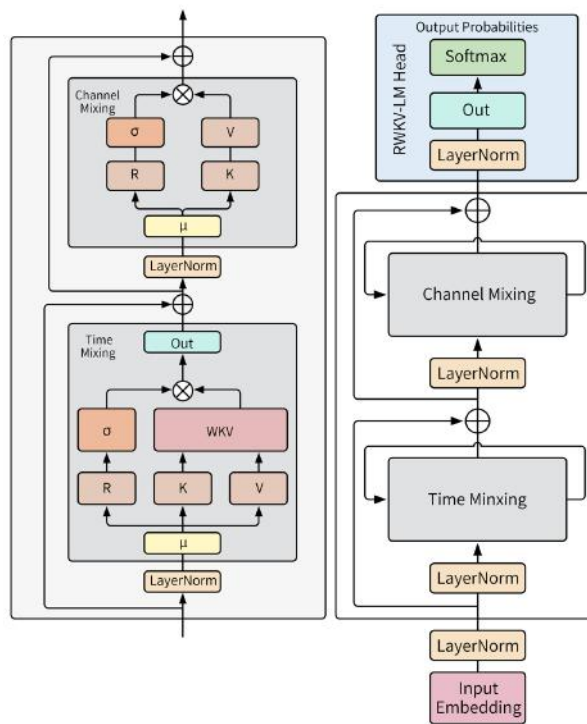


图 5 RWKV 网络架构

2.2.3.2 大模型的训练目标多样化

基础大模型是对世界知识的压缩，从基础模型到通用模型，模型的性能的构建主要来源于包含四个阶段：预训练、指令微调、奖励建模和对齐微调。这四个阶段分别需要不同规模的数据集，采用不同的训练目标，从而训练得到不同尺寸的模型，因此所需要的计算资源也有非常大的差别。

- 预训练，在将大规模语料库中的通用知识编码到庞大的模型参数中起着关键作用。对于训练大模型，有两种常用的预训练任务，即语言建模和去噪自编码。
- 指令微调，目标是增强（或解锁）大语言模型的能力，是一种提高大语言模型能力和可控性的有效技术。使用格式化的实例以有监督的方式微调大语言模型（例如，使用序列到序列的损失进行训练）。指令微调后，大语言模型展现出泛化到未见任务的卓越能力，即使在多语言场景下也能有不错表现。
- 奖励建模，目标是构建一个模型，用于进行文本质量评价。在使用场景中，指令微调模型会根据一个提示词，生成多个不同结果，然后由奖励模型进行质量排序。
- 对齐微调，目标是将大语言模型的行为与人类的价值观或偏好对齐。与初始的预训练和指令微调不同，语言

模型的对齐需要考虑不同的标准（例如有用性，诚实性和无害性）。已有研究表明对齐微调可能会在某种程度上损害大语言模型的通用能力，这在相关研究中被称为对齐税。对齐微调是一项具有挑战的工作。现有的很多开源大模型只做到指令微调，都没有做到对齐微调。

2.2.3.3 Scaling Law 的指导意义

OpenAI于2020年最先引入了语言模型缩放法则，他们认为,增加模型大小比增加数据大小更重要。DeepMind于2022年提出几乎完全相反的观点:以前的模型明显训练不足,增加训练数据集的大小实际上会带来更好的性能提升。

影响模型性能最大的三个因素：计算量、数据集大小、模型参数量。当其他因素不成为瓶颈时，这三个因素中的单个因素指数增加时，Loss会线性地下降。

- OpenAI观点：最佳计算效率训练是在相对适中的数据量上训练非常大的模型并在收敛之前Early Stopping。影响模型性能的三个要素之间存在幂指数的关系，每个参数并受另外两个参数影响。当没有其他两个瓶颈时，性能会急剧上升，影响程度为计算量 > 参数 >>数据集大小。训练要同时增大参数规模和数据集大小。大模型比小模型的样本效率更高，能以更少的优化步骤和使用更少的数据量达到相同的性能水平。

- DeepMind观点：模型太小时，在较少数据上训练的较大模型将是一种改进；模型太大时，在更多数据上训练的较小模型将是一种改进。

可以通过 Scaling Law进行模型性能的预测。随着模型规模和复杂性的大幅增加，很难预测模型性能的变化。通过开发更好的模型性能预测方法，或提出一些新架构，使资源的利用更加高效，训练周期加速缩短。一些可能的方法包括：训练一个较小的“种子”模型并推断其增长，模拟 Increased Scale 或 Model Tweaks 的效果，在不同规模上对模型进行基准测试以建立 Scaling Laws。使用这些方法可以在模型构建之前就洞察到模型的性能。

2.2.3.4 模型可解释性具有重要意义

模型的可解释性是指以人类可理解的方式解释或呈现模型行为的能力。随着大模型的不断进步，可解释性将变得极其重要，以确保这些模型具有透明性、公平性和益处。大语言模型内部机制仍然不明确，这种透明度的缺乏给下游应用带来了不必要的风险。因此，理解和解释这些模型对于阐明其行为、消除局限性和降低社会不利影响至关重要。

模型的可解释性从技术角度分为传统微调范式的可解释和提示范式的可解释。传统微调范式的解释，用于

解释个体组件所学习的知识或语言属性，解释大语言模型如何为特定输入做出预测。提示范式的解释，目标是用于理解大语言模型如何能够迅速从有限示例中掌握新任务，阐明对齐微调的作用，分析幻觉产生的原因。

为提高模型预测的理解度，帮助用户建立适当的信任，同时也有助于研究人员和开发者发现模型的潜在问题并改进性能，需要制定评估生成解释的度量标准，研究如何利用解释来调试模型和提高性能。

2.2.3.5 模型幻觉是一个高价值的研究方向

当模型生成的内容不遵循原文(与给定的输入或源内容不一致)或者和事实不符，就认为模型出现了幻觉的问题。数据质量、数据重复、数据不一致及模型对自身能力的高估是导致幻觉产生的重要原因。在文本生成等大模型应用中，减少幻觉是一个亟待解决的重要问题。

为减少幻觉，可从预训练、微调、强化学习等多个阶段对模型训练进行干预。预训练阶段可关注语料质量；微调阶段可人工检查数据；强化学习阶段可惩罚过度自信的回复。此外推理阶段，也可通过解码策略优化、知识检索、不确定度测量等方式缓解幻觉。尽管减少幻觉取得一定进展，但可靠评估、多语言场景、模型安全性等方面仍存在诸多挑战。总体来说，大模型幻觉的评估与缓解仍有待深入研究，以促进大模型的实际应用。

2.2.3.6 超级对齐

一些研究表明大语言模型能与人类判断高度对齐，在某些任务上甚至优于人类判断，让我们看到了超级智能实现的曙光。超级智能是一把双刃剑，有助于解决许多重要问题，同时也可能削弱人类的权力并威胁我们的安全。为了治理这些风险，急需建立新的治理机构并解决AI模型的对齐问题。

OpenAI于23年7月首次提出超级对齐的概念，认为人类目前无法可靠地监督那些比人类还聪明的人工智能系统。其将投入20%的计算资源，花费4年的时间全力打造一个超级对齐系统，意在解决超级智能的对齐问题。

构建超级对齐系统，由一系列的关键性工作构成：1. 开发一种可扩展的训练方法；2. 构建超级对齐系统，进行系统验证工作；3. 在构建超级对齐系统的过程中，对整个对齐流程进行压力测试。

虽然当前的技术进展与这个理想仍有差距，但我们有理由相信研究者们能开发出具有超级对齐能力的AI系统。

2.2.3.7 多模态大语言模型

多模态大语言模型(MLLM)是近年来兴起的一个新的研究热点,它利用强大的大语言模型作为大脑来执行多模态任务。其中,OpenAI宣布 ChatGPT新增了图片识别和语音能力,使得 ChatGPT不仅可以进行文字交流,还可以给它展示图片并进行互动,这是 ChatGPT向多模态进化的一次重大升级。OpenAI联合创始人,ChatGPT架构师 John Schulman认为,添加多模态功能会给大模型带来极大的性能提升,“如果扩展出现边际收益递减,那么添加多模态就能让模型获得文本中无法获得的知识,并有可能掌握纯语言模型无法完成的任务。例如,通过观看与物理世界甚至是与电脑屏幕互动的视频,模型能获得巨大收益。”

从发展通用人工智能的角度来看,MLLM可能比 LLM向前更近了一步。MLLM更符合人类感知世界的方式,人类能够自然地接受多感官输入,这些输入往往是互补和合作的。因此,多模态信息有望使 MLLM更加智能。得益于多模态输入的支持,用户可以用更灵活的方式与智能助手进行交互;MLLM是一个更全面的任务解决者。虽然 LLM通常可以执行 NLP任务,但 MLLM通常可以支持更大范围的任务。

目前的 MLLM在感知能力方面仍然有限,导致视觉信息获取不完整或错误,这可能是由于信息容量和计算负担之间的折衷产生的。MLLM的推理链很脆弱,改进多模态推理的主题值得研究。MLLM的指令跟随能力需要升级,指令调整可能需要涵盖更多的任务来提高泛化能力。幻觉问题很普遍,这在很大程度上影响了 MLLM的可靠性,需要通过更高效的参数训练优化。

2.2.4 大模型的计算与推理

2.2.4.1 分词算法与分词器

分词算法与分词器作为大语言模型的基础组件,是将字符序列转化为数字序列,起到文本与模型间桥梁的作用。分词器决定了大语言模型的词表大小、文档压缩率,并直接影响模型的训练和推理效率。

分词算法大致经历了从 Word/Char到 Subword的进化,当前的主流分词算法是 BPE、WordPiece、Sentencepiece和 Unigram等算法。

理想的分词器具有如下特性:无损重构,分词结果应该可以无损还原为输入;高压缩率,词表大小相同时,同一批数据的 tokens数应该尽可能少;语言无关,基于统计、训练和分词过程都不应引入语言特性;数据驱动,可以直接基于原始语料进行无监督训练;训练友好,能够在合理的时间和配置上完成训练过程。

2.2.4.2 注意力机制及计算

注意力机制是Transformer的关键组成部分。它允许序列中的标记相互交互，并计算输入和输出序列的表示。自注意力机制的时间和存储复杂度与序列的长度呈平方的关系，占用了大量的计算设备内存并消耗大量计算资源。因此，如何优化自注意力机制的时空复杂度、增强计算效率是大语言模型需要面临的重要问题。

- 全注意力。在传统的Transformer 中，注意力机制以成对的方式进行，考虑序列中所有标记对之间的关系。同时，Transformer使用多头注意力而不是单一注意力，将查询、键和值分别投影到不同头部的不同投影上。每个头部输出的连接被视为最终输出。

- 稀疏注意力。全注意力的一个重要挑战是二次计算复杂度，在处理长序列时会带来负担。因此，提出了各种高效的稀疏注意力来减少注意力机制的计算复杂度，每个查询只能根据位置关系关注标记的子集，而不是整个序列。

- 多查询/分组查询注意力。多查询注意力是指不同头部在键和值上共享相同的线性变换矩阵的注意力变体。它可以显著减少计算成本，只牺牲少量模型质量。具有多查询注意力的代表性模型包括PaLM 和StarCoder 。分组查询注意力在多查询注意力和多头注意力之间进行权衡，头部被分配到不同的组中，属于同一组的头部将共享相同的变换矩阵。特别地，分组查询注意力在LLaMA 2模型中得到了采用和经验验证。

- FlashAttention。与大多数现有的近似注意力方法不同，这些方法在提高计算效率的同时牺牲了模型质量，FlashAttention从IO感知的角度优化了GPU上注意力模块的速度和内存消耗。FlashAttention作为CUDA中的融合核心实现，已经集成到PyTorch、DeepSpeed和Megatron-LM 中。更新的FlashAttention-2 进一步优化了GPU线程块和warp的工作划分，相比原始FlashAttention，速度提高了约2倍。

- PagedAttention。将每个序列划分为子序列，并将这些子序列的相应KV缓存分配到非连续的物理块中。分页技术提高了GPU利用率，并实现了并行采样中的高效内存共享。PagedAttention解决了因输入长度经常变化，导致碎片化和过度预留问题。

- 扩张注意力。设计原则是随着token之间距离的增长，注意力分配呈指数级下降。因此具有线性的计算复杂性和对token之间的对数依赖性，可以解决有限的注意力资源和token可访问性之间的矛盾。

2.2.4.3 大模型预训练

预训练在大语言模型编码一般知识方面起关键作用，是大模型获取能力的基础。通过在大规模语料库上进行预训练，大语言模型可以获得基本的语言理解和生成能力。在这个过程中，预训练语料库的规模和质量对于

大语言模型获得强大的能力至关重要。

从一些数据上可以看出，模型预训练是一项高成本的工作，需要不断进行优化。例如：GPT-3 175B单次训练花费460万美元；训练PaLM两个月左右耗费约3.4Gwh；GPT-3 175B 训练了4990亿个Token；OpenAI训练集群包括285k CPU和10k High-End GPU。

随着语言模型参数量和所需训练数据量的急速增长，单个机器上有限的资源已无法满足大语言模型训练的要求。需要设计分布式训练系统来解决海量的计算和内存资源要求问题。在分布式训练系统环境下需要将一个模型训练任务拆分成多个子任务，并将子任务分发给多个计算设备，从而解决资源瓶颈。但是如何才能利用包括数万计算加速芯片的集群，训练参数量千亿甚至是万亿的大规模语言模型？这其中涉及到集群架构、并行策略、模型架构、内存优化、计算优化等一系列的技术。

训练数十亿参数的大语言模型通常是一个高度实验性的过程,需要进行大量的试错。随着模型和数据的规模增加，有限的计算资源下高效地训练大语言模型变得具有挑战性。有两个主要的技术问题需要解决，即提高训练吞吐量和加载更大模型到显存中。当前的优化方案包括3D并行，ZeRO 和混合精度训练。

2.2.4.4 大模型的推理优化

大语言模型推理面临计算资源的巨大需求和计算效率的挑战。大语言模型的推理速度每提高 1%，都将比谷歌搜索引擎推理速度提高 1% 的经济价值（-- 英伟达 Jim Fan）。优化推理性能不仅可以减少硬件成本，还可以提高模型的实时响应速度。

大模型推理主要是考虑延迟和吞吐量。模型推理一般是自回归型的任务，往往是显存密集型的任务，除了模型占用显存外，KV cache本身也会占用大量的显存；大模型太大的时候，单机无法存放，这时候需要分布式推理。

主流推理框架有vLLM、Text Generation Inference、FasterTransformer。推理的计算优化有算子融合，高性能算子编写。推理的分布式优化有Tensor并行，Pipeline并行等。低精度优化有FP16、INT8、INT4量化推理，Weight Only量化等。推理算法优化可以通过去除无效算子，减少不必要的算子执行等方式；批量推理优化可以使用Continuous Batch，Dynamic Batch等方式；解码方式优化有投机解码，多解码头解码（美杜莎）等。

2.2.4.5 上下文窗口扩展

上下文长度是大模型的关键限制之一。大型上下文窗口可让模型更加准确、流畅，提升模型创造力。大部

分模型的长度为 2k, Claude扩展到了 10k, LongNet更是将上下文长度扩展到了 10亿。

增加上下文长度, 可以从不同角度进行实现: 更好外推能力的位置编码, 注意力计算优化, 模型参数的条件计算, 增加 GPU显存等。

通过位置插值扩展大语言模型上下文窗口相对容易。位置插值通过小代价的微调来显著扩展大模型的上下文窗口, 在保持原有扩展模型任务能力的基础上, 显著增加模型对于长文本的建模能力。另一个优势是, 通过位置插值扩展的模型可以充分重用现有的预训练大语言模型和优化方法, 这在实际应用中具有很大吸引力。使用原模型预训练数据规模大约 0.1%的代表性样本进行微调, 就能实现当前最佳的上下文窗口扩展性能。

2.2.4.6 模型压缩

以 GPT-175B模型为例, 它拥有 1750亿参数, 至少需要 320GB(以 1024的倍数计算)的半精度(FP16) 格式存储空间。此外, 为了有效管理操作, 部署该模型进行推理至少需要五个 A100 GPU, 每个 GPU配备 80GB内存。巨大的存储与计算代价让有效的模型压缩成为一个亟待解决的难题。

大模型压缩技术的最新进展, 主要分布在模型剪枝、知识蒸馏、低秩因式分解、模型量化等领域。在进行大模型压缩时, 会采用其中一种或多种组合方案。其中, 将低秩因式分解应用于压缩大模型方面可能会取得进展, 但似乎仍需要进行持续的探索 and 实验, 以充分利用其对大模型的潜力; 并非所有的量化技术都适用于大语言模型。因此, 在我们选择量化方法时需要谨慎考虑。

大模型压缩评估的主要衡量标准, 是对比未压缩大模型的压缩有效性、推理效率和准确性。模型压缩的核心指标包括模型的型号尺寸、压缩率、推理时间、浮点运算等, 分别从模型的磁盘或内存空间占用, 性能不变时的有效压缩占比, 推理时处理和生成输入数据的响应时间, 处理输入数据时浮点数的运算量这些方面进行指标衡量。

2.3 大模型与训练、评测数据

大模型与数据的相互作用确保了模型的初始性能, 并且可以通过数据对大模型进行微调以使其适应新的任务, 这同时驱动了整个 AI生态系统, 包括硬件、优化技术和数据处理等领域的不断进步。大模型和训练数据共同塑造了 AI的性能、适应性和实际应用价值。

2.3.1 训练用数据

在 AI 领域的百模大战中，大型语言模型的训练成为了一个关键的竞争领域。数据、算法和算力作为大模型训练的三驾马车，在这场竞争中发挥着至关重要的作用。其中，数据集作为大模型训练的基石，对于模型性能 and 创新能力具有关键影响，尤其是数据质量问题更是不可忽视。在当前的技术背景下，大模型的训练数据通常汲取于多种渠道，具体如下：

1、开源数据集：各个研究领域都存在一些广为人知的开源数据集，如图像领域的 ImageNet、MNIST，或文本领域的 Wikipedia 数据集。这些数据集由学术界、研究机构或企业提供，为大模型提供了丰富的基础数据。

2、数据合作交流：众多的企业、研究机构和学者掌握着宝贵的数据资源，并在某些情境下愿意与外部实体共享。以医疗领域为例，一些医疗机构持有大量医疗影像数据，这些数据可以被用于图像解析或特定疾病的检测。

3、互联网规模数据：在我们日常使用的大型网络服务中，服务提供者往往会收集用户数据，这包括但不限于搜索记录、浏览活动、地理位置以及社交互动数据。

同时，为了满足大模型的高质量数据需求，弱监督技术和数据合成技术被引入：

- 弱监督技术利用少量标注数据和大量未标注数据生成训练样本：弱监督学习位于有监督学习和无监督学习之间，主要利用不完全或模糊的标签，而不是完全标注的数据。与传统的有监督学习需要为每个样本提供明确标签不同，弱监督学习可以利用少量标注数据和大量未标注数据。这种方法在现实中尤为重要，因为获取大量标注数据既昂贵又耗时。其工作原理包括使用启发式或基于规则的方法为未标注数据生成标签，利用半监督学习的方式，或通过多实例学习在知道集中存在正例的情况下进行学习。

- 通过数据合成技术模拟或生成新的数据点，增强数据集的多样性和规模：数据合成技术通过算法生成数据，而非传统收集方式，尤其适用于原始数据难获得或涉及隐私的场景。它能解决数据不足问题，增强数据集多样性，提升模型泛化并保护敏感信息。常用方法包括：模拟技术，如在医学图像中模拟病理情况；使用生成对抗网络 (GANs) 让两网络竞争生成数据；以及数据增强，对原数据进行变换如旋转或缩放以创造新样本。

