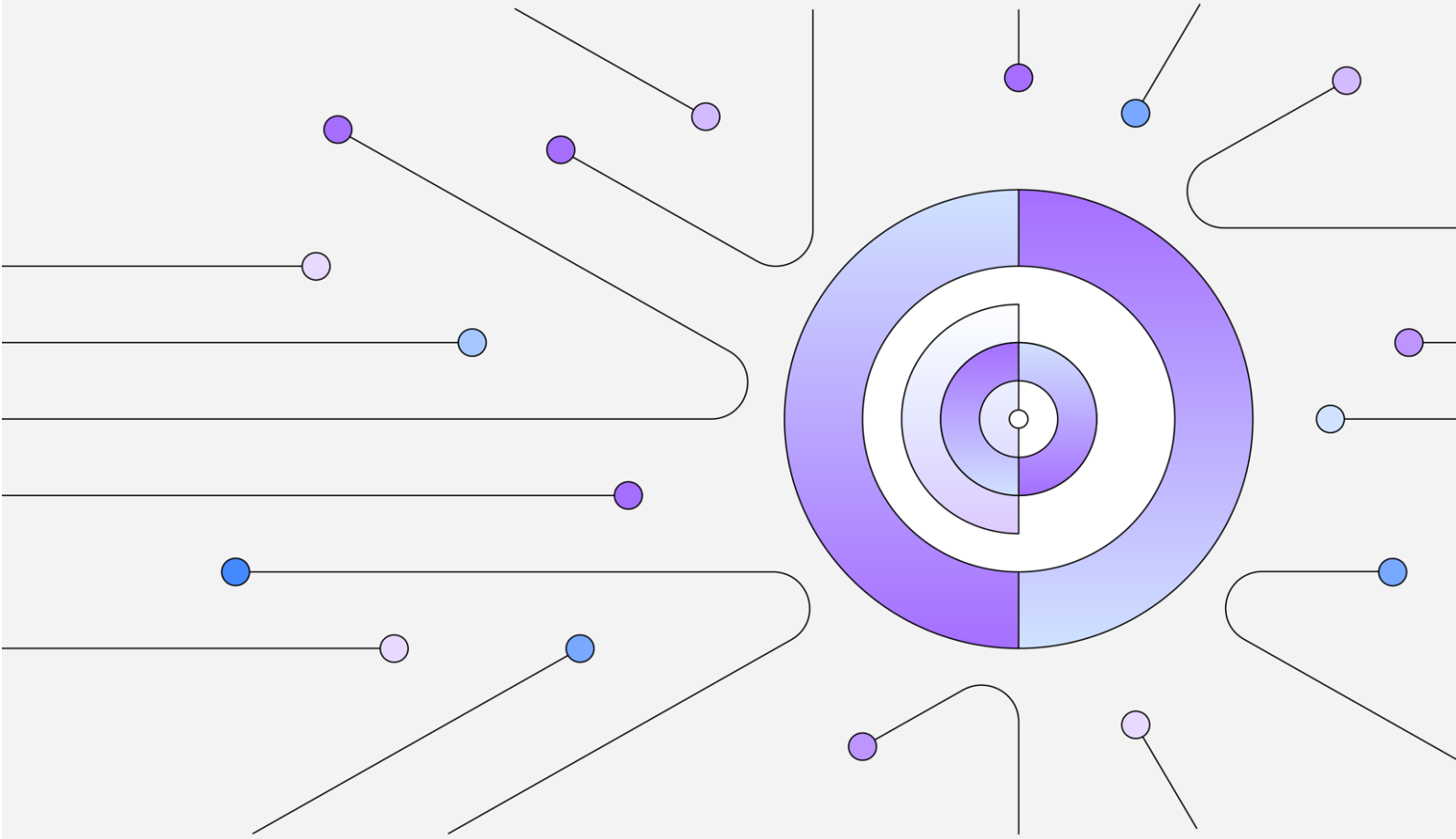
中国开源软件(OSS)推进联盟
China OSS Promotion Union

IBM

可信赖的企业级生成式人工智能白皮书



版权声明

可信赖的企业级生成式人工智能白皮书

编写委员会

顾问：陆首群

策划：谢东 程海旭 刘澎 梁志辉 孟繁晶

主编：程海旭 刘泽宇 石延霞 罗东文 张颖 刘晓金 孟迎霞 鞠东颖

工作组：（按照姓氏首字母排列）

白默涵 程文杰 初德高 董琳 樊斐 冯媛 葛巍 韩艳艳 姜朋慧 荆琦 李博文 李青 廖文静 刘佳怡 刘默驰 隆云滔 田忠 徐斌 徐孝天 杨军 杨悦 元中方 袁怿 原雪洲 臧倩 张侃 张玉明 赵则名 朱茉 庄雪吟

贡献者：（按照姓氏首字母排列）

曹岚 陈栋 丁伟 都娟 何蕾 李变 李玲 刘俊 刘胜利 倪栋 聂锦程 庞文峥 沈海军 孙盛艳 王彩彩 王积杰 王君 吴敏达 杨继辉 姚勇 张家驹 赵登科 赵蓉 郑维珺

序 言

生成式人工智能触发了新一轮人工智能浪潮，人工智能（AI）正在以前所未有的速度和规模，重塑着我们的生活和和工作方式，在推动经济转型和社会进步中展现出巨大的潜力。

企业是技术与创新转化为核心生产力的重要载体，那么企业在 AI 时代，如何打造新生产工具形成新生产力，帮助企业产销的产品持续的迭代与进化？ 可信赖的 AI 的重要性不言而喻。2019 年，我发表了“评人工智能如何走向新阶段”？ 触发了业界对人工智能发展方向的热烈讨论。同年 8 月份，COPU 提出研发 XAI 的任务，倡议机器学习、深度学习必须克服其自身的缺陷，打破黑盒子痼疾，建立可解释的机器学习模型，实现可解释、可信赖的人工智能，这在国内乃至全球都是最早提出这个任务的少数机构之一。2020 年 6 月，COPU 主办《第 15 届开源中国开源世界高峰论坛》，邀请 IBM 副总裁 Todd Moore 在会上作“可信任人工智能（反欺诈、可解释、公平性）”的报告，IBM 程海旭团队与 COPU 在此话题方面也进行多次研讨，并且应 COPU 要求写了三篇文章回应 COPU 提出的问题。并且，IBM 开源了针对反欺诈、可解释性和公平性的 AI 工具套件，也标志着可解释性 AI（XAI）的重要进展。IBM 作为全球 AI 治理平台的领导者，致力于将前沿科技转化为生产力，为企业提供开放、可信、有针对性的 AI 解决方案，共同开启企业级可信 AI 的新时代。

在如何帮助企业采用 AI 新技术形成新质生产力方面，尤其是当前 AI 技术日新月异、百模大战，技术重塑业务有其复杂性、差异性与多样性，在模型的选择、训练与调优、数

据的准备等技术问题，乃至场景价值、投入与产出等策略性问题上，都有着不同企业的疑虑与困惑。白皮书对于企业关注的 AI 模型及平台、数据治理以及 AI 治理等重点领域都有先进经验与理念的分享。在场景价值方面，白皮书通过深入分析汽车、金融等行业的成功案例，展示了 AI 技术如何助力企业实现转型和创新。在未来，人工智能的发展将继续以可信、安全为目标，依托算法、算力、数据为核心，帮助企业在 AI 智能时代持续进化，进而推动社会智能化的全面发展。

本白皮书也强调开源在推动 AI 发展中的重要作用。开源不仅促进了技术的透明性，还加速了研发进程，为构建开放、共享、协同、自由的 AI 生态提供了坚实基础。相信《可信赖的企业级生成式人工智能白皮书》的每一位读者都会开卷有益。

陆首群教授　中国开源软件推进联盟名誉主席

前 言

2024 年 3 月李强总理代表国务院在十四届全国人大二次会议上作的《政府工作报告》中，首次提出了开展“人工智能+”行动，这表明国家将加强顶层设计，加快形成以人工智能为引擎的新质生产力。

在企业端，人工智能产业的发展已驶入快车道，“让 AI 成为核心生产力”已经成为企业领导的迫切需求。据中国信息通信研究院公布的数据，2023 年中国人工智能核心产业规模达到 5784 亿元，增速 13.9%[1]。根据麦肯锡研究报告，到 2030 年前，生成式 AI 有望为全球经济贡献约 7 万亿美元的价值，其中中国有望贡献其中约 2 万亿美元，将近全球总量的 1/3[2]。

AI 不仅可以推动整体经济和 GDP 的大幅增长，还将为那些善用 AI 的个人和组织带来前所未有的竞争优势。放眼全球，生成式 AI 对高科技行业将产生最为显著的影响；在中国，先进制造、电子与半导体、消费品、能源、银行将是受影响最为显著的 5 大行业。

基于此，IBM 联合中国开源软件推进联盟(COPU, China OSS Promotion Union)，结合双方对企业应用生成式 AI 的深刻洞察、技术研究和业务实践，共同发布此报告，致力于推动企业高效、可信、负责任地应用生成式 AI，帮助企业打造新的竞争力，成为 AI 时代的真正受益者。

本报告首先阐述了生成式 AI 的演进和现状、全球立法和治理概况、应用前景和商业价值、风险与挑战、企业应用的关键因素；其次，对企业级生成式 AI 的参考架构进行了全面介绍，包括 AI 模型平台、数据平台和服务、治理、基础支撑平台、AI 应用，并展示了具有代表性的企业级应用生成式 AI 的真实案例和实施价值；最后提出企业应用生成式 AI 的战略规划方法及步骤，并对生成式 AI 的未来发展进行了展望。

目录

一 引言与背景

1.1 生成式人工智能的定义与演进

1.1.1 生成式人工智能的定义

生成式人工智能（Generative AI）是人工智能技术从上世纪 50 年代开始后，经过专家系统、机器学习两个发展阶段，演进到 2010 年代初出现的一种深度学习模型（如图 1）。它通过学习数据分布模式和规律，生成高质量的文本、图像、音频、视频四大基础模态，以及跨模态内容生成。

例如，通过学习大量文本数据，生成式 AI 可以生成具有类似风格的文章、小说、诗歌等文本作品。通过学习图片数据分布规律，生成式 AI 可以生成符合该分布规律的全新图片。通过对音频的深度学习，生成符合不同场景需求的数字人播报、语音客服、智能家居。使用深度学习模型对图像或视频进行分析和理解，再根据特定算法生成新的视频。最后，这些不同的模态还可以实现跨模态转化和生成，如将文本转化为图像、音频或视频，将图像转化为文本、音频或视频，应用于艺术创作、广告营销、教育培训、医疗诊断等领域。

生成式 AI 与之前传统 AI（也可称为判别式 AI Discriminative AI）最根本的不同在于：创造。生成式 AI 具有更大的灵活性和创造力，可以更好地模拟人类的想象力和创造力，生成更加多样化和全新的数据内容。而判别式 AI 则主要专注于已有数据的分类和预测，通过学习数据特征和标签之间的关系，进行模式识别和预测。例如判别式 AI 只可以区分出猫和狗的图片，而生成式 AI 则可以生成逼真的狗的图片。

基于这样不同的技术路径，生成式 AI 与判别式 AI 的成熟程度与应用方向也不同。判别式 AI 的底层技术相对成熟，在各个领域都有广泛的商业应用，包括人脸识别、推荐系统、风控系统、机器人、自动驾驶等。而生成式 AI 则在 2015 年前后才开始迅速发展，主要应用在内容创造、人机交互、产品设计等全新领域。

1.1.2 生成式人工智能的演进

生成式 AI 技术从 2010 年代初出现后，发展到 2022 年底，主要经历了三波浪潮：

第一波浪潮：2010-2015年。小型模型蓬勃发展。

变分自动编码器（variational autoencoders, VAEs）是第一个广泛用于生成逼真图像和语音的深度学习模型，为当今的生成式 AI 奠定了基础，也是当今大语言模型（large language models, LLMs）的基础。VAEs 基于编码器和解码器块构建而成。具体来说，编码器将数据集压缩为密集表示形式，在抽象空间中将相似的数据点排列得更紧密。解码器从这个抽象空间中进行采样以创建新内容，同时保留数据集的最重要特征。VAEs 不仅增强了重建数据的关键能力，而且还可以输出原始数据的变化形式。

这种生成新数据的能力引发了一系列小型模型的快速发展，其中 2014 年出现的生成式对抗网络（generative adversarial networks, GANs）具有突破性影响。GANs 由生成器和判别器组成，通过同时训练生成器和判别器来学习生成新的数据实例，以及更具创造性和多样性的文本。

第二波浪潮：2015年-2017年。模型规模竞赛风起云涌。

这个阶段，生成式人工智能领域出现了越来越多较大规模的模型。特别是基于循环神经网络（recurrent neural networks, RNN）和卷积神经网络（convolutional neural

networks, CNN）的生成模型，能够更好地捕捉上下文信息，生成更连贯、准确的文本，生成更加逼真的图像。

例如，2015 年，在计算机视觉领域，残差网络（residual network, ResNet）取得了突破性进展，这是一种深度卷积神经网络，能够在图像识别任务中取得更好的效果。2016 年，谷歌推出的 AlphaGo 成为第一个在围棋比赛中战胜人类职业选手的人工智能程序，这标志着人工智能在游戏领域的重大突破。

第三波浪潮：2017年-2022年。基础模型横空出世。

2017 年，里程碑式论文“Attention is all you need”提出一种全新的神经网络架构：Transformer。Transformer 使用一种全新的自注意力机制来处理序列数据，与之前传统的循环神经网络需要手动设计或学习完全不同。具体来讲，Transformer 将“编码器-解码器”架构与文本处理机制相结合。编码器将原始文本转换为“嵌入”表示。解码器将这些嵌入与模型之前的输出相结合，并连续预测句子中的每个单词。通过填空猜谜游戏，编码器可以了解单词与句子之间的关系，而无需任何人标记词性。Transformer 甚至可以在未指定特定任务的情况下进行预训练。学习这些强大的表示之后，就可以使用更少的数据来增强模型的专业化水平，以便执行给定的任务。Transformer 因其全面多样的功能而被称为基础模型。

同时，这个阶段的算力出现爆发式增长，并随着互联网、移动互联网的快速发展，数据也迎来指数级增长。这为大规模自监督或半监督的学习方法提供了强大的数据和算力保障，从而使得基础模型获得巨大成功，大大加速和扩大了生成式 AI 在企业中的应用领域，如自动驾驶、机器人流程自动化等。根据 IBM 发布的《2022 年全球 AI 采用指数》，全球企业采用 AI 科技的比例持续成长，达到 35%，比 2021 年上升 4%[3]。

1.2 生成式人工智能应用的现状

2022年底至今，生成式AI进入到第四波浪潮：更好、更快、更便宜的生成式AI产品。

2022 年可以说是生成式 AI 发展的又一个重要里程碑。继 2022 年 11 月 30 日 OpenAI 打响 chatGPT 第一枪后，全球领先厂商都快速地发布了自己的生成式 AI 产品，包括亚马逊云科技的 Bedrock，微软 Azure 的 OpenAI Service，IBM 的 Watsonx，谷歌的 Bard，阿里的通义千问，腾讯的混元，百度的文心一言。如果说之前的 AI 模型都是工具，这波浪潮的 AI 模型因为有接近全人类所有数据的支撑，而成为大脑。据不完全统计，截止到 2023 年 10 月，中国的生成式 AI 产品已超 300 个。从产品类型来看，主要包括文本生成、图像生成、视频生成三大类，其中文本生成的市场规模最大，占到了整个市场的 60%以上。图像生成市场增长迅速，视频生成市场尚处于起步阶段。

随着生成式AI技术的快速成熟，将出现第五波浪潮：杀手级应用程序（Killer APP)的出现。

随着大模型产品日益增加，大模型行业竞争将从比拼参数阶段，过渡到比拼落地应用阶段，会出现杀手级应用程序。Google 推出 Gemini 1.5 和 GPT-4 你追我赶，竞争激烈。百度在 2003 年 10 月 17 日的百度世界 2023 大会上，发布了文心达模型 4.0 版本，实现了基础模型的全面升级，综合能力比 GPT-4 毫不逊色。百度同时发布的十余款 AI 原生应用，涉及搜索、地图、文库、网盘，以及用 AI 原生思维打造的国内第一个生成式商业智能产品——百度 GBI，可以通过自然语言交互，执行数据查询与分析任务，还支持专业知识注入，满足更复杂、专业的分析需求 [4]。2024 年 2 月 Sora 产品的发布，更是让视频生成实现了代际跃迁，让虚拟现实成为可能。

1.3 生成式人工智能的风险及全球立法、治理概况

1.3.1 生成式人工智能的风险

生成式人工智能在快速发展的同时，也存在着一些潜在的风险，其所带来的与隐私保护、生成内容错误和幻觉、网络安全、偏见与伦理、知识产权等相关的风险已经显现。具体而言：

- **数据隐私保护：**训练数据如涉及商业秘密、保密信息等，或者未经用户同意，则可能涉及非法收集数据、侵犯个人隐私、侵犯他人知识产权或其他合法权益的情形。
- **生成内容错误和幻觉：**生成式 AI 依靠输入的数据进行预测和生成输出。但是，它有时会产生不准确或完全捏造的输出结果——即“幻觉”。这些幻觉可能会导致错误的决策或行动，从而可能给企业带来重大问题。
- **网络安全：**与任何数字工具一样，生成式 AI 系统也不能免受网络威胁。如前所述，这些人工智能系统有可能会被诱骗泄露敏感信息。因此，显然要制定强有力的网络安全协议。另一种新出现的威胁是“提示注入”，这是技术会利用提示来哄骗人工智能模型泄露本不该泄露的信息。更重要的是，实施这种技术并不一定需要高级技术技能。因此，首席安全官一定要全面掌握生成式 AI 可能遭到破坏的所有方式。只有了解每一种可能的攻击途径，他们才能真正保护自己的系统并保持强大的防御能力。
- **偏见与伦理：**人工智能的公正性取决于训练人工智能所依据的数据。如果这些数据中存在偏差，模型可能从训练数据中学习到偏见，进而生成带有种族、性别、宗教等方面的偏见内容。此外，还可能会出现其他伦理问题，如使用生成式人工

智能伪造艺术品、生成虚假文件、虚假新闻、伪造声音、网络钓鱼攻击、自动化的网络欺诈等，所有这些都是企业需要考虑的问题。

- **知识产权。**随着知识产权领域的不断发展，2023年或许将迎来生成式AI大规模应用的“Netscape时刻”。而随着公共数据和内容所有权的公平使用的相关政策、规则和诉讼不断增加，生成式AI也有可能迎来“Napster 时刻”（指行业的知识产权在互联网上公开、低成本地传播）。事实上，这些风险可能会促使企业更加关注专有数据和AI模型。

为了应对这些风险，很多国家正在努力制定伦理准则、监管政策，鼓励技术改进，以确保生成式人工智能的安全和道德使用。

1.3.2 全球立法、治理概况

生成式人工智能的飞速发展为各国立法和监管带来了新的挑战。由于人工智能技术的复杂性，与人工智能的开发、销售和使用相关的法律问题涉及范围很广，包括网络安全、数据安全、隐私、算法、内容、人工智能治理、知识产权、市场准入、反垄断与竞争、技术进出口等。因此，与人工智能有关的立法亦包括一系列的法律法规，不仅包括专门规范生成式人工智能的立法，还包括治理网络安全、数据安全、隐私保护和上述其他方面的立法。

本章节简要介绍截止 2024 年 3 月中国、美国和欧盟关于生成式人工智能的立法概况。

- **中国**

2023 年 7 月，中国国家互联网信息办公室（“网信办”）等七部门联合发布了中国首部关于生成式人工智能的规定，即《生成式人工智能服务管理暂行办法》（“暂行办法”）。该办法自 2023 年 8 月 15 日开始执行。利用生成式人工智能技术向中华人民共和国境内公众提供生成文本、图片、音频、视频等内容的服务属于暂行办法规制的范畴。但暂行办法明确将从事生成式人工智能技术研究、开发和应用的行业组织、企业、学术研究机构、公共文化机构等非公共服务提供者排除在其范围之外。

除此之外，中国现有的网络安全、数据安全和隐私保护相关法律法规，连同与人工智能相关的算法管理、深度合成管理、伦理准则等相关规定，均与暂行办法一起，共同建立我国生成式人工智能服务的法律框架。

- **美国**

在联邦层面上，白宫、国会和一系列联邦机构，包括联邦贸易委员会、消费者金融保护局和国家标准与技术研究所，已经提出了一系列与人工智能相关的举措、法律和政策。在短期内，美国的人工智能监管将更多地利用现有法律来对人工智能技术进行监管，而不是通过新的针对人工智能的法律 [5]。

- **欧盟**

2024 年 3 月 13 日，欧洲议会以压倒性票数通过《人工智能法案》。该法案预计将在 5 月或 6 月在走完所有审批程序后正式生效。法案中的相关条款将分阶段实施 [6]。该法案旨在保护基本权利、民主、法治和环境可持续性免受高风险人工智能的影响，同时兼顾 AI 技术的发展和创新 [7]。人工智能法案根据风险级别对人工智能的使用进行分类，禁止人工智能在特定方面的使用，并对高风险应用实施严格的监测和披露要求 [8]。

尽管全球对生成式人工智能的立法和监管措施在不同国家和地区有所不同，但一般来说，一些共同的趋势和原则逐渐出现，很多国家的立法重点通常集中在数据隐私、透明度和可解释性、网络安全、内容审核、知识产权保护、伦理审查和反垄断等领域。

总的来说，生成式人工智能的立法和监管仍在不断演进，以适应不断发展的技术和社会挑战。各国政府和国际组织都在努力寻找平衡，旨在确保技术的发展与社会、伦理和法律价值相一致。相关法律和政策仍在不断发展和完善过程中。

二 企业应用人工智能的机遇与挑战

2.1 生成式人工智能的应用前景与商业价值

2.1.1 生成式人工智能的应用前景

生成式人工智能的最终浪潮：世界模型的通用人工智能（AGI），全新的人机协同时代。

随着人工智能被投喂的大数据变为一切与我们的生产、生活息息相关的世界万物时，它会成为基于世界模型的通用人工智能。这个人工智能将会带来理解、生成、逻辑、记忆能力的突破，会出现独当一面的各类专业人才：数字艺术家、数字设计师、数字程序员、数字工程师、数字供应链专家等等。我们预计，到 2030 年，全能型、多模态的人工智能将进一步普及，人类的生产生活将进入全新的人机协同时代。生成人工智能有潜力彻底改变现有的经济和社会框架，就像电力和互联网一样。

2.1.2 生成式人工智能的商业价值

当下，“让 AI 成为核心生产力”已经成为企业领导的迫切需求。预计到 2030 年，AI 将提升人类生产力，带来高达 16 万亿美元的巨大价值 [9]。AI 不仅可以推动整体经济和 GDP 的大幅增长，还将为那些善用 AI 的个人和组织带来前所未有的竞争优势。不仅如此，AI 还可以帮助人类应对和解决诸如研发新药、改善制造业及食品生产效率、应对气候变化等最为紧迫的挑战。

IBM 商业价值研究院最新发布的《2023 年全球 CEO 调研》发现，四分之三（75%）的 CEO 认为拥有最先进的生成式 AI 的组织能够在竞争中获胜，43% 的 CEO

表示他们的企业已经在使用生成式 AI 来为其战略决策提供信息。企业级 AI 对企业最直接的价值是帮助优化业务流程，从而实现降本增效、提高生产力、以及提升客户体验。为了对生成式 AI 的商业价值进行更加客观的评估，IBM 商业价值研究院 (IBV) 联合牛津经济研究院，在 2023 年 5 月针对美国、澳大利亚、德国、印度、新加坡和英国的近 600 名企业高管开展了一项调研，其中包括美国的 200 位企业 CEO。在此次调研中，我们发现企业高管对生成式 AI 商业价值的观点，可以总结为以下三点：

第一：对生成式AI的投资回报积极乐观，但仍存谨慎态度。

受访企业高管预计，到 2025 年，基于过去几年开发的基准 AI 能力，生成式 AI 的投资回报率将从 2022 年的 7.1%增长到超过 10%[10]。因此，许多企业都计划在未来两年内继续推动生成式 AI 的采用。在 2022 年，只有 23% 的受访高管表示其组织对生成式 AI 进行了试点、实施、运营和优化，但预计到 2024 年这一比例将上升至 62%[11]。另外，在未来两到三年内，企业高管对生成式 AI 的投资预计将增长四倍。但是，即使这样，生成式 AI 项目的投资仍然仅占 AI 总支出的一小部分 [10]。说明受访高管对生成式 AI 的投资还是持谨慎态度。

第二：对生成式AI的加速采用面临巨大压力，但仍在努力掌握中。

首先，根据 IBV 的调研，64%的受访 CEO 表示正面临着来自投资者、债权人和贷款人的巨大压力，要求他们加速采用生成式 AI。超过一半的受访 CEO 表示，他们的员工要求加速采用生成式 AI（如图 1） [12]。

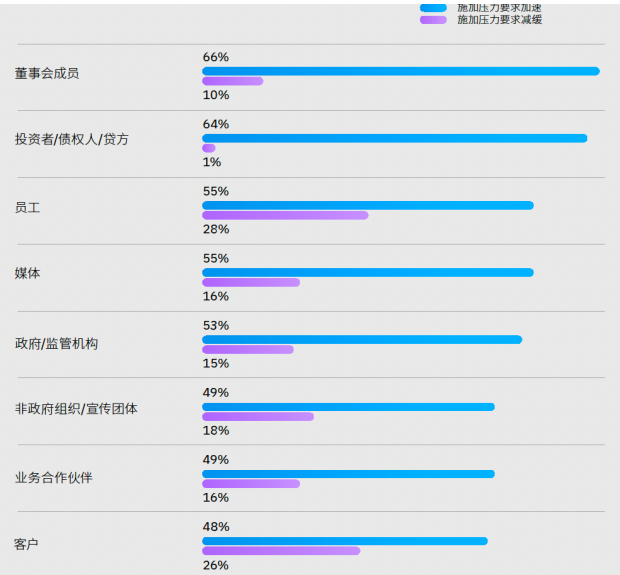


图 1 实施生成式 AI 的压力来源

面对这样的压力，企业高管快速掌握生成式 AI 技术。他们如今对生成式 AI 的认知水平远高于 2016 年传统 AI 出现第一波发展浪潮时的认知水平（如图 2）[10]。

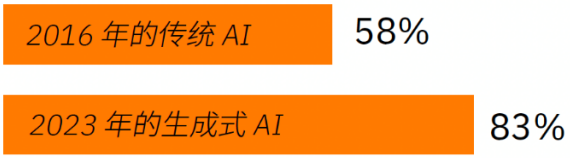


图 2 企业高管对 AI 的认知水平变化

第三：生成式AI的应用领域比较集中，但仍需与企业战略保持一致。

我们的调研数据显示，目前受访企业高管主要关注生成式 AI 在三个关键领域的应用：信息安全与信息技术，客户服务、营销与销售，研究与创新和产品开发（如图 3）[11]。

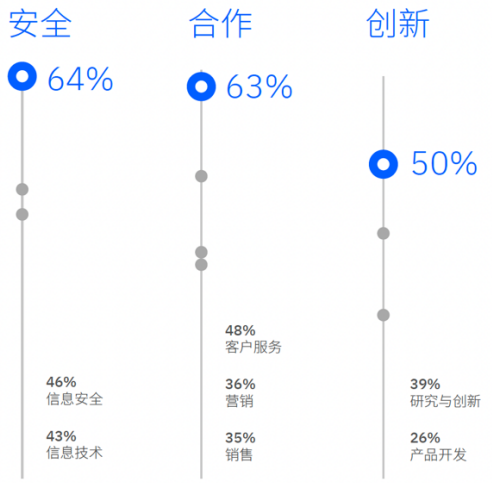


图 3 企业高管关注生成式 AI 的应用领域

同时，我们也看到，高管目前关注的这些优先领域大多是那些拥有最成熟 AI 能力的领域，而并不一定是战略痛点。因此，组织需要根据自身的战略能力和业务优先事项来明确 AI 的应用领域，确保 AI 的使用符合企业的长期战略，而不是将 AI 视为解决所有问题的“灵丹妙药”。

2.2 生成式人工智能带来的技术与非技术挑战

尽管生成式 AI 具有无比广阔的前景和潜力，但同时也带来了一些新的挑战。与其他颠覆性技术一样，企业在采用生成式 AI 的过程中，也需要做出适当的权衡，经过持续不断的实验和迭代才有可能取得成功。

2.2.1 生成式人工智能带来的技术挑战

生成式人工智能主要包括两大核心要素：海量数据、大规模算力。

首先，海量数据会带来以下8大技术挑战：

- **隐私安全性**：人工智能大模型处理大量的个人数据，隐私和安全性是一个重要关注点。保护数据的隐私，防止数据泄露和滥用是一个挑战，特别是在跨组织或跨边界数据共享的情况下。采用隐私保护的机器学习方法和安全数据分析技术，以便在保护隐私的同时实现机器学习的任务。
- **数据可得性**：海量、多源、动态更新的数据是训练模型和进行数据挖掘的必要条件。然而，对于某些领域和特定任务，获取足够量和高质量的数据是一项重大挑战。例如，某些领域的数据可能高度稀缺，或者数据的标注非常困难和耗时。在这些情况下，使用大量数据训练大模型可能不切实际。
- **数据准确性**：人工智能大模型的训练需要大量高质量的数据，并且通常需要对数据进行标注。数据质量和标注的准确性是一个挑战，因为错误或不一致的数据可能导致模型训练不稳定或性能下降。此外，对于某些任务，如图像识别和自然语言处理，数据的标注通常需要人类专家参与，这使得数据标注的成本变得非常高昂。
- **模型泛化性**：人工智能大模型在训练数据上表现出色，但在未见过的数据上可能泛化能力不足。过拟合是一个常见的问题，即模型在训练数据上过度拟合，而在新数据上的表现较差。选择适合的模型非常重要，这需要仔细地选择模型的超参数和架构，以便提高模型的泛化能力。
- **模型解释性**： 人工智能大模型通常被视为黑盒，即很难理解模型的决策和推理过程。这在某些应用场景中是不可接受的，如医疗和金融领域，因为解释模型的决策对于决策的可信度和可解释性至关重要。为了解决这个问题，研究人员正在研究可解释性的机器学习模型和方法，以便更好地理解模型的决策过程。

- **模型适配性**：在人工智能大模型的开发中，选择合适的算法和模型架构是关键。然而，从众多的算法和模型中选择最合适的一个可能是具有挑战性的，因为不同的任务和数据可能需要不同的模型来实现最佳性能。
- **模型可扩展性**：随着模型规模的增大，人工智能大模型的可扩展性和效率成为挑战。大模型需要更多的计算资源和存储空间，对于实时应用或边缘计算等资源受限的场景是否能高效运行是一个问题。
- **模型高效性**：优化模型的架构和参数，减少模型的计算和存储需求。采用模型压缩和量化技术，减小模型的规模，提高计算效率。使用分布式训练和模型并行化技术，提高模型训练和推理的速度和效率。

其次，大规模算力同样也会带来3大技术挑战：

- **算力强大性：**生成式 AI 需要处理海量的数据，这就需要强大的计算能力和存储能力。根据《2022-2023 全球计算力指数评估报告》，生成式 AI 计算市场规模将从 2022 年的 8.2 亿美元增长到预计的 2026 年的 109.9 亿美元，其占整体 AI 计算市场的份额也将从 4.2%增长到 31.7%[13]。
- **算力可用性：**对于人工智能大模型的训练和应用，算力可用性是一个重要的因素。由于大模型需要大量的计算资源，包括高性能的计算设备和大型存储空间来存放数据和模型，这对于许多组织和研究人员来说是一大挑战。除了硬件资源外，网络带宽和延迟也是影响大模型应用的重要因素。在分布式系统中，训练大模型通常需要将大量的数据从一个节点传输到另一个节点，这需要高带宽的网络连接和低延迟的通信。如果网络连接的速度很慢或者存在大量的延迟，那么训练大模型的时间将会大大增加，这可能会使得组织和研究人员难以承受。

- **算力优化性：**生成式 AI 的训练和推理过程需要大量的计算资源，因此需要不断优化算法和模型，降低计算复杂度和内存占用，提高计算效率。同时，在处理大规模数据时，如何提高单芯片算力、突破算力利用率、实现更高能效比，是算力基础设施需要面对的重要挑战。

2.2.2 生成式人工智能带来的非技术挑战

除了技术挑战之外，生成式 AI 还会带来一些非技术挑战，主要包括以下 4 个方面：

- **监管必要性：**生成式AI从诞生之日起，已迅速实现了“消费化”。这种大规模采用意味着一些用户可以在没有正式指导的情况下使用生成式AI。他们在没有护栏的情况下使用生成式AI，其行为可能无法受到监管，并且可能会导致不可预测的后果。如果缺乏适当的监督，组织就无法正确识别、量化或管理采用新兴技术的相关风险。在全球范围内，只有不到60%的受访高管认为其组织已经为AI监管做好了准备，69% 的受访高管预计会因采用生成式AI而受到监管罚款 [14]。因此，组织需要安全、负责任地利用强大的生成式AI，明确想要实现什么样的目标，以及实现这一愿景所需做出的改变。
- **社会伦理性：**人工智能大模型的发展和应用引发了许多伦理和社会问题，包括公平性、透明度、责任和权益等方面的考虑。因此，需要制定相应的政策和规范来确保模型的公正和可接受性；制定合适的法律法规和伦理准则，确保人工智能大模型的使用符合道德和法律要求；开展公开和透明的讨论，促进社会对人工智能技术的理解和参与；注重公平性和权益保护，进行数据脱敏和去偏倚处理，避免对特定群体的歧视和偏见。

- **环境保护性：**基础模型需要大量的计算、存储和网络资源，而这会消耗大量能源，产生高碳排放，给环境保护和气候变化带来了巨大挑战。据研究，训练一个大型自然语言处理模型的碳足迹与5辆汽车在其整个生命周期中的碳足迹大致相同。因此，企业应该适当考虑相关环保性。同时，社会各界正在研究如何加快大模型推理速度、降低算力成本、减少能耗，以此来突破预训练模型的发展制约。
- **人机协同性：**随着 AI 时代到来，企业需要快速实现人员技能的转型和提升，来拥抱 AI 浪潮。技术加速使每个人都变成了“超级个体”，人和机器的协作关系重新被定义和划分。人才需要合理地借助工具和技术，审时度势，提升自身价值与战斗力。而人才技能的转型往往伴随组织文化的更新，优秀公司早已把鼓励创新和学习的基因扎根在企业文化之中。

2.3 生成式人工智能在企业应用中的关键因素

企业在应用生成式 AI 时，需要重点关注三大关键因素：

第一个关键因素：组织和技能。

根据 IBV 调研，多达 80%的受访高管认为，由于生成式 AI 的兴起，劳动力角色和技能正在发生变化。展望未来，受访高管表示人才和技术技能至关重要，组织将优先建立和发展既能帮助员工使用生成式 AI，又能完成只有人类才能胜任的工作技能。随着生成式 AI 的日益普及，57%的受访高管预计创造力技能将变得更加重要。超过一半的受访高管认为技术技能、时间管理和优先级规划能力的重要性也会随着生成式 AI 的普及而大幅增加[15]。

另外，我们从调研中也发现，87%的高管预计生成式 AI 将更加广泛地增强员工的能力，而不是取代他们（如图 4）[15]。

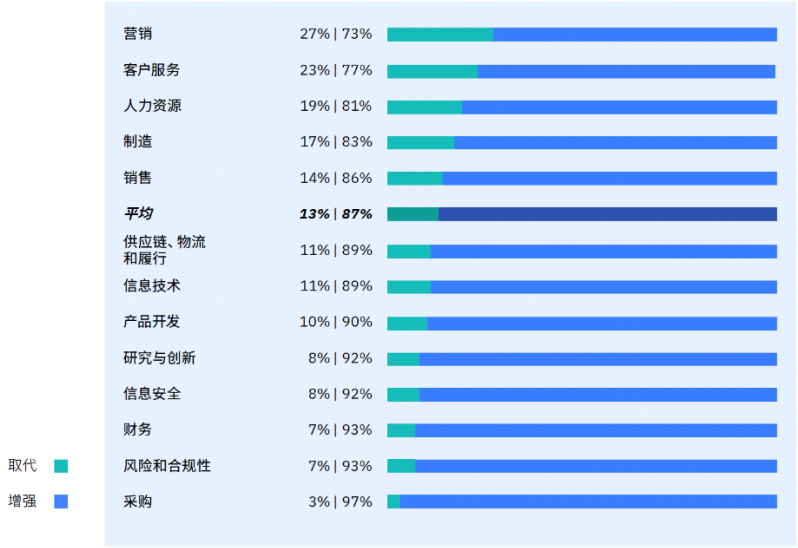


图 4 企业高管预计生成式 AI 对员工技能的影响

但是，并非所有职能的员工都会受到同等程度的影响。从上图我们可以看到，受访高管预计会用生成式 AI 取代的最多的三项职能是：营销、客户服务、人力资源。而最不可能取代的三项职能是采购、风险和合规、财务。一线员工可能会受到最大的影响，但也可能受益最多。

因此，为了帮助企业全员更好地适应和承担在不断变化的工作场所中的新角色和新责任，企业高管应全面领导并推动生成式 AI 转型。

首先，在组织层面，从转变观念、设定目标、建立原则、营造文化入手。

从“+AI”的被动思维转变为“AI+”的主动思维，即在设计之初就以 AI 为中心，这将有助于更深入地理解生成式 AI，增强响应市场形势变化的敏捷性，并确保投资和资金分配与整个组织各个层面的支持相一致。定义生成式 AI 采用的财务和非财务目标，并确定具体、可量化的措施，包括希望员工积极接受的变革。为 AI 的伦理道德使用设定界限。生成式 AI 模型很强大，但必须负责任地使用它们。这包括尊重隐私、透明度、公平性和问责

制。积极营造试验文化，认识到生成式 AI 对所有人都是新事物。鼓励团队使用生成式 AI 进行测试、迭代和改进，并跟踪成功指标。

其次，在人才层面，从选用育留着手。

了解人才资源的来源和分布情况，认识到潜在技能短缺，并将顶尖人才分配到竞争优势最关键的领域。评估生成式 AI 对员工团队的潜在影响，跨职能重新定义或重新部署角色，增强技能互补，依靠团队合力，以更好地利用生成式 AI。并成立 AI 技能学院，对具有相应资格的员工进行再培训和/或技能提升培训，不仅优先发展技术技能，还应优先增强协作、沟通和同理心。课程还应涵盖基础模型的合理使用和不当使用，从而促进负责任的 AI 使用。在培训的基础上，启动激励计划以推动职业发展。

最后，从运营层面，为了加快AI的采用，企业需要重塑和重建运营模式。

具体举措包括：促进跨职能理解，简化 AI 部署流程，并确保在整个组织中实现生成式 AI 和基础模型的优势；建立 AI 集成框架，以便在整个运营中无缝部署 AI；建立符合监管标准和最佳实践的稳健型数据与 AI 治理实践；在不同业务部门、技术团队、数据科学家和决策者之间营造一种协作式环境等。

第二个关键因素：负责任AI与伦理。

生成式 AI 如同当年的“西部淘金热”，对财富的追逐已经超过了规则和法规。但是如果组织太急于求成，而未考虑复杂的 AI 伦理问题，就可能会因短期利益而损害长期声誉。

根据 IBM 商业价值研究院调研：58%受访高管认为采用生成式 AI 存在重大伦理风险，如果没有新的治理结构或者至少更加成熟的治理结构，就无法管理这种风险 [16]。然

而，许多高管都难以将原则付诸实践。尽管 79%的受访高管表示 AI 伦理对其企业级 AI 方法很重要，但只有不到 25%的受访高管实施了 AI 伦理的共同原则 [17]。

因此，企业可以从以下三个举措入手，更好地构建企业负责任的 AI 和伦理体系：

首先，CEO 不能在 AI 伦理问题上推卸责任。根据 IBM 商业价值研究院调研，80%的受访高管表示，企业领导者（而不是技术领导者）应当对 AI 伦理负主要责任 [17]。CEO 必须掌控全局并为其他人开辟道路。除了决策以外，CEO 还必须负责向其他领导者普及关于新兴伦理问题的知识。通过将关于可信 AI 的对话提升到其他高级管理层和董事会的层面，CEO 可以确保这些关键利益相关者不会被边缘化。这样组织可以加快行动速度，同时保持领导层协同一致。

其次，通过满足客户期望来赢得信任。建立一个值得信任的品牌需要数十年的时间，而摧毁它只需要几天的时间。在数据泄露和不信任的时代，消费者、员工和合作伙伴对不以诚信行事的企业毫不宽容。根据 IBM 商业价值研究院的调研，37%的消费者曾为了保护隐私而选择更换了品牌 [18]。69%的受访员工表示，他们更愿意接受那些他们认为具有社会责任感的组织的工作机会 [19]。组织内需要建立自下而上的协作信任文化，让 AI 伦理成为每个人的责任，并让 AI 治理成为一项集体共同目标。同时，组织从内而外，需要广泛、透明地传达企业的伦理价值观。在内部对员工进行再培训，确保在工作中合理运用 AI，避免不当运用 AI。在外部，针对合作伙伴开展 AI 伦理和偏见识别培训，强调可信 AI 的重要性。

最后，为所有 AI 和数据投资做好伦理和监管准备。超过一半 (56%) 的受访 CEO 推迟了重大投资，等待对 AI 标准和法规建立清晰的认识 [20]。72%的组织将因伦理顾虑而放弃生成式 AI 带来的收益 [21]。企业掌舵者应做好准备，随时根据监管风向的转变和新出台

的法规做出调整。确保应用场景易于解释，AI 生成的工件清晰可识别，AI 训练保持透明且接受持续批判。建立归档文化，持续记录组织中使用 AI 的所有实例和相关治理，有效管理风险。通过清单来记录使用 AI 的每个实例，确保 AI 生成的资产可以追溯到基础模型、数据集、提示或其他输入。同时将这些源信息植入到数字资产管理和其他系统中。

第三个关键因素：数据和平台。

生成式 AI 模型需要大量数据，而负责任地提供数据则需要整个组织的协作。根据 IBV 最近开展的一项调研，60%的组织尚未建立一致的企业级生成式 AI 方法 [15]。

在混合云旅程中走得更远的组织更有可能发挥出生成式 AI 的优势，因为云转型需要更全面的数据方法。但是，主要利用云来降低各孤立领域成本的组织，可能需要重新审视其方法，通过打通孤岛实现互联互通。IBM 商业价值研究院的研究表明，大约五分之三的受访高管表示混合云和生成式 AI 在创造价值方面是相互关联的。另外 40%的受访高管仍在竭力让其多个不同平台保持协同一致 [15]。

统一数据可能是一项艰巨的任务，但如果缺乏明确的目标，那么可能会得不偿失。不过，基于可靠数据构建的混合云和生成式 AI 平台，可以开启通往更有价值的全新生态合作的大门。近三分之二的受访高管表示，生成式 AI 可以改善并加速与生态系统合作伙伴的数据共享 [15]。

因此，企业可以从以下三个方面打造协同一致的数据和平台：

首先，企业应评估并了解创建生成式 AI 用例的数据和混合平台需求。这就需要了解企业所拥有的数据类型，以及处理和分析此类数据的计算要求。依据这些需求，设定平台的选择标准，以支持使用生成式 AI 和相关基础模型。这些标准可能包括：用例特异性，成本（模型开发和运营费用），相关数据的可用性和可访问性，预测精度与计算效率之间的

平衡，安全措施和协议，所需的定制化程度，系统整体性能，跨不同环境的可移植性，法律和监管标准合规性。

其次，需要评估当前和潜在合作伙伴的实力，从中甄选出能够有效满足混合平台需求，并能为创建差异化优势助力的生态系统合作伙伴，共创成功。企业需要联合这些生态系统合作伙伴，确立共同的目标，使用一致的指标，并采用零信任安全实践，全方位提高整个生态系统的安全性。企业可以利用开放式混合技术，为组织和合作伙伴生态系统创建一致、可扩展和优化的通用平台。

最后，将基础模型集成到运营中，推动大规模部署时，需要确保可以扩展这些模型，而不会影响业务成效或导致运营中断。这就需要强大的模型管理、性能监控和持续改进机制。同时，由于基础模型需要访问大量、多样化且可能敏感的数据集，因此要建立稳健的数据治理实践。这包括符合监管标准和最佳实践的数据收集、存储、访问、处理和安全协议。

三　企业级生成式人工智能的技术、产品与解决方案

3.1 企业级生成式人工智能参考架构

为更好的应对企业生成式人工智能所面临的挑战，我们从技术要素、治理要素和规划实施方法三维度进行企业级生成人工智能参考架构的讨论，并在后续的章节中详细展开。

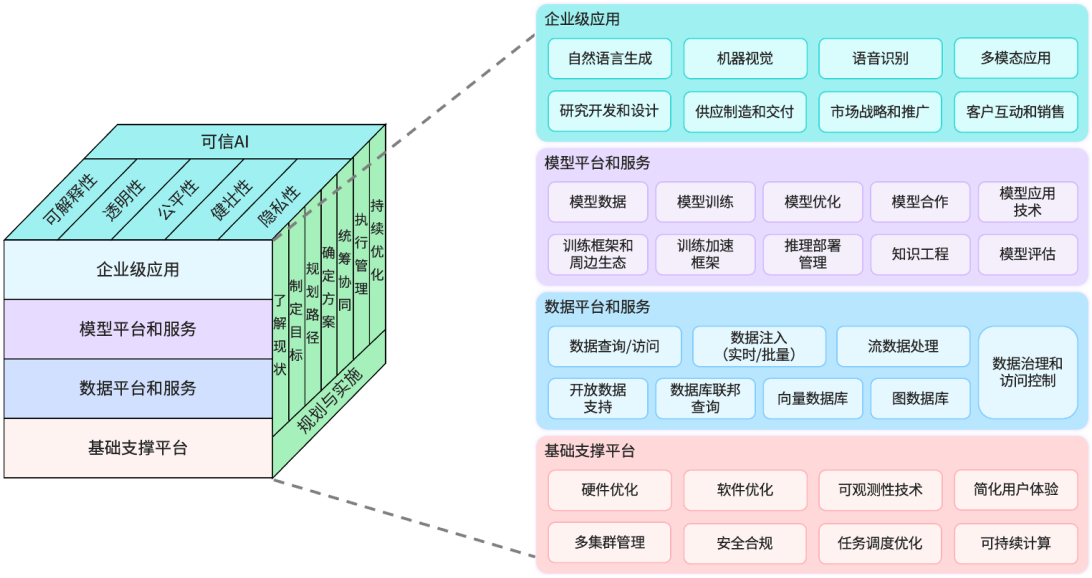


图 5 企业级生成式人工智能参考架构

3.1.1 企业级生成式人工智能参考架构的技术要素

从技术角度出发企业级生成人工智能架构的重点技术要素包括：模型平台和服务、数据平台和服务、基础支撑平台、企业级应用四大部分。

章节 3.2 将具体讨论模型平台和服务部分。依托基础支撑平台层所提供的基础设施服务，模型平台和服务部分为上层的人工智能应用提供全面的支撑，其内部又可细分为四个技术层面：训练框架和周边生态、训练加速框架、推理部署管理，以及模型与数据。这几个层面的功能实现了从模型训练到部署和应用的完整链条，可以从容应对大型模型应用于

企业实际场景中需要克服的诸多挑战。基于这些技术功能的支撑，我们进一步深入探讨了模型平台和服务在企业生成式人工智能实施过程中的模型的评估与选择，数据准备，模型训练、优化以及典型应用等方面所扮演的角色。

章节 3.3 将具体讨论数据平台和服务。数据是生成式人工智能的另一大基石，是企业的重要资产，为更好的满足企业对于专有数据的安全合规需要，数据平台和服务在落地实施过程中支持考虑多种部署方式相结合，这其中本地部署的场景具备一定优势。为实现高质量可信可靠的数据内容，规避“垃圾进，垃圾出”的风险，数据治理必须贯彻相关业务活动的始终。随着人工智能的发展，在数据平台和服务层面涌现出一些新需求，为了更好的管理多样化海量数据和知识，实现全方位的数据管理，新一代数据管理平台演进出了湖仓一体的架构。在文章中除湖仓一体的技术要素之外，我们还注意到开源开放的数据管理技术生态能够加速企业创新，快速适应市场变化。

作为承载生成式人工智能落地的基础支撑平台，我们以应对大规模数据处理，应对算力利用率，增强人机协同三个典型挑战为例，在 3.4 章节探讨了如何在企业数字化转型的过程中更好的应对生成式人工智能应用带来的挑战，或现行企业级数字化平台如何更高效稳定的服务于企业级生成式人工智能的落地实施。

基于以上技术要素，在 3.5 章节继而展开探讨了生成式 AI 的企业级应用。文中参考 IBM 的组件业务模型作为方法论，从业务赋能，研究开发和设计，供应制造和交付，市场销售，客户互动各个方面进行了阐述。随着生成式 AI 技术的到来，企业对 AI 的应用开启了一个新的篇章，也将迎来新的“黄金时代”。尽管“让 AI 成为核心生产力”已成为企业日益迫切的需求，但实际的落地应用却非一日之功。面对各不相同的应用场景和复杂

需求，企业管理者们也产生了诸多的困惑。文中重点分享了汽车、金融两大行业领域在生成式 AI 的成功经验。

3.1.2 企业级生成式人工智能参考架构的治理要素

从企业的角度出发，对于生成式人工智能的治理应该融入业务周期的各个环节，同时贯穿从 AI 应用到基础设施各个技术层面。可解释性、透明性、公平性、健壮性、隐私性是企业级生成式 AI 治理的五大关键特征。第四章将讨论如何将治理与 AI 全生命周期相结合，介绍不同架构层级的相关技术手段和工具，通过引入对应的评估技术和一系列量化指标矩阵，从而确保在企业级生成式人工智能的可信可靠，帮助企业实现和维护高标准的治理水平。

3.1.3 企业级生成式人工智能参考架构的规划与实施

企业级生成式人工智能架构的成功，离不开统筹的规划和合理全面的实施。第五章将展开规划时企业需重点考虑的组织要素，并结合生成式人工智能的特点展开了全面实施的方法步骤：了解现状、制定目标、规划路径、确定方案、统筹协同、执行管理、持续优化。

3.2 人工智能平台和服务

3.2.1 人工智能平台和服务的总体功能架构图

如图 6 人工智能平台和服务的总体功能架构图所示，人工智能平台的具体功能又可以分为训练框架及周边生态、训练加速框架、推理部署管理和模型与数据四个层次，实现了从模型训练到部署和应用的完整链条，为各种人工智能应用提供了全面的支持和服务。

图 6 人工智能平台和服务的总体功能架构图

- **训练框架及周边生态**：主要涉及各种人工智能模型的训练框架和相关的生态系统，包括各种开发工具、库和框架，以及数据处理、模型评估等辅助工具。
- **训练加速框架**：主要关注如何提高模型训练的速度和效率，涉及分布式训练框架，以及各种加速算法和优化技术，旨在提高平台的整体性能。
- **推理部署管理**：主要涉及模型的部署、管理和运行，推理引擎的选择和配置，以及运行时的监控和管理等工作。
- **模型与数据**：这一层可分为知识工程和基础模型两大部分。知识工程介绍构建、管理和利用知识库或知识图谱的技术，旨在支撑和增强基础模型功能。基础模型部分围绕

将模型应用于企业实际业务场景中的关键步骤和技术挑战，从模型的评估与选择，数据准备，微调与训练、合作、优化以及典型上层应用等方面进行展开。

3.2.2 人工智能平台第一层：模型训练框架及周边生态

在模型训练领域，有许多成熟的开源软件和工具可供选择，它们通常被组合使用以构建完整的模型训练流程，并逐渐形成了丰富的开源生态系统。下面主要介绍一些常见的开源软件和工具。

3.2.2.1 PyTorch

PyTorch 是一个用于机器学习领域的人工智能研究和商业生产的开源框架。它用于构建、训练和优化深度学习神经网络，用于图像识别、自然语言处理和语音识别等应用。 它为 CPU、GPU、多 GPU、多节点上的并行和分布式训练提供计算支持，同时它还拥有许多可用于不同领域的特定库和工具，具有灵活且易于扩展的特点，所有这些都使 PyTorch 成为机器学习领域的领先框架 [22]。

3.2.2.2 TensorFlow

TensorFlow 是一个开源深度学习框架，截止发稿时，它已成为世界上采用最广泛的深度学习框架之一。TensorFlow 为开发者提供了即时执行、计算图模型、简单易用的 API、灵活的架构和分布式处理等功能，可以在多架构和多核系统以及将计算密集型处理作为工作任务进行分配的分布式进程上运行。由于其灵活、可扩展和模块化的设计，TensorFlow 并不限制开发人员只能使用特定的模型或应用程序，开发人员不仅可以实现机器学习和深度学习算法，还可以实现统计和通用计算模型 [23]。

3.2.2.3 Keras

Keras 是一个基于 Python 的深度学习库，与其他深度学习框架不同。该项目易于学习和使用，并且具有在框架之间轻松移植模型的额外优势。Keras 尝试定义神经网络的高级 API 规范，提供用户界面，同时可以良好的兼容不同低层框架。基于 Keras 前端可以在研究中快速构建神经网络模型的原型。Keras 通过项目自身的图数据结构实现，摆脱了对于底层后端框架的图数据结构的依赖，使开发者无需精通后端框架实现细节 [24]。

3.2.2.4 Transformers

Transformers 为用户提供了可以轻松下载和训练最先进的预训练模型的 API 和工具。这些模型支持不同模态的常见任务，包括：

- 自然语言处理：如文本分类、命名实体识别、问答、语言建模、摘要生成、翻译、多项选择和文本生成。
- 计算机视觉：如图像分类、目标检测和分割。
- 音频：如自动语音识别和音频分类。
- 多模态：如表格问答、光学字符识别、从扫描文档中提取信息、视频分类和视觉问答等。

Transformers 支持 PyTorch、TensorFlow 和 JAX 之间的框架互操作性，这为用户提供了在模型的生命周期的每个阶段使用不同框架的灵活性，模型也支持导出 ONNX 或 TorchScript 等格式，可以方便地在生产环境上部署 [25]。

3.2.3 人工智能平台第二层：训练加速框架

训练加速框架主要关注如何提高模型训练的速度和效率，涉及分布式训练框架，以及各种加速算法和优化技术，旨在更快地完成模型训练过程，从而提高平台的整体性能。

3.2.3.1 Ray

Ray 是一个开源的分布式计算框架，由 UC Berkeley RISELab 开发，旨在为大规模、复杂的分布式应用程序提供高效、可扩展和易于编程的解决方案。相较于传统的分布式框架（比如 Hadoop、Spark 等），Ray 在 API 和工具集上有更丰富的支持，使得开发者可以轻松地构建分布式应用程序，且支持主流深度学习框架例如 TensorFlow、PyTorch 等。其核心优势在于其简洁的 API 和高度可扩展的架构，提供了一种简单而强大的方式来并行化和分布式计算，使得用户可以轻松地将单 di 机程序扩展到大型集群 [26]。

3.2.3.2 Colossal-AI

Colossal-AI 是一个分布式深度学习框架，它是一种用于高效训练大规模深度学习模型的开源软件框架。它旨在解决在训练过程中由于模型和数据规模庞大而遇到的各种挑战，例如内存限制、计算资源不足和训练速度缓慢等问题。Colossal-AI 通过使用一系列优化技术和并行计算方法，使得在有限的硬件资源下，能够更快地训练出更好的模型。Colossal-AI 的核心优势在于其灵活性和可扩展性。它支持各种深度学习框架，如 PyTorch、TensorFlow 和 MXNet，并提供了丰富的 API 和工具，使用户能够轻松地构建、训练和部署模型。此外，还能够根据不同的硬件资源进行自适应调整，以充分利用计算资源并提高训练效率 [27]。

3.2.3.3 DeepSpeed

DeepSpeed 是一个分布式深度学习优化库，由微软研究院开发，旨在提高深度学习模型的训练速度、减少资源消耗，同时保持模型精度。DeepSpeed 支持多种深度学习框架，如 PyTorch，并通过一系列技术实现高效训练。由于大模型动辄需要几十上百 GB 的显存来支持训练和推理，在现有的通用 GPU 上很难实现单卡运行（如英伟达 V100，A100，H100 等型号），所以必须用到多机多卡的架构，而 DeepSpeed 就为解决这些问题应运而生，它具有高效、易用和可扩展等特点，同时为用户提供了详细的文档和示例，方便用户快速上手 [28]。

3.2.4 人工智能平台第三层：推理部署管理

一旦模型训练完成，就需要将其部署到实际的应用场景中进行推理。推理部署管理层主要涉及到模型的部署、管理和运行，包括模型的优化和压缩、推理引擎的选择和配置，以及运行时的监控和管理等工作。

3.2.4.1 Kubeflow

Kubeflow 是由 Google 主导的一个开源项目，旨在简化机器学习工作负载在 Kubernetes 上的部署和管理。它将机器学习领域的各个组件整合到一个统一的平台中，使得用户能够更轻松地构建、训练和部署模型。它充分发挥了 Kubernetes 的弹性和可扩展性，用户能够轻松在多个节点上运行大规模的机器学习工作负载。Kubeflow 还提供了一个统一的开发环境，整合了多个流行的机器学习框架和工具，这使得团队成员能够使用他们喜欢的工具，并在一个共享的平台上协同工作。Kubeflow 通过集成 KServe 等组件，

使人工智能模型能够无缝地从研究和开发阶段转移到生产环境。这种平滑的过渡可以大大加速模型的部署过程 [29]。

3.2.4.2 Caikit

Caikit [30]是一个开源的人工智能工具包，通过一组开发人员友好的 API，使用户能够通过统一的格式管理模型。它为创建和使用针对各种数据领域和任务的人工智能模型提供了一致的数据接口。Caikit 通过让人工智能模型作者专注于使用新技术解决已知问题，简化了应用程序使用的人工智能模型的管理。Caikit 具备以下功能：

- 将不同社区的模型（例如 Transformers、TensorFlow、Sklearn 等）合并到通用 API 中管理。
- 从用户数据创建模型并运行训练作业。
- 以数据结构调用数据 API 来运行模型推断，无需转为 tensors。
- 实现了从静态正则表达式到多 GPU 分发等多种训练技术，以帮助用户正确的拟合模型。
- 将来自不同 AI 社区的模型（例如，transformers、tensorflow、sklearn 等），合并到一个通用 API 中。
- 可根据特定任务，使用新模型更新应用程序，而无需更改客户端。