在数据的使用上，企业必须确保数据的合法性、隐私和安全。数据收集、处理和使用应遵守法律，保护用户隐私，在追求技术创新的同时，确保数据的合法性、隐私保护和伦理问题的考量也应当得到足够重视，数据来源的知识产权已经成为大模型发展的一个问题。同时，数据和隐私的平衡是大模型应用面临的一个重要问题，用于生成式人工智能大模型的预训练、优化数据，应符合相关法律法规的要求，不含有侵犯知识产权的内容，包含个人信息的应符合“告知-同意”原则等要求。此外，企业应遵循国家和地方的数据法规，尤其是在数据跨境传输时，需要定期审查数据管理活动的合规性。

与此同时，大模型的参数量与训练数据量两者之间存在一种微妙的相互依赖：大模型拥有广阔参数空间，可以揭示数据背后的信息，但恰是因为这种广阔性，它们需要大量、多样化且有代表性的数据来防止过度拟合并确保模型的泛化能力。一个充足且高质的数据集是大模型真正发挥潜力、避免误导和实现真正业务价值的关键。

训练大模型各阶段所需的数据有着不一样的要求。在**预训练阶段**，数据需要广泛和多样，以促进模型对多种结构和模式的学习，为大模型打下良好全面的基础。进入**微调或任务特定训练阶段**，数据需要高度相关和有代表性，确保模型能够专注于特定任务的细节和特征。在**验证与测试阶段**，数据集应当独立、多样且真实，可以全面评估模型在未知数据上的性能。在整个流程中，数据的质量、时效性和完整性始终是关键，一个模型无论结构多么先进，输出的质量将始终基于输入数据的质量。

2.3.2 大模型评测及评测数据

评估大模型的通用能力不仅是对其在特定任务上的性能进行度量，还应当探究大模型在广泛、多样化的任务和场景中的适应性和鲁棒性。**多任务学习评测**能够检测模型是否能在多种任务上保持其性能，从而真实地反映其泛化能力；**零样本或少样本学习**评估可以揭示模型在面对少量或没有标注数据的任务时的快速适应性；**对抗性测试**可以评估模型对输入扰动的鲁棒性；**跨域和跨语言评测**可以考察其在不同环境或文化背景下的表现；**模型的解释性和可视化评测**可以提供模型决策过程的透明度，确保其不仅仅是一个‘黑箱’。这些评测维度共同构成了大模型通用能力的全景。

大模型的评测模式也有多种：通用数据集的选择题评分；GPT4 更优模型评分；竞技场模式评分；单项能力的评分；通用测试的场景测试评分。评测数据分为通用数据和场景数据。但目前大模型的评测任务仍然缺少统一标准：

- **评测榜单的数据多样性**：国内众多评测榜单，如SuperCLUE、OpenCompass和智源的FlagEval，虽然在某些数据集上有所交集，例如C-Eval、CMMLU和MMLU，但它们也都有各自独特的数据集特点。这种多样性意味着每个榜单都可能对模型的某些方面进行更深入的评估。
- **评测策略的多变性**：同一数据集可能因为评测策略的不同而导致模型得分的巨大差异。例如，OpenCompass和智源的FlagEval在Qwen-Chat数据集上的评测方法可能存在细微差别，从而导致了不同的评测结果。
- **评测得分的真实性挑战**：一些评测题目，特别是选择题，可能并不完全反映模型的真实理解能力。模型可能因为在预训练阶段接触过相关内容，或者掌握了某些应试策略，而在这些题目上获得高分。

● 人工评测的主观性：尽管基于竞技场的评测方式试图实现公正性，但其仍然受到人工评价的影响。人的评价往往带有主观性，这可能会对评测结果产生不同程度的偏见。

大模型评测的核心目的是确定模型的“聪明”程度，深入探讨其性能、特点和局限性，为行业应用提供方向。通过评测，我们可以更好地了解模型的性能、特点、价值、局限性和潜在风险，并为其发展和应用提供支持。

尽管评测是基于涵盖广泛类别的综合测试集，如《麻省理工科技评论》的600道题目，涵盖了多个类别和子类别；或是IDC的多层次评测方法，将大模型分为服务生态、产品技术和行业应用三层，去评测每一层的能力。但由于大模型领域和应用广泛，不同的领域和应用需要不同的评测标准和方法，大模型的评测仍面临着诸多挑战。例如，评测结果可能会被用于营销的工具，从而导致测评的真实意义被忽视；开源和闭源之间的选择权带来的公正性问题，开源测评可能会导致受试模型提前训练以提高分数，而闭源测评可能会引发对评测的公正性的质疑；并且，目前行业内缺乏统一评测标准，尚未出现一个广泛认可的大模型评测标准或方法，各评测机构和组织可能会提出不同的评测标准和方法。

尽管如此，行业普遍认为评测为用户提供了选择大模型的选择参考，并期待大模型的评测技术可以综合评估大模型，在技术性能、行业应用、安全性和行业认知等多个维度。

2.4 生成式 AI应用开发技术

生成式 AI 技术落地企业需要围绕大模型进行应用的开发，随着大模型相关应用开发流程逐渐标准化，应用中不同功能的组成部分逐渐被抽象成大模型的应用组件，这类模块化的组件易于添加和更改，能够快速敏捷地根据场景需要进行组合及适配，每种组件都有与之对应的技术。一个大模型应用在设计时除大模型本身外可能会用到三类技术：提示工程类，企业私域知识管理和应用类(包括向量库、知识图谱、微调、文本处理等)，以及应用框架类(包括思维链 CoT、智能体 Agent 等技术)。

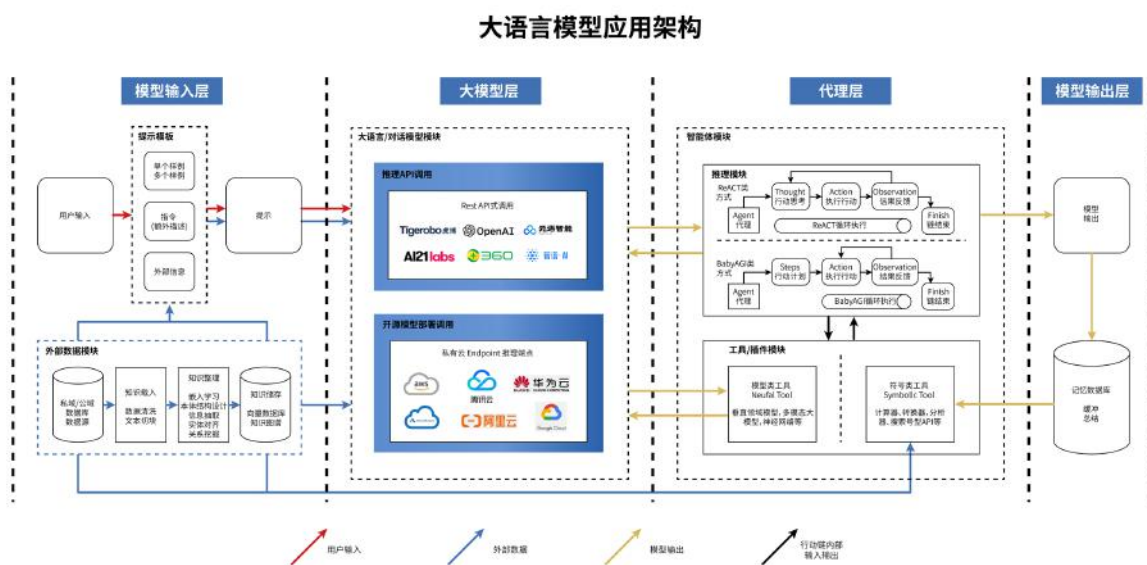


图 6 大语言模型应用架构

2.4.1 提示工程

提示工程 (Prompt Engineering) 是一门较新的学科，关注提示词开发和优化，帮助用户将大语言模型用于各场景和研究领域。提示是一种自然语言文本，要求生成式 AI 执行特定任务，其每个组成部分都会对最终输出产生影响，因此提示需要模板来进行设计与编排，用于帮助模型生成理想的输出。提示模板整合了模型任务指令与额外描述、外部知识、样例、以及用户输入来生成最终输入模型的提示文本内容：

- **指令**：指令是对任务的明确描述，激发模型解决对应任务的能力。通过对指令的理解，大模型才能生成准确的回答，因此指令是提示模板中最为重要的组成部分；
- **额外描述**：在复杂的应用场景下，除了指令之外，还需要对任务补充更多的额外描述，对模型输出需要添加一些额外的限制和规则，从而避免模型生成不符合意图的输出，也能对模型生成物的安全性进行控制；

- 外部信息：在提示模板的设计中，可以通过提供外部信息来增强大模型的知识，让大模型有能力回答模型自身认知范围外的问题。外部知识的使用不需要改变大模型自身的权重参数，因此相比于通过微调来增强模型的知识节省了训练用的资源；

- 样例：在提示模板中，可以加入输入和期望输出的样例组，让模型模仿其输出的格式。相比于不添加样例的零样本(Zero-Shot) 上下文学习，有效的单样本和少样本学习(One-Shot / Few-Shot Learning) 的输出更加规范；

- 用户输入：即使意图相同，不同的用户的输入之间也可能存在较大的差异性。在构建 AI应用的工作流程时，有时需要先判断用户输入所对应的意图，然后再根据用户的意图选择对应的工作流程，用于整合编排出模型的提示；

提示工程能帮助模型生成规范且准确的输出，让不熟悉大模型交互机制的用户简单高效地应用大模型。通过设计提示模板，大模型也可以作为专家模型解决特定场景下的任务和问题，所消耗的资源成本相对较少，帮助企业敏捷并高效地进行大模型应用开发。使用提示模板在未来可能会成为根据场景设计大模型应用的主要方式。

2.4.2 企业私域知识管理和应用

企业落地生成式 AI应用时，常见的阻碍包括幻觉问题、知识欠缺问题以及数据安全问题：**1. 幻觉问题**：“幻觉”是指模型生成不正确，无意义或不真实的文本的现象。大模型的底层原理是基于概率，如果大模型的专业领域知识不足，就会出现“一本正经的胡说八道”现象；**2. 知识欠缺问题**：预训练模型所掌握的知识基于训练数据，除此之外的信息模型无从知晓。同样的，一些非公开数据，比如企业内部数据，应用内数据等，也是无法被预训练模型所知晓；**3. 数据安全问题**：企业的经营数据、合同文件等都是机密数据，机密数据的外泄将严重影响企业的经营。

针对这些阻碍，一种高效的解决方案是将企业的知识语料全部放在本地，适时适当的给模型注入所需要的知识素材，从而在保证数据安全的基础上控制输出内容。以下将分别介绍三种企业私域知识管理和应用的方法：向量化及向量数据库、结合知识图谱、和模型微调。

2.4.2.1 向量化及向量数据库

图像、文本和音视频等非结构化数据都可以通过数据提取 (Loading)、分块 (Chunking)和嵌入学习转化为语义向量 (Embedding)，变成计算机可以理解的格式，这一过程被称为向量化。由此一来，两个语义的相似度

就转换成了两个向量的相似度，可以通过汉明距离、欧式距离或者余弦距离等数学方法计算，在查找知识素材时就能根据语义以及上下文含义查找最相似或相关的数据，而不是使用基于精确匹配或预定义标准查询数据库的传统方法。

向量数据库负责存储向量化之后的数据并执行矢量搜索，其主要特点是高效存储与高效检索。一方面，向量数据库也是一种数据库，除了管理向量数据外，还支持对传统结构化数据的管理。另一方面，向量数据库在向量检索方面支持多种近似最近邻搜索算法的使用，通过索引预先构建，加快检索速度。

在实际场景中，设计大模型应用时，向量数据库有两种使用方式：

一是用做外部知识库，存储企业中的非结构化数据。使用时，通过检索所需要的企业内部知识，以提示的方式注入给模型，就能通过大语言模型的上下文学习能力，加强模型的知识与信息范围。如何保证召回知识的完整性和准确性仍是该领域的难题；

二是用作记忆库，存储模型的输出结果。将大模型的交互记录存储起来，在需要的场景中，作为外部知识提供给大模型，就能让大模型学习到用户的行为和环境状态，获得记忆能力。

值得注意的是接入外部知识时，由于现阶段大模型对输入文本长度有一定的限制要求，所以对于长文本，不能一次性将所有内容输入给大模型。需要对文本进行分割，分别处理后合并。目前常见的三种方法包括 Map-Reduce方法，Refine方法和 Map-Rerank方法。

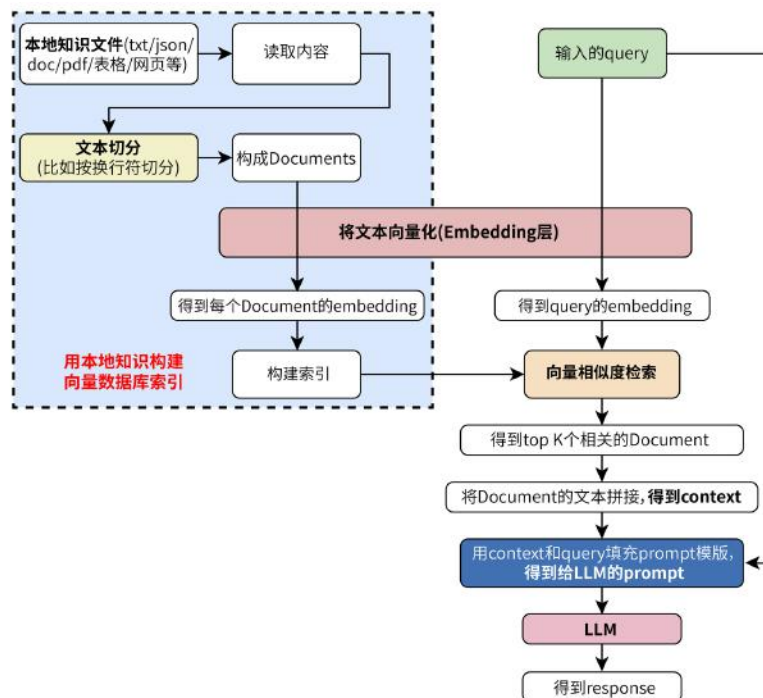


图 7 私域知识问答过程的简单实现

2.4.2.2 结合知识图谱

在人工智能的早期发展流派中，存在“符号主义”和“连接主义”两种主要流派。符号主义认为人脑的本质是一个符号推理系统，侧重模拟人的心智，研究怎样用计算机符号表示人脑中的知识并模拟心智的推理过程，其中知识图谱就是其典型代表。连接主义则认为智能活动是由大量简单的单元通过复杂的相互连接后并行运行的结果，侧重于模拟人脑的生理结构，也就是人工神经网络，大语言模型正是属于这个派系。

图谱概念最初被提出时，由于在设计、维护和标准化等方面的成本过高，这项技术并没有受到人们的广泛关注。而在大模型时代，人们在使用中发现大模型性能强大但难以被控制，而知识图谱这种结构化，高密度的知识表达形式，恰好适合弥补大模型推理与输出结果的低可解释性。同时，大模型可以在知识图谱的设计、维护和标准化等方面提供解决方案，降低成本。由此，大模型与知识图谱的组合，渐渐被引入各个应用场景。知识图谱与大模型结合主要有两种方式：

1、大模型辅助知识图谱生成：大模型既蕴含自然语言处理的能力又包含了大量的通用知识，因此可以辅助知识图谱搭建时必需的本体结构设计、信息抽取、实体消歧、实体对齐、关系挖掘等需要处理文本的地方，并减少构建、扩展和更新知识图谱过程所消耗的人工工作量。

2、知识图谱增强大语言模型：在给定文本输入的情况下，用大模型生成逻辑查询，然后在知识图谱上执行该查询以获取需要的子图，生成结构化上下文。最后，将结构化上下文与文本信息融合以生成最终的输出。如何精确完整的检索所需的所有子图仍然是一个巨大的挑战，这一点和向量化部分所面临的问题类似。

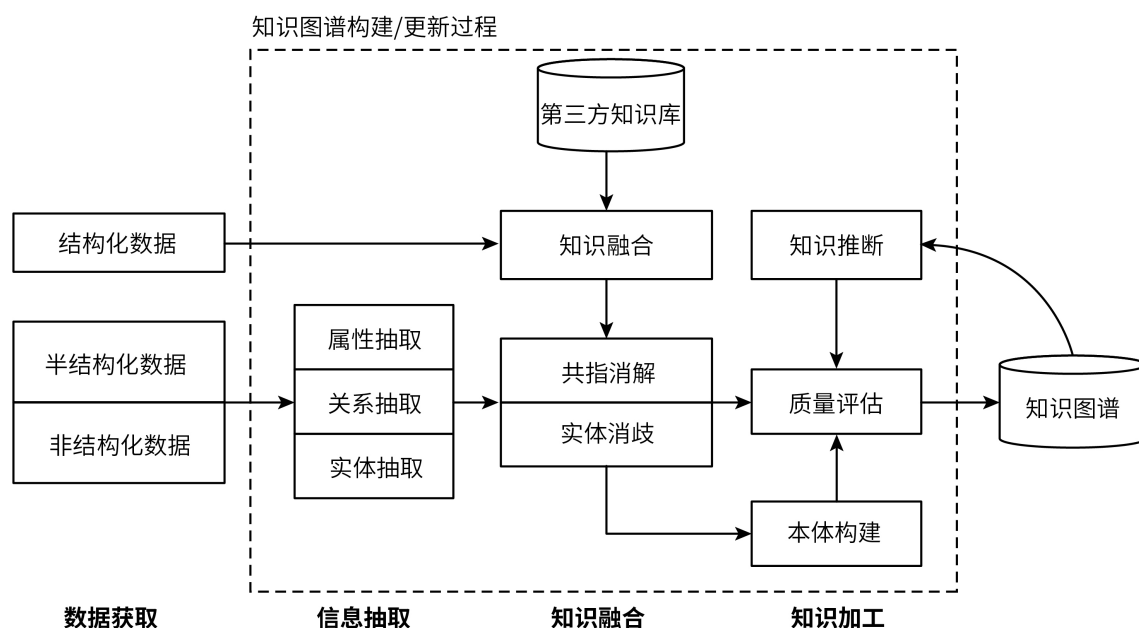


图 8 知识图谱的技术架构

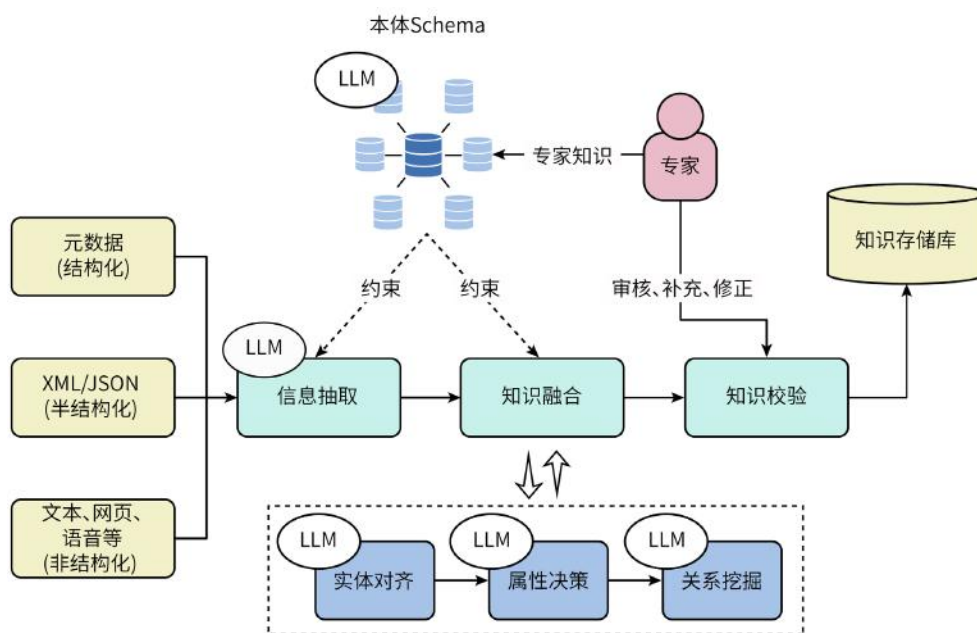


图9 大语言模型辅助知识图谱生成

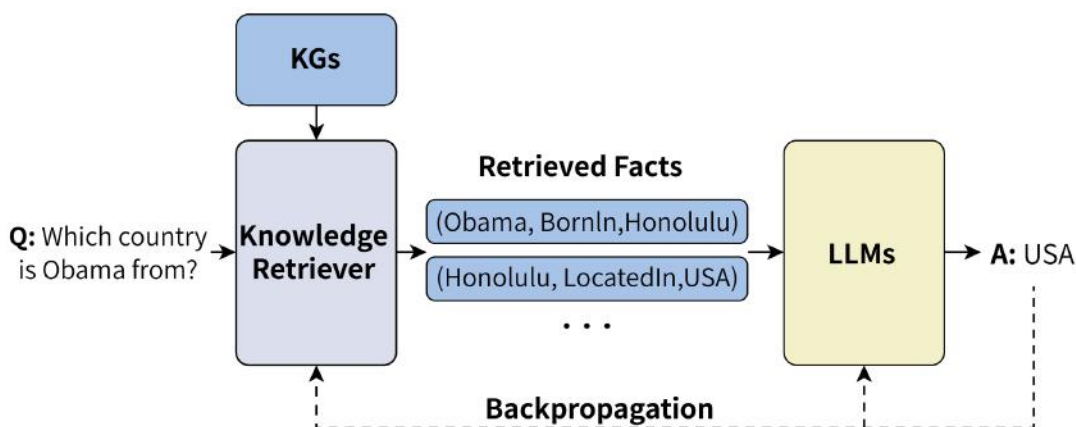


图10 用知识图谱检索增强大模型知识

2.4.2.3 模型微调

上下文学习让大模型能够通过检索知识库来获得某一特定知识或者拥有某一特定记忆。而如果通过监督微调的方式改变模型参数，就可以让模型长久学会某一知识或者适应特定领域场景任务。