特别的，Caikit 为应用程序开发人员提供了一个抽象层，他们可以通过 API 使用 AI 模型，而无需了解模型的数据形式。 换句话说，模型的输入和输出采用易于编程且不需要数据转换的格式。 这有助于模型和应用程序彼此独立地发展。

3.2.4.3 Nvidia Triton

Nvidia Triton Inference Server 是由 Nvidia 开发的开源推理服务器，旨在简化和加速深度学习模型的部署和推理过程。它支持多种深度学习框架，包括 TensorFlow、PyTorch、ONNX 等，使用户能够在一个统一的平台上部署和管理各种类型的模型，是一种分布式且合作的缓存架构，可以加速数据密集型应用的 IO 性能 [31]。

3.2.4.4 NVIDIA TensorRT

TensorRT 是一个用于高性能深度学习推理的平台，可用于优化训练好的模型。在使用 TensorRT 优化模型之后，仍然使用传统的 TensorFlow 工作流进行推理，兼容包括 TensorFlow Serving。TensorRT 还可以进行较低精度（FP16 和 INT8）的模型校准，几乎不损失准确性。使用较低精度模型减少了对 GPU 内存的需求，且能达到更快的计算速度，同时还能使用 Tensor Cores 进行计算加速 [32]。

3.2.5 人工智能平台第四层：知识工程

3.2.5.1 嵌入（Embedding）

嵌入（Embedding）是一种将对象（如文本、图像和音频）表示为连续向量空间中的点的方法，其中这些点在空间中的位置在语义上对机器学习（ML）算法具有意义。结果上，嵌入使得机器学习模型能够找到相似的对象。与其他机器学习技术不同，嵌入是通过各种算法（例如神经网络）从数据中学习而来的，而不是明确要求人类专家进行定义。它们允许模型学习数据中的复杂模式和关系，这是人类很难识别的。嵌入的使用使得模型能够捕捉词汇和概念之间的语义关系，从而提高了模型的语义理解和生成能力。

3.2.5.2 向量数据库

向量数据库旨在高效存储、管理和索引大量高维向量数据。这些数据库正在迅速引起关注，为生成式人工智能（AI）用例和应用程序创造额外价值。与传统的关系数据库不同，在向量数据库中，数据点由具有固定维数的向量表示，并根据相似性进行聚类。这种设计实现了低延迟的查询使其能够有效地处理高维向量数据，成为以人工智能驱动的应用程序的理想选择 [33]。

向量数据库支持对相似性进行快速查询，这对许多企业级生成式人工智能应用中是非常重要的。对于搜索相似模式或实例的任务，如图像识别、语义搜索和推荐系统。向量数据库的发展满足了企业对于在其业务决策中利用高维数据的不断增长的需求。通过将模型服务与向量数据库相结合，企业能够更好地处理大规模、高维的数据集，为其 AI 应用提供更准确、更快速、更灵活的支持。关于向量数据库的更多细节详见 3.3.2.8。

3.2.5.3 知识图谱

知识图谱，又称为语义网络，表示了现实世界中的一系列实体，如对象、事件、情境或概念，并展示了它们之间的关系。这些信息通常存储在图数据库中，并以图结构可视化，因此得名为知识“图”。知识图谱由三个主要组成部分构成：节点、边和标签。任何对象、地点或人都可以是一个节点，边定义了节点之间的关系。

3.2.5.4 GenAI Engine

GenAI Engine 是一种引擎，使用户能够轻松训练、验证、调整和部署生成式 AI 基础模型以及机器学习能力，并且可以在短时间内使用少量数据构建 AI 应用程序。该引擎构建

在现代生成式 AI 和机器学习能力之上，支持多种关键用例，包括高级问答（Q&A）、内容摘要、内容分类以及针对特定目的生成内容。GenAI Engine 的灵活性和高度集成的特性使其成为构建多种 AI 应用程序的理想选择，为用户提供了快速、高效地利用生成式 AI 和机器学习能力的平台。

3.2.5.5 检索增强生成（Retrieval Augmented Generation，RAG）

大型语言模型（LLMs）通常对各种主题有着惊人的了解，但它们仅限于它们训练时使用的数据。这意味着希望将 LLMs 用于私有或专有业务信息的客户无法直接使用 LLMs 来回答问题。检索增强生成（RAG）是一种架构模式，它使基础模型能够为未包含在模型训练数据中的专业或专有主题生成事实上正确的输出。通过在用户的问题和提示中加入从外部数据源检索的相关数据，RAG 为模型提供了“新的”（对模型而言是新的）事实和细节，以此为其响应提供基础 [34]。

3.2.5.6 图数据库

图数据库是一种以图结构存储数据的数据库类型，其中数据以节点（实体）和边（关系）的形式表示。图数据库可以使用图算法有效地查询和分析复杂且相互连接的数据。在人工智能领域，图数据库的概念可用于构建知识图谱，将实体和关系表示为图中的节点和边，有助于 AI 系统理解复杂的领域知识，并支持更智能的推理和决策。其次，图数据库可用于自然语言处理（NLP）任务，通过存储语义信息提高文本理解和生成的质量。此外，对于推荐系统、模型解释、机器学习工作流程管理和语义搜索等任务，图数据库都提供了强大的支持。关于图数据库的更多细节详见 3.3.2.9。

3.2.6 人工智能平台第四层：基础模型

在实际应用中，基础模型的人工智能解决方案在企业业务场景中的落地并非仅是单一的模型问题，而是基于业务需求构建的系统性问题。如图 7 所示，企业在落地基础模型通常需要经过几个关键步骤，以确保模型的有效性和可用性。首先是在众多模型中选择并评估最为适宜的模型，利用企业内部可信数据对选定的模型进行训练、调优和增强，以确保其在企业应用场景任务中表现良好，监控模型在实际应用中的表现，并根据反馈信息对模型进行调整和优化，在部署过程中，需要考虑到安全性、可扩展性和可维护性等方面的因素，以确保模型能够稳定可靠地运行。这些阶段相互关联，形成了一个循环迭代的过程，帮助企业不断优化和改进基础模型的性能和效果。本章节围绕基础模型，对模型评估、模型数据准备、模型微调与训练、模型合作、模型优化、模型应用等关键技术进行介绍。

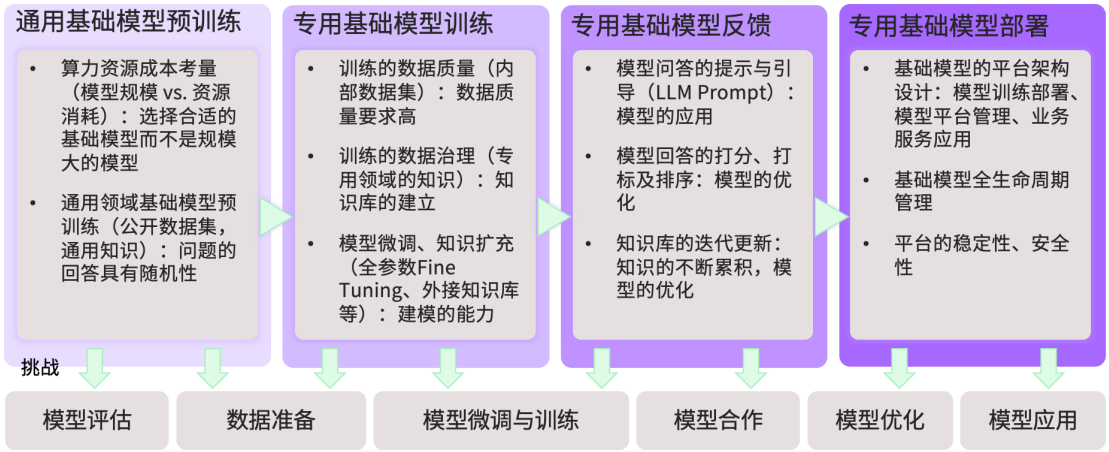


图 7 企业基础模型落地成功的要素与挑战

3.2.6.1 模型评估

模型评估一直以来都是人工智能领域的重要议题。从机器学习，到深度学习，再到现在的生成式 AI，不同阶段的模型评估指标也呈现不同的特点。在机器学习和深度学习阶

段，模型的主要任务是分类(分类模型)和预测(回归模型)，模型结果是否正确是明确的。分类模型的主要评估指标是准确率、召回率、精确率、F1 等等 [35]。回归模型的主要评估指标是均方误差(MSE)、平均绝对误差(MAE)和 R-squared 等等。这些方法按场景和侧重点的不同，以不同的角度和方法统计计算值和真实值的差异，从而评估模型的优劣。但是到了生成式 AI 阶段，基于通用大型基础模型，模型的主要任务变成了文本生成和图像生成等。生成的文本与图片是否“正确”具有强烈的主观性，计算维度也与之前不同。针对以上新出现的问题，在文本生成领域，提出了 BLEU 和 METEOR 等评估指标。图像生成领域则提出了 Perceptual Loss 和 Fréchet Inception Distance 等方法 [36]。

随着大语言模型的广泛应用，评估大语言模型的方法也变得越来越重要。大语言模型的评估有很多不同的侧重点。较为重要的是知识和能力评估以及对齐评估。知识和能力是基础模型一切能力的基础。知识补全是评估模型知识能力的主要手段，它基于现有的知识库，比如 Wikidata、LAMA 等，通过将这些知识库中提供的主题——关系——对象三元组置空，然后用语言模型填入缺失的部分来进行评估。推理能力是另一个重要能力，包括常识推理、逻辑推理、多跳推理和数学推理四个方面。每个方面都有特定的数据集用于基准测试。比如：

常识推理可以使用 CommonsenseQA [37]问答数据集以及关于社交常识的 Social IQA [38]问答数据集。

逻辑推理是通过给定一段文字和一个问题，模型需要从候选答案列表中选择最适当的答案。相关的数据集包括 ReClor [39]、LogiQA [40]和 LSAT [41]等，它们都是由标准化测试(学位考试和公务员考试)提供的多项选择逻辑问题组成的。

多跳推理是指通过多个环节的信息得出最终答案的能力，是更复杂的推理能力。HybridQA [42]是目前最有代表性的多跳测试基准数据，它的每个问题都与异构的多个信息源(表格和文本段落)相关联，模型需要同时利用表格和文本信息才能回答，缺少任何一种信息都无法完全回答问题。

数学推理的数据集则主要来自人类综合性考试的数学部分以及数学竞赛试题。对齐评估则更像是某种软性能力的评估，评估模型是否具有伦理价值对齐能力，以及它们是否生成可能违反伦理标准的内容。

评估模型的对齐能力目前已有商用产品，比如 IBM 的 OpenScale。同时，一些数据集也可以用来测试模型的能力。比如，PROSOCIALDIALOG 是一个大规模的多轮对话数据集，教导对话系统如何应对有问题的对话内容，数据集涵盖了各种不道德、有偏见的情况，它可以提供基于社会规范的建设性反馈，对话的过程往往需要人工参与校准。

在选择应用程序的模型时，还需要考虑以模型性能，模型大小和计算资源需求，语言支持，协议许可，社区活跃度等关键因素。例如，以采纳一种支持中文 RAG 应用的 Embedding 模型为例：

第一步，将模型选择范围缩小到有中文支持的模型。

第二步，可综合衡量 Chinese Massive Text Embedding Benchmark（C-MTEB）和 Hugging Face Massive Text Embedding Benchmark （MTEB） [43]等公共测评榜单，选中一些性能靠前的模型，例如 Baize General Embedding（BGE）系列的 bge-large-zh-v1.5 模型。

第三步，结合应用的实际场景制定和采纳相关评测指标。关于更多模型评估指标可以参考附录二　人工智能指标。

第四步，结合实际数据进行综合测评，择优选取。

3.2.6.2 模型数据准备

当企业场景需要超越原始大语言模型的能力时，通常需要对企业内部的数据进行收集和整理，对模型进行微调和训练以满足特定场景的需求。这个过程可能涉及多个阶段，包括数据的收集、标注和预处理。在这个过程中，企业需要充分了解自身的业务需求和数据特点，以便选择合适的数据收集方法和工具，从而更好地满足业务需求。

3.2.6.2.1 数据收集

数据收集的目的是从各种来源获取与问题或任务相关的数据，以便后续的数据清洗、预处理。以下是一些常见的数据收集来源包括但不限于 [44]：

- **公开数据集**：公开数据集是基础模型训练数据的重要来源之一，通常由学术机构，企业等组织公开发布，涵盖了各种数据类型，例如 UCI 机器学习库、Kaggle 竞赛数据集等。
- **企业内部数据**：通常来自公司内部各个部门和业务领域的运营活动，这些数据对于企业内部决策、业务优化、产品改进等方面具有重要意义。
- **合成数据**：在某些情况下，难以获得足够多样化的真实数据，可以考虑使用合成数据，通过模拟或生成数据来模拟真实数据的分布和特征。
- **数据爬取**：如果没有合适的公开数据集，可以考虑从互联网上爬取数据，但需要注意遵守网站的使用条款和法律规定，以及尊重隐私和版权。

- **实验设计和数据采集**：对于某些特定的问题，可根据需要设计实验并收集数据，通过实地观察、实验调查、传感器收集等方式来完成，需要考虑数据的多样性、覆盖范围和质量等因素。

3.2.6.2.2 数据清洗

数据清洗在机器学习中涉及到识别数据中的缺失值、异常值、重复值等问题，并进行相应的修正和处理，包括：处理缺失，异常值，重复值，不一致的数据格式，特征选择和转换（如数值化、标准化、归一化等），类别型数据（如 One-Hot Encoding 或者 Label Encoding 等），时间序列数据，数据不平衡等方面。在实际应用中可能需要根据数据集的具体情况和需求进行适当的调整和扩展。数据清洗的目标是确保数据的质量和可靠性，为后续的机器学习建模和分析提供可靠的基础。

3.2.6.2.3 数据标注

数据标注，又称为数据注释，是在开发机器学习（ML）模型时的预处理阶段的一部分。这个过程涉及到对原始数据（例如图像、文本文件、视频）的识别，然后为这些数据添加一个或多个标签，以指定其上下文，使得机器学习模型能够做出准确的预测。在数据标注的过程中，人工标记者或专业工具被用来为数据集中的每个样本分配适当的标签。这些标签可以是对图像中物体的识别、文本的分类、视频中事件的描述等。通过为数据集中的每个样本添加标签，为机器学习模型提供有监督学习所需的训练数据。

3.2.6.2.4 数据划分

数据划分通常需要将数据集分为训练集、验证集和测试集三部分进行处理。常见的数据划分的方法主要包括：随机划分、分层划分、时间序列划分、K 折交叉验证等。在实际应用中，可以根据具体情况对上述方法进行调整。需要注意的是，数据划分应该尽可能保证各个子集的数据分布一致，以保证模型在各种情况下都能表现出良好的性能。

3.2.6.2.5 数据增强

数据增强(Data Augmentation)是一种用于改善模型性能和泛化能力的技术，它通过创建原始数据的修改版本来增加用于模型训练的数据量。这些修改可以包括旋转、缩放、翻转或其他形式的变换，目的是增加数据的多样性，以帮助模型学习更多的特征和规律，提高模型的泛化能力。此外，数据增强也可以帮助防止模型过拟合，提高模型的健壮性。

常用的数据增强技术包括：图像数据增强（如旋转，缩放，剪裁，翻转，改变亮度，对比度，饱和度等，文本数据增强（同义词替换，随机插入，随机交换，随机删除等），音频数据增强（改变音调，音量，速度，添加背景噪声等），数据插值等。随着生成式人工智能的技术发展，模型的尺寸越来越大，还可采用自监督数据生成，领域数据/专家数据等方式来实现增强。用户可结合模型的实际应用场景、具体需求和数据类型来选择合适的数据增强方法 [45]。

3.2.6.3 模型微调与训练

一般来说，大语言模型可以通过构造良好的提示激发模型的能力，一种典型的提示方法是将任务描述或示范以自然语言文本的形式表达的上下文学习（in-context learning,

ICL）。此外，采用思维链提示（chain-of-thought prompting）可以通过将一系列中间推理步骤加入提示中来增强 ICL。有的场景通过提示无法解决问题或者需要过长的上下文提示，这种情况下就会涉及模型微调，常见的微调方式包括使用无标签数据进行继续预训练、使用标签数据对模型进行指令微调、通过强化学习对模型进行对齐微调等。

继续预训练优势是可以容易获取到无标签数据，常见的使用场景包括对 LLM 进行进行多语言支持的扩展、垂直领域知识的增强，增加 LLM 文本长度的支持等。

指令微调是以有监督的方式微调 LLM（例如使用序列到序列的损失进行训练），指令微调后 LLM 可以展现出泛化到未见过任务的卓越能力，为了进行指令微调，首先需要收集或构建指令格式的实例。构建指令数据集可以通过人工方式、利用基础模型自动生成、结合使用开源指令数据集。由于指令微调涉及多种任务的混合，因此在微调过程中平衡不同任务的比例非常重要，一种广泛使用的方法是实例比例混合策略，即将所有数据集合并，然后从混合数据集中按比例采样每种实例。

LLM 有时可能表现出预期之外的行为，例如编造虚假信息、追求不准确的目标，以及产生有害的、误导性的和有偏见的表达，因为模型预训练使用了语言建模的目标，即用单词预测进行预训练，但这没有考虑到人类的价值观或偏好。为了避免这些预期外的行为，一些研究提出了人类对齐，使得 LLM 的行为能够符合人类期望，对齐微调使得 LLM 的行为能够符合人类期望。基于人类反馈的强化学习（RLHF）使用收集到的人类反馈数据对 LLM 进行微调，有助于改进对齐的指标（例如，有用性，诚实性和无害性）。RLHF 采用强化学习（RL）算法（例如，近端策略优化（Proximal Policy Optimization，PPO）通过学习奖励模型使 LLM 适配人类反馈。这种方法将人类纳入训练的循环中来开发对齐得

良好的大语言模型，如 InstructGPT。对齐微调的数据集通常由人工进行精细的设计，成本较高，一些开源的数据集有 HH-RLHF [46]、SHP 等。

由于 LLM 包含大量的模型参数，进行全参数微调将会有较大开销，于是提出来参数高效微调（parameter-efficient fine-tuning），旨在减少可训练参数的数量，同时尽可能保持良好的性能。常见的用于 Transformer 语言模型的参数高效微调方法有适配器微调（adapter tuning）、前缀微调（prefix tuning）、提示微调（prompt tuning）和低秩适配（LoRA）等 [47]。

3.2.6.4 模型合作

在实际企业模型应用中，面对复杂的业务场景需求，通常可以将基础基础模型和领域模型进行合作。通用基础模型具有许多优势，其中包括强大的自然语言理解能力、内置大量世界知识、以及具备任务拆解和总结能力等特点。这些基础模型可以解决多个下游任务，为企业提供了广泛的应用可能性。相比之下，专业领域模型则更为精细，虽然部署所需资源较少，但其优势在于经过专业领域的长期训练和优化，表现出千锤百炼的能力。然而，专业领域模型的适配性较窄，一种领域模型通常只能对接一种具体任务，相较于通用基础模型，其应用范围相对有限。在实际情况中，企业往往已经开发了一些专门针对其特定领域或业务需求的领域模型，因此，将两者结合起来，可以形成更灵活、高效的解决方案，并充分利用已有的能力，最大程度地发挥模型的优势。

一些常见的模型合作的方式包括：

- **模型组合**：将领域模型的预测结果整合到基础模型中，从而扩展基础模型的知识和提高精度。例如，基础模型可用于任务框架拟定和任务分解，又领域模型对分解任务进行处理，最终由基础模型对所有步骤的答案进行组织整理。
- **模型堆叠**：将领域模型和基础模型串联起来，形成一个更为复杂的模型。通过增加模型深度，可以提高模型的复杂度。例如，基础模型可对任务从不同维度进行定义，领域模型则从不同角度回答问题，最终由基础模型整理所有步骤的答案。
- **模型分工**：将用户任务分解，让大、领域模型各自专注于不同的任务。例如，基础模型处理开放式自然语言处理任务，而领域模型则专注于特定行业的语言任务。
- **模型调整**：调整基础模型的参数，使其更好地适应特定行业的语言数据。例如，通过小模型的对基础模型的结果进行校正，调整基础模型以适应特定领域的数据集。

3.2.6.5 模型优化

基础模型优化是指在设计、训练和部署大型 AI 模型时所采取的一系列技术和策略，旨在提高模型的效率、性能和可扩展性。这一过程涵盖了模型训练优化，模型压缩，推理优化等多个方面。

3.2.6.5.1 训练优化

由于目前的大型模型往往包含数十亿、数百亿、甚至数万亿个参数，这意味着在模型训练阶段需要频繁进行大量的浮点运算，对计算能力的需求是巨大的。例如，要训练一个规模为 Llama-70B 的模型，需要在庞大的计算集群上进行数月之久。这涉及到的时间成本和电力成本不容忽视，因此需要考虑如何加快速度，优化整个训练流程。

一般而言，大型语言模型的结构主要基于 Transformer，其中每一层的结构相对固定，因此，常见的训练加速方法主要集中在如何实现 Transformer 的并行化。在本质上，训练过程涉及大量的矩阵乘加运算，因此必须思考如何降低计算参数的矩阵操作。通常情况下，可考虑采用多个计算设备进行并行计算，同时优化设备间的通信负载以及单个设备上的计算时间，以期望实现并行计算时的线性加速比。此外，大型模型的训练时间也与训练数据量密切相关，所以也可通过数据分割来实现并行计算，运用分治法的思维处理数据内部特征。另外，由于目前主流计算设备为 GPU，单个 GPU 的显存相对有限，难以支撑整个模型训练过程中所需的模型和数据存储，因此大型模型的训练也必须依赖多 GPU 卡并行进行。

模型训练加速的优化方案目前一般包含数据并行，模型并行，流水线并行等，并伴随有使用低精度浮点数来降低单次计算所需要的计算力或计算时间，且还能降低 GPU 显存的使用量。

数据并行化比较直观，在多个计算硬件上分别加载同一个模型结构，然后将数据分割成不同的子数据集分发到不同的计算硬件上分别计算，寻找数据内部特征，最后将每个计算硬件上的模型寻找到的数据特征进行整合，糅入一个模型，从而达到数据的并行化计算。

模型并行化在分割模型时的力度取决于整体训练模型的所需算力的大小，需保证每个计算单元都能分配到足够的计算量。例如，将模型结构进行二维（2-dimensional）分割，同时考虑计算硬件的拓扑结构，尽可能让模型间的通信量少，且信息在网络中传递的路径最短，这样就能在加速计算的同时减少网络负载，让整个训练过程整体最省时，硬件资源利用率最大化。在模型并行化的时候，还会有更细粒度的张量并行（tensor

parallelism）和流水线并行（pipeline parallelism），这些并行化需要深入理解模型的结构，才能将模型按照不同的切分方式来分割并行。

在将模型训练分割并行化的加速优化能力考虑到极致后，还可以通过模型量化，即前文提到的使用低精度的浮点数来进行计算加速，此外还有算子融合等方式来进行计算的优化。有很多开源项目正在研究这些领域，如 vLLM，bitandbytes 等。值得注意的是，由于这些方法一般牵涉到对模型的修改或是对数值的修改，故可能会存在模型收敛出现问题，或者模型精度出现偏差，往往都需要针对数据集做进一步的模型结果调优。

3.2.6.5.2 模型压缩

模型压缩技术是指的是一系列旨在减小深度学习模型的体积和计算复杂度，同时保持其性能的方法。这些技术对于在资源受限的环境下部署模型、提高推理速度或降低能耗都非常有用，常见的大语言模型压缩技术包括 [48]：

模型剪枝通过去除网络中不必要的连接或参数来减小模型的大小。剪枝技术可分为非结构化剪枝和结构化剪枝两种形式。非结构化剪枝是指在不考虑模型结构的情况下，去除模型中相关度较低的参数，从而达到减小模型尺寸的目的。而结构化剪枝则是通过剪除模型的整个部分，例如神经元、通道或层，来进减小模型的大小。非结构化剪枝不改变模型的结构，剪枝力度细致，潜力大，但需要搭配相关的硬件。相比之下，结构化剪枝的粒度较粗，剪枝后会改变模型结构，对模型性能影响较大。因此，结构化剪枝可剪枝比例通常较非结构化剪枝低，但实现技术简单，不需要相关的硬件配合。不论是非结构化还是结构化剪枝，在剪枝后通常需要进行后续的微调，以弥补剪枝带来的模型性能下降。

模型量化通过将浮点参数转换为单字节或更小的整数，从而显著减小大语言模型的大小，它通常包括离线阶段（offline stage）和在线阶段（online stage）两个主要阶段。离线阶段的量化过程通常在训练后进行，此时模型已经通过了训练并获得了较佳性能。在此阶段，将训练好的模型来分析其权重分布和激活响应等信息，以确定适当的量化策略。这涉及到选择合适的量化比特数（如 8 比特、4 比特等），以及确定量化的范围和方法（如线性量化、非线性量化、对称量化、非对称量化等 [49]）。在线阶段是指将已经量化的模型部署到实际环境中，以进行推理或应用。根据硬件支持的精度不同，可能需要对参数进行反量化操作来进行推理计算，以适应硬件的特定要求和限制。此外，在线阶段还涉及到模型的部署、配置和优化，以确保在实际应用中能够达到预期的性能和效果。对大语言模型的量化技术又可以分为后训练量化（PTQ，Post-training quantization）和量化感知训练（QAT，Quantization-aware training） [48]。在后训练量化中，模型在完成训练后通过将参数转换为低精度数据类型来实现压缩，如 GPTQ。相比之下，量化感知训练将量化过程集成到模型的训练过程中，如 QLoRA。后训练量化在模型训练完成后应用，简单直接可快速实现，但可能无法充分考虑到量化对模型性能的影响，导致性能损失较大；量化感知训练则集成了量化到训练中，可以更好地优化模型参数以适应低精度的量化，但可能增加训练开销，适合对性能要求高的场景。

知识蒸馏：通过训练一个小型模型来近似一个大型模型的输出。在这个过程中，大型模型（教师模型）的“知识”被传递给小型模型（学生模型）。学生模型通常比教师模型要简单，因此更适合在资源受限的环境中部署，这种蒸馏方式又可称为传统知识蒸馏（或白盒知识蒸馏） [48]。对于大语言模型来说，涌现能力蒸馏（或黑盒知识蒸馏）着重于从教师模型（即 LLM）中提取某种特定的涌现能力，并将其转移给学生模型。大语言模型的

涌现能力（Emergent abilities）[50]指的是这些大模型具备的某些能力，这些能力在较小的模型中并不存在或表现较弱。这些能力可能是由于大规模数据训练和模型结构的复杂性而产生的。根据学习的能力的不同，涌现能力蒸馏又可以细分为不同的类型[48]：In-Context Learning（ICL）蒸馏采用结构化的自然语言提示，包含任务描述和可能的任务示例，旨在蒸馏大语言模型的上下文学习能力。Chain of-Thought（CoT）蒸馏将中间推理步骤融入提示中作为学生模型的训练数据，以培养学生模型的推理能力。Instruction Following（IF）通过阅读任务描述来增强语言模型在执行新任务时的能力，而不依赖于少量示例，旨在蒸馏大语言模型的指令学习能力。

3.2.6.5.3 批量推理

目前，大语言模型推理过程主要受到内存 IO 的制约，而不是计算资源的限制。换言之，将 1MB 的数据加载到 GPU 所需的时间超过了这些 GPU 计算单元在相同大小数据上执行 LLM 计算所需的时间。这意味着 LLM 推理吞吐量的主要瓶颈在于能够将大批量的数据装入高带宽 GPU 内存中。因此，为了提高 LLM 推理的效率，除了优化计算速度外，还需要关注如何更有效地管理和利用 GPU 内存。根据 NVIDIA 的报告显示[51]，随着并发数的增加，推理吞吐量通常会有显著的增加。这意味着优化批处理和并行性能是提高 LLM 推断效率的关键策略之一。

批量推理的技术又可分为静态批处理（Static Batching）和连续批处理（Continuous Batching）两种。批处理的传统方法称为静态批处理，即批次的大小在推理完成之前保持不变。与传统的深度学习模型不同，由于 LLM 推理的迭代性质，批处理可能会变得棘手。由于批处理中不同序列的生成长度与批次的最大生成长度不同，GPU 的利用率较低，如果

输入序列也具有相同的大小，那么每个静态批处理才可实现最佳可能的 GPU 利用率。相比之下，连续批处理 [52]不再等待批处理中的每个序列都完成生成，而是实现了迭代级别的调度，其中批处理大小是根据每次迭代确定的。结果是，一旦批处理中的一个序列完成生成，就可以插入一个新的序列，从而实现比静态批处理更高的 GPU 利用率。

3.2.6.5.4 推理引擎

推理引擎是指用于执行机器学习模型推理（即模型的预测或输出）的软件组件或系统框架。在深度学习领域，推理引擎通常是指能够有效地将训练好的神经网络模型应用到实际数据上，以产生所需结果的软件组件。这些引擎通常优化了模型的计算和内存使用，以提高推理速度和效率，并且通常针对特定的硬件架构进行了优化，如 CPU、GPU、TPU 等。常见的推理引擎有 Text generation inference（TGI），vLLM，DeepSpeed-MII，OpenLLM，MLC LLM，Ray Serve，CTranslate2 等 [53]。