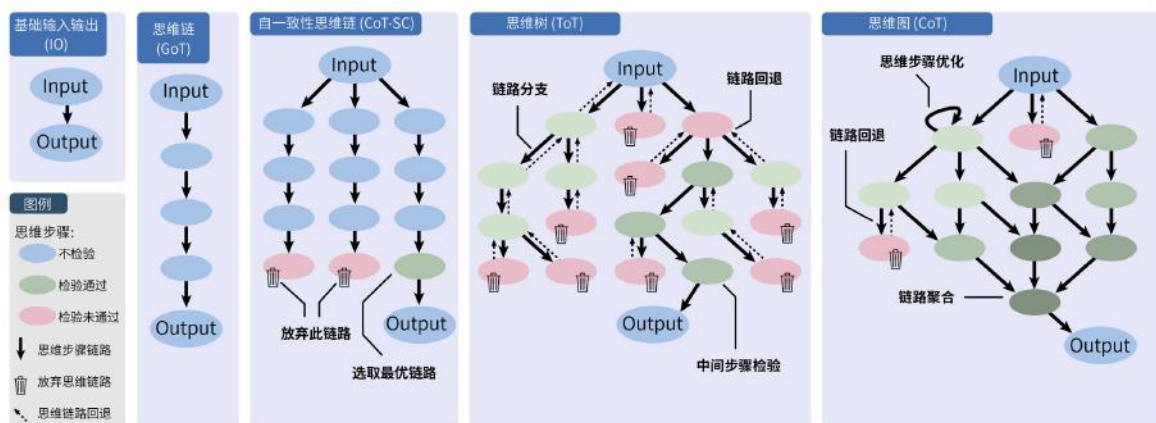
随着模型变得越来越大，在消费级硬件上对模型进行全部参数的微调变得不可行。此外，为每个下游任务独立存储和部署微调模型变得非常昂贵，因为微调模型与原始预训练模型的大小相同。参数高效微调（PEFT）方法旨在解决这两个问题，对比全参数微调，PEFT以微调少量可训练参数的方式，降低微调的硬件要求和计算成本。当下流行的PEFT方法包括Adaptor Tuning、Prompt Tuning、Prefix Tuning、P-Tuning等。但需要注意的是，如果微调过程中过于拟合微调数据，就会让大模型丧失原本的预训练知识能力，产生灾难性遗忘。

2.4.3 AI应用框架

2.4.3.1 大模型思维链

强大的逻辑推理能力是大模型“智能涌现”出的核心能力之一，让 AI 如同具备人的意识，而其中的关键在于思维链(Chain of Thought, CoT) 技术。在思维链中，问题被拆解为多个推理步骤，每一步的输出都作为下一步的上下文输入，从而使模型能够逐步推导出答案。此外，思维链可以基于零样本或少样本的方式来设计，提供有效样例的思维链能生成更稳定的推理步骤，适用于自由度较小的应用场景，但同时也受到模型理解能力和上下文窗口长度的限制。

思维链有助于解决各类复杂的问题，但受限于模型性能，模型并不总能得到正确的中间推理步骤，从而难以稳定地生成准确的最终结果，由此需要更加稳定的思维过程。基于思维链的概念，近几个月提出了许多基于思维链的大模型思维增强策略：



● **自一致性思维链 CoT-SC (Self-Consistency)**：单一路径的思维链难免会生成失败的中间推理步骤，并导致推理的失败。因此，生成多条推理路径，并从众多的推理结果中选择最一致的答案，能够提高模型输出的准确性与稳定性。这种方式需要大模型对推理结果进行有效的集成与统计。

● **思维树 ToT (Tree of Thoughts)**：通过将树状推理结构引入思维链，从而得到思维树。思维树的特点是在每个中间推理步骤都会引出不同的分支并检测其正确性，在失败的时候结束这个分支并在需要时退回到上一正确节点。在众多分支推理完成后，和多链路的 CoT 相似，从多个推理结果中选择最优的或者最一致的答案。

● **思维图 GoT (Graph of Thoughts)**：思维图是思维树的一个变种。思维树的每条分支推理路径和其他路径互不关联，是一种相对更加独立的推理方式，而思维图中每个的中间推理步骤都会结合上一层其他的推理步骤。这种思维方式的优点在于会综合考虑多方面的推理，从而在最终输出中能够得出更全面的分析。

以上这些思维链的加强策略主要是通过增加推理的路径和分支来提升模型推理的准确性和稳定性。但由于使用了多条思维路径，大模型的推理次数也会有所增加，在设计应用的时候需要根据场景选择性能与成本合适的推理方案。

2.4.3.2 大模型工具链

在大模型解决对应的任务时，有时候大模型并不具备完成特定任务的能力，或者难以通过自身获得准确有效的答案。因此在实际的应用落地时，大模型会经常需要外部工具的辅助来解决对应问题。通过将工具使用和思维链技术结合，从而得到以 ReAct为代表的基于思考+行动(Reasoning + Acting)的工具链。工具链通过观察在行动中使用工具得到的输出，结合宏观任务目标来思考并计划下一步骤的行动计划。在此处，工具也可以被视为处理特定任务的专家模块，大模型作为路由器将任务的每个步骤路由到最合适的专家模块来完成，这种模块化推理、知识和语言的系统最早由 AI21 Labs公司所提出，被称作 MRKL式神经符号架构，适用于复杂推理场景。在这种架构中，工具链所使用的工具可被大致分为两类：符号类(Symbolic)工具和神经类(Neural)工具。

- 符号类工具通常是函数或者 API接口，例如数学计算器、联网查询器、计时器等等。通过将代码封装成函数或者 API并对这些工具进行功能描述，就可以让大模型在行动中调用对应的工具，从而解决模型本身难以解决的问题。

- 神经类工具是一些参数规模通常更小、用于完成特定任务的小模型或者专业知识领域的大模型，例如 OCR模型、关键词提取模型、意图识别模型、图像识别模型等等，这些模型通用性不强，但在专业领域上的表现强于通用大模型。

对于实际的应用场景来说，工具链让模型拥有了解决自身能力外任务的可能性。工具的使用不仅可以拓展模型能力，也可以提升模型输出的可靠性，解决模型输出的稳定性问题、幻觉问题、可解释性问题等。

2.4.3.3 智能体 Agent

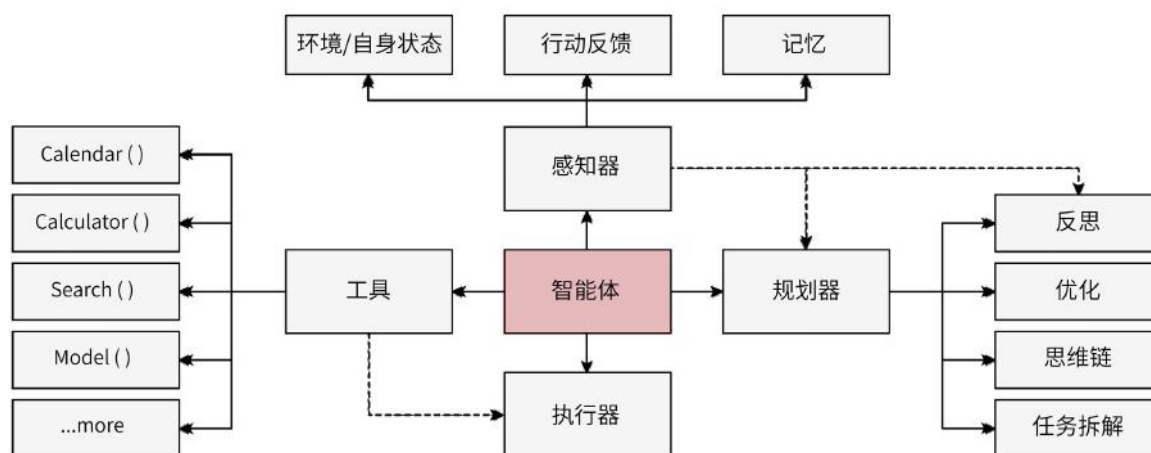


图 12 智能体 Agent 组件与结构设计

在大模型时代之前，智能体(Agent)的概念就已经出现在强化学习领域，能通过感知环境的状态并根据已有的行动策略选择对应的动作，并影响环境状态的变化。在大模型时代，智能体得到了大模型强大的生成能力的加持，对自身与环境的理解相比于传统的神经网络更加深刻，能够在自由度更高的情况下生根据宏观目标生成行动规划，从而应用于环境更加多变的应用场景中。

基础的智能体结构包含了三类主要功能组件：规划器、感知器、以及执行器。这三个组件的相互结合为智能体带来了强大的任务解决能力以及环境适应能力。

规划器：大语言模型驱动的规划器通过对宏观任务的理解、环境与自身的感知、以及先前行动结果反馈的记忆与解析，综合思考并规划出下一步骤的行动。其背后的设计方法类似于思维链的设计思路且有多种，例如：ReAct式的逐步规划、BabyAGI式的统一规划、多路径式的最优路径规划等等。在计划生成之后，后续可以通过观察行动执行结果对计划进行反思及优化(Reflection and Refinement)。规划器对模型文本理解与生成能力的要求较高，需要有良好的上下文理解能力、指令遵循能力和步骤推理能力，以便综合考虑所有相关信息、让思考过程易于解析以及将宏观目标解析成合理的中间步骤。

感知器：感知器为规划器提供了信息来源，这些信息可包括：

1、环境与自身的状态：对环境与自身的感知除了常规的文本模态、也可以是视觉、听觉或者其他模态的感知方式。为了让规划器能够充分利用得到的信息，需要将不同模态的信息统一成规范的文本模态。

2、行动的反馈：规划器通过解析行动结果来判断出行动是否成功达到预期，从而规划出下一步的行动。在这个过程中，需要感知器观察并解析每一步的输出，并从中提取关键的信息传递给规划器进行判断。

3、记忆：对于具有学习能力的智能体，通过收集先前步骤的计划、观察与反思结果能够帮助模型更好地适

应环境。通过先前步骤执行的完成度来辅助下次相似场景下的行动规划，能让后续行动的计划更加有效。

基于感知器所能提供的不同种类的信息来源，规划器才能不断地在环境中做出正确的判断与决策，生成适应环境的行动规划。

执行器：在智能体的工作流程中，计划中每步的行动都基于工具的使用。执行器需要将来自规划器的行动指令转化成工具的选择以及符合规范的工具输入，因为大部分工具难以直接解析模型输出的非结构化的文本。在行动完成后，需要执行器及使用的工具生成行动结果的反馈，并传递至规划器用于生成或调整下一步的行动规划。

智能体的组件在设计的过程中可以根据应用场景的需要选择不同的设计方法，就例如规划器可以使用逐步规划或者统一规划、单一路径或者多路径的方式，感知器可以基于不同的模态或不同的解析方式，执行器则可以使用现有的 API 接口或者自己设计的工具。设计者能根据场景需要选择合适的组件设计方案。

在自由度较高的应用场景中，可以通过让智能体获得学习能力来提升其环境适应性。智能体通过记录与对象交互的方式和结果，在未来处于相似的场景时能够做出更好的判断与行动。目前，智能体可以通过长期记忆和编排新工具的方式来进行学习。长期记忆的实现是通过向量数据库，记录多轮对话的模型输出与用户输入，智能体会回忆先前相似场景下的会话记忆，并综合用户的反应生成更加符合用户意图的输出。编排新工具则是根据任务、环境和自身状态，通过利用现有工具来构建有效的工具使用流程。将这些流程连同其功能描述一起储存到工具库中，以便智能体在未来的类似情境中能够重复使用这些新工具，这种方法也是智能体学习的一种方式，工具在此处可以由代码或者提示组成，都是通过编排基础的工具来创造新的工具。先前英伟达和多所美国大学联合研究的 Voyager 项目中就提到了这种能够让智能体在自由环境中根据宏观目标终身学习的方式。

2.4.3.4 多智能体的交互与协作

根据应用场景的不同，智能体的交互对象也会有所变化。现阶段智能体的交互对象大致有三种：环境、人类、智能体：

- 与环境交互：从环境中，智能体能够获得环境状态的信息，并规划出对应的行动并执行。同时，智能体的行动也能对环境造成影响。通过给予智能体操作权限，智能体能够根据自身对环境的判断做出对应的改变环境的动作。
- 与用户交互：人机之间的交互是智能体重要的应用方向之一，用户在其中扮演的角色在不同的应用场景下也有所不同。用户可以作为任务发布者，给智能体发布任务或者向智能体提问，而智能体需要把任务或问题作为有待完成的宏观目标并尝试完成。另一类应用场景是人类用户作为环境中的一部分，而不影响智能体的宏观目标，

在这种情况下智能体通常需要收集并总结多个用户的行动作为环境信息提供给智能体并做出相应的行动。

● 与智能体交互：在与智能体的交互场景中，智能体的交互对象包括智能体自己以及其他智能体。与智能体自身的交互更加类似于思维链，通过自问自答的方式引导自身达成复杂的任务目标。另一种则是多智能体间的交互，是目前关于智能体的重点研究方向之一，其中根据有无共同的宏观目标划分成两种交互方式：一种情况是，当智能体之间没有统一的宏观目标时，每个智能体都处于相对自由的状态，智能体的行动通常不会过多受到其他智能体的影响，例如斯坦福大学AI小镇的相关研究。这类场景通常有助于观察智能体间的协作，偏向于游戏场景以及生物生态与社会学的科学研究。而另一种则是智能体之间有着统一的宏观目标，近期热门的多智能体项目MetaGPT的核心实现方式就是通过多个智能体之间相互的协作来完成用户输入需求的理解，虽然每个智能体所分配到的任务会有所不同，但宏观的目标都是相似的：完成用户输入的需求。这对于涉及生产流程的场景而言有着极大的应用潜力。

在设计应用时，设计者需要根据场景制定智能体的交互对象与方式。其中人类用户也可以被视为特殊的智能体，智能体可以和人类协作共同完成由人类或智能体所发布的任务，主要根据场景的需要和智能体的性能而决定。

2.5 生成式 AI安全和监控

生成式 AI在近年来已经取得了显著的进展，但随之而来的是一系列的安全痛点和挑战。以下是一些主要的痛点及挑战：

1、对抗性攻击：

当我们谈论深度学习模型，特别是在图像识别、语音识别或自然语言处理等领域，它们通常被认为是可靠的。但是，这些模型也存在一种特定的脆弱性，即对抗性攻击。

对抗性攻击是一种特定的攻击方式，其中攻击者会对模型的原始输入进行微小的修改，这些修改对人类来说几乎是不可察觉的。但是，这些微小的扰动可以导致模型产生完全不同、甚至是错误的输出。这种现象是由于深度学习模型在高维空间中的决策边界可能存在微小的不规则性，对抗性攻击正是利用这些不规则性。

例如，在图像识别任务中，攻击者可以向图像添加几乎不可见的噪声，使得原本被正确分类的图像被误分类为另一个完全不同的类别。这种攻击的关键在于选择和产生对抗扰动，这些扰动是经过精心设计的，目的是使模型产生错误的判断。

对抗样本攻击的目标是向输入数据添加最少量的扰动，从而导致所需的错误分类。这种攻击方式对于验证

模型的鲁棒性和安全性非常重要，因为它揭示了模型在面对微小扰动时的脆弱性。

为了增强模型对抗对抗性攻击的鲁棒性，研究者和工程师采用了多种策略。其中，对抗性训练是一种策略，它在训练过程中同时使用正常数据和被扰动的数据，使模型能够识别并抵抗这些扰动。此外，模型集成方法通过组合多个模型的预测来提高鲁棒性，即使某个模型受到攻击，其他模型也可能做出正确预测。在数据输入模型前，可以进行预处理如去噪或平滑来降低对抗性扰动的影响。在训练过程中，加入正则化或约束也可以使模型更鲁棒。动态防御策略，如随机化技术，可以在模型推断时引入随机性，使攻击者难以精确预测模型行为。同时，模型也可以在预测前检测输入数据，以确定其是否为对抗样本，并据此决定是否处理。

2、模型窃取与反向工程：

模型的内部结构和工作原理往往是其核心竞争力。但随着技术的发展，攻击者也开始尝试复制或了解这些未公开的模型。以下是攻击者可能会采取的几种方式：

(1) 模型窃取：攻击者会尝试复制模型的功能，而不必知道其具体的结构或参数。这通常通过向模型发送输入并观察其输出来实现，然后使用这些输入/输出对来训练一个新的模型，使其模仿原始模型的行为。

(2) 黑盒攻击：在这种情况下，攻击者无法直接访问模型的内部结构或参数，但他们可以访问模型的输入和输出。攻击者可以利用这些信息来推测模型的工作原理。例如，基于查询的方法和基于模型迁移的方法是两种常见的黑盒攻击方法。

(3) 反向工程：攻击者尝试通过分析模型的输入/输出或其他可用信息来重建或了解模型的内部结构和工作原理。

(4) 隐私攻击：攻击者可能利用模型的输入/输出来计算出用户的训练数据或其他隐私数据。

这些攻击方法对于模型的所有者来说都是一个安全风险，因为它们可能导致知识产权的丧失、模型的滥用或隐私泄露。

为了抵御模型窃取与反向工程的威胁，专家提出了多种策略。比如通过引入差分隐私技术，在模型训练中加入随机性，确保模型的输出不会暴露过多关于单个数据点的信息，从而维护用户隐私。此外，通过模型混淆，对模型的权重和结构进行细微调整，虽然不影响其性能，但使攻击者难以复制。为了进一步加强安全性，他们还限制了对模型的查询次数，避免攻击者收集大量数据，并对模型的输出加入随机噪声，使其难以推断模型的内部结构。利用同态加密技术，能够在加密数据上执行计算，无需解密即可进行模型推断，确保了数据和模型的双重安全。同时，在模型中植入独特的水印，一旦模型被非法复制，都可以通过这个标识进行追踪。最后，通过定期监控模型的使用和输出，可以确保能够及时发现并应对任何异常或攻击行为。

3、数据隐私泄露：

随着机器学习和人工智能的广泛应用，模型的安全性和隐私保护成为了一个重要的研究领域。攻击者可能会利用模型的输出来推断其训练数据，从而获取敏感信息。这种攻击方式通常被称为信息提取攻击。

在实际应用中，例如，一个人脸识别模型可能会返回与训练数据相关的结果向量。攻击者可以通过这些向量来恢复或预测训练数据中的人脸图像，从而导致用户的肖像隐私泄露。此外，医疗数据、个人偏好和其他敏感信息也可能在模型中被泄露。

为了防止这种隐私泄露，研究者提出了多种防御方法，如模型结构防御、信息混淆防御和查询控制防御。但随着攻击技术的进步，新的隐私威胁也不断出现。模型在提供强大的功能的同时，也带来了隐私泄露的风险。为了保护用户隐私，需要在模型设计、训练和部署过程中采取适当的安全措施。

4、模型偏见与不公平性：

模型的预测和决策主要依赖于其训练数据。当这些数据存在偏见或不平衡时，模型可能会反映这些偏见，导致如招聘模型因性别或种族偏见而不公平地选择候选人，或金融和医疗领域的模型对某些群体产生歧视性待遇。此外，如果训练数据对某些群体的代表性不足，模型的决策可能不准确。更为严重的是，模型还可能放大数据中的社会偏见，导致更大的社会问题。为了解决这些挑战，研究者和开发者正在努力采取措施，如优化数据质量、预处理数据、以及进行公平性测试，确保模型决策的公平性和无偏性。

5、模型的不确定性与不可预测性：

模型在分析数据和预测时可能面临不确定性或不可预测的输出，这些不确定性可能因模型的训练数据、结构或其他因素而产生。特别是当模型在训练时没有接触到某些特定的输入数据时，其输出可能会变得不确定。这种不确定性可能导致用户对模型的信任度降低，尤其是在关键领域如医疗和金融中，不确定的输出可能导致严重的后果，如误诊或金融损失。尽管研究者和开发者正在尝试多种方法来减少这种不确定性，如优化模型结构、使用更多样的训练数据或采用不确定性估计方法，但要完全消除模型的不确定性仍然是一个巨大的挑战。

6、模型的可解释性问题：

大模型因其复杂的结构和众多参数，经常被看作“黑盒”。虽然它们在多种任务上有卓越的表现，但其决策机制往往难以解读和理解。这种模糊性可能引发多方面的问题：例如，当用户或决策者无法明确模型的决策逻辑时，他们可能对其输出持怀疑态度，特别是在如医疗和金融等关键领域。同时，对于某些需要遵循法规的应用，模型的决策可能必须受到审查或验证，而模型的不透明性使这一过程变得困难。这种不透明的决策可能掩盖了模型的偏见或错误，可能导致不公或不准确的结果，进而触发伦理和安全问题。为了应对这些挑战，研究者们正在努力研究提高模型可解释性的方法，希望使其决策过程更加清晰和易于理解。

2.6 生成式AI应用设计

2.6.1 AI应用的呈现方式

在设计完整个 AI应用的工作流程之后，需要将工作流的最终输出以某种方式呈现给用户并进行交互。基于应用场景的需求不同，AI应用的交互方式也会有所不同，目前有几种常见的呈现方式：软件、搜索框、弹窗与数字人、应用插件、虚拟账户等等。

- 软件 - 以软件为载体的 AI应用通常是将大模型能力以及配套的工作流程打包成一个 APP，并直接上架于各平台的应用市场当中。由于大模型目前对于算力仍然有较大的要求，所以以软件为载体的 AI大模型应用大多数是基于云端提供服务。

- 搜索框 - 搜索框通过用户的输入来检索对应的信息，并给予相应的反馈。通过将大模型的能力接入搜索框，大模型能够综合多条返回的搜索结果，并生成更加符合用户意图的答案。其中举例来说，LangChain官方使用文档网站上就应用了 AI智能搜索，通过一个搜索框将大模型与用户的问题对接，并通过查询知识库反馈出对应的答案。

- 弹窗与数字人 - 弹窗通常在各个网站上被用作客服，能够让大模型驱动的 AI智能客服机器人与用户进行交互。随着生成式 AI技术的发展，AI数字人技术随之出现且日益完善，通过一个虚拟形象与用户交互，能提供更加人性化的交互体验。

- 虚拟账户 - 在企业微信、钉钉、飞书等办公协同软件中，可以通过添加虚拟账户的方式来交互，虚拟账户背后接入的则是生成式 AI大模型。与用户交互之外，虚拟账户还可以设置在群聊当中，对群聊中的信息进行管理和分析，提供群聊信息的监控和摘要等等。

- 应用插件 - 大模型不一定要替代现有的工作流程，也可以在现有流程的基础上提供增值服务，或者帮助优化或增强现有的功能。应用插件通常可以设置在现有的软件或者平台中。

随着生成式 AI技术的日益成熟，其在各种应用场景中的交互方式也变得多样化，这些交互方式不仅增强了现有工作流程的效率和功能，还为企业和开发者打开了无限的可能性，使 AI更好地融入日常生活和工作中。

2.6.2 AI应用的职责范围

目前生成式AI仍然难以避免地会出现幻觉等影响模型生成可靠性的问题，所以为了让应用的使用更加稳定，在设计不同场景下的AI应用时，需要考虑AI是否拥有直接对环境操作的能力。在企业场景中，除了提出计划、分析与建议之外，通过设置大模型是否拥有执行行动的权力，可以将AI应用分类成超级助理和超级员工，

并处理不同种类的任务。

- 超级助理：现阶段大部分的大模型应用都是以助理的方式，以一种“辅助驾驶”的方式辅助员工进行分析，但最终的选择、执行和判断仍然基于员工。超级助理通常不直接执行任务，而是为用户提供信息、建议或分析，帮助用户做出决策。这种AI应用的目的是辅助员工，而不是替代员工。超级助理在企业的应用场景十分丰富，例如数据分析、计划生成、智能文档总结、信息结构化提取等。

- 超级员工：超级员工指的是可以自主执行任务，并对环境造成影响与改变的AI应用，例如邮件智能回复、工单自动处理、智能客服机器人等。相比于超级助理，这种AI应用需要更多的权限和决策权，整个工作流程自动化程度高，人工仅在输入和输出的部分提供信息和监管。在设计超级员工类型的AI应用时，更需要考虑模型安全性和输出稳定性对流程的影响。

AI应用流程的高自动化程度意味着在资源充足的情况下，整个流程运作更高效，能够敏捷地应对快速变化的场景。但由于减少了人工的参与，需要更加稳定的模型来驱动整个工作流程，并通过设置输出检测等方式来提升模型的安全性。目前AI大模型大部分难以直接胜任超级员工的职责，因此更多地是做为超级助理为员工提供分析建议与辅助决策。

2.6.3 AI应用的输出模式

不同的场景需要模型生成不同模态的生成物。根据模型的模态，AI大模型不仅能生成文本，同时也能够根据需要生成代码、指令、图像、视频以及其他种类的序列数据。

- 文本 - 大语言模型能够根据提示来生成对应的文本生成物，其中最为普遍的是非结构化的文本内容，例如文本续写和对话。非结构化的文本通常作为最终的输出呈现给用户，或者通过人工或者大模型解析其中的信息并生成下一步的输出，例如思维链的中间步骤。但为了能够更好地和后续工作流程结耦，在工作流程当中更多会生成符合一定规范的文本、以及结构化的文本内容。在LangChain所采用的ReAct链中，每一步中间步骤都遵循了一定的输出格式，将思考、观察、行动等步骤名称写在中间步骤的输出中，更加方便后续步骤的解析。

- 代码 - 代码的分析与生成也是大模型领域重点的研究方向之一。对于代码模型，通过单个大模型或者多个智能体的方式，能完成代码编写、代码补全、代码注释等任务。CodeGeeX就是国内知名的代码生成应用，能够以插件的方式加入到VS Code中，并以助理的方式帮助用户编写代码，生成的代码可以自动在环境中测试与运行。

- 图像与视频 - 通过文字与图像多模态的模型，用户能够通过自然语言的描述来生成图片。处理生成图片之外，多模态的模型也可以通过图像来生成对应的文字描述。对图像的解析能够为工作流程提供部分视觉的理解

能力，能够完成更多领域的工作任务。同时，将时间作为维度之一，通过结合先前的图片来生成下一帧的图片，模型就拥有了生成视频的能力。

- 其他序列数据 - 大模型能够预测一条数据序列中的下一个数据。大模型通过训练能够预测并生成各种种类的序列数据，例如蛋白质序列数据、天气预报数据等。序列数据预测生成能够应用于多种场景，例如金融、科研、医疗等。

AI大模型在多模态数据生成方面展现出了广泛的应用潜力。不仅可以生成文本，还能产生代码、图像、视频和其他序列数据。这种多模态生成能力使得AI更为灵活，能够适应各种场景需求。此外，多模态模型通过文字描述生成图像或反之，为AI赋予了视觉理解能力。这种跨模态的交互为金融、科研、医疗等领域带来了巨大的应用前景。

3 生成式 AI 企业应用落地实践探索和总结

3.1 生成式 AI 与企业数字化转型

数字化转型是近十多年来促进经济社会发展和企业创新成长长盛不衰的主题，其重要的推动力量，就是数字化领域方面的技术迭代持续不断，创新浪潮一浪接一浪。从大的趋势来说，如我们在第一章提到的，最初推动数字化转型的，是移动互联网和云计算这两项重大的技术范式变革，它在企业乃至全社会的快速普及，使得数字技术资源的获取的便易程度比以前大幅提升，带来的是企业的业务数字化的速度大幅度提升，企业的边界得以大幅拓展，接触客户的方式从线下的实体渠道为主转变为每个人的手机中的 app(to C)，亦或是云端的 API(to B) 为主。

在业务活动大量数字化之后，海量的数据积累和云化大数据平台以及机器学习技术的快速成熟与普及，使得企业数字化转型进入到了数云融合的第二阶段，这也是当前大多数企业正在经历的数字化转型阶段。在这一阶段，基于数据的客户画像、精准营销、服务定制、故障预判、工艺优化、风险控制、供应链优化等等场景，形成了数据对业务流程的反向优化支持。而基于海量数据形成的各种机器学习 AI 业务模型，或者通过隐私计算等技术手段提供的数据本身，使其具有了越来越强大的价值生成能力，即数据越来越具有了资产的属性。

而这一轮的生成式 AI 的技术浪潮，可以看作是对企业数字化转型的又一轮重大的技术推动力。如我们在第一章提到的，企业的大量的非流程化的高价值知识，是以非结构化的文本、图像和视频数据格式海量积淀的。以具有泛化能力的生成式 AI 技术对其进行加工处理，能够系统化地提炼出散落在员工脑中的知识体系，甚至生成人工难以觉察和萃取的更高价值的知识（这一点可以比照 AlphaGo 在和人类顶级棋手对弈中走出人类无法解读其意图的落子）。而这些是之前大数据的数据分析、建模技术、Specific AI 技术等所无法实现的技术创新。

随着生成式 AI 技术的不断的进步和完善，它对企业知识的处理能力会从精度、广度和深度这三个维度上不断提升，最终我们认为有可能出现可以深刻洞察企业运营全局的“企业超级运营大脑”。当然，在这之前的几年时间，GenAI 可以在不同的岗位上提供知识辅助的超级员工助理，会是 GenAI 技术在企业落地的主要形态。

事实是，云计算 + 互联网，大数据 + BI/SAI (Specific AI)，以及生成式 AI + 企业知识治理，这三股技术浪潮在大多数企业内部是交织叠加在一起的。对于一个具体的企业的数字化转型的节奏，不会是按照这三股技术浪潮的先后顺序逐一实现。因此，企业就非常需要“从后天的视角来审视今天的现状，从而规划明天的行动”这样的策略。比如对云能力的引入，需要考虑算力资源和 MaaS 战略选择；数据治理，则需要扩展到对非结构化数据的治理、管理和利用；而 MLOps 的探索，需要立刻思考和 LLMOps 的整合；至于业务需求的满足，场景的创新，

生成式 AI 的相关技术尝试和引入，则成为一项重要的举措，而支持生成式 AI 技术的应用开发框架和平台，则需要提早设计。

总体来说，逐步构建出数据 / 语料和业务智能之间的飞轮效应，将是企业在即将到来的数字化转型的智能时代中，获取决胜竞争力的关键举措。

本白皮书虽然聚焦于生成式 AI 企业应用落地的技术相关问题，但我们深知技术的迅速发展将要求组织的转型变革和人才的升级赋能。企业的 AI 战略不仅仅关注技术的开发和应用，而更应强调在整个组织中实施人工智能，需要确保其实现业务战略制定和运营过程中保持有协同和兼容。

生成式 AI 的核心特征，有助于我们理解对人才和组织的要求。它有这样的一些特征：

- 数据驱动：大模型依赖于大量的数据进行训练和优化，实现深度学习。数字化的能力既是要求人，也是要求组织的。
- 自我迭代能够不断优化自己的模型，提高决策和预测的准确性。这是一种演化机制，自适应机制。
- 多领域应用：不局限于特定的行业或任务，可适用于多个领域和场景。生成式 AI 适应领域多，带来的变革是从点状创新到平台化创新。

在生成式 AI 的战略实施过程中，人才的角色和需求正经历显著的变革。这主要体现在以下三个方面：

- 多学科交叉人才需求：AI 的应用价值发挥，需要结合多个领域的知识，因此需要拥有跨学科知识和经验背景的人才。
- AI 与业务融合能力要求：人才需要能够深刻理解业务，并在其中发掘 AI 技术的应用场景。
- 人才持续学习能力要求：鉴于技术的快速发展，员工需要具备持续学习和掌控新技能的能力。

AI 技术的引入还可能会对现有的组织结构和协作模式形成冲击。比如：扁平化管理要求，推动组织更加灵活，加强跨部门协作；敏捷发展要求，要求快速响应市场变化，及时调整战略和产品方向等。

另外在推进生成式 AI 的过程中，组织需要确保技术的发展和应用符合道德伦理和法规要求。要确保 AI 系统的决策过程透明可追溯。要防止算法偏见，确保决策的公正合理。

总之 AI 战略的实施，对人才组织提出了新的要求和挑战。在技术的驱动下，组织需要不断地进行结构、文化、以及策略的变革，以适应不断变化的市场环境和技术发展。通过构建一个灵活、学习型的组织，我们能够充分发挥生成式 AI 的潜力，引领组织走向一个更加智能和创新的未来。

3.2 企业应用落地的关键问题与应对方法

随着生成式 AI 技术的蓬勃发展，许多企业纷纷展示了强烈的关注和兴趣，积极拥抱这一新兴技术。尽管我们在市场上见证了各式各样的生成式 AI 模型在企业的应用实践，至今仍未见到所谓的“爆款”应用诞生，同时，

建设的方法论和技术架构仍在不断变化和调整中。这部分的原因归因于生成式 AI 技术自身的固有限制，例如 token 的经济性、可控性、和准确性等问题；另一部分则是由于生成式 AI 技术快速的迭代更新，速度之快让人目不暇接，众多的学术论文堆积如山，使得企业需要组建专门的研究团队以跟进最新的技术进展。更为关键的一点在于，生成式 AI 项目的落地实施有许多未曾遇到的挑战，也是本文讨论的焦点。在生成式 AI 技术的实施过程中，企业面对着从模型的选择与定制，场景价值的深入挖掘，到成本的优化、算力配置、知识权威、用户体验优化、安全合规保障、现有技术的整合利用和学习成本控制等一系列挑战。因此，企业急切需要一个有 Whole Picture 的生成式 AI 架构，在这个架构 Picture 下，不仅能够看到快速搭建模型、算力、数据和场景四大层面的能力，还能通过持续迭代产生实现高业务价值和低技术边际成本的效果，而且能够展示出大模型相关的协作架构，职责角色区隔。这样才能在企业内有效打通从生成式 AI 技术到业务场景通道。

为解决生成式 AI 落地企业的一系列问题，神州数码推出了一站式大模型集成平台：神州问学。神州问学（参考下图）在一个平台上，为企业提供模型、算力、数据和应用的连接能力，它既是企业的大模型集成平台也是企业的大模型运营平台。神州问学从模型、数据、算力、应用四个角度打通各项资源，屏蔽繁复的技术细节，协助企业投产和运营自己的大模型应用。

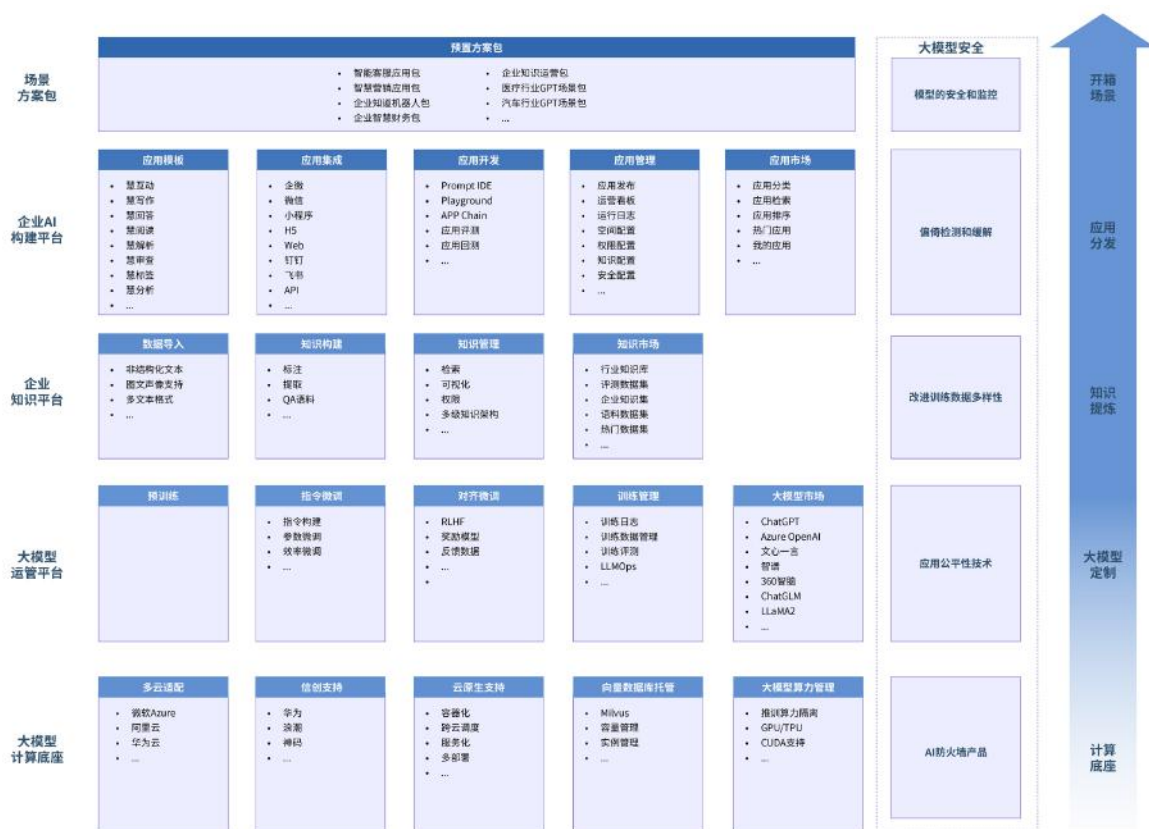


图 13 神州问学—企业生成式 AI 应用落地参考架构示意

神州问学不仅在神州数码自身的数字化转型过程中，成功实施了“神州数码超级员工”（如下图），也指导了其他客户落地生成式AI应用场景。



图 14 神州问学落地神码超级员工

神州问学着眼解决企业生成式AI落地过程中的六大核心问题，打通生成式AI技术落地到企业应用的最后一公里。

3.2.1 创新的切入点，场景的选择

企业在落地生成式 AI应用时，面临的第一个问题就是“用大模型来做什么？”，也就是场景选择的问题。在下图中我们可以看到，生成式 AI的技术能力从问答会话到文本生成，再到智能体、多模态，可以用的地方很多。这时我们需要一个思考框架，在企业选择场景的时候，做出决定。