3.2.6.6 模型应用技术

基础模型的上层技术赋予了模型更加智能、灵活的特性，使其能够更好地适应不同的任务和环境。它们帮助企业的基础模型开发人员更快地开发和部署模型应用。通过这些技术手段，企业能够更灵活地应对不断变化的业务需求和环境变化，提高模型的适用性和性能。

3.2.6.6.1 思维链提示

2022 年，大语言模型的效果越来越好，并涌现出了强大的逻辑推理能力。同时随着模型规模的不断变大，模型也变得更容易被“提示”。但是基础模型在做数学推理和知识推

理时的表现还不尽如人意。在这样的背景下，出现了思维链（Chain-of-thought，CoT）的概念。思维链（CoT）的概念被首次提出。这是一种改进的提示策略，用于提高大语言模型在复杂推理任务中的表现。简单来说，CoT 给基础模型提供了一些相关的上下文学习，让基础模型更容易给出最终正确的答案，通过把问题分解为多个中间步骤，为模型的行为提供一个可以解释的窗口，给出如何得出答案的具体分析方法，并提供可以被用来调试的路径，实现可验证性。对于足够大的模型，甚至可以把思维链推理的步骤作为示例包含在 few-shot 提示中 [54]。

3.2.6.6.2 由少至多提示

思维链提示在各种自然语言推理任务中表现出了显著的效果，但是，对于那些比提示中示例更难的问题，表现往往不太好，比如组合泛化。为了克服这种问题，由少到多提示（least-to-most prompting）提示策略被提出 [55]，其关键思想是把一个复杂问题分解成一系列更简单的子问题，然后依次解决，以前解决的子问题的答案有助于解决每个子问题。它包含两个阶段，第一个阶段把一个复杂问题分解成一系列更简单的子问题，这个阶段的提示包含演示分解的固定示例，然后是要分解的特定问题。第二个阶段依次解决子问题，这个阶段的提示由三部分组成，第一个是演示如何解决子问题的恒定示例，第二个是之前回答的子问题和生成的解决方案的潜在空列表，第三个是接下来要回答的问题。原问题作为最后一个子问题追加。

3.2.6.6.3 LangChain

LangChain 是一个开源编排框架，用于使用大型语言模型（LLM）开发应用程序 [56]。LangChain 的工具和 API 在基于 Python 和 Javascript 的库中使用，可以简化构建聊天

机器人和虚拟代理等 LLM 驱动型应用程序的过程。LangChain 几乎可以作为所有 LLM 的通用接口，为构建 LLM 应用程序并将其与外部数据源和软件工作流程集成提供集中式开发环境。LangChain 基于模块的方法允许开发人员和数据科学家动态比较不同的提示，甚至比较不同的基础模型，而无需重写代码。这种模块化环境还允许程序使用多个 LLM：例如，应用程序使用一个 LLM 解释用户查询，并使用另一个 LLM 编写响应。著名的 LangChain 工具示例如：Wolfram Alpha 提供强大的计算和数据可视化功能，实现复杂的数学功能；Google 搜索提供 Google 搜索访问权限，为应用程序和代理提供实时信息；OpenWeatherMap 获取天气信息；维基百科支持对维基百科文章信息进行高效访问等。

3.2.6.6.4 MiniChain

MiniChain 旨在在一个小型库中实现核心提示链接功能，它利用函数装饰器和 YAML 模板来实现链式操作，用户只需要 20 行左右代码，就可以编写一个简单的聊天机器人，向量数据库等等。MiniChain 不管理文档和嵌入，可使用内置 FAISS 索引的拥抱面部数据集库。 MiniChain 可以自动生成一个提示头，旨在确保输出遵循给定的类型化规范 [57]。

3.2.6.6.5 AI Agents

AI Agents 是一种软件程序，旨在与其环境交互，感知接收到的数据，并根据该数据采取行动以实现特定目标。AI Agents 能够模拟智能行为，可以像基于规则的系统一样简单，也可以像高级机器学习模型一样复杂。AI Agents 使用预先确定的规则或经过训练的模型来做出决策，并且可能需要外部控制或监督。相对于传统的 AI Agents，自主 AI Agents（Autonomous AI Agents）是一种先进的软件程序，可以在没有人类控制的情况

下独立运行。它们可以自主思考、行动和学习，无需人类不断输入。这些代理广泛应用于医疗保健、金融和银行等不同行业，使事情运行得更顺畅、更高效。它们可以适应新情况，从经验中学习，并利用自己的内部系统做出决策。

AI Agents 的内部结构可以根据具体的应用和任务而有所不同，它的内部结构由四个关键部分组成，分别是 Environment（环境）、Sensors（传感器）、Actuators（执行器）以及 Decision-making mechanism（决策机制）。AI Agents 会通过传感器或其他数据源感知环境。传感器可以包括视觉传感器（如相机）、听觉传感器（如麦克风）、物理传感器（如触摸传感器）等。这些传感器帮助代理获取环境中的信息，例如图像、声音、位置等。AI Agents 使用适当的知识表示方法来组织和存储从环境中获取的信息。这些信息可能包括先验知识、学习到的模式或规则。基于感知到的环境信息和存储的知识，AI Agents 使用决策制定机制来生成适当的行动。这可能涉及使用逻辑推理、统计分析、规划算法或机器学习技术来评估不同行动的可能结果和潜在风险。决策制定过程旨在使代理能够选择最佳行动以实现其目标。然后，Agents 制定计划或一系列步骤来实现其目标。一旦决策制定完成，AI Agents 将执行行动并与环境进行交互。这可能涉及控制执行器（如机器人的电机）、发送指令（如语音助手的语音合成）或与其他代理进行通信。执行行动后，Agents 会观察执行结果，并将其用作反馈以调整下一步的决策。最后，在完成上述的执行行动后，AI Agents 通过与环境的交互获得反馈。这些反馈可以来自环境中的直接观测结果，也可以来自人类用户或其他代理的指令和评估。Agents 使用这些反馈来学习和改进自己的行为。这可能包括使用监督学习、强化学习或迁移学习等技术来调整决策制定和行动执行过程，以提高代理的性能和适应能力。在现实的业务场景中，AI

Agents 在自然语言处理、机器人技术、个性化推荐、还在医疗诊断、金融风险管理、智能城市管理等领域都展示出了广泛的应用，对日常生活产生了重大影响。

3.2.6.6.6 多模态

在人工智能领域，随着深度学习和神经网络技术的发展，多模态大语言模型成为了研究的热点之一。传统的自然语言处理模型主要关注文本数据的处理，而多模态大语言模型则将文本、图像、声音等多种形式的数据进行整合，实现了多模态信息的联合学习与应用。这一模型的出现，为机器在不同感知模态下进行跨模态的语义理解提供了新的思路和解决方案。多模态学习具体可以划分为几个研究方向[58]：多模态表示学习（Multimodal Representation），模态转化（Translation），对齐（Alignment），多模态融（Multimodal Fusion）和协同学习（Co-learning），常见技术如多模态指令调优（Multimodal Instruction Tuning ，M-IT）、多模态上下文学习（Multimodal In-Context Learning，M-ICL），多模态思维链（Multimodal Chain of Thought ，MCoT）以及构建任务解决系统的通用框架（LAVR）[59]等。多模态算法可分为基础模型和大规模多模态预训练模型两类。基础模态是多模态的基本框架，在此基础上改进了许多新的大规模多模态预训练模型。

3.2.7 IBM 人工智能平台 watsonx.ai

IBM watsonx.ai 是 IBM watsonx 人工智能与数据平台的一部分，它将基础模型支持的生成式 AI 功能和传统机器学习整合至一个贯穿 AI 生命周期的开发平台，利用企业数据调整和指导模型，并通过易于使用的工具来构建和完善高性能提示，从而满足企业客户的需求。利用 watsonx.ai，使用一小部分数据，用户能够在短时间内构建 AI 应用程序。IBM

watsonx.ai 提供了多种能力，包括模型多样性和灵活性，用户可选择开发的模型、开源模型和第三方模型，或构建自己的模型；IBM 对 IBM 开发的模型提供支持，并针对第三方知识产权索赔向客户提供赔偿；IBM watsonx.ai 提供端到端的 AI 治理，企业可以通过整个公司的可信数据来扩展和加速 AI 的影响，无论数据位于何处；同时 IBM watsonx.ai 支持混合式多云部署，提供将企业 AI 工作负载集成并部署到所选混合云堆栈中的灵活性 [60]。

3.2.7.1 基础模型支持

IBM watsonx.ai 用户可以访问 IBM 选择的 Hugging Face 开源模型和其他第三方模型，包括 Llama 3 and Mixtral 8x7b，以及经过 IBM 开发的不同规模和架构的基础模型，包括开源的 Granite 模型和 IBM 定制的 Granite 模型等，以支持不同的企业领域和用例（如 RAG）。watsonx.ai 当前支持的模型可参考附录一 ，可用的基础模型支持自然语言和编程语言的各种用例，并支持多种语言，用户可以在 Prompt Lab 中查看这些模型的可以执行的任务类型和 Prompt 样例 [61]。

IBM 的基础模型的 Granite 系列包含一系列 decoder-only 模型，可以高效地预测和生成语言。这些模型是使用来自优质数据集的可信数据构建，涵盖领域包括金融（SEC 提交）、法律（Free Law）、技术（Stack Exchange）、科学（arXiv、DeepMind Mathematics）、文学（Project Gutenberg (PG-19)）等，符合严格的 IBM 数据清理和治理标准，经过清理，包括去除仇恨、滥用和亵渎、数据重复以及黑名单网址等。

3.2.7.2 Prompt Lab

通过 IBM watsonx.ai，AI 构建者可以使用其中的基础模型，并使用提示工程构建提示。用户可以使用聊天、自由形式或结构化模式在提示编辑器中与基础模型进行交互。多

种交互方式使用户可以制定最佳的模型配置，支持不同的自然语言处理（NLP）任务，如问答、内容生成和摘要、文本分类和提取等。

Prompt Lab 是一个基于图形用户界面的无代码工具，可快速测试不同的模型和提示。使用 Prompt Lab，用户可以快速比较使用了不同代码格式和指令的提示之间的输出差异。以 llama-2-chat 为例，用户可使用 Prompt Lab 对模型进行 Prompt 调优，将 Prompt 保存成为模板或回话，支持查看、导出 Curl、Python 的调用代码等操作 [62]。

3.2.7.3 Tuning Studio

IBM watsonx.ai Tuning Studio 通过提示微调（Prompt-tune）基础模型，有助于利用标签数据对基础模型进行调优，以获得更好的性能和准确性。提 v 示微调是一种高效、低成本的方法，可以在不重新训练模型和更新其权重的情况下，让基础模型适应新的下游任务。调优完成后的模型，可以在 Prompt Lab 中被使用。IBM watsonx.ai Tuning Studio 的后续版本还将提供模型微调等功能。

使用 Tuning Studio，用户可以通过调优较小的基础模型，提高其在自然语言处理任务（如分类、摘要和生成）上的性能，使其在同一模型系列中实现与较大模型相似的结果。调优可以基础模型的多种能力，如生成特定风格的新文本，以特定方式生成总结或提取信息，文本分类等。调优的基本流程包括设计与使用模型良好配合的提示（可借助 Prompt Lab 进行提示工程实验）、按照格式创建用于模型调整的训练数据、创建调整实验以调整模型、评估调整后的模型以及部署调整后的模型等，Tuning Studio 为这个过程提供了基于图形用户界面的无代码工具 [63]。

3.2.7.4 数据科学与 MLOps

由 IBM watsonx.ai 基础模型提供支持的工具、流程和运行时环境，可以帮助数据科学家自动构建 ML 模型，通过连接到各种 API、SDK 和资料库，自动化从开发到部署的整个 AI 模型生命周期流程。MLOps 支持用户以可视化或或使用代码的方式构建模型，以公平和可解释的方式部署、监控完整的生命周期，利用 MLOps 简化任何工具的模型生成，并提供自动模型重新训练，其具体功能如表 1 所示 [60]：

表 1 watsonx.ai MLOps 功能

功能	目标	描述
管道编排	创建自动化管道	供数据科学家构建、训练和部署 ML 模型的单一协作平台，支持广泛的数据源，使团队能够简化其工作流程。借助自动化 ML 和模型监控等高级功能，用户可以在整个开发和部署生命周期中管理其模型。
CPLEX 优化引擎	解决优化问题	使用 CPLEX 优化器揭示提示性分析以改善用户的业务决策，例如规划、调度、定价、库存或资源管理。CPLEX 决策优化引擎应用专业的数学算法和基于约束的编程来解决用户业务目标。在 CPLEX 求解器中，可共享表格或视图，以增强合作并加快洞察力。
可视化建模	直观地开发预测模型	借助易于使用的工作流程，在统一的数据和 AI 平台上将可视化数据科学与开源资料库和基于笔记本的界面相结合。

自动化开发	加速完成整个 AI 生命周期	初学者可以利用 AutoAI 快速入门，专家级数据科学家则可以加快 AI 开发的实验。AutoAI 会自动执行数据准备、模型开发、特征工程和超参数优化。
合成数据生成器	生成合成表格数据	利用现有数据或定制数据模式，生成合成表格数据集。可以连接到现有数据库、上传数据文件、对列数据进行匿名处理、根据需要生成尽可能多的数据，以解决数据缺口或训练经典 AI 模型。

3.3 数据平台和服务

3.3.1 生成式人工智能数据管理的挑战

生成式人工智能训练过程中需要大量的数据，这些数据既有原来传统数仓（如企业内部现存的关系型数据库）中积累的数据，也有来自文本，图片，音频，视频等多样性数据的训练要求。企业要把内部积累多年的数据资产变成人工智能，需要一个数据平台打通各个数据，打破数据孤岛，以统一的方式提供给模型训练使用。因此，新一代平台要在满足接入传统数仓的同时支持新的数据格式，进而构建满足模型平台和服务层数据访问要求的知识库。企业需要对数据进行不同程度的预处理以满足模型训练的要求，这一过程需要多种数据处理工具的支持。在使用数据的过程中：

- 贯彻数据治理以满足保护隐私，安全规范的相关法律法规要求。
- 甄别高质量的数据，提高训练的效率。
- 实现数据甚至知识的生命周期管理，满足数据，知识不断更新，不断迭代的需求。

新一代数据湖仓技术正是为了应对目前不断发展的分析和人工智能需求而生的，解决海量多样数据的管理难题的同时保证数据质量（准确，公平等）和数据安全。

3.3.1.1 数据管理技术的发展

随着企业数字化的发展进程，数据管理系统不断面临新的挑战，回顾数据技术管理发展的历史，有助于我们更好的从发展的眼光看待企业级人工智能对数据管理系统的新需求。详见图 8 数据管理发展历史。

图 8 数据管理发展历史

在 90 年代中后期，传统的数据仓库技术开始出现，主要以关系型数据库组织结构化数据。数据通过转换、整合、清理后导入到数据仓库，其中数据存储的结构与定义的模式（schema）强匹配。这种技术主要用于决策支持和商业智能，通常绑定在特定供应商，可扩展性有限，对非结构化和实时数据处理能力有限。

进入 21 世纪初，随着数据量和种类的增长，数据湖技术应运而生，以满足企业对多样化原始数据、全量存储和全生命周期管理的需求。数据湖从企业多个数据源获取原始数据，可以是任意类型，从结构化到非结构化。这降低了大量数据清理的成本，具有灵活可扩展的特点。然而，数据湖项目也面临一些挑战，包括维护的复杂性、数据质量不佳、对数据科学家的高要求以及性能有限。存在数据治理缺失、数据孤立和碎片化的问题，有时甚至形成数据沼泽。此外，数据湖的巨大挑战之一是单一结构的架构问题。例如，Hadoop 以低成本存储大量数据、支持开放的数据格式和自动复制高可用性等方面表现优异，但是 Spark 作为大数据处理框架由于其支持数据转换、流式处理和 SQL 等功能而得到广泛认可，但不能与现有数据湖环境友好共存，必须外挂专有的计算集群。

随着云计算技术的进步，云数据仓库得以发展。具体而言，引入了计算和存储的分离，有效解决了传统数据仓库在可扩展性方面的挑战。通过增加计算资源，可以确保在处理大数据量时仍能保持高性能。其中，Snowflake 是一个具有代表性的例子。它的优势在于易于管理，但相对于本地数据仓库而言成本较高，仍然存在供应商锁定的问题，同时也需要进行数据迁移，仅能够满足一些有限的人工智能（AI）/机器学习（ML）用例。

虽然数据湖已经在特定的应用场景中已经被证明是成功的。然而，随着生成式人工智能应用的企业级落地，企业迫切需要对这些部署进行现代化升级，以保护在这些系统中的基础设施、技能和数据的投资，从而满足业务增长带来的数据需求：

- 数据格式的多样性需要支持更多的开放的数据结构。
- 数据的快速增长需要可扩展的存储，大量数据的处理需要可按需扩展的计算资源。
- 数据的运维管理，安全需要传统数仓的事务能力。

很明显，一种有效的方法是将传统数据仓库或数据集市的关键特性与数据湖的优势结合起来。以下几个关键要素迅速浮出水面：

- 具备弹性和可扩展的存储，以满足不断增长的数据规模需求。
- 采用开放的数据格式，使数据对所有人都可访问，同时对高性能进行优化，并具备良好定义的结构。
- 开放的元数据（可共享），能够支持多个消费引擎或框架。
- 支持数据更新（ACID 特性）和事务并发处理。
- 综合的数据安全和数据治理，包括数据血缘、完整的数据访问策略定义和执行，以及地理分布等。

这些要素共同导致了湖仓一体的出现。湖仓一体是一种数据平台，将数据仓库和数据湖的优点融合在一起，形成统一、协调的数据管理解决方案。

以 Databricks，Dremio，Starburst 等为代表的第一代数据湖仓提供者通常只提供了单一引擎，只擅长处理商业智能（BI）或者人工智能（AI）单个工作负载，他们依托公有云部署，来支持计算和存储资源的弹性扩展，数据治理能力相对薄弱。对于很多企业而言，数据资产是他们的核心资产之一，他们需要更多的部署选择以保证数据被安全合理的访问以实现数据价值。

3.3.1.2 数据湖仓 vs 数据仓库

传统数据仓库没有实现计算和存储分离，新一代云数据仓库实现了计算和存储分离，数据湖仓原生支持计算和存储分离。传统数据仓库主要是为了结构化和半结构化数据设计的，需要打开额外功能或者使用特殊方法来支持开放数据文件和开放的表格式。数据湖仓从设计之初就支持结构化和非结构化，内部很多文件也同样以开放数据文件格式存储。传统数据仓库绑定了专有提供商的查询引擎，数据湖仓可以根据需要切换不同的查询引擎。

图 9 数据湖仓对比图

3.3.2 数据湖仓技术介绍

数据湖仓技术的开源生态非常活跃，在本章中我们将分章节，从数据格式、元数据管理、查询引擎、知识库和联邦查询等方面，介绍开源技术实现。图 10 湖仓开源技术是一个开源技术的概览。

图 10 湖仓开源技术

3.3.2.1 数据存储

对象存储服务、块存储服务和文件存储服务是云计算和分布式存储中常见的三种存储模型 [64]。

3.3.2.1.1 对象存储服务

对象存储服务是一种在云计算环境中存储和检索大规模非结构化数据的模型。

在对象存储中，数据被组织为对象，以对象为基本存储单元，每个对象包含数据、元数据和唯一的标识符，并通过唯一的标识符进行检索。对象存储通常提供松散的一致性，并支持分布式架构，使其成为云存储和大数据分析的理想选择。

对象存储通常适用于需要存储、检索和管理大规模非结构化数据的场景，例如图片、视频、文档等。

典型的对象存储服务提供商包括：IBM Cloud Object Storage (COS)，Amazon S3 (Simple Storage Service)，Microsoft Azure Blob Storage(ADLS)，Google Cloud Storage (GCS)。开源对象存储服务包括 Ceph，MinIO。

3.3.2.1.2 块存储服务

块存储服务将数据划分为固定大小的块，并将这些块存储在独立的设备上，每个块都有唯一的地址，允许直接读写单个块。

由于块存储提供了低延迟、高性能和随机访问的优势，因此它特别适用于对存储性能有较高要求的应用场景，比如数据库存储，虚拟机镜像存储等，这些特点也使其成为许多企业应用的首选存储模型。

典型的块存储服务提供商包括：IBM Cloud Block Storage，Amazon Elastic Block Store (EBS)，Microsoft Azure Managed Disks，Google Cloud Persistent Disks。开源实现如 Ceph，GlusterFS，MinIO。

3.3.2.1.3 文件存储服务

文件存储服务为用户提供了类似传统文件系统的层次结构，以文件和目录的形式组织数据，并通过网络协议（如 NFS、SMB）提供对这些文件的访问。

由于文件存储允许多个用户或设备同时访问相同的文件，支持文件的共享和协作， 因此它非常适用于需要共享数据和支持多用户协同访问的场景，如企业共享文件、应用程序配置文件等。

典型的文件存储服务提供商包括：IBM Cloud File Storage，Microsoft Azure Files，Amazon Elastic File System (EFS)，Google Cloud Filestore。开源实现如 GFS, HDFS，Ceph。

3.3.2.2 大数据中常见的文件存储格式

二进制形式存储的格式因为其更小的文件体积，更快速的序列化，支持跨语言等种种特性，成为了大数据首选的存储格式。评判一个文件格式是否适合二进制形式存储，可以从以下几点去分析：

- 倾向更快的写入还是更快的读取速度。
- 是否支持文件分割，并行处理数据。
- 压缩算法的支持，压缩性能的比较。
- 模式演变（ Schema evolution）的支持。
- 查询引擎的适配（例如：Spark 倾向于 Parquet, Hive 倾向于 ORC）。
- 数据本身是扁平化的，还是嵌套的。
- 数据读取是整体读取，还是少数字段的读取。
- 数据是否有频繁改动，对 ACID 的需求。

3.3.2.2.1 行存储与列存储

文件格式按存储方式可以分为行式存储和列式存储。

行式存储是以行为单位进行存储， 一条数据所有字段都存储在同一个块上。其写性能较高，保证事务特性更容易，压缩效果较差。

列式存储是以列为单位进行存储，将同一列的内容连续存放在一起。其写性能方面效率低，当读取少数几列时，性能较高，此外列存储的压缩效率高，比较难实现事务特性。

3.3.2.3 开放数据文件格式

本节将介绍三种主流的文件格式，Parquet，Avro，ORC，并分析各自的优缺点和适应场景。

3.3.2.3.1 Parquet

Apache Parquet 基于列存储的文件格式，并支持嵌套格式数据。Parquet 在大数据领域的应用场景包括 Apache Spark，Apache Hive 和 Apache Impala 等分布式计算框架。此外，作为 Apache Arrow 的底层存储格式，Parquet 还提高了数据交互的效率。

Parquet 文件格式是自解析的，其 schema 信息以及其他元数据信息一起存储在文件的末尾。Parquet 文件是可分割的，因为它在 Footer 中存储了文件块边界信息。系统通过读取这些信息，可以确定是跳过还是仅读取文件的特定部分，从而实现更高效的读取，或并行处理。对于模式演进，Parquet 支持自动模式合并，可以从简单的模式开始，根据需要逐渐添加更多列。Parquet 的优点包括:

- 列裁剪：只读取需要的列，实现高效的列扫描，减少 IO 操作；
- 谓词下推：因为 Parquet 中记录了每一个 Row group 的列统计信息，包括数值列的 max/min，字符串列的枚举值信息。这样可以从源头过滤掉不符合条件的数据，只读取需要的数据，进一步减少 IO 操作。
- 更高效的压缩与编码：因为同一列的数据类型相同，所以可以针对不同列使用更合适的压缩与编码方式，降低磁盘存储空间。

3.3.2.3.2 Avro

Apache Avro 是基于行存储的文件格式。它可以支持动态类型、嵌套数据结构和快速的二进制编码。Avro 将数据定义和数据存储在一个文件中，其中数据定义（Schema）以 JSON 格式存储，使其便于阅读和解释，详情可参考 IBM 网站 [65]。Avro 的优点包括：

- 支持模式演进。 它可以处理类似缺少字段、添加字段和更改字段等的模式更改。
- 支持跨编程语言实现。
- 支持复杂的数据结构，如数组（arrays），枚举类型（enums），maps 和 unions。

3.3.2.3.3 ORC

ORC 是基于列存储的文件格式。和 Parquet 类似，它并不是一个单纯的列式存储格式，仍然是首先根据行组分割整个表，在每一个行组内进行按列存储。和 Parquet 不同，ORC 原生是不支持嵌套数据格式的，而是通过对复杂数据类型特殊处理的方式实现嵌套格式的支持。在 ORC 文件中保存了三个层级的统计信息，并实现谓词下推。ORC 提供了 3 级索引，并利用这些索引规避大部分不满足查询条件的文件。ORC 格式的表还支持事务 ACID，详情可参考 Apache 网站 [66] 。OCR 的优点包括：

- 有多种文件压缩方式，并且有着很高的压缩比。
- 提供了多种索引，row group index、bloom filter index。
- 支持复杂的数据结构。
- 支持事务 ACID。
- 支持谓词下推。

3.3.2.3.4 开放数据文件格式总结

Avro 是行存储格式，最大的优点是可以解耦数据的生产者和消费者，实现快速的数据接口升级和兼容性。还有一些系统也会选用 Avro 格式去存储 log 文件。说到列式存储，Parquet 目前是大数据分折领域使用最广的列存格式，也是使用 Spark 推荐的存储格式。而 Hive 对 ORC 的支持更好。ORC 文件通常比 Parquet 文件小，ORC 索引可以加快查询速度。对于 ORC 和 Parquet 的选择问题，具体还要看其依赖的计算引擎，我们不能脱离了整个生态环境去进行评判。

3.3.2.4 开放表格式

Table Format 是表的抽象，将数据集文件组合起来，以单个“表”的形式呈现，允许人和工具与表数据高效交互，它本身并不存储数据，只是定义了表的元数据信息以及数据文件的组织形式、统计信息以及上层引擎读取和写入的相关 API。

开放式表格式提供了额外的类数据库功能，简化了数据湖的优化和管理开销。这些功能包括 [67]

- ACID 事务：保证操作的原子性，保证数据的一致性
- 记录级别的操作：允许单个行的插入、更新或删除
- 索引：提高性能，如分区技术
- 并发控制：允许多个进程同时读写相同的数据
- 模式演化：允许在表的生命周期内添加或修改表的列
- 时间旅行：让您能够查询过去某个时间点的数据

本章将介绍三种主流的表格式：Iceberg、Hudi、Delta Lake，并比较它们的异同点，更多对比可以参考 [67]。

表 2 Iceberg、Hudi、Delta Lake 的对比

	Iceberg	Hudi	Delta Lake
ACID	支持	支持	支持
多版本控制	支持	支持	支持
时间旅行,snapshot 回滚	支持	支持	支持
模式演变	支持	有限支持	支持
数据变更	Insert，Merge into，Delete，Merge on read	Upsert, Delete, Insert, Merge on read，Copy on write	Update，Delete，Insert，Merge into，Merge on write
分区演变	支持	不支持	不支持
索引管理	否	支持	否
文件格式支持	Parquet，ORC，Avro	Parquet，Avro	Parquet
依赖 Hive	是	是	否，自有元数据管理

Apache Iceberg 可以适配 Presto，Spark 等引擎提供高性能的读写和元数据管理功能。具有以下特点：

Apache Iceberg 相较于 Delta Lake 和 Hudi 是更加通用化的设计，它完美的解耦了计算引擎底下的存储系统，便于多样化计算引擎和文件格式，很好的完成了数据湖架构中的 Table Format 这一层的实现，因此也更容易成为 Table Format 层的开源事实标准。Delta Lake 的定位是流批一体的存储层，其一大优点就是与 Spark 的整合能力，尤其是其流批一体的设计，配合 multi-hop 的 data pipeline，可以支持分析、Machine learning、CDC 等多种场景。另外，开源的 Delta Lake 是 Databricks 闭源的一个简化版本，它主要为用户提供一个 table format 的技术标准，闭源版本的 Delta Lake 基于这个标准实现了诸多优化。Hudi 强调了其主要支持 Upserts、Deletes 和 Incremental 数据处理，另一大特色是支持 Copy On Write 和 Merge On Read。具体选择那种技术架构要结合业务需求来考虑。

3.3.2.4.1 Hive MetaStore

Hive Metastore（HMS）是 Apache Hive 中负责存储和管理元数据的组件。元数据就是描述数据的数据，例如表名、表类型、存储路径等信息。当我们存储一张表，它的数据部分会存在文件系统中，它的元数据部分通常存储在 Hive Metastore 中。Hive Metastore 会将这些元数据存储在所关联的关系型数据库（例如 MySQL、PostgreSQL）中。在 IBM watsonx.data 中，HMS 使用 PostgreSQL 来持久化数据。从 Hive 3.0 开始，Hive Metastore 已经完全独立于 Hive， 无需安装 Hive 的其余部分即可运行，不限于 Hive，其他第三方服务也可以使用其作为元数据库服务。换句话说 Hive Metastore 就像是一个图书管理员，分门别类地记录了书籍的名称，目录，摆放位置等信息，当读者需要借一些书籍，图书管理员可以快速地定位并给与这些书籍的详细信息。总的来说 Hive Metastore 的

重要作用之一，是帮助底层计算引擎高效地定位并访问分布式文件系统中的数据源。计算引擎可以通过这些元数据来确认如何解析、授权和高效执行用户查询。Hive Metastore 中的元数据与数据湖中的数据一样重要。这意味着其元数据必须是持久的、高可用的，并应该具备灾难恢复能力。Hive Metastore 功能架构图参考 hive 官网 [68]。Hive Metastore 的主要功能：元数据存储，元数据管理，元数据查询优化。

Hive Metastore 作为元数据和数据文件之间的桥梁。提供了数据抽象和数据发现两个核心的功能。当您创建一个新表时，与模式相关的信息，如列名、数据类型等，会存储在 Hive Metastore 的关系数据库中。Hive Metastore 并不是完美的，也存在着架构本身的缺陷，例如存储性能瓶颈与容灾备份的需求。Hive Metastore 也在不断完善，例如引进了缓存机制。IBM cloud 也提供了完全托管的高可用的 Hive Metastore 功能供广大用户选择，详见 IBM blog [69]。

3.3.2.5 数据联邦查询

3.3.2.5.1 数据联邦查询技术的介绍

数据联邦查询技术是一种先进的数据库或数据存储系统的查询方法，广泛应用于湖仓一体化架构。它允许在分布式环境中跨多个分散的数据源执行复杂的数据查询，使不同的用户能够无缝地在这些分布式数据源上通过标准 SQL、JDBC 或 ODBC 等统一查询方式高效地访问数据并且无需移动或者集中存储数据，从而节省了建立集中数据仓库的成本，避免了海量数据复制的工作量和资源浪费。对于企业构建统一数据平台，大量迁移数据成本太高，通过数据联邦查询技术可以接入已有数据系统，加速为生成式 AI 提供的统一数据平台接入现有的企业数据资产。

3.3.2.5.2 开源湖仓架构中的数据联邦查询技术

基于实现跨多个数据源进行查询和分析时的工作原理不同，我们可以将数据联邦查询技术分成三类：开源的联邦查询引擎、数据虚拟化平台、分布式数据处理工具。虽然它们的工作原理不同，但是它们都可以支持复杂的联邦查询操作。

3.3.2.5.3 开源的联邦查询引擎

开源的湖仓架构通常需要能够支持联邦查询的开源引擎，以实现在分布式环境中查询和整合多个数据源。

当联邦查询引擎接收到联邦查询的 SQL 语句时，通常会解析查询计划，并根据各个目标数据源的要求转换成和目标数据源相关的 SQL 语句；在转换后，联邦查询引擎可能会进行一些优化步骤，以确保生成的查询在性能和效率上都能得到优化；然后，联邦查询引擎生成与目标数据源相兼容的原生 SQL 查询语句，直接发送到各个数据源进行查询；最后，联邦查询引擎将把从各个目标数据源上得到的结果整合到一起，最终提供一个统一的查询结果。

以下是一些常用于湖仓架构的开源数据联邦查询引擎：

- Presto（即 PrestoDB）：PrestoDB 是由 Facebook 开发的一个开源、灵活、可扩展的分布式 SQL 查询引擎，支持连接多种数据源，包括关系型数据库（如 MySQL、PostgreSQL）、NoSQL 数据库（如 Cassandra、MongoDB）、数据湖（如 Apache Hive、Amazon S3）等，这种多数据源的支持使得 PrestoDB 成为一个适用于复杂数据生态系统的查询引擎。同时它也支持用户在多个数据源中执行联

邦查询。因此，PrestoDB 的灵活性和高性能使其成为企业和用户在 Open Lakehouse 架构中的一个强大选择。

- Trino（即原 PrestoSQL）：Trino 是 一个开源的分布式 SQL 查询引擎。它是 PrestoDB 的分支，继续发展和维护 Presto 的开源项目，并提供了许多改进和新功能。Trino 内置了多种 Connector 支持多种数据源连接，Trino 的灵活性和性能使其成为大数据处理和分析领域的一个重要工具，特别适用于需要在分布式环境中查询各种数据源的场景，在 Lakehouse 架构中被广泛应用。
- Dremio：Dremio 是一款开源的新一代自助服务的数据湖引擎。它是一款完整的产品，通过界面化的 SQL 输入查询数据湖的数据。Dremio 支持连接多种数据源，包括数据湖（如 Amazon S3、Azure Data Lake Storage）、关系型数据库（如 MySQL、PostgreSQL）、NoSQL 数据库（如 MongoDB、Cassandra）等，也支持多数据源的联邦查询功能，使用户能够轻松访问和整合不同类型的数据。
- Apache Drill：Apache Drill 是一个开源的分布式 SQL 查询引擎， 具有敏捷性、灵活性和易用性，专为 Hadoop，NoSQL 和云存储设计。它支持多种类型的 NoSQL 数据库（几乎可以查询任何类型的 NoSQL 数据库）和文件系统查询，它支持联邦查询，用户可以通过 SQL 查询语言整合不同类型和位置的数据。

3.3.2.5.4 数据虚拟化平台

数据虚拟化平台（Data Virtualization Platforms)是一种数据集成技术，通过创建一个抽象的、统一的数据访问层，使得用户可以从一个单一的接口访问或查询分布在多种数

据源（关系型数据库、NoSQL 数据库、文件系统、云存储等不同类型的数据源）中的数据，而无需了解底层数据源的具体细节。

数据虚拟化平台能够对数据进行抽象，隐藏了数据的物理位置和格式细节，使得用户可以以一种更简单、更统一的方式查询和操作数据 [70]。

当数据虚拟化平台接收到联邦查询的 SQL 语句时，首先会将查询请求转换为逻辑查询计划；根据虚拟数据视图和元数据信息，数据虚拟化平台进行查询优化，这可能涉及到重写查询计划，选择合适的执行计划，并利用缓存和索引来提高查询性能；然后，数据虚拟化平台将逻辑查询计划转换为和各个目标数据源相关的 SQL 语句，进而发送到各个目标数据源进行查询；最后，数据虚拟化平台将把从各个目标数据源上得到的结果整合到一起，最终提供一个统一的查询结果。

湖仓架构通常会利用多种数据虚拟化平台来实现数据的统一管理和查询。比如一些常用的数据虚拟化平台：Denodo、TIBCO Data Virtualization 和 IBM Cloud Pak for Data。 这些平台来自不同的供应商，为企业级应用而设计，提供了可靠的技术支持、管理功能和监控能力，我们可以根据特定需求和架构设计选择合适的一种或多种数据虚拟化平台来实现数据的统一管理和查询。

3.3.2.5.5 分布式数据处理工具

分布式数据处理工具是一类用于处理大规模数据集的软件工具，它们在多台计算机或服务器上处理可能分布在不同数据源、不同位置中的数据。联邦查询技术允许在多个分布式数据存储之间进行查询和操作，而无需将数据集中到一个单一的位置或系统。因此，分

布式数据处理工具常常与联邦查询技术结合使用，以支持在分布式环境中跨多个数据源进行查询和操作，从而实现数据的统一访问和管理。

当分布式数据处理工具接收到联邦查询的 SQL 语句时，首先从各个数据源中提取需要查询的数据，可以以分布式的方式分区加载到分布式数据处理工具的数据集中，这样数据将会在存储中的多个节点上进行分布；SQL 查询语句被转换成适用于分布式数据处理工具的查询计划，这个计划会分解查询操作，使得可以并行处理不同部分的查询；最终，分布式计算集群中的节点将各自的计算结果合并，整合成一个统一的查询结果。常用的工具有 Apache Spark，Apache Doris 等。

3.3.2.6 开源 SQL 查询引擎

3.3.2.6.1 SQL 查询引擎

SQL 查询引擎是一种软件组件或系统模块，用于解析、执行和处理 SQL 查询语句。这类引擎能够接收、解释和执行用户提交的 SQL 查询，并从数据存储中检索、操作和处理数据，最终返回符合查询条件的结果。关系型数据库都内置 SQL 查询引擎的支持，对于数据湖和湖仓，这就需要独立的查询引擎来实现用统一 SQL 对各种数据源执行操作。独立的查询引擎不依赖于特定数据库系统或数据存储技术，为用户提供了跨数据源执行查询和分析的能力，允许在不同数据存储系统中进行数据聚合、联接、筛选和分析，提供了更灵活的数据处理和查询功能。

在湖仓一体架构中，多个独立的开源查询引擎可以被使用，以便针对存储在数据湖中的数据执行不同类型的查询和分析操作。湖仓一体架构中常用到的一些流行的开源 SQL 查询引擎，如：Presto，Apache Spark，Apache Hive，Apache Drill 等。

3.3.2.6.2 Presto

Presto 是一个高性能、分布式的 SQL 查询引擎，用于处理大规模数据分析和查询。Presto 采用 MPP（Massively Parallel Processing 大规模并行处理）架构，支持分布式计算，能够运行在大规模的集群上，实现高并发性和高可扩展性 [71]。

Presto 分布式的架构和设计理念，让 Presto 具备非常快速的查询执行速度和低延迟，即使在 PB 级别甚至更大规模的数据量下也能表现出色。除此之外，Presto 支持标准的 SQL 查询语言，并且可以无缝查询多种数据存储系统，包括关系型数据库、NoSQL 数据库、云存储等。基于这些显著的优势，Presto 在各个领域都有着广泛的应用，从数据湖、数据仓库、实时分析到日志分析等，都能发挥出色的效果。它的高性能和灵活性使得企业能够快速且灵活地分析处理海量数据，为决策提供更可靠的数据支持。因此，Presto 已成为许多组织和公司进行数据分析和处理的首选工具之一。

3.3.2.6.3 Apache Spark SQL 模块

Apache Spark 作为一个开源的分布式计算系统，设计用于处理大规模数据，并支持复杂的数据处理和分析任务。Apache Spark 不是传统意义上的 SQL 查询引擎，尽管它最初是为支持复杂的数据处理任务而设计的（如机器学习、图分析、流处理等），但它也提供了功能强大的 Spark SQL 模块，用于执行 SQL 查询和操作结构化数据。这种 SQL 查询的功能使得 Spark 更易于使用，并且使得用户可以通过 SQL 来处理和分析数据，尤其是对于熟悉 SQL 查询语言的用户来说更加方便。Apache Spark 架构图参考 [72]。Spark 分布式计算架构，可以使用户在大规模数据上执行高性能的 SQL 查询和操作。此外 Spark SQL 提供了一个统一的 API，允许用户使用 SQL 查询和常规的 DataFrame API（类似于关系

型数据库表）来处理数据。Spark SQL 使用 Catalyst 查询优化器来优化 SQL 查询计划，并支持标准的 SQL 语法，包括 SELECT、JOIN、GROUP BY、WHERE 等操作。

除此之外，Spark 与 MLlib（Spark 的机器学习库）集成，可以无缝进行机器学习模型的训练和推断，并支持丰富的数据处理操作，包括数据清洗、转换和分析。

3.3.2.6.4 Apache Hive

Apache Hive[68]是建立在 Hadoop 之上的数据仓库软件，它提供了类似于 SQL 的查询语言，称为 HiveQL，用于查询和分析存储在 Hadoop HDFS 中的大规模数据集。它最初由 Facebook 开发，用于处理他们庞大的数据集。它于 2008 年作为开源项目捐赠给 Apache 基金会，并迅速成为 Hadoop 生态系统中广受欢迎的组件之一。

Hive 可以将数据存储在 Hadoop 的 HDFS（Hadoop 分布式文件系统）中，也支持其他存储格式，如 HBase 和 Amazon S3。对于熟悉 SQL 的用户来说，学习和使用 Hive 相对容易，可以直接与 Hadoop 生态系统无缝集成，利用 Hadoop 集群的强大功能；但是由于使用 MapReduce 等批处理作业，对于实时性要求高的场景，Hive 可能无法满足，对于一些复杂的查询或小规模数据集，性能可能不如其他实时处理引擎。

总体来说，Apache Hive 在处理大规模数据时是一个强大的工具，尤其适合用于批处理和对数据进行较复杂分析的场景。

3.3.2.6.5 Apache Drill

Apache Drill[73]是一个开源的分布式 SQL 查询引擎，最初由 MapR 公司开发，其目标是提供一种能够实时查询大规模分布式数据的解决方案。

Apache Drill 能够使用标准的 SQL 语法直接查询多种数据源，包括传统关系型数据库，文件系统数据，NoSQL 数据库和云存储等。它可以在查询过程中无缝地处理这些不同的数据源，无需预定义模式或进行数据转换。Drill 是为分布式环境设计的，能够在多个节点上并行执行查询，从而提高查询性能和可扩展性。用户能够在应用程序中嵌入 Drill 引擎，使得数据查询和处理能力可以被直接集成到应用程序中，从而简化了数据分析和应用开发的过程。虽然 Drill 主要用于批量查询和分析，但也支持实时查询，通过轻量级的执行计划和查询引擎，尽可能地提供快速的响应时间。

Apache Drill 的发展一直专注于提供更高的查询性能、更好的兼容性以及更广泛的数据源支持。其持续改进和发展使得它成为处理大规模数据查询和分析的重要工具之一，并且在数据格式多样性和无模式查询方面有着显著的优势。

3.3.2.6.6 SQL 查询引擎的选型

面对不同的场景和需求，如何选择 SQL 搜索引擎？可以先参考下面这张表，对这四种 SQL 搜索引擎有个更深入的了解。

表 3 四种开源 SQL 查询引擎的比较

场景/特性	Presto	Apache Spark	Apache Hive	Apache Drill
交互式查询	Presto 是专注于交互式查询的引擎，适用于需要快速响应用户查询的场景。	Spark 虽然可以执行 SQL 查询，但对于大规模数据的交互式查询，性能可能	Hive 在交互式查询方面性能较差，不太适用于需要即时响应的场景。	Drill 适用于需要实时查询的场景，能够在较短时间内完成对数

		不如专门的 SQL 引擎。		据的查询和分析。
批量处理	Presto 能够执行批量处理，但更擅长于交互式查询，不是最佳的批量处理引擎。	Spark 是通用的大数据处理引擎，适用于批量处理和流处理，具有广泛的用途。	Hive 专门用于批量处理和大规模数据分析，对于需要对静态数据集进行批量处理的场景较为适用。	Drill 也适用于批量查询和处理，但其重点是在无模式查询和实时性能上。
多数据源查询	Presto 非常擅长于查询多种数据源，支持各种数据格式和多种数据源的无缝查询。	Spark 也能够处理多种数据源，但在查询多种数据源方面可能不如 Presto 灵活。	Hive 适合于与 Hadoop 生态系统集成，能够查询 HDFS 等存储系统中的数据。	Drill 专注于无模式查询，支持多种数据格式和多数据源查询。
实时性能要求	Presto 能够提供较快的查询响应时间，适用于对查询响应速度要求较高的场景。	Spark 在一些场景下能够实现近实时处理，但在某些复杂查询下性能可能受限。	Hive 在实时性能方面表现较差，不适合需要即时响应的场景。	Drill 在一些场景下能够提供较好的实时性能，但对于复杂的查询或小规模数据集

				可能性能不理想。

这些引擎在不同的场景中表现出不同的特点和优势，选择哪一个取决于具体的使用场景和需求。例如，如果需要高性能、交互式分析，则 Presto 可能是一个不错的选择；如果需要一个通用的大数据处理引擎，包括流处理和机器学习，则 Spark 可能更适合。根据具体的业务需求和数据处理目标，选择最适合的引擎才是至关重要的。

3.3.2.7 数据处理和注入

Apache Flink 是开源的分布式引擎，用于对无界限（流）和有界限（批处理）数据集进行有状态处理。流处理应用程序旨在连续运行，最大限度地减少停机时间，并在摄取数据期间对其进行处理。Apache Flink 专为低延迟处理、在内存中执行计算、实现高可用性、消除单点故障以及水平扩展而设计。Apache Flink 专为流式传输优先而开发，为流处理和批处理提供了统一的编程接。Apache Flink 提供支持的一些常见应用程序类型包括：事件驱动的应用程序，数据分析应用程序，数据管道应用程序。

Apache Spark 如前面所述的设计用于处理大规模数据除了有强大的查询能力之外，也能很好的处理数据的 ETL，原生支持批处理和流处理，相对于 Flink 原生流，Spark 是通过微批处理，延时性略差于 Flink。Spark 更适合快速的批处理。

CDC 变更数据捕获是一种经过验证的数据集成模式，用于跟踪数据更改，并向必须响应这些更改的其他系统和服务发出警报。变更数据捕获有助于确保所有依赖数据的系统数据同步，功能正常。Debezium 是 Red Hat 开源的变更数据捕获工具，支持 Mysql，

MongoDB，PostgreSQL，SQL Server，Oracle，Db2，Cassendra 等，目前没有直接支持 Presto，需要去扩展。

3.3.2.8 向量数据库

3.3.2.8.1 RAG 和向量数据库

在专业领域生成式人工智能方面，企业往往用到 RAG 和向量数据库，参考 3.2.5。这里简单描述了 RAG 的主要组成: 依次是：数据提取 — embedding（向量化）— 创建索引 — 检索 — 自动排序（Rerank）— LLM 归纳生成。事实上，几乎任何企业都可以将其技术或政策手册、视频或日志转化为称为知识库的资源，从而增强 LLM。这些来源可以支持客户或现场支持、员工培训和开发人员生产力等用例。除此以外 RAG 还降低了 LLM 泄露敏感数据或产生不正确或误导性信息的可能性。同时也可以降低在企业环境中运行基于 LLM 的聊天机器人的计算和财务成本。IBM 推出的 AI 和数据平台 watsonx 就包括了 RAG 功能 [74]。

向量数据库是一种特殊的数据库，它以多维向量的形式保存信息。可以参考 3.2.5.2。向量数据库在跟 LLM 结合以后，可以有多种方式支持 LLM，包括：

- 提供提示工程的知识库。
- 做相似度搜索，分类等。
- 作为 LLM 模型的缓存。

同时向量搜索也在改变传统数据库的搜索能力，使数据库结合向量搜索具备相似度搜索的能力。不少传统的数据库比如 PostgreSQL 就支持向量搜索插件的方式支持向量搜

索。根据向量数据是否开源友好、是否是数据库，向量数据库可以简单划分为下图四项限。

图 11 向量数据库分类图 [75]

通常向量数据库有如下特性：

- 支持向量相似性搜索，它会找到与查询向量最近的 k 个向量，这是通过相似性度量来衡量的。矢量相似性搜索对于图像搜索、自然语言处理、推荐系统和异常检测等应用非常有用。
- 使用矢量压缩技术来减少存储空间并提高查询性能。矢量压缩方法包括标量量化、乘积量化和各向异性矢量量化。
- 可以执行精确或近似的最近邻搜索，具体取决于准确性和速度之间的权衡。精确最近邻搜索提供了完美的召回率，但对于大型数据集可能会很慢。近似最近邻搜索使用专门的数据结构和算法来加快搜索速度，但可能会牺牲一些召回率。
- 支持不同类型的相似性度量，例如 L2 距离、内积和余弦距离。不同的相似性度量可能适合不同的用例和数据类型。

- 可以处理各种类型的数据源，例如文本、图像、音频、视频等。可以使用机器学习模型将数据源转化为向量嵌入，例如词嵌入、句子嵌入、图像嵌入等。

向量数据库绝不仅仅是用来进行简单的向量检索，要想真正提升开发者的开发效率和使用成本，需要系统开发者深入理解硬件、存储、数据库、AI、高性能计算、分布式系统、编译原理、云原生等方方面面，以确保其稳定性、性能和易用性。除此以外，可扩展性、安全性、性能以及成本问题也是用户所关心的。

3.3.2.8.2 Milvus

Milvus 是 Zilliz 于 2019 年 10 月正式开源的基于原生向量设计的分布式向量云原生数据数据库。它集成了目前在向量相似性计算领域比较知名的几个开源库（Faiss，SPTAG 等），通过对数据和硬件算力的合理调度，Milvus 能够很好地应对海量向量数据。

Milvus 目前是最活跃热度最高的向量数据库，Milvus 2.3.x 提供了 GPU 版本，性能呈现比 CPU 版本快 3 - 10 倍。除此以为，Milvus 先后支持了范围搜索，Upsert、Kafka Connector、Airbyte，动态 schema 等种种特性。Milvus 已有应用场景包括：图片检索系统，视频检索系统，音频检索系统。分子式检索系统，推荐系统，智能问答机器人。Milvus 的特点包括：

- 支持 11 种索引类型，是目前支持索引类型最多的向量数据库。
- 支持 RBAC。
- 云原生支持，可伸缩。
- API 文档齐全。

3.3.2.8.3 Chroma

Chroma 是 AI 原生的基于向量检索库实现的轻量级开源向量数据库。作为后起之秀，Chroma 在 2023 年中发布了第一个面向生产的版本 V0.4，它的优点是易用、轻量，由于刚刚发布所以功能相对简单。Chroma 简化了构建 LLM 应用程序的过程, Chroma 下一个重要的里程碑是从单节点到分布式系统以及提供云服务能力 [76]。Chroma 的主要特点有：

- 功能丰富：支持包括查询、过滤、密度估计和许多其他功能。
- 支持 LangChain (Python 和 Javascript)、LlamaIndex。
- 在 Python notebook 中运行的相同 API 可扩展到生产集群。

3.3.2.8.4 Weaviate

Weaviate 是一个开源向量数据库。它可以无缝扩展到数十亿个数据对象。其凭借易用、开发者友好、上手快速、API 文档齐全等特点脱颖而出。Weaviate 更适合需要快速集成向量数据库的开发人员。Weaviate 的一些关键特性有：

- 速度：Weaviate 可以在几毫秒内从数百万个对象中快速搜索出最近的 10 个邻居。
- 灵活性：使用 Weaviate，可以在导入或上传自己的数据时对数据进行向量化，可以利用与 OpenAI，Cohere，Hugging Face 等平台集成的模块。
- 快速部署：从原型到大规模生产，Weaviate 都强调可伸缩性、复制和安全性。
- 搜索扩展：除了快速向量搜索，Weaviate 还提供推荐、摘要和神经搜索框架集成。

3.3.2.8.5 Qdrant

Qdrant 可以作为 API 服务运行，支持搜索最接近的高维向量。使用 Qdrant，可以将嵌入或神经网络编码器转换为应用程序，用于匹配，搜索，推荐等任务。以下是 Qdrant 的一些关键功能:

- 通用的 API：提供 OpenAPI v3 规范和各种语言的现成客户端。
- 速度和精度：使用自定义 HNSW 算法进行快速准确的搜索。
- 先进的过滤方法：允许基于相关矢量有效载荷的结果过滤。
- 不同的数据类型：支持字符串匹配、数字范围、地理位置等。
- 可伸缩性：具有水平扩展功能的云原生设计。
- 效率：内置 Rust，通过动态查询规划优化资源使用。

Qdrant 以 Rust 语言构建，提供 Rust、Python、Golang 等客户端 API，能够满足当今主流开发人员的需求。Qdrant 更适合追求低成本基础设施维护的开发人员。

3.3.2.9 图数据库

图数据库是一种以图结构存储数据的数据库类型，其中数据以节点（实体）和边（关系）的形式表示。图数据库可以使用图算法和遍历有效地查询和分析复杂且相互连接的数据。在人工智能领域，图数据库应用非常广泛，可以参考 3.2.5.6。开源的图数据库比较多，这里只例举其中几个：

- Neo4j 是目前最流行也是时间比较久的开源图数据库，原生支持图数据存储，提供集群，ACID 事务支持，支持单机部署，可以跟 Spark 集成。

- JanusGraph 用 Java 实现的分布式图数据库，支持 ACID，可以跟 Spark 集成，支持创建任意多图。
- Dgraph 用 Go 实现的分布式图数据库，支持用 GraphQL 查询，支持多跨数据中心复制，支持高可用和高可靠性。不支持 Spark 集成。

3.3.3 IBM 湖仓管理工具 watsonx.data

IBM 在 2023 年 7 月发布了湖仓产品 IBM watsonx.data. 做为新的数据管理的战略产品，IBM 在 watsonx.data 中投入了很多资源，在收购了 Ahana，成为 Presto 社区的重要贡献者之一来影响查询引擎的市场的同时，IBM 还在开源的基础上做了很多增强，详细架构参考图 12 IBM watsonx.data 架构图。

图 12 IBM watsonx.data 架构图

3.3.3.1 多云部署

大部分湖仓提供商只在部分或者自有的公有云上提供服务。而在实际使用中，企业的数据实际可能存在多种云提供商和企业内部数据中心。IBM 依赖于 Red Hat OpenShift 虚拟化这一层，屏蔽不同云的差异，实现了同一套实现支持混合云部署。这不仅节省了运维成本同时为了企业内部数据共享提供了便捷。数据在生成式人工智能中是企业的重要资产，所以支持本地部署，保护企业数据资产，是企业构建统一数据平台的重要考量之一。

3.3.3.2 数据治理

大部分的开源实现没有很好的数据治理，在整个 AI 或者数据分析中，可信的数据才能得到可信的结果。没有可信的高质量的数据输入，很难得到理想的 AI 赋能。质量差的数据是企业获得高质量人工智能分析的主要障碍，很容易导致“垃圾进“和”垃圾出”的问题。数据资产是企业最重要的核心资产，如何安全，合理合规的使用数据在整个数据生命周期管理中非常重要。如果数据使用过程存在泄密，不合规等情况，那么没法开展有效的智能分析。不同人工智能系统对于定量（结构化）、定性（非结构化）数据集的处理能力也不尽相同。IBM 在开源的 Presto 的基础上，不仅构建了内置的数据访问控制，用户可以通过不同对象级别做不同的访问控制来保证数据只有在被给定权限的用户才能访问。同时 IBM watsonx.data 跟 IBM Knowledge Catalog 集成完成数据质量控制、数据脱敏、和数据生命周期管理等数据治理场景，从一开始就把数据治理问题考虑到产品中。在 IBM watsonx.data 产品中可以一键集成 IBM Knowledge Catalog，并应用 IBM Knowledge Catalog 中制定的脱敏规则等，在后续通过 Presto 查询的过程中，数据就严格脱敏。

3.3.3.3 支持图形，图像，视频和 RAG

随着短视频，社交媒体的发展，企业的数据格式不仅仅包含传统的数仓，数据湖里的数据，还有越来越多的图形，图像，视频数据产生。watsonx.data 产品内置了 Milvus 服务，用户可以直接在 watsonx.data 中启动 Milvus，通过向量化后对图形，图像，视频做搜索，同时也可以把 Milvus 做为 RAG 的知识库，完成知识库构建，知识搜索，再结合 LLM 完成最终的答案，而传统湖仓服务没有包含 AI 新挑战的向量数据库，需要企业单独管理。未来 watsonx.data 会有更多的数据资产到企业知识的场景和功能，一站式的服务企业分析和 AI 的数据需求，来满足生成式人工智能的统一数据平台要求。同时 watsonx.data 也会引入数据治理能力到知识库，完成整体的数据治理。

3.3.3.4 支持多查询引擎

watsonx.data 不仅内置了 Presto 做为查询引擎，同时还内置了 Spark 引擎跟 watsonx.data 无缝集成，用户可以根据自己的需要使用 Spark 引擎注入数据，使用 Presto 引擎查询数据，或者直接通过 spark 引擎做数据中间处理再插入回湖仓，还可以通过 Spark 跟企业内部图数据库集成。不仅如此，同样的数据 watsonx.data 可以通过共享 Hive Metastore 的方式使用 Db2 查询引擎或者 Netezza 做为查询引擎来满足不同的集成和业务需求。除此之外，watsonx.data 还提供了基于 C 语言的查询引擎，目前还处于技术预览阶段。

3.3.3.5 数据处理部分

watsonx.data 除了跟开源的数据处理产品集成，还可以跟 IBM Data Stage 集成，通过图形化任务编排的方式完成复杂的 ETL。未来 watsonx.data 还会跟 IBM CDC 等其他产品家族的集成，可以实现更多的数据实时注入能力，也可以通过内置的 Spark 引擎完成数据加工处理，提高分析和 AI 使用数据的效率。

3.3.3.6 安全可扩展

watsonx.data 不仅仅基于很多开源组建，做为企业级产品，watsonx.data 做了很多安全增强，修复了很多安全漏洞，同时做了更多跟第三方的集成增强，比如在原来 Hive Metastore 集成上提供了 Kafka listener 接口，允许外部去同步 Hive Metastore 的变更。作为企业级的产品，watsonx.data 解决很多企业基于开源的升级和运维的痛点，不仅有可靠的安全扫描，版本升级测试，而且有企业级服务团队提供补丁和升级包，这将大大节省企业运维的成本。

3.3.3.7 开放的生态

watsonx.data 基于开源组建构建，使用开放的数据格式和表格式，可以集成开源的各种报表、数据处理和机器学习工具。IBM 在 2023 年收购了 Presto 基金会两大创始成员之一 Ahana，成为 Presto 开源项目的主要贡献方之一，IBM 在不停地回馈社区。

3.4 基础支撑平台

3.4.1 基础支撑平台综述

企业通常需要结合实际业务拥有自己的模型能力，训练私有模型更好的为业务赋能。在这一过程中，会遇到来自技术和非技术领域的诸多挑战，如安全合规、大规模数据处理、算力利用率问题（章节 2.2.1），以及从日常体验上对于人机协同（章节 2.2.2）提出的新要求。