图 15 生成式 AI 现阶段常见技术场景

“场景-痛点-方案-价值”就是这样一个结构化的框架，可以帮助企业更系统、更有效地进行生成式AI应用的场景选择。生成式AI的技术能力使得落到企业的具体业务场景时非常碎片化，但我们可以看到当前生成式AI主要应用方向有：知识库问答、资料解读审查、知识萃取分类、办公文案创作四大类。其中的痛点也各不相同。办公文案创作相对比较标准，而且是个人能力增强，可以采用公有云大模型方案，其他三类，都是企业能力增强，更倾向于大模型私有化部署。下图是利用这个框架进行场景选择和创新的思考地图：

场景	知识库及问答	资料解读审查	知识萃取分类	办公文案写作
痛点	讲解与问询供需不平衡，导致响应质量低、服务成本高；占用专业人员时间资源，同时满意度不高；	海量非结构化文档输入，解读审查内容需要大量专业人员时间逐步分析梳理；人工处理专业水平参差不齐，容易出现遗漏或错误；	企业拥有大量资料，需要安全管控；但资料分散，寻找困难，花费大量资源梳理，同时资料还在不断输入，甚至历史资料无人管理；	标准办公文案写作耗时费力；新媒体宣传千篇一律；专业领域文章不符合要求，错误百出；
方案	私有化部署，在通用大模型基础上，结合企业内部私域知识库进行调整，形成具备私域知识能力的模型；并对用户提供知识问答服务	私有化部署，在通用大模型基础上，对典型类型资料进行标注，同时输入海量标准文档，调整训练大模型具备识别核心内容、审查异常信息的能力	私有化部署，在通用大模型基础上，对典型类型资料进行标注，使大模型具备根据企业需求识别关键信息能力，萃取后形成分类体系，分级管控；	在通用大模型基础上，如企业/行业/专业有标准文档模板结构，可以通过输入海量标准文档，训练大模型；用户可以通过核心内容提示引导大模型写作输出成果文档；
价值	- 降低专业人员时间占用； - 降低专业讲解培训成本； - 提高知识覆盖度和及时性； - 提高问询响应时效和准确性； - 有效管控安全合规，分级分类管理 - 有效跟踪问询需求，有助管理决策	- 降低资料解析审查时间周期； - 提高解析审查准确率； - 有效管控安全合规，尤其是涉密或有安全风险资料解析审查；	- 资料识别效率高，类型广泛； - 不必持续投入人员对资料分析分类； - 涉密或有安全管理要求资料内容泄露风险大幅降低； - 大幅提高资料搜索效率；	- 提高文档生产效率，用户更多集中在引导和优化上； - 提高创意文案的创意方向； - 标准文档编写更符合模板要求； - 预防文档错别字等错误，降低风险
客户特性	- 专业壁垒高行业，如医疗、法律等 - 组织管理复杂，流程工具较多 - 业务客服数量较多，人员流动大	- 标准文档数量多，如招标采购、合同诉讼、年报财报等 - 审查内容多，法律法规要求较高，如政府机关、银行金融等	- 企业资料多，有查阅历史资料需求； - 多出现在政府事业单位和金融机构； - 查询资料如诉讼证据、纪念活动等	- 企业有大量标准文档编写工作； - 创意型企业，如传媒、游戏等； - 专业性文章写作，如监管审批等；

图 16 快速场景创新启发

场景 (Scene):

确定业务场景，明确生成式 AI 可以在哪些领域或部门中应用，如销售预测、客服、产品推荐等。并且这需要对于每个潜在的应用场景，收集相关的数据和信息，以便更好地理解该场景。以及基于企业的业务目标和策略，为每个场景设置优先级。

痛点 (Pain Point):

在每个潜在的应用场景中，识别和分析主要的业务问题或痛点。这些痛点可能是流程繁琐、成本高昂、效率低下等。并与业务部门沟通，确认识别的痛点是否真实存在，是否真正影响业务效果。

方案 (Solution):

根据每个痛点，选择或设计合适的生成式 AI 解决方案。考虑模型的技术可行性、数据需求等因素。为每个痛点设计完整的解决方案，包括数据准备、模型训练、系统集成等步骤。通常也会在小范围内进行项目试点，验证所设计方案的有效性和可行性。

价值 (Value):

对于每个方案, 预估其潜在的商业价值, 如预期的收益、成本节省、效率提升等。对比不同方案的预期价值, 确定哪些方案可以为企业带来最大的价值。最后还需要在方案实施后, 定期跟踪和评估其实际效果, 与预估的价值进行对比, 以便进行调整和优化。

通过这个“场景 - 痛点 - 方案 - 价值”的框架, 企业可以更系统、更有目的地选择和实施生成式 AI 应用项目, 确保每个项目都能为企业带来实际的商业价值。

3.2.2 开发工具的使用

企业的生成式 AI 项目, 从基础大模型出发, 包含微调 / 对齐工程、运维工程、提示工程和数据工程 (知识工程) 四个维度, 如下图所示。微调工程调教大模型的基础能力, 通过参数微调直接校正基础大模型, 对齐微调通过强化学习, 对齐企业意志和要求, 约束大模型能力, 克服幻觉, 保留涌现能力。提示工程则是大模型能力挖掘的方法。

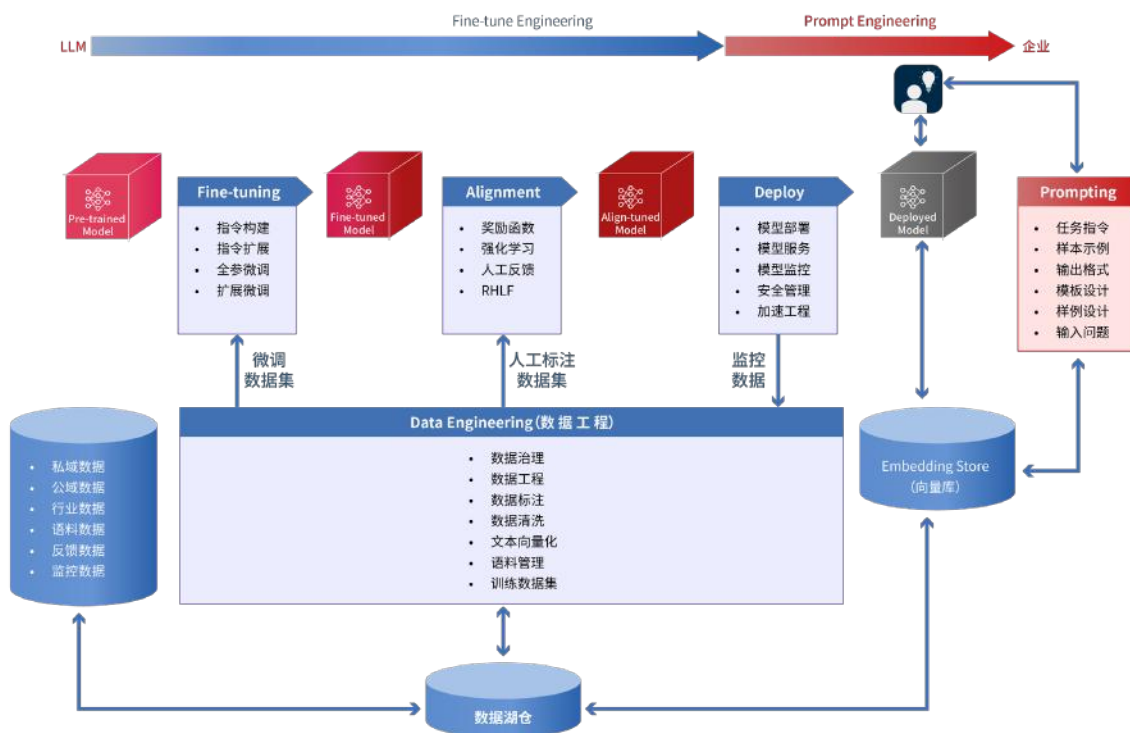


图 17 生成式 AI 应用工程数据流

无论在生成式 AI 项目落地实施的过程，还是在后期知识运营，模型运维，算力管理上，都依赖于一系列的开发工具来提升实施效率、保证模型性能和实现系统稳定。这对开发工具提出了以下关键特性要求：

● 深度学习框架支持

深度学习框架，不仅是开发的脚手架，降低开发难度，它还需要提供多模型支持，支持多种预训练模型和算法；解决高性能计算问题，建立软件到硬件计算桥梁，优化计算性能，支持 GPU/TPU 等硬件加速；深度学习框架也支持分布式训练，支持分布式和并行计算，这也是加速大模型的训练和微调的常用方法。

● 具备数据处理能力

数据处理能力是构建和部署大模型的关键环节之一，涵盖了数据清洗、特征工程和数据增强等多个方面。这些环节确保模型能够从整洁、高质量的数据中学习有效的信息，并进一步提高模型的泛化能力和预测精度。

数据清洗提升数据质量，构建准确、一致和可信的数据基础，包括处理丢失值、异常值、重复值，转换数据格式，统一不同来源数据的标准，以及验证数据的准确性等，通过数据清洗消除噪声，确保模型训练的准确性和稳定性，并降低错误预测的风险。

特征工程从原始数据中抽取、生成和选择对模型预测最有价值的特征。涵盖特征抽取、特征转换(如归一化或标准化)、特征选择(如通过相关性分析或特征重要性排序)等步骤，以提高模型的预测性能，减少计算负担，加快模型训练和预测的速度。

数据增强和生成扩充训练数据，增强模型的泛化能力和鲁棒性，它通过各种技术手段(如旋转、平移、噪声注入等)创造新的训练样本，或对现有样本进行重抽样，以防止模型过拟合，提高模型在面对新数据或略有偏差的数据时的预测稳定性和准确性。

总之，在数据处理的不同阶段中，企业需要细心审视数据质量，并通过科学的方法和工具来提取、转换和增强数据，为模型训练和应用打下坚实的基础。处理好的数据不仅有助于提高模型性能，还能帮助企业避免因数据问题带来的业务风险和损失。

● 模型开发与管理工具

模型开发与管理工具在 AI 项目中扮演着极为关键的角色，它涉及到模型的开发、部署、运行和优化等多个阶段。

自动化流程指的是将模型的训练、验证和部署等环节整合成一个自动化的、连贯的工作流。通过减少人工干预，确保过程的顺畅运行和减少潜在错误。它包括模型的自动训练、参数调优、验证测试、以及将模型从测试环境推送至生产环境的过程。

版本控制确保模型的各个版本及其相关配置、数据和参数得到合理管理。通过追踪模型的变化，可以确保

在出现问题时能够迅速定位问题并进行修复或回滚，它记录模型的所有变更、存储模型的各个版本、及时回滚至之前的稳定版本等。

模型监控保证模型在部署后的稳定运行，确保在模型性能下降或发生其他问题时及时发现，并通知到相关人员。它实时监控模型的多种性能指标(如准确率、延迟、吞吐量等)、及时识别潜在问题并通过报警功能通知到团队。

模型开发与管理工具强化了 AI项目的稳定性和效率，其中自动化流程、版本控制和模型监控分别关注项目的流程管理、版本管理和性能监控，共同构成了模型管理的框架。在实际应用中，高效地整合这三大要素，能够显著提升项目的成功率，确保大模型在整个生命周期中的稳定运行和持续优化。这有助于企业在保证服务质量的同时，能够快速响应市场的变化，不断调整和优化大模型以满足业务的发展需求。

- 部署与服务化工具

利用容器化技术，使用 Docker, Kubernetes等支持模型的灵活部署和伸缩；微服务架构，将模型部署为微服务，确保系统的解耦和可扩展性。提供 API管理，提供稳定的 API接口，支持高并发访问。

- 良好的开发环境

支持交互式开发，如 Jupyter Notebook，支持快速的代码迭代和数据探索，提供代码检查和测试框架，保证代码质量。可视化 UI，提供模型训练和预测的可视化工具，方便理解模型状态。

- 大模型安全能力

数据安全能力，确保数据在存储和传输过程中的安全性；模型防护能力，防止模型的不正当访问和攻击；访问控制能力，管理用户和服务的访问权限。

- 监控与优化工具

监控系统和模型的性能指标，及时发现问题；优化计算资源的调度和利用，完成资源调度；成本管理方面，能够监控和优化开发和运营的成本。

在具体的实践中，企业需要选择和整合适合自己的开发工具和技术栈，形成一套高效、稳定的大模型应用解决方案。并不断地优化和更新开发工具和流程，是保持和提升研发和 IT竞争力的关键。

3.2.3 企业知识工程构建

构建知识工程涉及到大量的数据处理、信息获取和知识输出。在这一过程中，生成式 AI可以发挥重要作用，因为它具有理解语言、生成文本和学习新知识的能力。下面是一个基本的框架，展示了企业如何利用生成式 AI

技术构建自己的知识工程：



图 18 基于生成式 AI 技术的知识管理和应用架构

● 数据收集和预处理

通过数据采集可以获取和整理与企业核心业务相关的各类数据，这可能包括文本、图像和视频等。而数据清洗是通过各种技术手段去除噪声，确保数据的质量和准确性。另外通过数据标注为数据打上相关的标签或者类别，帮助模型更好地理解数据。

● 模型训练和优化

生成式 AI 项目需要根据企业的具体需求，选择合适的大模型，这可能是预训练的大模型或特定领域的模型。再利用收集到的数据对模型进行微调，使其更加符合企业的实际需求和应用场景。这个过程中的重要抓手是：模型验证，在实际应用中不断验证和测试模型的效果，根据反馈进行进一步的优化。

● 知识抽取和组织

这是指利用模型从各类数据中提取有价值的信息和知识。并利用抽取的信息构建知识图谱，组织关系和逻辑结构。当然也要求不断更新和完善知识图谱，使其始终保持最新状态。

● 监控和维护

企业知识系统也需要实时监控模型的性能和输出质量，确保其稳定运行，并保障知识工程在各个环节的数据安全和隐私保护。

在这个框架上，企业可以根据自身的需求和特点进行调整和优化，发挥生成式 AI 的最大潜力，推动知识工程的成功实施。

3.2.4 模型选择和部署

在生成式 AI 应用的开发过程中，模型选择是一个关键步骤(如下图)，它直接影响到最终的应用效果。一方面模型选择不仅仅是选择一个预训练模型，还包括了对模型架构、规模、能力等多个方面的综合考虑，二是据我们观察，对企业来说，单一的一个大模型并不是能够“包打天下”的好的方案，而是合适的场景采用合适的大模型。从实践来看，大模型存在四种部署模式：公有云 API 部署、专有云部署、私有云部署、一体机部署，不同的部署模式也为企业提供了不同选择，建议企业根据情况选择一个或两个主模型，再加若干辅助模型。由此可见模型选择是个综合决策的过程，以下是一些可以考虑的关键因素：

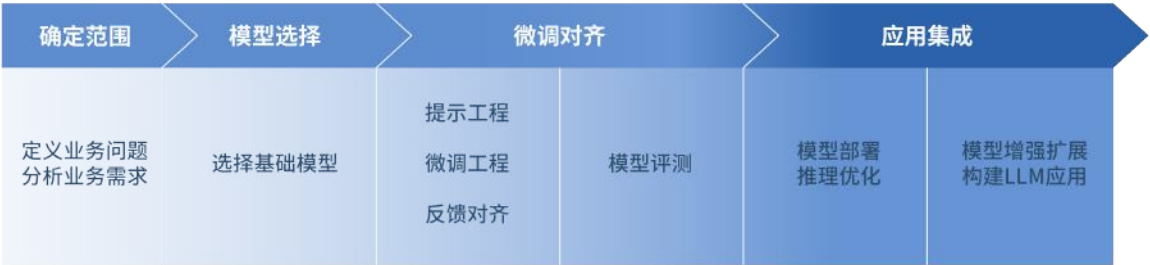


图 19 生成式 AI 应用开发流程

- 明确业务需求

准确理解和定义你试图解决的问题是关键，比如文本生成、图像识别、或者其他，确定模型需要满足的性能标准，比如在某个任务上的准确率、响应时间等。考虑企业的资源(例如计算资源、预算和时间)限制。

- 理解模型特性

评估不同模型的能力，如处理不同类型任务的潜在效果、泛化能力等，关注模型的规模和复杂性，考虑到部署的可行性和效率。调研了解模型的预训练情况以及它在特定任务上的微调潜力。

- 评估和验证

模型通常都有一些在标准数据集上对不同模型进行基准测试数据，可以对比它们的性能。但是那还是不够的，我们还需要通过具体实验验证模型在实际业务场景下的表现。模型选择也需要特别考虑模型的可解释性，这在某些业务场景(比如金融、医疗)中可能特别重要。

- 技术和资源要求

考虑模型实现和部署的难度以及所需的专业知识，评估模型的计算资源需求与企业的资源是否匹配；关注模型的长期维护成本，包括硬件、软件和人力资源等等。

- 法规和伦理考虑

确保模型的使用和应用符合相关的法律法规要求。评估模型是否符合社会伦理和企业价值观；同时也要保证模型在数据处理上符合隐私保护的要求。

- 持续优化防止漂移，为避免数据特性的缓变，带来的模型漂移问题，在模型部署后，需要持续监控其性能和输出质量，并及时更新模型，以适应业务的变化和技术的进步，也包括利用用户反馈和使用数据不断优化和调整模型。

虽然大模型在许多应用场景中表现出色，可以处理多种类型的任务，但其他算法模型在特定场景或应用中可能更加高效和适用。在实际的 AI 项目落地过程中，大模型往往需要和其他算法模型(小模型)进行互补，共同构建更强大、更灵活的解决方案。

大模型不会是企业唯一需要的模型，大小模型混布才是常态。我们可以在以下几点看到大小模型的混合才是出路：

- 资源优化，在某些应用中，小型模型或专用模型可能在计算效率上要优于大模型；较小的模型通常需要较少的计算资源和存储空间，更适用于资源有限或边缘计算的场景。

- 专用任务或领域可能存在已经充分研究和验证的专用模型，这些模型在特定场景下可能比大模型更加精确和高效。并且小模型往往更易于定制，以满足某些特殊的业务需求或适应特定的数据特点。

- 可解释性，在某些应用(如医疗诊断、金融分析等)中，模型的解释性至关重要。专用模型或较小的模型通常比大模型更容易解释和理解。特别是在有模型审计要求的场景下，可解释性和审计能力也是选择模型的重要考虑因素。

- 技术栈整合，企业可能已经有现成的模型和解决方案，整合这些现有资源而非全面替换，通常是更经济、更稳妥的选择。现有的小型或专用模型可能已经与业务流程紧密集成，保留这些模型可以避免繁琐的变更管理。

- 模型组合事实上在某些情况下，可以获得比单一模型更优的性能和泛化能力，并且利用集成学习方法(如 Bagging、Boosting)结合多个模型的预测，通常能够得到更稳健和更精确的预测结果。

大模型具有强大的学习和表达能力，非常适用于多种任务和广泛的应用场景，但在实际的项目实施中，应结合具体的业务需求、数据特性、技术背景等因素，灵活选择和利用各种算法模型。大模型与其他模型并非互斥，它们在实际应用中可以相互补充，发挥各自的优势，共同构建更完善的 AI 解决方案。

3.2.5 算力资源规划和管理

在企业级应用中，大模型要求强大的算力支持，其算力架构设计需要通盘考虑模型的训练、微调和推理阶段的资源需求，并兼顾到性能、可伸缩性、成本和效率等特性(参考下图算力架构技术栈)。

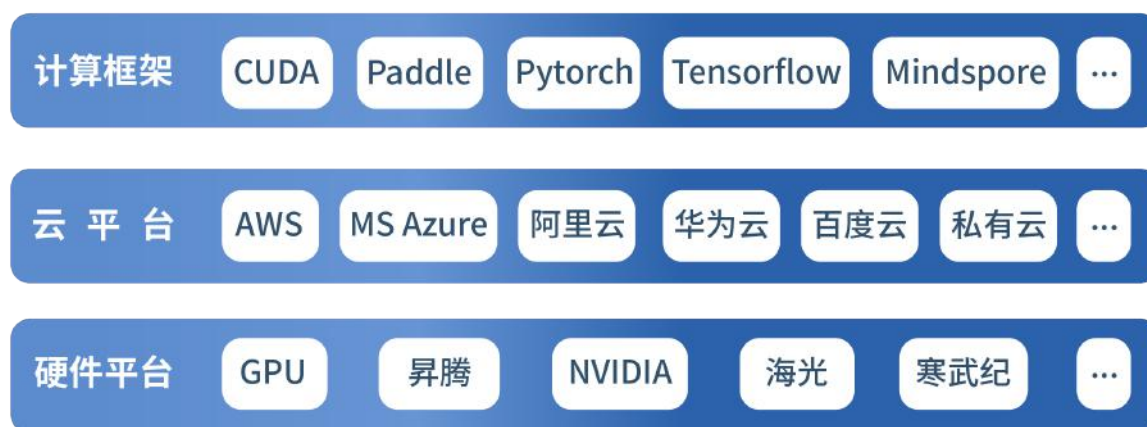


图 20 算力架构技术栈

算力架构设计原则上需要考虑：训练、推理和微调，算力需求：训练 > 微调 > 推理。企业的不同生成式 AI 场景不同，对算力需求也各不相同，原则上可以根据模型的用户数、交互频率、知识库大小进行测算。训练时算力是实施期算力需求，原则上外租模式更经济；推理和微调是生产和运维期算力需求，原则上自采模式更经济。以下是算力资源规划和管理中的关注要点：

- 硬件资源

GPU 集群，合适的 GPU 硬件，支持模型的高性能计算需求，甚至有些场景需要多 GPU 并行计算的能力。高速的存储和 I/O 能力，是支持大规模数据的快速读写和传输必然要求，集群网络传输能力也是分布式计算和数据、模型高速传输的基本保障。

- 分布式计算

利用分布式并行训练策略，如数据并行、模型并行或流水线并行，可以加速模型的训练和微调。而在模型推理阶段，负载均衡策略，合理分配计算资源，避免负载倾斜现象，能有效保证响应速度。

- 计算和软件框架

除了需要分布式计算和高效运算的深度学习框架(例如 TensorFlow, PyTorch)，采用容器化技术(例如 Docker, Kubernetes) 支持的模型，可以提供强大的灵活部署和管理能力，完成还能自动化部署、监控、报警和恢复。

- 弹性计算架构

弹性伸缩的计算架构根据实际计算需求动态调整资源，最大化资源利用率。加上混合云策略，还可以充分利用公有云和私有云的资源 and 优势。

- 数据管理

最佳实践中优化数据的输入 / 输出流，减少数据预处理和传输的时间，能解决很多性能问题；而数据备份和恢复能力是保证数据安全兜底手段。

- 安全和合规

除了需要确保数据的安全存储，还要避免传输过程中的风险，避免外部和内部的安全威胁。而且在算力架构设计时，还要符合行业要求和标准，特别是政企和银行。

- 监控和优化

不仅需要建立全面的系统监控，覆盖硬件、网络、模型性能等层面，而且监控数据也是持续优化模型和系统性能，降低成本，进行性能调优的基本手段。

在大模型的落地过程中，企业只有通过综合考虑硬件资源、分布式计算、软件框架、数据管理和系统安全等多方面的要求，才能设计出既满足业务要求、又高效稳定的算力架构。

3.2.6 生成式 AI 应用的安全工程

大模型因其深度学习和数据驱动的特点，能够处理和生成大量的信息，但也带来了一系列的安全和伦理挑战，这方面也是实践中，企业特别优先考虑的方面，使得保障大模型应用的安全性变得至关重要：

- 数据安全和隐私保护

数据脱敏，确保用于训练模型的数据已经进行了充分的脱敏处理，保护个人隐私；数据加密，运用数据加密技术来保护数据在传输和存储过程中的安全；访问控制，严格控制和监督哪些人、何时、如何访问数据。

- 模型鲁棒性和安全性

攻击对抗能力，可以部署对抗样本的检测和防护机制，确保模型不会被恶意攻击所欺骗；错误纠正能力，通过增强模型的错误检测和纠正能力，及时识别和修正不准确或者有害的输出；安全审计措施，要求对模型进行定期的安全审计，确保其行为符合预期和规范。

- 算法伦理和合规性

我们不仅要算法的公平性，避免和减少算法可能导致的偏见问题，还要尽可能使模型的预测透明和可解释。实践中，也会进行合规检查，来保障模型的开发和应用过程合规合法。

- 用户体验和安全性的平衡

任何安全举措都是牺牲效率为代价的，恰当的平衡才是实际可行的解决方案。实施实时监控机制，可以在

模型输出不当内容时迅速干预。建立一个健全的用户反馈机制，保障用户在使用过程中的安全体验。引入敏感内容过滤，筛选和屏蔽可能的敏感、不当或有害信息。

● 灾备和故障恢复

以防万一，必要的备份方案，确保万一情况下的数据和模型有充分的备份，以便在发生问题时能够迅速恢复。这需要制定明确的故障检测和应对机制，才能在模型出现问题时能够迅速定位并解决。而且还需要定期的进行压力测试和安全模拟，不断验证系统的可靠性和安全性。

模型安全是个综合手段，在生成式 AI 技术日益普及的今天，这些安全方案的实施无疑将成为每个组织不可或缺的一部分。

大模型安全工程是一个体系工程，不是“单点”的方案，在生成式 AI 项目实施过程中，除了常见的私有化模型部署要求，还需要跟企业用户认证中心集成，打通账号和角色权限。大模型访问也需要留痕，使其可审计；同时，相关的知识分级权限控制，也是配套架构，让合适的人通过大模型获得合适的信息，让大模型知道它该知道的。大模型本身并没有身份识别，然而让大模型能够识人，识别身份、职位，在企业环境也是常见的，做到大模型开口之前先识人。大模型知识域控制，也是企业环境里要求，让大模型“知之为知之”。另外相应大模型访控策略是避免攻击的重要手段，特别是需要对公服务的环境。在企业环境还需要为大模型设置伦理价值观围栏，避免不必要的公众事件。在国内合法合规，还要考虑大模型备案问题。

3.3 企业应用落地的四类驱动模式

企业在落地生成式 AI 项目时，根据自身的战略举措、预算规划、认知深浅、行业特点，有不同的选择，形成了不同策略。从我们的众多生成式 AI 项目实践来看，有以下四种驱动模式：



图 21 企业大模型落地的四类驱动模式

3.3.1 点状场景实施驱动

企业在数字化和智能化的浪潮中，面临着无数的选择和决策。生成式 AI 技术作为一个前沿且高度发展的技术领域，已在多个行业和场景中展现出了巨大的潜力和价值。而当企业决定引入生成式 AI 技术时，采用单点场景进行尝试和探索，逐渐演化成为一种行之有效的模式。这背后，蕴含着一系列的逻辑、策略与深思熟虑。

首先，单点场景的探索能够帮助企业更为深入和细致地理解生成式 AI 技术的工作机制、能力边界和实际应用价值。例如，在知识问答或智能客服的场景中，大语言模型能够展现其在自然语言处理、知识理解和应答生成等方面的技能。企业通过这一场景的具体实践，可以观察和分析大模型在语言理解精度、应答准确度、交互流畅度等多个方面的表现，进而得到一个更为真实和客观的评价。

其次，单点场景的试点实施能够在较小的范围和较低的成本下，测试和验证生成式 AI 技术的适用性和效益。这一阶段，企业能够更加聚焦地投入资源，针对特定的业务流程或问题，设计合适的模型应用方案，实施定向的技术优化和调整，并通过观察业务指标的变化来验证模型的效用。这不仅可以避免在尚未充分了解技术特性

和业务适配度的情况下盲目大规模投入，也为后续的大规模推广和应用提供了宝贵的实践经验和数据支持。

再次，通过专注于单一场景，企业可以集中技术和业务团队的力量，进行更加深入的技术研究和业务创新。这包括了基于具体业务需求对模型的微调、基于实际数据和反馈的模型优化、以及基于场景特性的业务流程改造等。这一过程不仅能够最大化地挖掘和发挥生成式 AI 技术的价值，也能够推动企业的业务团队和技术团队之间的深度协作和共同创新。

此外，基于单点场景的探索还有助于推动企业内部的学习和能力建设。通过实际的项目实施，企业的技术和业务人员能够获取宝贵的实践经验，提升对生成式 AI 技术的理解和应用能力，并在此基础上培养出一批能够在更多场景中推广和应用生成式 AI 技术的关键人才。

在单点场景获取到成功的经验和效益验证后，企业可以基于此进行更为广泛的应用推广和模式创新。这包括了将生成式 AI 技术应用到更多的业务场景和流程中，进行更为复杂和高级的业务创新和价值创造。同时，企业也可以在此基础上不断优化和拓展大模型的技术体系，通过集成更多的数据源和业务逻辑，提升模型的智能化水平和业务价值。

总之，在引入生成式 AI 技术进行场景创新时，企业通过专注于单一场景的深入探索和应用实践，既能够在较小的范围和较低的风险下验证技术的价值和可行性，也为后续的更为广泛和深入的技术应用和业务创新打下了坚实的基础。这一模式的实践也将为企业在数字化和智能化的道路上积累宝贵的经验，推动企业在更多场景和领域中实现技术创新和业务突破。

案例

某地方政府部门一直致力于提高其服务效率和民众满意度。面对众多的社会保障事项和巨大的民众服务需求，部门决定引入先进的生成式 AI 技术，以期在信息处理、业务响应和服务提供等方面实现显著提升。在多个可能的应用场景中，部门选择“数字柜员”作为生成式 AI 技术的首个落地项目。

数字柜员将首先被设计为能够解答公众关于社会保障的常见问题，并能够引导用户完成一些基础的自助服务操作，如查询社保缴费记录、退休金发放情况等。

模型训练与优化方面，通过收集和整理大量的社保相关的问答样本、业务流程信息和政策文件，部门与技术团队合作训练和优化大模型，以确保其在相关场景中能够提供准确、流畅的服务。在模型的训练和优化过程中，项目团队特别注重与业务部门的协作，确保模型在解答问题时能够充分符合实际的政策和流程。同时，也通过模拟用户进行多轮的测试和验证，确保其在真实场景中能够提供稳定可靠的服务。

用户体验优化方面，在数字柜员的界面设计和交互逻辑上进行精心打磨，确保其既能够清晰明了地展现信息，也能够提供友好易用的交互体验。

服务上线与推广上，在数字柜员上线初期，通过多种渠道(如官网、社交媒体、线下服务中心等)进行宣传和推广，并在一定期间内提供线上线下的引导支持。

效果评估与持续优化，数字柜员在上线后，通过简单易懂的语言和流畅的交互吸引了大量用户的使用。用户可以轻松获取社保缴费、待遇领取等方面的信息，并在部分常规服务上实现自助办理。而且通过收集用户的使用反馈和数据日志，系统评估数字柜员的服务效果和痛点问题，以便进行持续的优化升级。

该项目，不仅提升了部门的服务能力和民众的服务体验，也积累了丰富的生成式 AI 技术应用经验。这一切，都源自于部门在项目初期聚焦于单一场景、注重实际效果的策略和方法。在验证了生成式 AI 技术的价值和可行性后，部门也逐步计划将这一技术在更多的场景和业务中进行推广和应用。

3.3.2 大模型平台建设驱动

企业大模型是新的 IT 操作系统，对整个 IT 能力的建设有着全面而深刻的影响，其折射出的是一个信息化、智能化极其密集的企业运营景象。在这个景象中，数据不仅是决策的基础，更是创新、优化、探索的源泉。尤其是在大型企业、集团型企业和正经历数字化转型的企业中，生成式 AI 技术不仅是一种技术选择，更成为了推动企业进步的关键动力。这其中，构建大模型平台不仅仅是实施一种技术策略，更是在塑造一种全新的 IT 能力和业务协同模式。

- 数字化转型的核心是智能化数据运用

在数字化转型的过程中，数据的处理和利用成为衡量转型效果的关键。如何高效挖掘数据价值、如何在实际运作中应用数据分析结果、如何在策略层面体现数据驱动，这些问题的答案很大程度上依赖于企业大模型的建设应用。

在数据决策方面，大模型可以整合企业内外的众多数据，通过算法模型做出更精准的业务预判和决策建议。

在数据创新方面，基于大模型挖掘出的洞见，可以为产品创新、服务优化提供方向，推动企业形成数据驱动的创新机制。

在数据协同方面，大模型平台可以实现企业内部的数据协同，打破数据孤岛，促进各业务部门之间的信息流通和协同工作。

- 统一的大模型平台是 IT 能力的强大支撑

统一的大模型平台完成了能力整合，大模型平台整合了数据处理、模型构建、算法应用等多种 IT 能力，形成了一个全面、强大的支撑体系。

统一的大模型平台体系化的建立技术赋能机制，它将先进的 AI、大数据技术赋能于各业务部门，通过提供技术服务支持业务的智能化升级。

统一的大模型平台是最经济的成本控制策略，统一的平台减少了重复建设，降低了企业在大模型获取能力上的投入成本，提升了 IT 投入的产出比。

- 分布式模式，建立弹性的数字化转型路径

大型企业组织结构复杂，业务范围广泛，甚至跨行业部署，组织部门间协同困难，统一的大模型平台解决这些问题：

灵活响应，各业务部门在大模型平台上进行自己的业务场景的探索 and 实现，更加灵敏和贴合实际需求。

自主创新，业务部门可以依托平台自主进行数据分析和模型应用的创新实验，推动业务模式的不断优化。

风险分散，分布式的操作模式分散了中心化决策的风险，通过多元化的尝试找到最适合的发展路径。

- 大模型平台能够建设出与业务高度契合的平台

在构建大模型平台时，需要将其与企业的实际业务高度契合，这需要在平台建设的各个环节中充分考虑业务的特点和需求。这期间既需要进行数据源整合，深入理解各业务部门的数据需求，将其所需的数据源整合至大模型平台中，又需要进行模型定制，根据各业务部门的特定场景，定制模型和算法，满足其精准分析和决策的需求，而且建立了业务协同，通过强化大模型平台与业务部门的协同机制，形成快速响应的支持体系。

- 构建持续演进的大模型平台

数字化转型是一个持续演进的过程，大模型平台的建设也应具备良好的拓展性和适应性，以便支持企业未来的发展和变革。它不仅能够持续优化，根据业务的发展和变化，持续优化模型和算法，保持其与业务的高度契合，而且保持集中的技术升级，关注前沿的技术发展，及时将新的技术和方法引入平台，保持其技术的领先性。同时它也是能力拓展平台，随着企业的拓展和转型，不断丰富和拓展平台的能力，支持企业探索新的业务模式和创新路径。

在这样的理念和实践指导下，企业大模型平台的建设将更加贴合企业的数字化转型需求，更有效地支持企业在数字化时代的持续发展和竞争。

案例

某大型航空集团，作为一家拥有庞大飞机机队和多元化业务线的航空公司，涉及到航线规划、航班运营、维修管理、客户服务、货物运输等多个业务领域，面临着大量、多样的数据处理和分析需求。在数字化转型的大背景下，如何充分挖掘和利用这些数据，成为提升竞争力的关键。

平台架构设计方面，该集团大模型平台需要集成来自不同业务部门的数据、模型和算法，构建一个支持多业务场景的统一 IT 平台。包含数据层，整合包括航班数据、乘客数据、货物数据、机场数据、气候数据等多源数据；模型层，设计和部署用于预测、优化、决策支持的各类算法模型；服务层，提供给业务部门的数据服务、模型服务、分析服务等。

技术选型方面，技术层面需要综合考虑数据处理、模型训练、算法部署、实时计算等多种需求。其中数据处理，采用了 Hadoop、Spark 等分布式计算框架；模型训练采用了 TensorFlow、PyTorch 等深度学习框架，模型部署采用了 Kubernetes、Docker 等容器化部署方案。

数据安全和合规方面，设计数据权限管理、数据加密、数据审计等安全保障机制，确保平台的稳定可靠运行。

关键业务场景与应用方面有：

- 优化航线规划，通过模型预测不同航线的需求，分析旅客流动，动态调整航线和航班的规划，以最大化航班的运营效率和利润。
- 提升客户体验，分析乘客的旅行、购物、用餐等数据，推送个性化服务和优惠，增强乘客的飞行体验。
- 高效航班运营，利用实时气象数据、飞行数据，优化飞行路径和高度，减少能耗，提高安全性。
- 精准维修，分析飞机的运行数据，预测潜在的故障和维修需求，制定更精准的维修计划和备件管理策略。
- 货物运输优化，通过模型预测和优化货物的运输和分拣，提升货物运输的效率和准确性。

基于大模型平台，该集团能够在各个业务领域实现数据的深度利用，推动业务的智能化升级，形成数据驱动的决策机制和服务模式，提升企业的整体竞争力和运营效率。

3.3.3 企业知识治理驱动

在当前的信息社会中，数据被称为新的“石油”，它充斥在各行各业的每一个角落。企业的运营生产大量的数据，其中不仅包括了表格、数字这样的结构化数据，更包括了大量的文本、图像等非结构化数据。正如所述，“大

模型的神经网络训练机制压缩了大量自然语言语料，浓缩出其中的知识，体现了智能”，进一步指出了非结构化数据中蕴含的巨大价值。这些非结构化数据，通常来自企业的公文文档、合同章程、产品说明、操作手册、作业规范等，它们是企业运营的能力结晶，代表了企业的经验、战略和价值观，是值得被深入挖掘和利用的“宝藏”。

● 非结构化数据宝藏

这些非结构化的文本数据，能够为企业的各个层面提供重要的决策支持。例如，产品说明中可能包含了产品研发的关键思路，合同章程可能揭示了企业合作的风险和机会，作业规范则可能反映了企业内部的操作效率和潜在的改进空间。这些文本数据一旦被恰当地挖掘和利用，将极大地推动企业的创新和效率提升。

● 知识库建设的必要性

知识库的建设，实质上是对企业历年来积累的大量文本数据进行深入的挖掘和智能的管理。通过构建知识库，企业能够实现：

知识的有效沉淀，将分散在各类文档中的知识进行提炼和总结，构建一个系统的知识体系，防止知识的丢失和浪费。

知识的高效检索，当企业员工在面对问题和决策时，能够迅速地从知识库中检索到所需的信息和知识，提高工作效率。

员工的培训赋能，新入职的员工可以通过学习知识库中的内容，迅速地熟悉企业的业务和文化，提高自身的业务水平。

潜在的风险管理，通过分析知识库中的合同、协议等文档，企业能够及时发现和防范潜在的法律风险。

● 用大模型建设知识库

当我们谈到利用大模型构建企业知识库时，我们指的是运用语言模型对文本数据进行深入的理解和智能的处理。大模型有天然的文本处理能力，包括：文本理解，大模型能够理解文本数据中的逻辑和含义，发现其中的关键信息和隐藏的关系；也能进行知识提炼，通过分析和总结，大模型能够将文本中的关键知识进行提炼和表达；而且容易构建智能回答系统，构建好的知识库能够回答员工在工作中遇到的问题，为决策提供数据支持。