目前构建一套企业级生成式人工智能平台需要具备丰富的人工智能相关知识以应对上述挑战，这些知识涉及对于模型和业务的理解，混合云平台上部署应用时硬件集成，资源优化在内的方方面面。具体而言在本章节，我们将从技术栈的角度讨论生成式人工智能如何更好的与混合云相结合，在治理相关章节会讨论随着大模型技术在混合云上应用，如何利用现有混合云领域的技术和大模型相结合解决大模型在混合云架构落地中的治理能力的相关趋势（章节 4.6）。

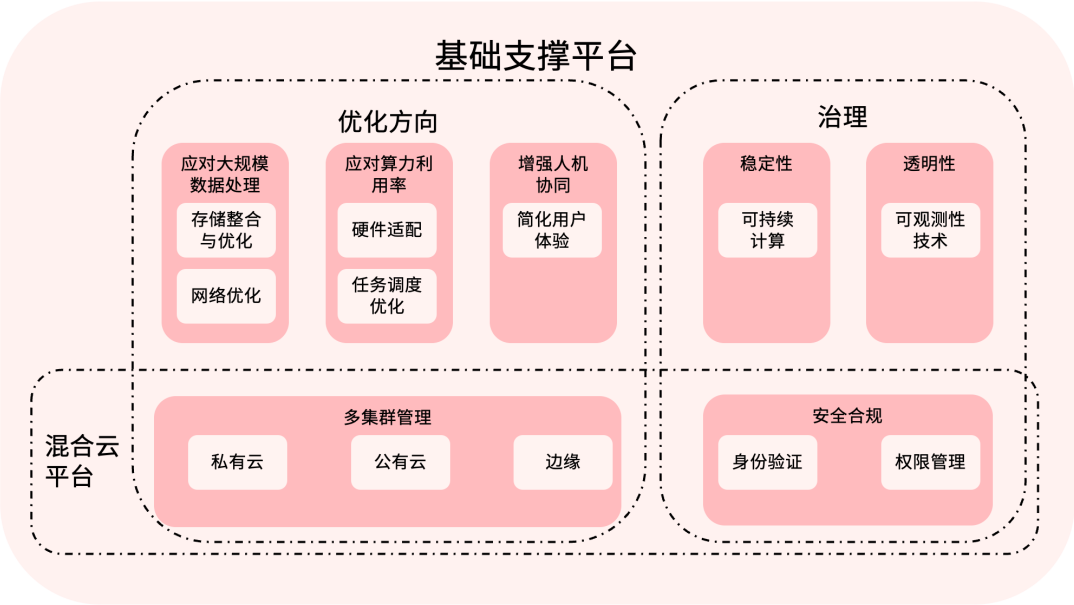


图 13 基础支撑平台概览

容器编排平台为上层应用（如数据服务平台，AI 平台）提供运维管理层面的支撑。同时实现通过一致性的运维管理方式将容器部署在异构的环境中（如私有云和公有云之间，或不同公有云平台之间），如 Kubernetes [77] 或 OpenShift [78]。为更好的应对大模型所带来的挑战，我们收录并整理了以下几种常见的措施或技术方向。请注意，这些措施或技术方向在实施阶段需要结合实际环境的硬件支撑情况。

3.4.2 基础支撑平台应对大规模数据处理的常见措施

由于海量、多源、动态更新的数据是训练模型和进行数据挖掘的必要条件。为更好的应对大规模数据带来的可扩展性挑战（章节 2.2.1），在基础支撑平台部分可以对数据使用的各个环节进行优化，常见的优化方向包含数据静态存储和数据流动（如网络传输）。在应对可拓展性挑战的同时，提高算力可用性。

3.4.2.1 存储整合与优化

作为对数据平台和服务（章节 3.3）的硬件支撑，对于数据存储而言，提供的全局数据平台能力，支持多种应用访问协议互通（如对象、容器、HDFS 等等）适配不同存储环境，实现数据的整合和调度，结合多种存储介质（包括磁带）实现分层存储环境降低数据总体拥有成本，提升端到端的数据处理效率 [79]。

3.4.2.2 网络优化

为了减少大规模数据运算时产生的网络开销，网络优化成为基础支撑平台层的常见优化措施之一 [80]。以 multi-nic-cni [81]项目为例，在支持云基础设施在运行期间动态变化的

同时，减少了手工维护成本，提高带宽利用率，通过对于应用完全透明的技术实现了底层网络接口的最优配置。

3.4.3 基础支撑平台应对算力利用率问题的常见措施

基础支撑平台为应对在处理大规模数据时，提高单芯片算力、突破算力利用率、实现更高能效比，这一领域的重要挑战（章节 2.2.1）。通常而言，首先会实现自动化硬件适配工作，将计算任务和硬件调度在混合云管理平台上统一调度，并在此基础上，实现优化计算任务的调度方案，提高能效比。

3.4.3.1 硬件适配

在硬件适配方面，通过自适应的硬件驱动配置，混合云通过设备扩展框架可以对多种算力设备进行支持。在实际使用中，需要考虑配置多个软件组件，如驱动支持，容器权限等，是困难且容易出错的。诸如 NVIDIA GPU Operator[82]这样的项目就很好的解决了这一难题。

3.4.3.2 任务调度优化

以 Multi-Cluster App Dispatcher（MCAD）[83]和 InstaScale[84]项目为例，这类项目实现了包括作业优先级在内的资源调度逻辑来更好的利用硬件资源。通常提供作业排队、作业优先级和抢占、超时以及系统用户之间资源共享的编排的能力，甚至包括动态扩展云托管混合云集群的能力[80]，从而实现设备利用率最大化。在 IBM 研究院的博客中[85]，分享了通过这种方式在分布式训练运行中有效地使用 GPU 的实践。

3.4.4 基础支撑平台增强人机协同简化用户体验的常见措施

基于控制台和图形界面引导用户在混合云平台上执行生成式人工智能相关任务，从而简化的用户体验（章节 2.2.2），以便用户有效地完成包括训练、测试和监控在内的任务是一项挑战 [80]。相对于在本地部署复杂的开发环境，同步大量训练数据进行训练的做法，这类简化用户体验的优化，显着降低了人工智能研究者进入云原生技术堆栈的门槛，开源项目如 CodeFlare [86]就很好的解决了这一问题。

3.5 生成式人工智能的企业级应用

3.5.1 生成式 AI 的五大模态

根据内容生产模态，生成式 AI 能够被分为四大基础模态，包括文本、音频、图像、视频，每一种模态技术都有着独特的应用场景和特点。此外，这四类模态的融合还带来第五类模态——跨模态内容生成模式，支持创造出更为丰富多彩的生成内容 [87]。

3.5.1.1 文本生成

文本内容生成可以大致分为非交互式和交互式两种。非交互式文本生成包括摘要/标题生成、文本风格迁移、文章生成、图像生成文本等技术。这些技术可以根据不同的使用场景，自动生成符合要求的文本内容，提高文本生成的效率和质量。交互式文本生成是一种更加智能化的应用方式，可以根据用户的需求和反馈，生成更加贴近用户需求的内容，主要包括聊天机器人、文本交互游戏等应用。

【代表性产品或模型】：JasperAI、copy.AI、ChatGPT、Bard、AI dungeon

3.5.1.2 音频生成

音频生成技术是一种通过算法和模型生成人工音频的技术。音频生成技术可以应用于特定场景下的文本生成语音，如数字人的播报、语音客服等。这些场景化的应用可以根据用户和场景的需求，通过算法生成符合要求的语音，提高用户体验和效率。此外，该技术在 C 端产品中也十分常见，如智能家居、车载音响、虚拟助手等。

【代表性产品或模型】：DeepMusic、WaveNet、Deep Voice、MusicAutoBo

3.5.1.3 图像生成

图像生成技术是一种通过算法和模型生成人工图像的技术。图像生成技术可根据使用场景分为图像编辑修改和图像自主生成。图像编辑修改技术可实现对图像的重构和修复，提高图像的质量和清晰度，满足用户对图像处理的需求，如图像修复、人脸替换、图像去水印等方面。图像自主生成技术通过算法和模型实现对图像的自主生成，可以为用户提供更加多样化的图像服务，如参照图像生成绘画图像、真实图像生成素描图像、文本生成图像等。

【代表性产品或模型】：EditGAN，Deepfake，DALL-E、MidJourney、Stable Diffusion，文心一格

3.5.1.4 视频生成

视频生成技术是一种通过算法和模型生成人工视频的技术。视频生成技术可以根据使用场景分为视频编辑和视频自主生成。视频编辑技术可应用于视频超分辨率、视频修复、视频画面剪辑等方面。视频自主生成技术的核心原理是使用深度学习模型对图像或视频进行分析和理解，再根据特定算法生成相应的视频。可应用于图像生成视频、文本生成视频等方面。

【代表性产品或模型】：Deepfake，videoGPT，Gliacloud、Make-A-Video、Imagen video

3.5.1.5 跨模态生成

跨模态生成是指通过组合不同模态的 AI 技术，实现模态间的转换和生成。跨模态生成通过实现不同媒介之间的转化和生成，拓展了人工智能应用的领域和应用场景，支持将不同的信息形式转化为人类可理解的其他形式，例如将文本转化为图像、音频或视频，将图像转化为文本、音频或视频，从而实现更加自然、直观、高效的交互方式。跨模态生成技术同时也可以应用于各个领域，如艺术创作、广告营销、教育培训、医疗诊断等，提升 AIGC 的产业化和工业化应用能力。

【代表性产品或模型】：DALL-E、MidJourney、Stable Diffusion, watsonx

3.5.2 业务赋能

企业在利用生成式 AI 进行业务赋能的过程中，需要构建基本准则，通过开放创新和柔性监管协同发展，达到有针对性地赋能。在具体实施过程中，有效的行动举措为生成式 AI 的落地提供指引，AI 联盟的成立将成为可靠的第三方，为业务赋能保驾护航，同时，生成式 AI 与各行各业的深度融合，为业务赋能带来更多的机遇和价值。

3.5.2.1 构建基本准则

首先是**开放**，企业应该积极拥抱领先的 AI 技术，并且借助开源社区、开源技术加速创新；其次是**针对性**，比如帮助企业使用自己的数据，开发针对特定场景、能快速产生收益的 AI 模型（如 HR 流程自动化、客服系统智能化、IT 应用现代化等），同时确保符合内部规章；第三是**可信**，这不仅涉及数据的治理、模型的监管，也包括各国、各行业的不同的

合规要求；第四是**赋能**，企业需要一个上手快、可扩展的工具平台，基于自己的数据来训练、调优、部署 AI 模型，而不只是当一个大模型的消费者 [88]。

3.5.2.2 推动开放创新

在生成式 AI 即将颠覆创新格局之际，当下正是组织重新评估其创新方法的绝佳时机。传统创新是一种封闭的内部流程，仅利用组织的内部资源，在严格保密的环境中创造惊喜和竞争优势。但传统的“封闭式”创新已不能满足当下基于合作的生态系统经济，开放创新是推动业务发展的明智决策。开放创新是一种需要共同投资和携手共创的生态系统游戏。平均而言，每投入 1 亿美元的创新支出，组织要与大约四家生态系统合作伙伴开展合作。开放创新的核心是基于共享数据和洞察建立合作伙伴关系 [89]。

随着生成式 AI 崭露头角，改变创新方式已经成为一项尤为紧迫的任务。企业高管们期望生成式 AI 在整个创新生命周期中发挥重大影响力，从构思、发现、评估、执行到商业化，以及应用于生态合作和成果衡量（见图 14 生成式 AI 在创新生命周期中的影响力）。他们不仅将生成式 AI 视为创新工具箱中的一件新利器，更坚信生成式 AI 将颠覆现代化企业创新的本质 [89]。

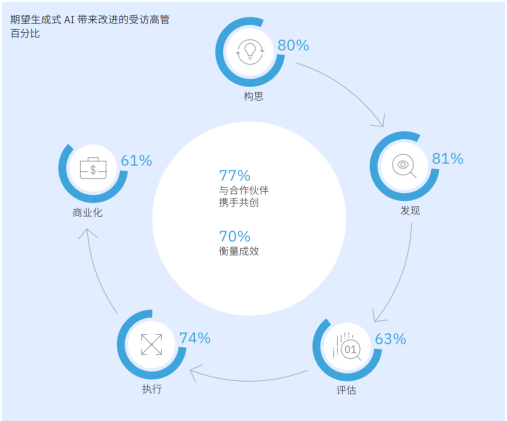


图 14 生成式 AI 在创新生命周期中的影响力

为什么许多企业都无法充分把握开放创新的商机？简而言之，因为这太难了。从网络安全问题、技术障碍到缺乏灵活性，多重挑战都将阻碍生态合作伙伴之间的创新合作。协同内部部门并消除创新中的职能孤岛已经够困难的了。再要引入外部合作伙伴，并通过“合纵连横”让它们发挥能力为共同目标而努力，这难度实在令人望而却步。

生成式 AI 可以帮助企业克服所面临的一些挑战。事实上，大多数组织表示目前正在评估生成式 AI 是否可作为开放创新工具或正在开展相关试点，主要就是因为生成式 AI 能够改善生态合作。

但仅靠生成式 AI 无法播散开放创新的种子。在技术指数级发展的时代，要将愿景转化为现实，组织必须明确可从与生态合作创新中获得哪些业务价值，以及实现这一目标需要哪些条件。

3.5.2.3 推进柔性监管

制度的合法性对于全球创新网络的风险有一定的解释力，许多学者的研究表明组织合法性是组织发展壮大的重要资源。2023 年 5 月，OpenAI 公司 CEO Sam Altman 在美国国会的人工智能监管听证会上表示需要建立一个新的立法和监管体系以应对 AI 的潜在风险。随着《生成式人工智能服务管理办法（征求意见稿）》《网络信息内容生态治理规定》《网络数据安全管理条例（征求意见稿）》《互联网信息服务深度合成管理规定（征求意见稿）》等政策法规的相继出台，我国正积极开展生成式人工智能的治理实践。

AIGC 技术的应用涉及多个行业，例如医疗、金融、媒体等，行业协会、企业和政府需要通力合作，制定适应各领域需求的指导方针和标准，鼓励 AIGC 技术在社会中的广泛应用，协同推动创新。

3.5.2.4 制定行动举措

编写人工智能行动手册，支持员工将其作为实践。行动手册应是动态文档，根据成功和失败经验以及 KPI 列明工作清单和工程原则。创建在设计中心和数据中心交汇点运行所需的架构和团队结构 [90]，这是真正的变革推动因素。

坚持文档记录。让数据科学家参与工作。必须深刻认识到，部署人工智能模型不是唯一的目标，也不意味着项目的终结。为扩展人工智能，在模型投入生成环境后，仍需评估并不断改进。如果模型无法重复运行，则意味着不可靠，而文档记录是实现可重复性的重要保证。

注重道德观念。持续监控人工智能模型的可解释性、公平性和强健性。开发检测算法（道德“机器人”），作为搜索无意偏见及其他问题的虚拟“显微镜”。

不仅要实现规模化运行，还要进行大规模创新。采用并整合深入而强大的自然语言处理能力，以及符合独特用例的其他前瞻性人工智能要素，从而明显提升商业价值。整合各种内部和外部数据源，为“最新尖端”技术分配资源，采用人工智能初创企业的思维方式。

3.5.2.5 加入 AI 联盟

通过与生态系统合作伙伴合作，寻求帮助。考虑与其他企业开展合作，共同制定和/或影响用于治理人工智能模型的相关标准，提高透明度并增进信任。与学术机构、智库、初创企业以及其他值得信赖的第三方开展合作 [90]。

日前，IBM 和 Meta 与全球 50 多个创始成员和协作者宣布成立 **AI 联盟（AI Alliance）** [91]，AI 联盟由来自业界、初创公司、学术界、研究和政府的领先组织构成，共

同支持人工智能领域的开放式创新和开放科学。以行动为导向，具有明确的国际性，旨在通过广泛而多样性的组织在各个领域和地区创造机会，从而在塑造人工智能发展的过程中，能够更好地反映社会的需求与复杂性。更多信息参阅 [92]。

AI 联盟致力于培育一个开放的社区，使开发人员和研究人员能够加速人工智能领域负责任的创新，同时确保科学的严谨性、信任、安全、保障、多样性和经济竞争力。通过汇聚顶尖的开发人员、科学家、学术机构、公司和其他创新者，我们将聚合资源与知识来解决安全问题，同时提供一个平台，共享和开发符合世界各地研究人员、开发人员和采用者需求的解决方案。

随着大模型技术的突破，新一轮人工智能浪潮正在引领各行各业快速发展。数据作为此轮变革的主要驱动力，已成为人工智能发展的关键战略要素。但国内人工智能行业正在面临高质量训练数据供给不足、训练数据治理水平不高、数据供需流通机制不畅等挑战，制约了我国生成式人工智能创新发展。

为破解 AI 数据短缺难题，中国人工智能产业发展联盟（AIIA）成立“数据委员会”。AIIA 数据委员会拟定 2023 年 10 月中旬举办成立仪式，成立后将与人工智能关键技术和应用评测工信部重点实验室、中国通信标准化协会大数据技术标准推进委员会（CCSA TC601）等组织加强协同，共同推动产业研究、标准研制、技术应用等相关工作 [93]。

3.5.2.6 关注行业机遇

随着生成式 AI 与各行各业深度融合，其赋能重构的行业将会持续增加。根据罗兰贝格的评估分析，生成式人工智能将率先对互联网与高科技、金融和专业服务等知识密集型行业带来较大影响，分别带来 6.5%、6.8%、11.3%的成本下降；其次将赋能教育、通信、

医疗、公共服务、零售、文娱和传媒等服务型行业；对当前数字化程度不高的农业、材料、建筑业、能源等传统行业影响相对较小。总的来看，生成式人工智能的价值发挥需要坚实的信息化、数字化支撑，有望在相关行业的研发设计、生产制造、运营管理方面创造巨大价值。

生成式人工智能深入赋能数字经济，为各行业领域带来新一轮发展机遇。伴随着生成式人工智能影响规模的不断扩大，赋能各行各业实现数字化变革与发展。金融业领域，生成式人工智能能够帮助绘制金融风险图，协助打击洗钱等金融犯罪。汽车业领域，生成式人工智能能够提高车载智能语音交互效率，还能为自动驾驶模型训练提供高质量合成数据，帮助解决自动驾驶系统开发过程中的数据和测试难题。更进一步，多模态生成模型正有望加速推动“多模态感知到决策规划”的端到端自动驾驶落地应用。传媒业领域，生成式人工智能可以根据文本提示生成文字、图片、音频、视频等，为广告配上引人入胜的视觉内容。制造业领域，生成式人工智能可以应用于机器视觉、数位分身和自主导航系统等，实现生产线和仓储物流等环节的无人化和智能化。农业领域，生成式人工智能可以通过遥感大模型测量农作物的长势情况，监测作物病害，预测农作物产量。生成式人工智能的进步性价值将持续推动各行业领域质量变革、效率变革、动力变革，推动经济高质量发展。

3.5.3 研究开发和设计

生成式人工智能在各个领域广泛应用，目前在研究开发和设计环节有很多业务场景，为复杂产品的研发以及相关设计职能带来极大的效率提升。生成式人工智能通过代码、图

像自动生成能力，可以提供基础性、重复性的初步设计，提升设计生产效率，缩短研发设计周期。

3.5.3.1 提升研究开发效率

人工智能在研发领域的应用范围非常广泛。例如，在药物研发领域，利用人工智能工具提高候选药物质量、优化临床试验设计、降低临床试验成本和时间。目前一些企业利用大模型赋能药企新药研发，通过与跨学科研究团队合作，推动医疗领域 AI 的可解释性。

3.5.3.2 提升代码开发效率

智能编码助手可以提升生产效率。旨在提升编码效率、减少错误，简化测试用例的编写过程，以及提升软件开发过程的效率和可靠性。例如，根据用户提供的自然语言描述或注释，自动生成相应的代码片段，从而提升编码效率，减少因手动编写代码而产生的错误。根据用户选定的代码片段，自动生成相应的单元测试用例，节省开发人员编写测试用例的时间，确保测试覆盖全面，提升代码质量 [94]。

3.5.4 供应制造和交付

在供应链中规划生成式 AI，需要对数据、AI 和自动化有整体的认识，增强供应链的运营模式，打造智能化的工作流。

3.5.4.1 提升数据可视化

在日益注重可持续发展的世界中，客户期望供应链提供从第一公里到最后一公里的完全透明度。如果能够引入对数据和 AI 管道的良好治理，智能工作流可以让这种可见性成为可能。

但可见数据并不总是等同于可消费的数据。而这就催生了对数据可视化的需求——实际上就是将数据转换为易于理解的格式并进行传递。与 AI 和分析相结合，数据可视化有助于模拟决策影响、预测运营挑战、对前瞻性的新战略进行建模，以及在没有可用历史数据的情况下对选项进行评估，尤其是应对一些前所未有的情形。可视化和模拟已成为最高管理层的关注点——超过一半（52%）的受访高管希望这些模型能够提高预测性运营的透明度和可见性 [95]。

3.5.4.2 实现供应链自动化

随着实时数据推动提高模拟效率和预测分析的准确性，企业可以更轻松地制定未来规划。CEO 正在迅速投资发展生成式 AI，以实现供应链的自动化和简化。事实上，89%的受访高管表示，自动化领域的关键投资将包括生成式 AI。而且，19%的受访高管表示生成式 AI 对于其供应链自动化的未来至关重要 [96]。

数据、AI 和自动化是相互依赖的。可以说，没有数据就没有 AI。而 AI 则是自动化的基础。正是因此，66%的受访者表示，如果没有整合于一体的数据和 AI 战略，其组织的数字化转型计划就无法成功 [97]。

通常，这种整体思维需要超越企业本身。为了提高透明度和可见性，越来越多的企业高管开始将智能工作流与其生态系统合作伙伴整合在一起。事实上，53%的受访高管预计

新兴技术将能够通过这些生态系统和网络数字连接来提高透明度和可见性。与现在相比，超过两倍的受访高管预计，到 2026 年，扩展至生态系统合作伙伴的工作流将通过智能自动化实现数字化。

3.5.4.3 增强供应链运营模式

创建自学习的模拟系统，以便积极识别、可视化并主动纠正关键运营异常。实现事务工作高度自动化，从而提高运营效率 [96]。

首先，要抢占先机。预测并拥抱颠覆。部署分析、数据可视化和仿真模型，以及用于模式识别的生成式 AI 功能。在竞争激烈的形势下，冷静而坚决地采取行动，确保供应链正常运转。

其次，将业务关键型接触点置于首要位置和中心位置。将最关键且最具差异化优势地供应链工作流与早期地预测性生成式 AI 用例进行协同整合。引入关键合作伙伴，通过协作加强预测能力。确保生成式 AI 驱动地工件可清晰识别且可审计。

此外，衡量预先建模地积极影响。定期评估生成式 AI 驱动式预测分析地绩效和投资回报率。设定明确地目标，确保这些工作能达到预期成效，并根据需要进行调整以实现持续改进。

3.5.4.4 打造智能化工作流

企业需要开发敏捷的智能化工作流程以快速应对日益严峻的形势。

首先，在各种计算环境中组装数据以配置工作流，从而支持人工智能和高度自动化。增强人工智能，打造更加智能化的工作流。

其次，管理 API 以在应用之间共享第三方数据源。API 管理可在需要的时间将数据移至所需位置。

再次，建立事件驱动的架构，以便在检测到特定情形时可自动由数据触发工作流。

3.5.5 市场战略和推广

长期以来，企业一直拥有打造高度个性化体验所需的数据。但这些数据存储在多个部门的不同数据集中，而营销团队直到现在仍然缺乏整合和利用这些数据的能力。生成式 AI 推动了高度个性化内容创建和实时数据分析，从而为营销团队提供实现个性化客户沟通所需的动力。

3.5.5.1 引领营销团队

超过四分之一（27%）的受访高管预计，在采用生成式 AI 之后，营销角色将实现自动化。尽管这对营销专业人士来说听起来很可怕，但全球大型广告组织 WPP 的首席执行官 Mark Read 指出，这其中蕴藏着巨大的机遇 [98]。

为了充分发挥其价值，生成式 AI 模型需要访问从营销、销售到服务的整个互动链的客户数据。这意味着，营销团队面临着巨大的增长机会，但也需要拓宽数据隐私与治理视野，以管理品牌风险并维持客户信任。然而，只有 24%的受访 CMO 表示其营销部分正在与销售部门和客户服务部门合作实施生成式 AI。

重新思考营销运营模式，实现更有效的人机偕行关系，让人类专注于更高价值的工作。增强创造力、创新力、战略思维、决策力、产品定位和营销能力可帮助营销团队提升技能并加速学习曲线。当营销团队取得进展之后，CEO 就可以将从中学习到的知识和经验提炼成路线图，以帮助其他智能部门在整个企业中更有效地整合这项技术。

3.5.5.2 专注营销内容

将营销材料与客户旅程中的接触点和关键时刻联系起来，打造更加优质的营销材料。简化内容创作流程，让人类专注于更具价值的工作，从而提高生产力 [98]。

告别写作困境。向团队展示生成式 AI 如何加速内容制作过程。利用基于您组织的数据而定制化大语言模型来协助构思主题、标题、社交帖子以及适合不同受众的消息变体。通过三重检查消除生成式 AI 或人类创作的任何内容中的偏见。

弥合营销内容与客户需求之间的差距。确定在哪些环节需要用内容来推动期望的客户行为和结果，并利用生成式人工智能制作能够缓解客户旅程中特定痛点的作品。

发现主动适应未来工作的人才。密切关注一线人员，探索生成式 AI 所创造的新角色。那些从一开始就接受生成式 AI 的人才将获得洞见、领先的实践和经验教训，这有助于定义未来的 MarOps 模型。

3.5.5.3 实现高度个性化

每一位客户都是独一无二的，但在传统的营销仪表盘中，这些个性化数据会被淹没在聚合数据的海洋中。建立个人关联所需的细节信息都被淹没了 [98]。

生成式 AI 可以将复杂的客户偏好与行为整合为营销人员所需的切实可行的洞察。通过更加迅速、动态地分析来自各种来源的客户数据，营销团队可以了解哪种方案最适合特定客户，并相应地调整外联工作。从个性化内容和体验到定制化聊天机器人支持，生成式 AI 可以帮助团队实时满足客户需求。

潜在应用的清单不断增长，CMO 应当专注于建立强大的分析能力基础，以帮助他们跟上变革的步伐。例如，78%的受访 CMO 预计到 2024 年底将使用生成式 AI 进行数据分析

并从数字 / 社交渠道中获取洞察，而目前这一比例为 36%。86%的受访 CMO 表示预计到 2025 年将使用生成式 AI 来分析客户洞察。

统一的数据将在高度个性化营销中发挥关键作用。为 CMO 赋予对所有接触点（包括销售和服务）的营销技术体系的自主权。建立多学科营销和 IT 团队。协同 CMO 与 CIO 的优先事项，激励两者建立合作伙伴关系。利用生成式 AI 建立真正一对一营销所需的基础架构、系统和数据集成。

全面了解客户的需求。消除职能孤岛，整合来自营销、销售和客户服务的数据，建立客户在贵公司业务中的完整个体旅程全貌。

利用客户数据增强开放模型。将您的客户数据打造为最强大的品牌差异化因素并防范错误信息。同时，利用开放和公共模型的速度和可扩展性优势来打造个性化体验与产品，并持续保护敏感数据。

3.5.6 客户互动和销售

生成式 AI 打造的高度个性化体验有望彻底变革企业与客户及员工之间的交互方式。利用来自销售、营销和服务智能的真实 360 度客户数据，生成式 AI 可以打造定制化体验，并确定“下一步最佳行动”，从而帮助企业吸引特定客户。

3.5.6.1 重新设计客户体验

首先，让同理心成为客户体验的指导设计原则。根据客户的关切点来开发生成式 AI 伦理，赢得客户信任，同时要求生态系统合作伙伴也遵守相同的标准。其次，通过为客户提供值得信赖的体验，从而获得数据回报。持续迭代以改进和个性化产品与服务，从而实现增长和更高的投资回报率，将数据来源转化为数据财富。最后，从客户首次接触品牌开

始，将生成式 AI 融入客户体验中。通过生成式 AI 推动个性化营销活动、定向广告和直接客户外联，并鼓励持续客户反馈 [99]。这种销售数字体验，将在各行各业的企业中被重新定义，个性化定制将成为一项标杆，仅仅提供定制化的互动体验式不够的，还要求体验必须是直观的，能够在用户提出要求之前就满足他们的需求。

因此，生成式 AI 有望提高这些期望，并为企业提供满足这些期望的必要工具。事实上，全球受访高管预计生成式 AI 将成为未来颠覆其组织的体验设计方式的首要趋势。

3.5.6.2 快速分析客户数据

金融服务公司可以使用生成式 AI 来快速分析其客户数据，以及来自社交来源和合作伙伴组织的数据，以确定哪些客户最有可能采取各种行动，从开设新的支票账户、投资资产到申请贷款等。然后，生成式 AI 可以帮助该金融服务企业的高管通过个性化策略和自动化、即时定制的优惠（翻译成客户的首选语言）实现真正的一对一营销。

3.5.6.3 简化在线搜索

在线零售商可以使用生成式 AI 来简化其搜索功能。顾客可以用自然语音（打字或语音）描述自己想要的产品，指定关键细节（例如颜色、尺寸或材料），而不必使用类别和过滤器。他们甚至可以包括预算和期望的交货日期，以进一步细化搜索结果。在这种情况下，顾客不仅可以轻松获得所需的产品，还可为零售商提供有价值的数据，以便用于指导未来的业务决策。