● 实施最佳实践

在具体的实施过程中，企业需要有明确的策略和方法论来指导知识库的建设，首先要完成数据整合，整合并清洗企业中现有的所有文本数据，保证数据的准确性和完整性；其次涉及模型训练，基于清洗后的数据，训练专属于企业的语言模型，确保模型充分理解企业的语境和特点；然后进行知识平台建设，构建知识库的线上平台，确保知识的检索和共享能够高效便捷的进行；同时需要保持企业知识的持续更新，随着企业的发展，知识库也需要不断的更新和升级，以保持其价值和准确性。

大模型的企业知识库建设不仅是企业知识管理的必然选择，也是提升企业核心竞争力的关键手段。它不仅
可以沉淀企业的宝贵经验，提高运营效率，更能在新员工培训、决策支持、风险防控等多个层面为企业带来持
续的价值。在数字化转型的浪潮中，企业必须充分认识到知识库建设的重要性，并采取实际行动，将其落到实处。

案例

某大型电器制造集团企业，拥有着数十年的历史和深厚的行业积累。公司的产品线丰富，
覆盖了全品类的家用电器和商用电器。在全球多个国家和地区也都有业务，每年要处理大量的
产品设计、制造、销售、售后等相关的文档和数据。

在业务快速发展的同时，该企业也面临着一些挑战，关键业务场景与应用方面有：

- 知识沉淀的问题，多年的业务积累让公司积累了大量的非结构化数据，如工程设计文件、
产品手册、市场分析报告、客户反馈等，这些宝贵的数据被散落在不同部门，难以形成有机的
知识共享。
- 员工效率的问题，员工在日常工作中，经常需要花费大量时间去查找历史文件和资料，降
低了工作效率。
- 创新能力的问题，虽然公司积累了丰富的经验，但是新员工很难快速获取和理解这些知识，
影响了公司的创新能力。

为了解决上述问题，该集团企业利用生成式 AI 技术进行知识平台建设：

- 整合数据，首先，公司对内部各部门的非结构化数据进行了整合和清理，包括工程设计、
市场报告、客户反馈等所有文本类数据。
- 定制大模型，基于上述数据训练定制了一个专属的大模型。这个模型能理解公司的产品、
业务、以及行业特有的术语和知识。
- 提炼知识与分类，对历史文档进行了深入的分析和提炼，自动提取出关键知识和信息，并
根据主题和业务领域进行了分类。
- 建设知识库平台，搭建了一个内部知识库平台，其中包含了通过大模型提炼出的知识和信
息。员工可以通过自然语言查询，迅速找到他们需要的信息和答案。
- 反馈闭环，持续优化，有专门的团队，负责知识库的持续优化和更新。定期收集员工的反
馈和建议，不断优化大模型的性能和知识库的内容，保持知识库的常用常新。

该知识平台的应用在：

- 新员工培训，新加入公司的员工可以通过知识平台快速了解公司的产品和业务，缩短入职
培训的时间。
- 市场分析，市场部门的员工可以在知识平台中快速检索到历史的市场报告和分析，为新产

品的市场定位提供数据支持。

- 客户支持，客服团队可以通过知识库迅速找到产品的使用手册和常见问题的解决方案，提高客户服务效率。

通过生成式 AI 技术和知识平台，该企业建立了内部知识的高效管理和利用，提高了员工的工作效率，也为企业的创新和发展提供了强大的支持。

3.3.4 战略型创新驱动

在当前数字化变革的大背景下，人工智能和生成式 AI 技术已经成为引领企业发展和创新的重要力量。对于那些生成式 AI 技术对其行业影响较深的企业来说，尤其需要关注和把握生成式 AI 技术的应用与实施。这其中，战略型创新模式的采用，往往是这类企业在落地生成式 AI 技术时的关键举措，他们有这样一些考量：

- 响应市场变化的敏捷性

在高度动态的市场环境中，分布式创新模式允许企业通过在不同业务部门并行推进生成式 AI 技术的实施，增强整个组织的响应速度和灵活性。

- 利用多元化的知识和专长

不同的业务部门往往拥有各自独特的专业知识和市场洞察能力。分布式创新模式能够充分利用这些多元化的知识和专长，推动生成式 AI 技术的更加精准和高效的应用。

- 降低实施风险

通过各个业务部门自主进行生成式 AI 技术的探索和实施，企业可以有效地分散实施风险，保证在某一方面遇到挑战时，其他业务依然能够正常推进。

- 创新成果的快速验证和推广

分布式创新使得每个部门都能根据自身的业务特性和市场需求进行生成式 AI 技术的实施和创新，并且能够快速在自己的业务领域验证其效果。一旦某一创新实践在某个部门取得成功，便可快速推广至其他部门，实现创新成果的快速扩散和积累。

- 引领行业标准和发展方向

对生成式 AI 技术应用较为成熟的企业往往有机会成为行业的引领者。通过分布式创新模式，企业能够在多个细分市场和业务领域同时推进技术创新，形成更全面和深入的行业影响力，进一步引领行业的标准制定和发展方向。

- 提高整体运营效率

通过生成式 AI 技术的分布式创新和应用，各个部门的运营效率将得到提升。无论是在市场响应、客户服务、产品设计或生产管理等方面，生成式 AI 技术都能够提供强大的数据分析和智能决策支持，大幅提升整体的运营效率。

总之，生成式 AI 技术通过深入到企业的各个业务部门和环节，能够极大地提升企业的创新能力和运营效率，强化其在行业中的竞争地位。分布式创新模式作为一种高效的组织和管理方式，能够有效地协助企业在生成式 AI 技术实施过程中充分挖掘和利用内部的多元化资源，促进技术创新的快速推进和广泛应用。

案例

某大型制药企业拥有庞大的研发和生产网络，涵盖多个疗法领域，并且涉足全球多个市场。随着生成式 AI 技术在制药研发、生产优化、市场推广等方面的应用逐渐成熟，该集团企业在整个组织中落地该技术，以提高研发效率、优化生产流程、精细化市场营销等，在落地过程中涉及以下一些关键举措：

- 知识沉淀的问题

多年的业务积累让公司积累了大量的非结构化数据，如工程设计文件、产品手册、市场分析报告、客户反馈等，这些宝贵的数据被散落在不同部门，难以形成有机的知识共享。

- 确立生成式 AI 技术战略方向

在企业战略层面，该集团明确了生成式 AI 技术主要聚焦于研发创新、生产优化、客户关系管理和市场分析等方面。

- 创新能力的问题

虽然公司积累了丰富的经验，但是新员工很难快速获取和理解这些知识，影响了公司的创新能力。

- 建立中央协调和支持平台

依托 IT 部门，构建一个中央大模型 AI 平台，提供统一的技术支持、数据管理和方法论指导，为各业务部门的分布式创新提供支持和协调。

- 部门层面的分布式创新

各业务部门根据自身特点和需求，确定各自的生成式 AI 技术应用方向和计划。其中研发部门利用大模型分析海量的科研数据，加速药物研发进程，减少不必要的实验，快速找到有潜力的药物候选分子；生产部门利用大模型进行生产流程优化，通过分析生产数据，预测设备故障，优化存储和物流，提高生产效率和降低成本；销售和市场部门，利用大模型分析市场趋势和消费者行为，为市场推广和销售策略提供数据支持，实现精细化营销；客户服务部门，部署生成式 AI 技术进行自动化客户服务，包括智能客服机器人，提高服务效率和客户满意度。

- 打通跨部门协同合作

鼓励不同部门之间的知识和资源共享，形成协同创新的良好氛围。例如，销售和市场部门的市场数据可以为研发部门提供有价值的市场需求信息。

从业务效果上看，加快药物上市速度，通过大模型分析科研数据，提高药物研发效率，缩短药物从研发到上市的周期；优化市场战略，通过精细化的市场分析，更加精确地定位市场和客户，优化营销策略；提升客户体验，利用智能客服提高服务响应速度和问题解决效率，提升客户满意度。

通过战略型创新模式，该制药集团企业成功实现了生成式 AI 技术的多点落地和应用，不仅推动了各部门的技术创新和应用，也提高了整个企业的运营效率和市场竞争力。同时，各个部门的实践经验和成果在整个组织中得到了共享和推广，形成了一个持续创新和学习的组织文化。

4 AI产业政策与发展趋势

4.1 我国 AI 产业政策

4.1.1 国家层面的 AI 政策

国家将 AI 发展作为建设创新型国家和世界科技强国，实现“两个一百年”奋斗目标和中华民族伟大复兴中国梦的强大支撑。国务院于 2017 年印发《新一代人工智能发展规划》，从构建科技创新体系、培育智能经济、建立智能社会、构建智能化基础设施体系等方面对我国 AI 发展做出战略性部署，明确到 2025 年 AI 产业进入全球价值链高端环节，新一代 AI 在重点领域得到广泛应用，到 2030 年 AI 产业竞争力达到国际领先水平，形成涵盖核心技术、关键系统、支撑平台和智能应用的完备产业链以及高端产业群。

工信部印发《促进新一代人工智能产业发展三年行动计划(2018-2020 年)》《新一代人工智能产业创新重点任务揭榜工作方案》，推动 AI 产品应用和产业培育。科技部等印发《国家新一代人工智能创新发展试验区建设工作指引》《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》《关于支持建设新一代人工智能示范应用场景的通知》，打造具有重大引领作用的 AI 创新高地，布局重大示范应用场景建设。教育部印发《高等学校人工智能创新行动计划》，提出到 2030 年，高校成为建设世界主要 AI 创新中心的核心力量和引领新一代 AI 发展的人才高地。国家标准化管理委员会等印发《国家新一代人工智能标准体系建设指南》，对 AI 重点急需标准研制及应用进行部署。国家互联网信息办公室等印发《生成式人工智能服务管理办法》，采取坚持发展和安全并重、促进创新和依法治理相结合的原则管理生成式 AI。在《国务院 2023 年度立法工作计划》中，《人工智能法草案》预备提请全国人大常委会审议。

4.1.2 典型地方和城市的 AI 政策

各地积极出台政策措施促进 AI 产业发展，典型地方和城市的 AI 政策如表 5.1 所示。

将 AI 作为经济社会发展的核心引擎。各地将 AI 作为驱动未来科技创新、产业发展、城市建设、社会治理的战略性技术。比如：上海市推动 AI 成为建设“四个中心”（国际经济中心、国际金融中心、国际贸易中心、国际航运中心）和具有全球影响力的科技创新中心的新引擎；安徽省打造全国重要的 AI 产业发展先行区和智慧产业新高地，为发展智慧经济、构建智慧社会奠定坚实基础，为加快建设现代化五大发展（创新发展、协调发展、绿色发展、开放发展、共享发展）美好安徽提供强大支撑。

致力于形成从技术研发到应用的促进体系。各地着眼于 AI 产业发展涉及的技术研发、产业化等关键环节，

从基础研究、核心技术突破、产业化、基础支撑、生态建设等方面对AI产业发展进行全方位支持。比如：广东省深圳市从算力供给、关键核心技术与产品、产业聚集、场景应用、数据和人才要素等方面部署具体行动方案；浙江省从核心技术突破、底层基础设施、创新场景赋能、特色聚集发展、产业生态构建等方面进行部署，打造AI产业发展新高地；陕西省从基础理论和关键技术研究、创新平台建设、优势产品培育、示范应用、企业创新发展、产业集聚发展六方面促进AI产业发展。

根据自身资源和产业基础确定AI政策重点。在促进AI产业发展的政策措施中，各地根据实际情况在技术、应用、场景等方面有所侧重。比如：北京市着眼于建设具有全球影响力的AI创新策源地，到2025年形成具有国际竞争力和技术主导权的产业集群；上海市注重拓展应用场景，政策措施中重视人才队伍建设、数据开放和利用等；重庆市以场景驱动AI产业高质量发展，全面支撑数字重庆、制造强市建设。

积极推动AI产业发展立法进程。广东省深圳市通过和发布了《深圳经济特区人工智能产业促进条例》，明确了AI概念和产业边界，建立了AI统计与监测制度，确定了市属各部门的职责，规范了促进和监管措施。上海市通过和发布了《上海市促进人工智能产业发展条例》，从基本要素、科技创新、产业发展、应用赋能、产业治理与安全等方面促进和管理AI产业发展。

地方 / 城市	时间	政策	重点内容
北京市	2017	北京市加快科技创新培育人工智能产业的指导意见	提出创新体系、产业集群、融合应用、发展基础等方面的主要任务。
	2023	北京市加快建设具有全球影响力的人工智能创新策源地实施方案（2023-2025年）	从理论研究突破、关键核心技术自主可控、产业规模提升、实体经济赋能等方面提出工作目标。
	2023	北京市促进通用人工智能创新发展的若干措施	推出算力、数据要素、大模型、场景、监管等方面的措施。
上海市	2017	关于本市推动新一代人工智能发展的实施意见	包括产业布局、人才高峰建设、数据资源开放、加强科研布局、营造创新生态等内容。
	2018	关于加快推进人工智能高质量发展的实施办法	出台了围绕人才队伍、数据资源、技术创新、空间生态、资本力量等五大发展要素推出 22条具体举措。
	2021	上海市人工智能产业发展“十四五”规划	包含创新、产业、企业、赋能城市数字化、空间布局、生态等方面的内容。
	2022	上海市促进人工智能产业发展条例	包含基本要素、科技创新、产业发展、应用赋能、产业治理与安全等内容。

广东省		2018	广东省新一代人工智能发展规划（2018-2030）	到 2025年在智能制造、智慧政府、智慧交通、智慧医疗、智慧教育、智慧金融、智慧安防等重点领域打造 250个全球领先的深度融合应用场景。
		2022	广东省新一代人工智能创新发展行动计划（2022-2025年）	从关键核心技术攻关、产业聚集发展、开源开放共享平台体系构建、与重点产业融合、创新生态系统构建等方面部署具体行动。
	深圳市	2019	深圳市新一代人工智能发展行动计划（2019-2023年）	从基础研究和核心技术、产品创新和产业发展、应用、基础设施和平台、人才、空间布局和创新生态等方面部署主要任务。
		2022	深圳经济特区人工智能产业促进条例	包括基础研究与技术开发、产业基础设施建设、应用场景拓展、促进与保障、治理原则与措施等方面的内容。
		2023	深圳市加快推动人工智能高质量发展高水平应用行动方案（2023—2024年）	从算力供给、关键核心技术与产品、产业集聚、场景应用、数据和人才要素等方面部署具体的行动方案。
	广州市	2021	广州国家人工智能创新应用先导区建设方案	着力培育核心产业集群，推广示范应用，建设平台载体，营造生态体系。
2021		广州市人工智能产业链高质量发展三年行动计划（2021—2023年）	推动建设 10 个人工智能产业园，开展 100 个人工智能典型场景应用示范，培育 1000 家左右人工智能企业	
浙江省		2017	浙江省新一代人工智能发展规划	到 2030年，形成较为完备的核心技术、产业发展、推广应用生态体系。
		2023	关于加快人工智能产业发展的指导意见（征求意见稿）	从核心技术突破、底层基础建设、创新场景赋能、特色聚集发展、产业生态构建等方面进行部署。
	杭州市	2022	建设杭州国家人工智能创新应用先导区行动计划（2022—2024年）	率先成为全国人工智能技术创新策源地、全国城市数智治理方案输出地、全国智能制造能力供给地、全国数据使用规则首创地、全国人工智能产业发展主阵地。
		2023	关于加快推进人工智能产业创新发展的实施意见	到 2025年，杭州成为全国算力成本洼地、模型输出源地、数据共享高地。

表 1 典型地方和城市 AI 政策措施

地方 / 城市		时间	政策	重点内容
江苏省	南京市	2018	江苏省新一代人工智能产业发展实施意见	从实际产业出发，构建新一代人工智能产业体系，建立协同发展机制，加快培育和发展人工智能产业
		2017	关于加快人工智能产业发展的实施意见	形成“两中心、三片区、一示范”的人工智能产业发展空间格局，到 2025 年，把南京打造成全球有影响力的人工智能创新应用示范城市。
		2023	南京市加快发展新一代人工智能产业行动计划（2023-2025 年）	到 2025 年，围绕钢铁、石化等重点行业和文旅、卫生等重点领域，打造 100 个可复制、可推广的标杆型示范应用场景。
安徽省		2018	安徽省新一代人工智能产业发展规划（2018-2030）	从关键技术、支撑平台、产品和服务、“AI+”行动、智能社会、产业集聚、人才队伍、基础设施等方面进行部署。
		2019	安徽省新一代人工智能产业基地建设实施方案	重点建设智能语音、智能机器人、智能家居、智能网联汽车等产业基地。
		2020	安徽省人民政府关于支持人工智能产业创新发展若干政策的通知	从创新能力、支撑平台、项目建设、应用示范、数据开放、产业集聚、试验区、基金支持、行业服务、人才支撑等十个方面推动 AI 产业发展。
	合肥市	2019	关于加快推进新一代人工智能产业发展的实施意见	到 2025 年，AI 在智能农业、智能制造、智能医疗、智慧城市等领域得到广泛应用，形成 60 个以上深度应用场景，建设 100 个以上应用示范项目。到 2030 年，形成较为完备的人工智能产业链和高端产业集群，产业规模和总体竞争力处于国内先进水平。
		2019	加快推进新一代人工智能产业发展若干政策	重点支持智能芯片、智能传感器、计算机视觉、智能语音处理、生物特征识别、自然语言处理、智能决策控制、新型人机交互等核心基础和关键技术。
		2020	合肥建设国家新一代人工智能创新发展试验区实施方案（2020-2023 年）	聚力开展前沿理论研究、关键技术研发、应用场景示范、创新企业培育、生态体系打造。
		2022	合肥市加快建设国家新一代人工智能创新发展试验区促进产业高质量发展若干政策	出台了技术研发、产品应用、企业招引培育、产业载体、平台、投资基金、生态体系等七方面的支持政策。

四川省		2018	四川省新一代人工智能发展实施方案	到 2030年，形成核心技术、关键系统、支撑平台、智能应用完备的产业链和高端产业群。
		2022	四川省“十四五”新一代人工智能发展规划	从基础理论、关键技术、重点产品、重大平台等十方面部署了主要任务。
	成都市	2018	关于推动新一代人工智能发展的实施意见	实施八项 AI+重点工程，促进智能经济和智能社会发展。
		2019	成都市加快人工智能产业发展推进方案（2019—2022年）	以平台聚合能力、以能力营造场景、以场景撬动市场、以市场壮大产业。
		2022	成都市新一代人工智能产业发展规划（2022— 2025）	从创新、产业提升、企业、产业协同、场景、生态等方面部署 AI产业发展重点任务。
		2023	成都市加快大模型创新应用推进人工智能产业高质量发展的若干措施	从强化智能算力供给、提升创新策源能力、提升产业发展能级、构建全域场景体系、加强生态要素聚集五个方面制定若干相关措施
重庆市		2020	建设国家新一代人工智能创新发展试验区实施方案	包含技术创新、基础支撑、产业应用、人才体系、开放协同等方面的内容。
		2023	以场景驱动人工智能产业高质量发展行动计划（2023—2025年）	围绕重点行业、重大项目、重大活动、新赛道布局重大场景，优化生态建设，探索 AI产业发展新模式新路径。
陕西省		2019	陕西省新一代人工智能发展规划（2019-2023年）	包含加强基础理论和关键技术研究、建设人工智能创新平台、培育 AI优势产品、推动人工智能示范应用、加速人工智能企业创新发展、促进人工智能产业集聚发展等六方面任务。
	哈尔滨市	2022	哈尔滨市加快新一代人工智能产业发展实施方案	到 2025年，AI在智慧农业、智能装备制造、寒地场景应用等领域深入融合运用。
		2022	哈尔滨市建设国家新一代人工智能创新发展试验区若干政策	提出了关键技术、平台、产业载体、应用场景、招引企业、国际合作、培育引进人才、金融、生态、数据开放等十方面的政策。