3.5.6.4 提升客户服务

在客户服务领域试点生成式 AI 有助于加速企业范围内的成功部署。在疫情封锁期间，人们清楚看到了客户服务可以实现的自动化水平，但同时缺乏人际接触也带来了一定的损失。借助生成式 AI，组织可以充分结合自动化与人性化的优势。通过将两者相结合，客户服务将成为一个概念验证项目，能够向企业的其他部门展示新技术工具如何提高员工满意度、影响客户参与度以及推动创造回报。利用生成式 AI 改善自动回应的质量和对话能力，可以快速演示如何利用 AI 的影响力来升级组织内部其他领域的服务。对于大多数组织来说，需求和机会都是广阔的。例如，大多数企业表示缺乏审查和重新训练客户服务机器人的能力，只有一半的企业能够在问题出现时主动提醒客户 [99]。

3.5.7 行业应用案例分享

随着生成式 AI 技术的到来，企业对 AI 的应用开启了一个新的篇章，也将迎来新的“黄金时代”。尽管“让 AI 成为核心生产力”已成为企业日益迫切的需求，但实际的落地应用却非一日之功。面对各不相同的应用场景和复杂需求，企业管理者们也产生了诸多的困惑。这里分别分享一个落地的企业专用模型问答系统以及汽车、金融两大行业领域在生成式 AI 的成功经验。

3.5.7.1 企业专用模型问答系统工程化落地

通用大模型对于专用领域的回答准确率通常低，需要构造企业专用大模型来满足准确率的要求。企业专用模型工程化落地主要包含下面几个阶段（详见下图 15 大模型解决方案）：

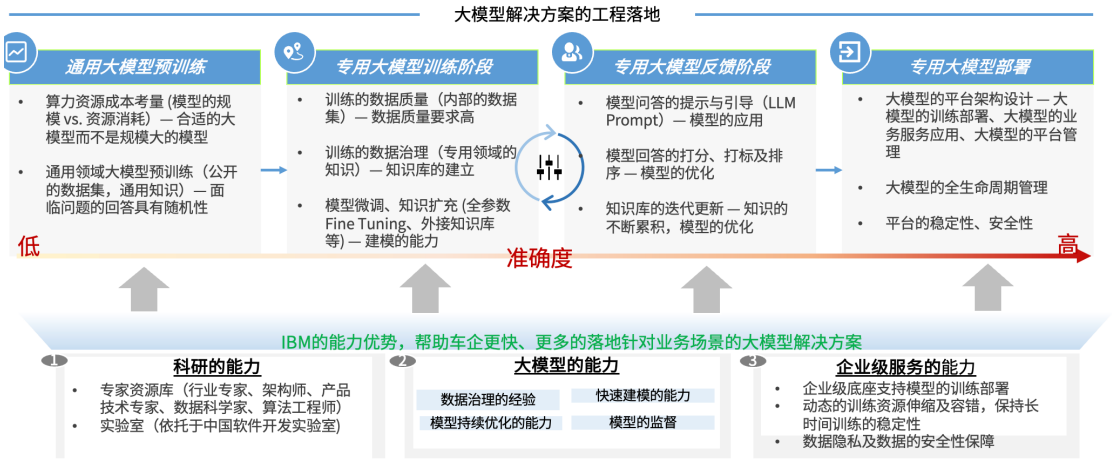


图 15 大模型解决方案

- 选取合适的模型。通用大模型使用了公开的数据集，通用知识，针对专有领域的准确度低，而且通用大模型预训练要花费大量的算力成本，对于企业而言，需要整合考量模型规模和资源消耗，选择合适的模型而不是一味追求规模大的模型。
- 训练专用模型。通常要构建企业的专有模型，企业需要提供高质量的内部数据集，对模型进行训练。这个阶段需要关注数据的质量，数据的治理，建立专有的企业知识库，对专有模型进行微调，知识扩充（可以外接知识库来做增强）。
- 专用大模型反馈阶段。通过给大模型问答提示与引导应用模型，对模型给出的回答进行打分，排序，进一步优化模型，不断更新知识库，随着知识的累积，继续优化模型。
- 专用大模型的部署。在整个过程需要一个整体平台覆盖模型训练，服务服务应用，模型部署等，需要对模型的生命周期进行管理，并且要求平台稳定，安全，可扩展。

在整个工程化落地的过程既需要科研能力，大模型能力，数据管理能力，还需要企业级服务能力。既需要对专业领域非常了解的专家，产品技术专家，数据科学家，算法工程师，架构师等等，可以依托 IBM 研发实验室进行共创。在训练数据部分要提供数据管理和

治理能力，模型训练过程中的快速建模，持续优化和监督能力。在平台部署要支持大模型的部署，有算力和其他硬件资源的支持，平台要有动态资源伸缩能力，保持长时间训练的稳定性。

目前专用模型应用中一个比较典型的案例就是专有智能问答，常见的智能问答系统架构可以参考下图，前端可以用问答交互系统（比如 watsonx Assistant）支持自然语言输入，并可以在里头定制一些问答流程。企业内部的领域知识通过 embedding 模型存储在向量数据库中（比如 watsonx.data）作为提示工程的知识服务。当问答系统把客户问答，以及上下文发给问答路由，到知识库做相似度搜索，把搜索结果作为提示，提示给基础模型（可以部署在 watsonx.ai 上开源模型或自有模型）最终把结果反馈给问答系统。

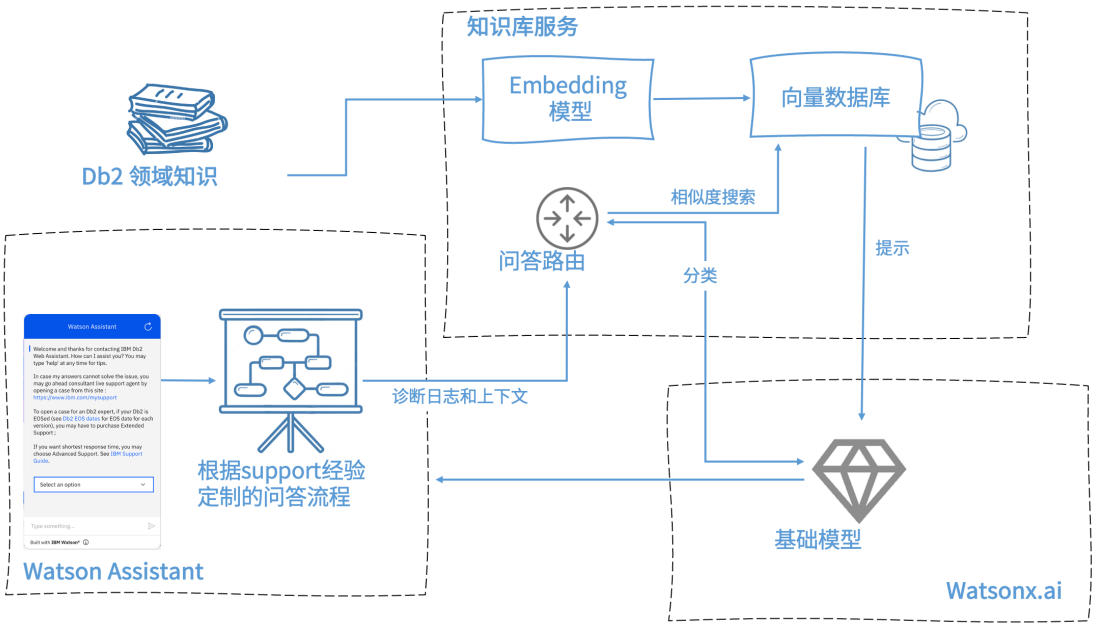


图 16 智能问答架构图

3.5.7.2 汽车领域

随着电动汽车、自动驾驶和先进安全功能等尖端技术的引入，催生了对更复杂、更智能的系统的需求，汽车行业发生了重大变化。在推动这些变革的工具中，生成式 AI 脱颖而

出，成为重塑汽车行业的一股迅速崛起的力量。事实上，生成式 AI 已经对汽车企业产生了明显的影响，可以在业务开发、生产管理、财务管理、业务管理、供应链、市场销售、服务售后等多个方面发挥积极作用。具体案例可参见章节 6.1。

图 17 汽车行业在生成式 AI 中的应用

3.5.7.3 金融领域

金融行业在使用生成式 AI 的时候，可以在业务开发与管理、渠道管理、客户互动、生产管理、运营支持、财务风险管理以及资源管理各方面得到很好的应用。具体案例可参见章节 6.1。

图 18 金融行业在生成式 AI 中的应用

四　生成式人工智能治理

4.1 生成式人工智能治理框架

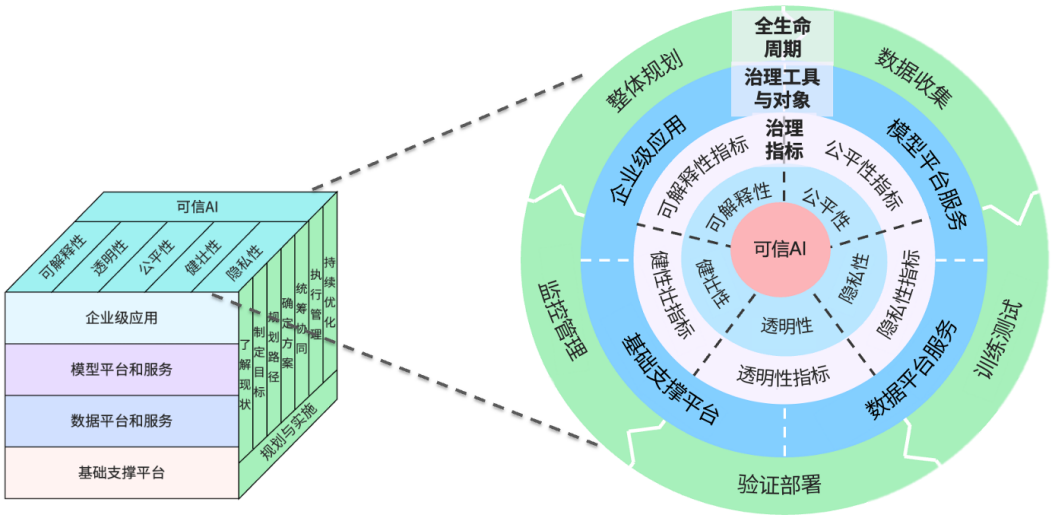


图 19 生成式人工智能治理框架

本框架的核心在于打造可信赖的 AI 系统，面向 AI 全生命周期的全时段，贯穿技术栈的全方位治理，围绕着五个关键特征，即可解释性、公平性、透明性、健壮性和隐私性展开[100]。

- 可解释性：AI 系统模型做出决策或预测的依据，这些解释应该可供具有专业知识和能力的人和公众所理解。在技术上可以通过知识工程，AIX360，数据地图，数据标准，元数据管理等方式提高模型和数据的可解释性。
- 透明性：AI 系统的相关数据（包括原始数据和使用过程中产生的元数据）应作为信息披露的内容，如出现在产品说明中或供审计使用。在技术上可以通过知识工程，模型生命周期，模型可视化，数据生命周期，数据地图，平台的可观测性等实现 AI 系统的透明性。

- 公平性：AI 系统应确保决策过程和结果不歧视任何个人或群体，其表现应与统计学规律以及业务内容相吻合，对所有用户均公平公正。例如，贷款审批模型对信用不良的人产生的“偏见”是合理的。在技术上可以通过偏见预防与检测，数据质量管控等来提高 AI 系统的公平性。
- 健壮性：AI 系统应对外界变化和潜在的攻击有抵抗力，能够稳定地运行，有效处理异常情况和蓄意的对抗攻击，降低安全风险。在技术上可以通过对抗攻击的检测与预防，数据质量管控，平台级的能源规划等来保障 AI 系统的健壮性。
- 隐私性：系统必须保护个人隐私，确保用户数据的收集、处理和存储安全，且遵循相关的隐私保护法律法规。在技术上可以通过联邦学习，多方安全计算，匿名化，差异隐私，数据脱敏，数据安全等技术来保障 AI 系统的隐私性。

基于这五个特征，本章节将会讨论如何将治理与 AI 全生命周期相结合，介绍在不同技术层面的相关技术手段和工具，通过引入对应的评估技术和一系列量化指标矩阵，从而确保在企业引入生成式人工智能的全生命周期内满足这些标准，帮助企业实现和维护高标准的治理水平。

4.2 融入 AI 全生命周期

Al 的治理不是一次性的任务，而是贯穿 AI 从引入、开发、部署到维护全过程的持续活动。这要求企业在 AI 系统的每个领域都实施相应的治理措施，以缓解企业在与 AI 系统协作时可能引起的风险，确保 AI 系统遵守所有相关法律和行业标准，保证 AI 系统的稳定性和可靠性，将 AI 战略与企业的业务目标对齐。以 IBM 的 Ethics by Design and the AI Lifecycle[101]和相关文章 [102]为例，通常将治理融入 AI 生命周期包括以下步骤：

图 20 将治理融入 AI 全生命周期

- 整体规划：设计 AI 治理的总体规划，分配治理职责，将治理和业务指标相统一，建立评估方式以及采纳或建立必要指标体系，手段。此阶段通常明确以下问题：
    - 企业生成式人工智能业务需要遵守哪些法律法规？
    - 企业生成式人工智能业务所涉及的数据需要遵守哪些隐私保护要求？
- 数据收集：获取训练数据，对数据进行探索性分析，创建相关数据和索引，数据脱敏，元数据管理，数据清洗，数据质量分析等。此阶段可以帮助识别明显数据的错误，有助于理解数据中的模式，检测异常值或异常事件，并找到数据之间的关系。需要注意的是，此过程需要遵循隐私数据保护相关的法律法规要求，如数据中涉及个人身份信息的部分需考虑匿名化处理。
- 训练测试：在此阶段团队将数据训练为模型，并对模型进行涵盖治理相关内容的测试，评估。MLOps 流水线的方式可以有助于将此过程自动化提高效率。以公平性治理为例，在模型构建工作开始前，可以进行数据偏差相关检验。相类似的，稳健性，可解释性治理等相关特性的检测工作也可以在此阶段完成。
- 验证部署：在模型正式部署到生产环境之前，需要验证其质量，验证其公平性、透明度、可解释性、稳健性和隐私性并生成相关报告，团队必须考虑是否适用于任何

监管或非监管要求。包括但不限于发布情况说明书、偏见结果、隐私声明或发布法律声明。如果模型通过验证，则将其部署到生产环境。

- 监控管理：基于在生产环境上收集到的相关指标数据，评估生产环境上模型的质量，公平性、透明度、可解释性、稳健性和隐私性等功能和非功能指标，如记录模型偏移情况。根据收集数据进行分析，如果监测到相关指标出现异常，采取对应行动，包括但不限于告警，人工纠正，重新训练模型等等。通过持续的优化行为以保证模型可信。

4.3 生成式人工智能模型治理技术

4.3.1 模型可解释

正如人工智能如何走向新阶段中提到的，当前需要打破 AI 的黑盒状态 [103]。现阶段，用一个可解释模型/算法来解释 AI 模型是一种尝试打破 AI 黑盒状态的手段。AI Explainability 360 项目 [104]构建了基于不同业务场景，范围（全局可解释，局部可解释），对象（数据，模型），数据格式（文本，视频）角度进行可解释算法选择的决策树 [105]。遗憾的是，目前这棵决策树上仍存在空白，这和业界对于如何打破 AI 的黑盒还处于起步阶段的状态一致，目前可以分为以下类别：

- 数据可解释：我们可以通过各种算法如（DIP-VAE，ProtoDash 等），来加强人们对于特征值或样本数据的理解。
- 模型可解释：

- 模型全局直接可解释：对于了解整个决策过程并保证其安全，可靠合规非常重要。通常可以用来处理决策树，布尔规则和广义线性回归等模型。
- 模型训练后全局直接可解释：通过后建一个解释模型的方式，在黑盒模型训练之后解释黑盒模型，以 ProtoDash 算法为例，可以更好的帮助人们建立黑盒模型决策结果和原始数据之间的联系，为本次模型决策的结果找到历史数据中的参考。
- 模型局部可解释：此类算法更多的侧重于对某一条具体数据的解释。如找到某条数据未能通过审核的依据。以 CEM 算法为例，此类算法更多的展现了某个样本的特征值对决策过程的影响。可以用于模型样本局部可解释或模型特征值局部可解释。

4.3.2 知识工程

参考章节 3.2.5，通过知识工程可以部分打开大模型的黑盒，大语言模型存在着局限性，例如幻觉问题，知识新鲜度问题，以及数据安全问题。为了解决这些问题，RAG（Retrieval Augmented Generation）检索增强生成技术成为很多企业的首选，通过这种架构，模型可以从外部知识库搜寻相关信息，然后使用这些信息来生成回应。具体的做法是把私域知识文档进行切片然后向量化后存储在向量数据库中，然后通过向量检索的方式找到最近似的结果，再将其作为上下文输入到大语言模型进行归纳总结。知识图谱，图数据库等技术也可以很好的反应实体和数据的关系，来作为输出的依据，提高可解释性。

4.3.3 模型可视化

对于训练结果，模型可视化可以增强（机器学习）模型的透明性，由于（机器学习）模型常常被视为“黑盒”，内部的工作机制对于最终用户不透明，这增加了在实际应用中建立对这些模型信任的难度。因此通过可视化手段增强模型的信任度，可视化可以揭示模型如何从输入到输出的具体处理过程，包括原始数据的质量和来源，数据的标注与特征工程，学习方法和算法，模型训练，以及模型的评估等，这种透明性帮助用户理解模型是如何工作的，通过不同阶段的透明性增加用户对模型的信任 [106]。

4.3.4 防范对抗攻击风险

模型的健壮性，主要体现在模型防范对抗攻击的能力，对抗攻击经常发生在模型分类的边界处，基于对抗攻击的原理，已知对数据和模型存在规避，投毒，推理和反演，模型提取等在内的多种安全威胁。通常，对于对抗攻击，我们要建立指标和验证机制来将专业领域的知识转化落地成为对于对抗攻击的预防机制。在 Adversarial Robustness 360 项目中收纳对抗攻击类型的列表、指标、验证标准等相关信息并对防御机制进行了开源技术实现 [107]。对抗攻击的预防可以根据采取措施的阶段大致分为预处理防御、后处理防御、训练防御、转换器防御方式。

4.3.5 模型公平

公平性算法，可以根据其在生命周期中的位置分为预处理，过程干预，后处理三大类，每个大类又涵盖了不同算法的实现。一般而言，预处理类型作用于原始数据，过程干

预作用于训练过程，后处理基于黑盒模型且无法修改数据或学习算法的情况，在不同位置对公平性进行处理有各自的优势和不足 [108]。

- 预处理算法：这类方法可以通过调整样本权重的方式生成一个新的数据集来同时解决群体公平和个体公平问题，但需要注意的是，由于偏见存在的方式可能会比较复杂，因此可能会影响转化后的数据集的质量和公平性。
- 处理中算法：通过在训练过程中添加正则化感知等技术来影响训练算法中的损失函数从而处理偏见。
- 后处理算法：对于只能访问黑盒模型的情况，只能采用后处理算法。并且可以避免对模型的二次训练，在实际过程中，需要考虑这类算法对模型结果准确性的影响。

4.3.6 隐私保护技术

从数据隐私的角度，我们看到了多种实现方式，如基于密码学同态加密，联邦学习，差异隐私，或数据匿名化与模型匿名化等手段。基于不同的密码学技术的能力和手段在 AI 生命周期的不同环节在进行数据隐私保护处理的同时完成了模型训练等任务。在实际使用中可以参考 AI Privacy 360 项目中建立的据隐私安全评估流程，和工具选择流程图。我们需要结合数据隐私保护的实际情况，找到合适的方式实现对于模型和数据隐私保护 [109]。

差异隐私：通过数学能力如随机噪声来保护个人隐私的同时保持数据的统计的准确性。但考虑到不同实现上的区别，有些实现方式可能很难与其他维度的算法同时生效。

匿名化：通过创建用于模型的定制匿名化方案，在使用训练数据训练模型之前对数据进行匿名化的方式来提高模型的隐私能力，但是这一过程中选择的标识符可能会影响到模型的识别能力。

同态加密：基于同态加密对密文的运算能力，通过对加密数据实施不同的分析和模型解决方案。

联邦学习/多方安全计算：各方通过协作训练模型的方式减少数据贡献和交互，增强了数据的隐私性。通常，该方法也会和其他方法如差分隐私，同态加密，多方安全计算相互结合。

4.3.7 模型漂移

模型漂移[110]是指由于数据变化或输入和输出变量之间关系的变化而导致模型性能下降。模型漂移会显著影响模型质量，随着时间的推移和漂移的积累，原本无偏见的模型可能会产生偏见。如果构建模型使用历史的数据和现行业务数据存在过大偏差，历史数据的模型可能无法正确对现行业务数据进行预测或判断，此时可解释性技术也无法生效。以下是漂移的典型情况：

- 元数据漂移 - Metadata Drift

当数据的元数据发生变化时，会发生这种漂移。元数据包含有关数据的上下文信息，如数据的架构、标准或类型定义。元数据的变化可能包括值范围的改变、新类别的引入或数据格式的变更，这些都可能影响模型的表现。应该持续监控和探查模型的输入和输出数据在这些层面的漂移，并定期评估模型表现是否随之发生变化。

- 上下文漂移 - Context Drift

收集数据的条件或它适用的环境也可能发生变化。即使数据本身没有变化，周围环境或适用场景的变化也可能使基于该数据的模型准确性降低。例如由于外部因素，市场条件或用户行为的变化可能导致上下文漂移。

- 置信度漂移 - Confidence Drift

这种漂移涉及模型随时间对其预测的置信度的变化。这可能是由于输入数据的变化或模型已学习的变量之间关系的变化。当模型的性能指标（如准确性或精确度）开始下降时，通常可以检测到置信度漂移。

- 数据分布漂移 - Distribution Drift

指输入数据的统计属性随时间发生变化。这种变化可以显著影响模型的性能，因为模型训练时的假设不再适用。例如，输入特征的均值或方差的变化或分类问题中类别比例的变化都是分布漂移的指标。

4.4 生成式人工智能模型治理工具

4.4.1 开源项目实现参考

4.4.1.1 模型可解释性 - AIX360

AI Explainability 360 [111]工具包通过多种方法和算法支持 AI 模型的可解释性。它包括直接可解释方法和事后解释方法以及相关评估指标，这些方法可以是局部的也可以是全局的。这些工具适用于不同的用户角色，从监管者到最终用户，为决策和合规提供适当的解释。此工具包为高风险应用设计，强调模型的透明性、可解释性以及消费者对 AI 决策的理解。它还提供了丰富的教程和资源，帮助开发者实施和理解这些方法。

4.4.1.2 模型公平性 - AIF360

AI Fairness 360（AIF360）[108]提供了一套综合工具，用于识别和缓解机器学习模型中的偏见。这包括各种数据和模型的公平性度量方法，以及用于减少数据集和模型偏见的算法。AIF360 支持对训练数据和模型进行公平性评估，并提供预处理、过程中处理和后处理的偏见缓解策略。工具集还包括教程和示例，帮助开发者理解和应用这些方法。相关案例可参见人工智能安全标准化白皮书（2019 版）附录 B.5 IBM 人工智能安全实践 [112]。

4.4.1.3 模型健壮性 - Adversarial Robustness Toolbox

AI 对抗健壮性工具包（ART）[107]提供了一套全面的工具和方法，用于增强机器学习模型抵抗对抗攻击的能力。ART 支持针对各种机器学习框架和任务类型的攻击方法，包括欺骗、数据投毒、模型提取和推断攻击。它不仅包括各种攻击技术，还提供了防御机制，如预处理、后处理、训练期间的防御和检测技术，以及对抗训练方法。此外，ART 也提供了一系列开发者教程，帮助开发者更好地了解和使用这些工具。

4.4.1.4 AI 模型全生命周期管理 - AI Factsheets

AI 模型全生命周期管理涉及从设计、开发到部署和维护阶段的全方位管理内容。所以需要一个系统以结构化的方式收集、记录和报告这些 AI 模型的关键信息和元数据。这些信息包括但不限于模型的训练意图、业务标签、训练数据、模型版本、性能指标、公平性和健壮性评估结果等，以帮助开发者、用户和监管者全面理解模型的行为和限制。AI Factsheets [113]项目旨在实现这些目标，增加 AI 模型的透明性，提升用户对 AI 系统的信

任，并应对法规和监管要求。同时 AI Factsheets 也有助于在整个开发和部署过程中实现更好的决策和监控。

4.4.1.5 模型开源项目 - InstructLab

大模型开源和开放通常围绕训练数据和训练过程，对于训练数据和过程，以 InstructLab 项目[114]为例，在实现模型通过 Hugging Face 平台开源共享的同时，公开模型训练所使用的数据集合。其背后通过 Large-Scale Alignment for ChatBots[115]技术，实现了模型训练所使用数据的公开透明。该项目基于 Apache 许可证发布的策略也允许使用者根据私有数据调整的模型，具备良好的商业兼容性。IBM 在博客中[116]分享了通过对这项技术的应用所取得的成果：

- 通过 IBM watsonx 和该技术显著改进了 Granite 模型。
- 有效提高了模型可解释性。
- 缓解了 GPT-4 等专有 LLM 生成合成数据的合法性风险。
- 出色的对齐数据可以为更小、更具成本效益的模型带来高级功能，以根据企业需求进行定制。

4.4.1.6 模型评估 – Hugging Face Evaluate

Hugging Face Evaluate[117]旨在为不同领域的机器学习和深度学习模型（如自然语言处理、计算机视觉、强化学习等）提供简单而一致的评估方式和评估指标。它允许用户通过一行代码在本地或分布式训练任务中评估模型，确保评估过程的一致性和可复制性。Hugging Face Evaluate 库支持广泛的评估指标，覆盖了多个机器学习和深度学习领域。

这些指标包括但不限于文本生成的准确性（如精确匹配）、分类任务的混淆矩阵、语言模型的困惑度（perplexity），以及自然语言处理任务的 ROUGE 和 BLEU 分数等。

4.4.2 基于开源的商业产品实现 – watsonx. governance

以 IBM watsonx. governance 平台为例，该平台集成了 IBM 现有的一些产品能力以及开源实现，如 AI Factsheets，除上文介绍的部分开源技术和工具之外，还包括以下部分：

4.4.2.1 AI 模型监控和评估 - IBM Watson OpenScale

AI 治理的核心任务是确保人工智能系统的开发、部署和使用过程中的透明度、安全性、公平性等，以实现可信 AI。这需要对 AI 模型的训练和部署后的表现进行持续的监控，并基于可信 AI 各维度的指标对数据和模型进行持续评估。附录二　将会详细介绍这些指标。IBM Watson OpenScale [118]产品提供了传统 AI 模型和 LLM 的可信 AI 监控和评估能力。 通过抓取模型在线调用的交易数据，实时评估模型的质量、数据和模型的漂移、各个特征和特征子组的公平性和可解释性，以此帮助用户持续提升 AI 模型的可信度。

4.4.2.2 AI 模型风险管理 - IBM OpenPages

AI 模型治理的另一重要维度是 AI 模型的风险管理。其重点在检查和报告模型的运行是否符合行业标准和法规要求。IBM OpenPages [119]产品可以帮助用户映射政策、度量和模型至多个监管要求，如欧洲 AI 法案，GDPR 等，并支持跨法规的模型风险评估。同时提供流程引擎和预警策略，满足风险事件的快速响应和审计需要。

4.5 生成式人工智能数据治理

数据治理主要包括主数据管理，元数据管理，数据质量，数据标准，数据安全与隐私保护，数据地图和数据生命周期等核心功能。数据治理在数据全生命周期中发挥着重要的作用，经过治理后的数据提供给消费者，才能最大化数据产生价值。详见图 21 数据治理功能图。

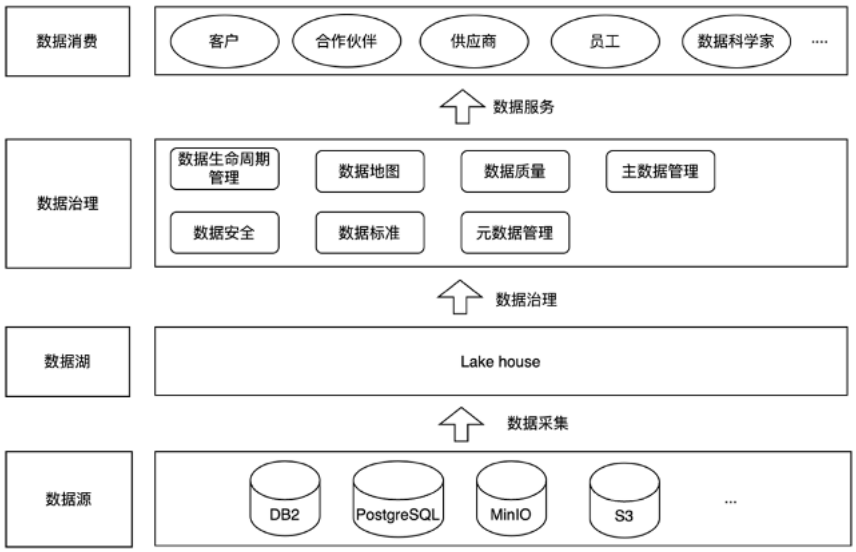


图 21 数据治理功能图

4.5.1 主数据管理

2018 年中国信通院牵头编写的《主数据管理实践白皮书（1.0 版）》[120]，主数据的定义如下：“指满足跨部门业务协同需要的、反映核心业务实体状态属性的组织机构的基础信息。” 从定义中可以看出主数据是企业中跨部门共享的核心业务数据，通过主数据管理，保证主数据的共享性、稳定性和可持续扩展性，解决企业在不同系统中存在的数据孤岛问题，降低沟通成本，提升跨部门协作能力。