续表 1 典型地方和城市 AI 政策措施

4.2 AI产业发展趋势

4.2.1 构建产业生态是各国 AI产业发展的“主战场”

美国全方位促进AI产业发展。美国国家科学基金会认为，推进AI发展不仅需要巨额科研投资，还需要建立与之相匹配的强大且可持续性的AI创新生态。美国时任总统特朗普2019年签署《美国AI倡议》行政令，要求集中联邦资源发展AI，并从研究投入、数据和模型、标准、技术工人、国际合作等五个方面促进AI发展。美国随后正式颁布《2020年国家AI倡议法案》，并设立国家AI办公室（NAIIO），作为美国创新生态系统中国家支持AI研究和制定实施政策的中心枢纽。美国2023年发布第三版《国家AI研发战略计划》，确立了九项战略，如表5.2所示。

	战略内容
战略 1	长期投资基础和负责任的 AI 研究
战略 2	开发有效的人类 -AI 协作方法
战略 3	理解并解决 AI 的伦理、法律和社会影响
战略 4	确保 AI 系统的安全性
战略 5	开发用于 AI 培训和测试的共享公共数据集和环境
战略 6	利用标准和基准衡量和评估 AI 系统
战略 7	更好地了解国家 AI 研发人才的需求
战略 8	加强公私伙伴关系，加速 AI 发展
战略 9	建立有原则和可协调的 AI 研究国际合作方法

表 2 美国国家 AI 研发战略规划

欧洲强调发展可信任的 AI。欧盟 2018 年通过《AI 通讯》，重点关注了 AI 伦理。2019 年启动“欧洲联盟 AI”（AI4EU）项目，汇集了 21 个国家的 79 家顶级研究机构和企业，促进欧洲 AI 生态系统建设。欧洲议会 2023 年通过了《AI 法案》谈判授权草案，对 AI 的使用和安全提出了新要求。英国 2021 年公布了为期 10 年的《国家 AI 战略》，主要包括：投资并规划 AI 生态系统的长期需求，支持向 AI 经济转型，确保对 AI 技术进行有效治理，保护公众和基本价值观等。法国 2018 年发布《法国 AI 国家战略》，内容包括吸引高质量研究人员、建立国际水平研究中心、加强人才培养等。2023 发布了《AI 行动计划》，重点支持法国和欧洲 AI 生态系统中的创新者，保护个人隐私。德国 2020 年发布了新版《AI 战略》，从专业人才、研究、技术转移和应用、监管框架和社会认同五个重点领域确定了新举措。

日韩强调发展为人服务的 AI。日本 2019 年发布《以人为中心的 AI 社会原则》，提出 AI 社会需要体现尊严、多元包容、可持续 3 个基本理念。日本 2022 年发布第三版《AI 战略》，从人才培养、赋能产业、技术体系、国际化等方面提出 AI 发展的战略目标。韩国 2019 年发布《AI 国家战略》，提出构建引领世界的 AI 生态系统、成为 AI

应用领先的国家、实现以人为本的 AI 技术等目标。韩国在 2023 发布有关 AI 发展的新计划，具体措施包括：在健康福利、卫生、教育、文化、抗灾应急、行政等各领域全面引进 AI 技术；同美国、加拿大、欧盟等的高校开展国际联合研究；将讨论数字化权利法案，强化 AI 伦理规范和可信赖性。

东南亚领先国家积极在 AI 领域进行布局。除新加坡之外的东南亚国家 AI 产业基础都比较薄弱，只有少数几个国家提出了 AI 政策和战略。新加坡 2019 年出台《国家人工智能战略》，计划在 2030 年成为人工智能广泛应用的智慧国家。新加坡还发布了《AI 治理框架》《AI 治理案例汇编》等，规范 AI 发展。印度尼西亚 2020 年发布了至 2045 年间发展 AI 的蓝图，AI 项目将聚焦教育和研究、医疗服务、行政体制改革、粮食安全和智慧城市等方面。马来西亚 2017 年颁布 AI 国家战略计划，已建设了 AI 创新生态系统，未来将开展 AI 立法、监督、道德等方面的研究。越南 2021 年发布《到 2030 年的 AI 研究、开发和应用国家战略》，目标是使 AI 成为越南第四次工业革命中的重要技术。

中东国家将 AI 作为“后石油时代”经济社会发展的核心驱动力。阿联酋 2017 年启动 AI 战略，投资交通、健康、太空、可再生能源、水、科技、教育和环境等九个领域的 AI 技术，促进经济多元化。沙特推出国家数据和 AI 战略，不但通过主权财富基金进行投资，还计划到 2030 年吸引 200 亿美元投资，培训 2 万名数据和 AI 专家。埃及的 AI 国家战略包括教育和培训、数据利用、AI 系统开发、技术应用等内容。约旦 2023 年召开国家 AI 项目启动仪式，公共部门将加快使用 AI 进程。巴林强调了投资现代技术和 AI 应用、实现可持续发展目标的重要性，力求成为该地区 AI 及其应用领域的先行者。阿曼 2023 年启动“AI 赋能国民经济国家倡议”，旨在促进在政府发展项目中使用 AI 技术，提供投资机会。土耳其发布 2021-2025 年国家 AI 战略，确定了培训、创新创业、数据和基础设施、国际合作等方面的战略重点。以色列在 AI 方面的国家计划包括为研究创造条件、创建行业生态、改善公民生活等。吉布提 2023 年公布了《吉布提智能国家：数字经济路线图》，旨在广泛使用数字技术和创新方案，为公民所面临的问题创造可持续的解决方案。

4.2.2 基础大模型+领域大模型将成为 AI 产业的底座

与之前的 AI 应用方案相比，统一的大模型基座具有较好的泛化能力，应用时需要的时间、成本和数据量较少，为 AI 应用提供了切实可行的路径，谷歌、华为、阿里等大型科技公司加快了基础大模型的研发应用进程。基础大模型通用性强但专用性弱，在基础大模型基础上，利用行业数据训练出的适用于特定领域的专用模型，能够很好解决具体行业场景中的问题。我国拥有庞大的产业基础和消费市场，并且处于产业转型和消费升级过程之中，制造、金融、政务、医疗等领域的头部企业凭借在特定行业的知识、业务和客户资源基础，分别开发利用特定

领域的大模型,能够获得商业价值和社会价值。少量基础大模型与众多领域大模型相结合,将成为 AI产业的底座,满足 AI场景应用的需求,促进 AI产业繁荣发展。

4.2.3 大型科技公司是 AI产业发展的重要力量

大型科技公司拥有技术、人才、资金、业务等方面的优势,能集中力量和资源研发大模型、搭建平台、提供工具等,在 AI产业发展中起着重要作用。

国外大型科技公司先在自身业务中引入 AI应用,再对外输出技术方案。微软在 AI领域深耕多年,2019年起又通过投资 OpenAI布局 AI发展,在 ChatGPT推出后,迅速将其能力应用于搜索、办公软件等产品中,并开始对外输出 AI能力和方案。谷歌自 2016年起一直将 AI作为发展重点,ChatGPT爆火后,谷歌全面升级 AI聊天机器人 Bard,并利用 AI技术优化升级智能助手、搜索、YouTube等产品。Meta将 AI引入其应用程序,为 30 亿用户提供类似于 ChatGPT的体验,并计划推出数十个具有独特个性的 AI聊天机器人,以吸引年轻用户。

国内大型科技公司利用国内数字化转型的契机,一方面推动自身产品的 AI化,另一方面将 AI能力对外输出,助力传统企业数字化、智能化转型。华为盘古大模型已经陆续推出矿山、药物分子、电力、气象、海浪等 10 多个大模型,支撑 400 多个业务场景的 AI应用落地。阿里推出了通义千问大模型,淘宝天猫、钉钉、高德地图等产品将有序接入。阿里还开放通义千问的模型能力,让外部企业结合行业知识和应用场景,训练自己的企业大模型。百度发布文心一言大模型,自身所有业务都将基于 AI进行重构。百度行业大模型将深入能源、金融、教育、办公、媒体等领域,带动千行百业降本增效。腾讯发布了混元大模型,该模型已经深度支持 50 多个腾讯内部业务。腾讯通过腾讯云对外开放混元大模型的能力,为不同产业场景构建专属应用。

4.2.4 产业化沿着 ToB和 ToC两条路径向纵深场景拓展

AI在 ToB领域的创新应用引导传统产业智能化发展。随着行业大模型的推出,AI在 ToB领域的商业化应用逐步深化,AI应用从企业内部智能化逐步扩展到产业智能化,从提升企业效率扩展到提升产业整体效率。在 AI与产业融合的过程中,各行业会出现新的应用场景和商业模式,创新性加速推动产业智能化进程。

AI在 ToC领域的深入应用将产生新的人机交互方式和 AI公司。目前 ToC领域的应用尚有很大的发展空间,摩根士丹利最近一项针对 2000 人的调研发现,有 80%的人没有用过 ChatGPT或谷歌的 Bard。结合个人日常工作和生活,未来自然会话、语音助手、数字人等将在 ToC领域得到重点应用,将改变当前 C端互联网应用格局,并会诞生一些新 AI公司。语音助手是 AI最适宜进行 ToC应用的功能,自然会话有望成为主流的人机交互方式,而大模型的人机交互界面有望成为新一代流量入口。机器人技术和功能完善之后,应用于家政、教育、健康等

服务业，在即将到来的老龄化社会中带给人们更好的生活方式。数字人应用门槛和成本将大幅降低，成为像手机 APP 一样便捷的应用，提供情感支持、资讯、生活和健康管理等服务。

4.3 促进我国 AI 产业发展的对策建议

4.3.1 优化和完善 AI 产业生态

加强基础研究，加大对高校和研究机构的支持力度，解决大模型发展中的科学和理论问题，为我国 AI 产业发展奠定坚实的理论基础。促进 AI 与科学与工程计算、云计算、物联网和大数据的融合发展，推动 AI 算法、数据集和模型的快速迭代演变。破除我国大模型基础研究和应用研究之间的鸿沟，强化我国头部企业和高校、科研机构研究之间的联系，提升知识和技术的流动性，促进 AI 研究成果的转化和产业化。加快综合性人才培养，加大对中小型生成式 AI 技术企业的支持力度，促进形成繁荣可持续的 AI 产业生态。

4.3.2 积极推动 AI 落地应用

重点推动行业模型 + 领域知识的发展模式，让应用 AI 的机构和企业能以较低的成本和风险使用 AI 技术和能力。着重解决经济社会发展和企业经营管理中的紧迫现实问题，让在 AI 应用中的投入能产生明显的经济和社会效益。先替代培训、运维等重复且标准化工作，再逐步过渡到替代业务流程优化、商业模式创新、技术创新等创新性工作，逐步积累经验和技能，有序推进 AI 落地应用。

4.3.3 推动 AI 关键核心技术领先发展

发挥我国的 AI 人才、研究成果、产业链和消费市场优势，保持大模型基础研究在国际上的领先性，促进更多核心技术达到世界领先水平，提升我国 AI 技术与产业在国际上的竞争力与地位。推动在国内建立国外主流开源代码托管平台镜像，稳定我国 AI 技术发展的源泉。芯片指令集及设计、AI 分布式框架、云计算体系既要兼容开源，又要支持国内 AI 算力生态，在不断使用和发展已有技术成果的同时，确保国内形成领先和完整的技术体系。鼓励头部企业联合中小企业搭建先进的基于国产软硬件的 AI 训练和服务基础设施，研发领先的全栈国产化 AI 软硬件产品。

4.3.4 完善数据治理体系

强化系统思维，利用法治和技术手段，建立覆盖数据处理全过程、弹性包容的数据要素治理制度，政府、

企业和社会力量多方协同完善数据治理体系。加强数据积累、数据标准化，促进数据线上化、及时化、准确化等，为生成式 AI 技术发展和应用提供数据资源基础。建立更加完善的数据安全体系，保障公有云中的数据的安全，打消在 AI 训练和使用中对数据安全的担心。建设 AI 时代下的高质量语料库，坚持政府引导、市场主导，提升高质量数据的开放程度，加强标准规范建设和监管，提高语料库建设和应用的效率。

5 写在最后

20 世纪最有影响力的哲学家之一维特根斯坦(Ludwig Wittgenstein) 对语言和认知之间的关系进行了非常深刻的思考。维特根斯坦在其早期著作《逻辑哲学论》中提出，语言的结构反映了世界的结构。命题的任务是描绘事实，这种描绘与其所代表的现实事物之间存在某种“逻辑形式”的对应关系。真实的命题描绘了现实的事实，而命题的结构和逻辑形式能够映射出现实世界的结构。当我们审视大语言模型，其工作原理可以与维特根斯坦的这一思想相类比。大语言模型通过分析大量的文本数据，捕捉和模拟语言的统计结构。它的训练目标是捕捉数据中的模式，这些模式反映了人们如何描述和认知他们的世界。换句话说，大语言模型尝试模仿这些模式，从而能够生成与人类描述相似的文本。而被大量语料训练形成的语言模式所体现的应该就是现实事务的“逻辑联系”，从而体现出某种程度的推理的能力。

不过，维特根斯坦在学术研究的后半期，对其早期的哲学思想进行了深刻的反思和修正。他从一个更为实用的和实践导向的角度来重新考虑语言、意义等哲学问题，对语言的本质和功能有了更为复杂和丰富的理解。在其著名的《哲学研究》(Philosophical Investigations) 中，关于语言与现实世界的关系，他强调语言的多样性，并提出了“语言游戏”的概念，表示语言并不总是描绘现实，而是在各种社会实践中获得其意义。维特根斯坦的晚期思想为我们提供了一个深刻的视角，来看待语言、意义和实践之间的复杂关系。从维特根斯坦的哲学观点，似乎也可以预见生成式大语言模型，尽管在技术上取得了令人震惊的进展，但在真正理解语言和其背后的深层意义方面，仍然有许多挑战需要克服。另一方面，6 月 9 日，在 2023 北京智源大会开幕式上，深度学习的重要开创者和卷积神经网络之父 Yann LeCun 在 Keynote 演讲中，再次提出他的观点，即基于自回归的生成语言模型无法获得关于世界真正的知识，要实现真正人类水平 AI 的这一宏伟目标，他再次提出了「世界模型」(World Model) 的研究方法论。

其实，我们完全不必去纠结于生成式大语言模型是否具有“真正的意义的理解能力和推理能力”，乃至它是否是走向“通用人工智能（或人类水平的智能）”的正确路径这样的“大问题”。我们现在从具体的实践中就可以清晰地预见到，生成式 AI 技术，在知识处理和应用领域具有巨大的潜力，而这一潜力，无论是

和行业知识、企业私域知识，乃至个人的知识体系（包括经历）进行深度融合，就将会爆发出巨大的场景。

例如，我们大胆预测，很快，每一个人都会拥有自己的数字分身，企业的每一个岗位，都会出现一个岗位数字助理的角色。

本节虽然是白皮书的收尾，但生成式 AI 所引发的技术变革和应用场景的爆发才刚刚开始，参照过去几次大的技术范式变革对社会生活和经济活动的影响方式和程度，目前最常见的“+大模型”的应用模式有可能还远不是生成式 AI 技术真正的“杀手级”应用场景。而生成式 AI Native 的应用，不仅仅在应用本身，还有人们的使用模式，或者是基于该技术的应用开发方式，都应该是“原生”形态的。

纸质形式编撰的白皮书的内容，显然是远远无法跟上生成式技术和应用的高速发展的，但正如开篇所说，我们希望这个白皮书对我们的企业客户朋友们快速构建新技术知识框架，能有少许帮助；对生成式 AI 技术应用领域里的技术生态伙伴朋友们，能有一个交流的“话题平台”，共同研究、总结、积淀生成式 AI 技术对企业数字化转型的推动的方法论和最佳实践。文中的疏漏和谬误在所难免，我们会根据朋友们的反馈，持续迭代更新，后续以电子版、印刷版结合的方式保持和各位的交流。

6 引用

1. Gartner's glossary: Digital Transformation,
<https://www.gartner.com/en/information-technology/glossary/digital-transformation>
2. Andrej Karpathy's Blog : Software 2.0,
<https://karpathy.medium.com/software-2-0-a64152b37c35>
3. 《2023年新型智算中心算力池化技术白皮书31页》 ,
<http://221.179.172.81/images/20230907/41591694069780486.pdf>
4. 民生证券: 国产 AI算力芯片全景图,
https://pdf.dfcfw.com/pdf/H3_AP202303201584401192_1.pdf
5. 亚马逊 AWS: 什么是基础设施即服务(IaaS),
<https://aws.amazon.com/cn/what-is/iaas/>
6. 国家信息中心 2020年发布的智算中心建设指南,
<http://scdrc.sic.gov.cn/News/339/10713.htm>
7. Scaling laws for neural language models,
<https://openai.com/research/scaling-laws-for-neural-language-models>
8. An empirical analysis of compute-optimal large language model training,
<https://www.deepmind.com/blog/an-empirical-analysis-of-compute-optimal-large-language-model-training>
9. Introducing Superalignment,
<https://openai.com/blog/introducing-superalignment>
10. Model Parallelism,
<https://huggingface.co/docs/transformers/v4.15.0/parallelism>
11. Jim Fan's twitter about speedup,
<https://twitter.com/DrJimFan/status/1666132981839437825>
12. Zhu, Xunyu, et al. "A survey on model compression for large language models."
arXiv preprint arXiv:2308.07633 (2023).
13. Zhao, Wayne Xin, et al. "A survey of large language models."
arXiv preprint arXiv:2303.18223 (2023).
14. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study, <https://dl.acm.org/doi/fullHtml/10.1145/3544548.3580688>
15. Lilian Weng's Blog: Prompt Engineering,
<https://lilianweng.github.io/posts/2023-03-15-prompt-engineering/>
16. Pan, Shirui, et al. "Unifying Large Language Models and Knowledge Graphs: A Roadmap." arXiv preprint arXiv:2306.08302 (2023).
17. Li, Xingxuan, et al. "Chain of Knowledge: A Framework for Grounding Large Language Models with Structured Knowledge Bases." arXiv preprint arXiv:2305.13269 (2023).

18.Baek, Jinheon, Alham Fikri Aji, and Amir Saffari. "Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering." arXiv preprint arXiv:2306.04136 (2023).

19.Besta, Maciej, et al. "Graph of thoughts: Solving elaborate problems with large language models." arXiv preprint arXiv:2308.09687 (2023).

20.Yao, Shunyu, et al. "React: Synergizing reasoning and acting in language models." arXiv preprint arXiv:2210.03629 (2022).

21.Lilian Weng' s Blog: LLM Powered Autonomous Agents, <https://lilianweng.github.io/posts/2023-06-23-agent/>

22.Liu, Yang, et al. "Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment." arXiv preprint arXiv:2308.05374 (2023).

23.之江实验室: 2023生成式大模型安全与隐私白皮书,
<https://www.zhejianglab.com/home>

24.可信 AI技术和应用进展白皮书,
<https://www.tsinghua.edu.cn/>

25.Chen, Bo, et al. "xtrimopglm: Unified 100b-scale pre-trained transformer for deciphering the language of protein." bioRxiv (2023): 2023-07.