实现主数据管理目标主要包括以下几点[121]：

- 建立组织机构。主数据的管理不仅仅是一个技术问题，由于主数据涉及的业务部门，业务流程繁多，主数据的管理需要各部门达成共识，共同推进，建立组织机构可以有效的推进主数据管理的执行过程。同时实现主数据管理往往需要得到高层领导的足够重视和授权。
- 制定主数据管理标准。只有建立统一的标准化数据模型，才能实现跨部门，跨业务流程的数据集成和共享。建立完善的主数据实施框架。主要包括系统现状的分析与评估，明确主数据实现目标，指定主数据实施方案等。

4.5.2 元数据管理

元数据，或称为“数据的数据”，是用来描述数据的信息，它提供了对数据的详细描述，使得数据的理解、使用和管理变得更加高效。元数据不仅记录了数据的基本信息，如数据的来源、格式和质量等，还包括了数据的结构、规则和约束等更深层次的信息。与元数据相对的是数据本身，即元数据所描述的对象，它可以是文本、图像、视频等任何形式的内容 [122]。

数据和元数据之间的关系可以类比于图书和图书目录之间的关系。就像图书目录通过记录图书的标题、作者、出版社等信息来帮助人们找到和了解图书一样，元数据通过记录数据的相关信息来帮助用户找到、理解和有效使用数据。

元数据有很多分类方式，一种被广泛接受的分类是：业务元数据，技术元数据和操作元数据。

业务元数据：业务元数据主要描述了数据在业务过程中的含义和用途，它关注数据如何支持业务活动。例如，业务元数据可以解释一个数据字段表示的是客户的姓名还是客户的账号，帮助业务人员理解数据的业务含义。

技术元数据：技术元数据关注的是数据的技术层面的描述，包括数据的结构、格式、存储位置等信息。它主要被数据工程师和 IT 专业人员使用，以支持数据的集成、管理和维护工作。例如，技术元数据可以描述一个数据库表的结构，包括表中的字段名、字段类型和约束等。

操作元数据：操作元数据记录了数据的操作历史和状态信息，如数据的创建时间、最后更新时间、访问记录等。它对于监控数据的质量、审计和遵从性管理非常重要。操作元数据使得组织能够追踪数据的生命周期，确保数据的准确性和可靠性。

元数据存储库是用于存储和管理元数据的系统。它不仅存储元数据本身，还支持元数据的搜索、查询、更新和维护等功能。通过元数据存储库，组织可以有效地管理其数据资产的元数据，提高数据的可发现性和可用性。

元数据管理的核心是要对元数据进行规划、控制和监督的过程，以确保元数据的质量和有效性。一个有效的元数据管理策略可以帮助组织实现更好的数据治理，提高数据的一致性、透明度和可信度。元数据管理通常包括元数据的收集、存储、维护、共享和使用等方面。

4.5.3 数据质量

数据质量在人工智能实施中至关重要的原因有很多，错误倾斜的数据会影响模型的结果。并不是所有的数据都是平等，有些数据在整个模型训练中占据更重要的地位，有些数

据更多的是辅助和补充，含金量并不高。所以对数据进行剖析，评估数据的质量的重要性不言而喻。数据质量的评估标准主要有 6 个方面，包括：[123]

- 准确性：准确性是指数据记录的信息是否存在异常或错误，是评估数据质量的首要标准。
- 一致性：一致性是指数据是否遵循了统一的规范，数据集合是否保持了统一的格式。
- 完整性：完整性是指数据是否存在缺失的情况，数据缺失的情况可能是整个数据记录缺失，也可能是数据中某个字段信息缺失，不完整的数据可能导致数据的错误倾斜，只有完整的数据才是有意义的。
- 及时性：及时性是指数据从产生到可以查看的时间间隔，也叫数据的延时时长。过时的数据可能会影响数据分析结果的准确性和可靠度，数据分析师需要定期检查数据并及时更新。
- 有效性：有效性是指数据要符合相关行业的业务规则。如银行卡、电话、邮箱的格式等。
- 唯一性：唯一性是指针对某个数据项或某组数据，没有重复的数据记录。

数据质量管理是从数据产生到消亡整个生命周期的管理，数据质量管理的目标是通过可靠的数据提升数据价值，帮助企业开展业务和获取更多的经济利益。

4.5.4 数据标准管理

数据标准是指企业为保障数据的内外部使用和交换的一致性和准确性的规范性约束。制定数据标准，实现数据标准化是开展数据治理的基础。

企业的数据标准来源非常丰富，制定数据标准之前，我们通常需要考虑企业本身制定数据标准的需求，为什么要制定数据标准，结合企业内部的实际情况，同时也要考虑外部行业的监管需求，是否有国家相关标准参考，结合之上的这些需求规划数据标准的制定。

DCMM 描述的数据标准包括 [121]：

- 业务术语：业务术语标准化保证企业内跨部门人员对某一具体技术名词理解的一致性，提高协同工作效率和沟通准确性。
- 参考数据和主数据：参考数据是对其他数据进行分类和规范的数据，主数据是跨部门之间共享的数据，参考数据和主数据标准化，可以有效的提高数据质量和数据可用性。
- 数据元和指标数据：数据元是指表示数据的最小单元，数据元标准化可以提高数据的一致性和准确性。指标数据通常用于统计分析，为管理层决策提供参考，指标数据标准化提高了决策的准确性。

4.5.5 数据安全

数据安全体系大致分为三个方面：数据安全战略，数据全生命周期安全，基础建设安全。

数据全生命周期安全包含采集，传输，存储，数据共享与使用安全，销毁安全等等。产品在实现时需要制定数据保护策略和数据保护规则，并根据这些策略和规则对数据进行脱敏，实现数据安全共享。

常见的数据保护策略包括 [121]：

- 数据分类分级：对数据进行分类并根据其敏感性和重要性设定安全级别。

- 数据访问控制：对数据访问设置权限控制，仅允许授权的用户能够访问相应的数据。
- 数据加密：对数据进行加密，将数据转化成密文显示或者存储，只有授权的用户才能查看原始数据。常见的数据匿名化技术包括：数据脱敏，泛化，数据置换，数据替换。
- 数据合规性合法性检查：确保数据在全生命周期中遵守相关法律法规和标准，参考 1.3 法律法规部分。
- 备份和灾难恢复：企业应制定合理的备份策略，定期对数据进行备份，进行异地备份等，同时还需要进行灾难恢复测试，以保证备份数据的可用性和备份恢复的及时性。对数据进行备份是保障数据安全和防止数据丢失的重要措施。

4.5.6 数据地图

亚信科技的数据事业部总经理高伟在《数据资产管理》一书中提到： 数据地图是一种图形化的数据资产管理工具，它提供了多层次的图形化展示，并具备各种力度控制能力，满足业务使用、数据管理、开发运维不同应用场景的图形查询和辅助分析需求。由此可以见，数据地图主要解决取数据和用数据的两大难题。

数据地图其实还衍生出三个非常重要的应用：全链路分析、血缘分析和影响分析。

全链路分析可以查找某个对象上下游所有数据链路的关系，用户能够清晰的看到数据从哪里来，被用到哪里。

血缘分析通过向上溯源，从而找到以某个数据对象为起点的所有相关元数据对象以及这些元数据对象之间的关系。通过血缘分析，可以帮助用户快速定位问题，进行差异化分析，指标波动分析等。

影响分析通过向下挖掘，找到下游关联的所有元数据，反应数据的流向和加工过程。通过影响分析，当用户想修改一张基础表时，可以快速定位这张表关联的所有下游表。

4.5.7 数据生命周期

数据生命周期管理是指对数据在整个生命周期过程中的管理。从静态看数据存在生成，活动，衰退，归档和销毁各个阶段，从动态看数据存在数据采集，存储，处理，交换，销毁等阶段。数据生命周期管理根据数据生成的不同阶段管理数据，确保数据在各个阶段的完整性和准确性，提高数据质量。数据生命周期管理大致分为 5 个阶段 [124]：

- 数据创建：大数据时代，数据的来源异常丰富，但并非所有数据对于企业都是必不可少的，必须评估数据与企业的相关性和价值，制定相关的收集策略。
- 数据存储：收集后的数据经过清洗和整理，需要根据不同的数据集选取不同的存储类型，还需要充分评估该存放方式的基础架构是否存在任何安全漏洞，以及是否可对数据进行各种不同类型的处理，例如数据加密和数据转换，同时经过整理后的数据还可确保敏感数据遵守隐私和政府政策的要求，如 GDPR。
- 数据共享与使用：数据经过处理以后，企业可以分享给组织内部或者外部的利益相关者。在数据共享时，可以采取相应的数据保护策略，以防止未经授权的数据被访问或者隐私数据泄漏。在数据使用过程中，企业还需要制定数据备份和恢复策略，防止数据丢失和灾难恢复。

- 数据归档：需要定义何时归档，归档到何处以及归档多长时间等等。
- 数据删除：当数据不在使用时，企业需要对数据进行安全销毁。

4.5.8 数据治理工具

4.5.8.1 开源实现

- Amundsen，托管在 LF AI and Data Foundation 的开源项目，主要包含元数据管理，通过结合 Neo4j 或者 Apache Atlas 提供元数据，支持从很多现有关系型数据库中以及湖仓引擎中抽取元数据，没有 RBAC 支持 [125]。
- DataHub，由 LinkedIn 和 Acryl Data 开源的项目，包含元数据管理、数据发现、数据血缘、数据质量控制，支持细粒度的数据访问控制，支持 RBAC，允许对元数据进行细粒度访问控制 [126]。
- OpenMetadata 提供了统一元数据发现、数据血缘、数据治理、数据质量、定义术语以及人员协同的开源平台，支持现有大部分关系型数据库、湖仓引擎甚至报表系统集成。其数据安全主要是各种 OAuth 认证 [127]。
- Apache Altas，支持元数据管理、数据发现、数据分类、数据血缘，通过跟 Apache Ranger 集成可以完成数据脱敏 [128]。

4.5.8.2 IBM 数据治理工具

IBM 提供了企业级数据安全治理方案 IBM Knowledge Catalog，支持制定统一的数据策略，对多种数据源的元数据统一管理，多维度的数据质量分析，实现企业级数据安全管控方案。在 IBM Knowledge Catalog 内部，内置了多样化的自动化的数据管理功能，使其

在增强数据治理方面发挥着关键的作用。其中包括：业务术语、定义数据类、业务和技术血缘、数据地图、数据质量规则、数据保护策略和规则、数据脱敏。尤其是其中提供了行业加速器、行业标准的术语库、数据类库和规则库。方便用户基于开箱即用的行业资产快速客制化，构建私有的行业数据目录。同时提供 GDPR 等法规相关术语和规则库，满足数据合规使用和数据安全的需求。IBM Knowledge Catalog 同时提供给了丰富的界面支持，可以很好的展示上下游数据加工和转换链路。在数据地图部分人工智能自动构建资产知识图谱，可以全面、直观的图形化展示系统、业务域、资产元数据的全貌。这些补充能力可以有效加速企业数据目录构建和相关数据治理工作的开展。借助自动化和人工智能能力，IBM Knowledge Catalog 不仅可以简化元数据存储库的建设流程，构建数据质量和安全管理的闭环，还能够加速业务团队获得更全面的、可靠的数据，使业务分析师和数据科学家能够无缝地利用它们开发规范性人工智能模型和生成人工智能模型框架。

4.6 生成式人工智能在基础支撑平台治理的新趋势

基础支撑平台通常可以通过硬件（如防火墙）和软件（如混合云平台通常包括的命名空间，控制组，身份验证，组角色等）实现多种安全策略管理机制。这些安全机制共同协作可以更好的应对生成式人工智能的风险（章节 1.3.1），尤其是在数据隐私保护，网络安全，知识产权保护等方面。（图 13 基础支撑平台概览）

随着生成式人工智能应用的落地与实践，我们收集到一些最新的研究成果和行业趋势资讯，这些信息反映了业界随着生成式人工智能应用在基础支撑平台部署实施所遇到的新问题和新趋势。

4.6.1 可观测性技术

通过结合云原生可观测性技术，从而实现贯穿应用层，AI 层，数据层的端到端的跟踪，以帮助理解和追踪模型参数，性能指标，业务 KPI 在内的数据变化，并进一步将这些观测到的结果融入模型改进的反馈循环中，参考 IBM Instana 团队的相关博客 [129]介绍了通过将可观测性技术和生成式人工智能相结合以实现将治理融入 AI 生命周期中监控管理的相关实践。

4.6.2 可持续生成式人工智能

正如章节 2.2.2.3 中提到的，生成式人工智能需要大量的能源资源。对数据中心而言能源的总带宽是有限的，如何在有限的能源带宽下，更好的规划和部署生成式人工智能应用，从而达到最优的服务可靠性，成为了一项新的挑战。根据国际能源署（IEA）的预测，数据中心是电力需求增长的重要驱动力，到 2026 年人工智能相关电力的增常需求会大幅增长 [130]。混合云所具有的灵活性，如支持多云部署和公有云私有云部署相结合的灵活部署方式，可以在成本规划方面提供更加灵活和丰富的选项。关于这方面的研究尚处于起步阶段，如 Kepler 项目可以指定负载的测量 GPU 的能耗 [131]为后续优化提供数据支持，CNCF TAG Environmental Sustainability [132] 正在开始和推进这方面的开源合作。同时人们发现，应用专业的人工智能技术平台作为基础指导企业应用级别的可持续计算，能够实现能源高效调度和利用，提升能源接入系统的可靠性和高质性 [133]。

4.7 生成式人工智能治理的指标矩阵

指标表参见附录二　人工智能指标。

4.8 生成式人工智能治理的小结与展望

生成式人工智能是一把"双刃剑",为我们带来巨大机遇的同时也存在不可忽视的风险。本框架为负责任地开发和使用该技术提供了指导原则和建议,但仍有待不断完善。我们呼吁所有利益相关方参与进来,共同推动生成式人工智能的健康、可持续发展。

五 企业级生成式人工智能的规划与实施方法

随着人工智能不断发展，企业高管们正在竭力应对生成人工智能对企业的影响。生成人工智能将很大可能颠覆传统业务模式，并改变员工的日常工作方式。根据 IBM 商业价值研究院 (IBM IBV) 最近开展的一项调研，五分之四的受访高管表示生成式 AI 将改变员工的角色和技能，正在合作或计划利用基础模型并采用生成人工智能。它拥有预测分析、机器学习和其他人工智能技术的能力，可以自动化重复和繁琐的任务和流程，并不断创新[134]。

在生成人工智能突然间成为瞩目焦点的同时，但只有 10%的组织已成功地在多个业务单位和流程中部署了人工智能。对于许多企业来说，在采用生成人工智能时，由于技术种类繁多，单个业务场景可能会涉及到多种技术，而多个项目间可能会共用某项核心技术，每个企业必须平衡使用这项强大的技术的投入以及其创造的价值，它需要全企业级别的政策、做法和指导方针帮助企业从技术角度明确未来主要的技术发展方向，同时能支持企业更有目的性的选择合作伙伴。

但是目前大多数企业缺乏全企业级别的生成人工智能的战略规划，政策、做法和指导方针。许多企业在实施相关项目时，并没有事先进行统筹规划，导致错误选择技术方案，项目无法落地，范围蔓延，管理复杂难度高，运营成本高昂，没有专业团队等问题，最终导致项目失败或者无法产生价值。

所以企业在制定合理、全面、可实施的人工智能战略时，重点可考虑以下要素：

- 根据企业自身现状，竞争优势以及未来战略规划设定总体人工智能应用的目标，成功衡量指标，制定相关规划以满足未来发展需求。

- 衡量人工智能对业务的影响，深入分析和流程优化，评估商业价值，在主动采纳新兴技术与商业价值之间获取平衡。
- 无论自建还是与合作伙伴合作，需要考虑评估并采取措施规避主要人工智能风险，包括算力可用性，数据准确性，模型泛化性，模型解释性，隐私安全性，模型适配性，模型可扩展性，模型高效性，社会伦理性，社会环保性。
- 确保实施的敏捷性和持续优化，让企业既可短期内获得阶段性商业收益、增强组织的信心和支持，又可以根据实施情况即使调整相应规划以更好适应业务需求和内外部变化。
- 建立有效的团队，一支由业务、技术（人工智能、数据）、风险管理、战略、法务等多领域人才组成的团队对于组织人工智能的发展至关重要。

企业可以依照以下步骤开展实施相关工作：

第一步：了解现状。

评估企业及所处行业当前数字化成熟度，预估现在和未来 5-10 年法规、客户、竞争对手和核心市场在人工智能领域的发展变化，在此基础上结合企业的未来战略设定相关人工智能应用方向。

各行业可以利用生成式 AI，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业现状，评估企业的数字化成熟度，前瞻性地预测行业未来的发展趋势。

第二步：制定目标。

确定变革的关键驱动因素 - 从行业驱动因素到市场因素。确定如何将人工智能纳入企业技术架构和业务结构，制定与现有企业战略叠加的人工智能目标 - 降低生产成本、减少人力、增加收入、创新产品和服务等。

当前，企业面临着激烈的竞争和客户不断变化的需求。为了生存和发展，企业必须寻求各种方式来提高效率和降低成本。通过生成式 AI 自动化生成内容，可以大大减少人工操作时间，从而提高员工的工作效率。通过优化运营流程、减少人力成本、节省时间成本等方式，生成式 AI 能够帮助企业实现成本下降。调查发现，75%的受访企业表示降本增效是企业应用生成式工具的首要目的，另外提高敏捷性与市场反应速度（36%）、满足差异化产品与服务创新（34%）、增强办公效率与内部协同（32%）同样是企业引入生成式工具的主要目的 [135]。

第三步：规划路径。

根据业务发展需求制定人工智能应用计划列表，运行测序活动以创建有序计划的初步视图。确定测试/迭代排序的时间表和相关高级资源要求。

第四步：确定方案。

对业务场景分析，业务部门需要与数字化部门一起对每一个应用场景进行分析与拆解，确定每个业务应用具体采用的人工智能相关技术以及方案。为所需的投资、回报、资源、品牌发展等创建总体业务评估。

企业目前正经历着迅速增长的数据量，随着业务需求的多样化，数据类型也呈现出多样性。在选择数据方案时，必须考虑以一种统一的方式来管理各种不同类型的数据，同时需要确保所选方案具备足够的可扩展性，以满足不断增长的数据需求。

良好的数据质量是确保模型调试获得精准输出的关键因素。因此，在制定数据方案的初期阶段就应当充分考虑数据质量和数据安全性，以确保所建立的系统能够稳健地支持企业的数据增长和多样化需求。

第五步：统筹协同。

根据不同业务应用场景对于人工智能技术的不同需要，分析各项技术间的协同性，并根据需求与能力匹配程度制定不同的技术实施策略。将关联度更强，技术、数据训练集、甚至方案复用性高，在实现基础设施层，服务组件层，生态应用层进行有效整合，实现良好的技术协同能很好地帮助项目有效衔接，减少重复投入和复杂管理。

第六步：执行管理。

制定详细的执行计划，包括人力资源、时间、资金等。在基础设施层，服务组件层，生态应用层三个技术层面确定哪些是战略核心，从中短长期确定哪些需要企业自建，招聘相应团队。哪些进行外包或合作，寻找最合适的技术伙伴。确保建立有效的人工智能管理体系，包括清晰的角色，流程，指标，质量管控过程。实现团队敏捷执行项目，与生态各种参与方的良好长期合作关系，持续管控人工智能相关风险。

第七步：持续优化。

定期结合审核每个项目的建设效果以及商业价值，结合外部行业在各个业务领域人工智能的最佳实践，以及市场的最新人工智能技术发展，对内部的人工智能应用和管理进行优化迭代，确保企业内的人工智能应用与时俱进。

在人工智能赋能的数字化创新与可持续发展帮助企业在未来发展创建新的竞争力，成为智能化时代的真正受益者。

六 企业应用生成式人工智能的参考案例与实施价值

当下，生成式 AI 技术突飞猛进，善用 AI 的企业获得了更大竞争优势。过去，在数据为先的发展阶段，聚焦数据与数据生命周期，IBM 提出人工智能阶梯（AI Ladder）的方法，从数据的收集、组织、分析、融合四个步骤为企业规模化部署 AI 奠定基础。这些工作在一个现代化的人工智能阶梯当中处于底层，也就是“+AI”的工作。今天，企业在积极探索如何将 AI 用于企业的应用，如何对企业的工作流实现智能自动化、甚至替换现有的工作流，最终让 AI 来完成工作——企业正从以数据为先的“+AI”阶段，步入以 AI 为先的“AI+” 的全新发展阶段。 纵观 AI 的发展历程，IBM 一直处于突破性 AI 科技的前沿，在 IDC 2023 年的市场调研中 IBM 被评为全球 AI 治理平台的领导者。IBM 致力于将 AI 嵌入企业的战略核心，并致力于将前沿科技转化为生产力。我们为企业提供开放、可信、有针对性和以实现价值创造为使命的 AI 解决方案。这些方案整合了 IBM 在软硬件以及咨询的全栈能力，并且在全球的汽车、电子、制造业、消费品、金融、医疗领域都有长足的实践经验。在此与您分享 IBM 的经典案例。IBM 愿成为您的转型伙伴，与您携手共创企业级可信 AI 新时代！

6.1 IBM 案例

6.1.1 IBM + 源卓微纳 + 艾科斯幂：以 AI 会友，共创制造业智能化故事

源卓微纳科技（苏州）股份有限公司是一家在业界处于领先地位的高科技公司，专注为高端电子电路、IC 载板、先进封装、微机电系统 (MEMS)、泛半导体、太阳能和微纳器件制造提供生产设备和工艺解决方案。艾科斯幂信息科技有限公司（X-POWER）是一家科

技创新公司，为客户定制化提供智能化数字化整体集成系统解决方案, 2023 年成为 IBM金牌合作伙伴。

在产品研发过程中，源卓微纳面临着做市场调研和市场评估，人力投入高、检索效率低的挑战、也不能保证技术调研的准确性、及时性和全面性。希望找到一种方式来帮助研发团队提高工作效率。另外，为了赢得客户的满意度，源卓微纳对客户的承诺是 7*24小时的技术支持和售后服务，远程服务 15 分钟内响应，驻点区域 4小时内到达。源卓微纳一直在寻找合适的智能手段来提升售后服务效率。

艾科斯幂与 IBM 合作根据源卓微纳的业务需求，选择了 watsonx Assistant 做为智能助手提供前端入口和语义理解的能力，Watson Discovery 做为文档存储和检索工具，并集成了 IBM 最新的 AI 开发平台 watsonx.ai，为源卓打造了企业级智能问答知识库。这个体系还利用IBM AI 驱动的应用集成方案 Cloud Pak for Integration (CP4I)进行应用集成。watsonx.ai 为 IBM 企业级 AI 开发平台，基于最新生成式 AI 功能，使数据科学家、开发人员和数据分析师能够利用开放直观的用户界面来训练、测试、调整和部署 AI。watsonx Assistant 提供面向业务的更智能的对话式 AI 平台。Watson Discovery 为 AI 支持的智能搜索和文本分析平台。CP4I 具备提升应用程序速度与质量的卓越优势。

项目实施之后，全面提高了源卓微纳的研发效率和售后满意度：

- 研发售后人员登陆 OA 系统，根据登陆 ID, 系统会判断登录者有哪些权限。之后到达基于 IBM watsonx Assistant 搭建的“智能问答界面”；
- 根据用户的问题进行语义分析、同时基于关键字在 Watson Discovery 知识库中进行检索返回到 watsonx Assistant；

- 透过 watsonx.ai 大语言模型进行深加工，使得答案更加准确和人性化，并将答案返回到 watsonx Assistant 智能问答界面上

源卓微纳与艾科斯幂选择 IBM watsonx 系列产品是看中了 IBM 方案在以下四方面的独特价值：

- 本地部署，数据安全
- 混合云部署云能力，容易迁移
- 一体化的平台集成能力，易上手
- 企业级技术支持能力

6.1.2 一汽-大众基于 IBM 业财一体化平台构建全面预算体系

一汽-大众汽车有限公司（以下简称一汽-大众）于 1991 年 2 月 6 日成立，是由中国第一汽车集团有限公司、德国大众汽车股份公司、奥迪汽车股份公司和大众汽车（中国）投资有限公司合资经营的大型乘用车生产企业，是我国第一个按经济规模起步建设的现代化乘用车生产企业。经过 32 年的发展，一汽-大众产能布局已覆盖东北长春、西南成都、华南佛山、华东青岛以及华北天津。累计产销汽车超过 2500 万辆，销量规模位列中国乘用车行业第一阵营。

在汽车行业高速发展的今天，其上下游产业链复杂多变。在转型升级的过程中，如何做到市场的快速应变、实现业财融合的精细化管理是车企所面临的挑战。一汽-大众在转型升级中，精细化管理的需求越来越旺盛，企业也面临着缺乏全面预算系统支撑、无法支持实时测算，以及和 Excel 集成不好等一系列挑战。

一汽-大众在进行了多个软件产品的对比后，选择了与 IBM 合作，基于 IBM 业财一体化平台 Planning Analytics with Watson 实现了精益化管理目标。

通过这一方案，一汽大众实现了：

- 以财务预算为基础建立覆盖公司、工厂、各部门的预算管理系统，建立关键指标监控与自动化分析体系；
- 提供预算编制全过程的目标下达、在线编制、提交汇总，多上多下的审批管控过程能力；
- 利用强大的分析能力来增强预算过程的管控和纠偏，确保经营目标的落实；
- 建立多维度的业财融合预算分析能力。

全面预算管理确保了业务可以按照目标组织运营，也保证了一汽大众有能力根据市场的变化及时做出调整。一汽-大众选择与 IBM 合作，不光因为 IBM 是一汽-大众的长期合作伙伴，同时也是因为 IBM Planning Analytics with Watson 业财一体化平台能够满足企业复杂业务需求，且具备快速建模、弹性应变，以及实时场景的分析、测算能力，从而实现了企业精细化管理的目标，提升了效益。

6.1.3 延锋汽车数智之旅

延锋汽车的总部位于上海，是一家全球性的汽车零部件厂商。延锋汽车在全球 20 个国家拥有 9 家研发基地、240 多个工厂与技术中心，员工总数超过 55,000 人，为全球整车制造商提供汽车零部件产品的设计、开发和制造。面对挑战，延锋汽车的做法是——携手像 IBM 这样技术领先且拥有丰富企业经验与技术专长的公司，共同探索数据为先的数智化之路，实现降本增效与创新发展。

场景一：AI赋能数据实时抽取——解决开源的数据抽取工具Kafka带来的运营瓶颈

延锋汽车在每一个分支工厂都部署了一套开源的 Kafka 集群，用作 MES 系统中多项实时生产数据的抽取，提供给各个工厂的 MI 看板系统进行查询和展示。

基于 IBM 的方案，延锋的样板工厂开始采用 IBM Cloud Pak for Integration 中的 Event Streams 组件来做实时数据的抽取。生成数据的应用程序从 MES 系统中抽取零配件生产班次、生产数量、需求数量、返修数量、排序以及其他相关的生产数据，发送到对应的数据主题频道。抓取数据的应用程序通过订阅 Event Streams 的相应主题频道，可以直接使用相应数据。MI Skynet 看板系统则可以选取指定的表字段，进行后续的仪表板展示和预警分析。

通过部署 Event Streams 这一企业级的数据抽取解决方案，延锋汽车可以实现"一键"部署，开箱即用，零宕机滚动升级，时刻拥有最新的 Kafka 稳定版本。同时，组件自带图形化操作界面，几乎不需要额外的技能培训。利用高安全性和异地复制功能，还能获取企业级灾难恢复能力。先进的模式注册表和丰富的 Kafka 连接器以及可扩展的 REST API，轻松扩展现有企业资产的范围。不仅如此，IBM 还提供配套的企业级售后服务、专家咨询和及时的问题排查，能够帮助客户获取所需的技术专业知识。

场景二：AI赋能海量数据高速传递——实现分支生产车间和总部之间海量数据的高速传输，为智能库存与预测夯实数据基础

为了实时掌握分布在全球 240 多个工厂众多车间的零部件库存使用情况，延锋汽车利用各工厂的监控摄像头将成千上万张的实时照片快速地传回总部。起初，智能制造部门采用传统的复制粘贴的方法来传输批量的照片文件，由于传输速度慢、网络延迟明显、丢包严重，需要多次分批次手工选择对应照片文件进行复制，这样既耗时又容易误操作，同时

无法断点续传、无法自动重连、无法自定义传输速度，主干网的传输带宽无法得到充分利用。

在 IBM 团队支持下，延锋汽车仅用一天时间就完成了小而美、轻量级的 IBM Cloud Pak for Integration - Aspera 的组件部署，构建起企业级的文件传输解决方案，使延锋汽车的文件传输速度平均提高了 10 倍，节约了人工等待时间，避免了人工误操作，实现了断点续传和自动重连，并且可以动态配置传输带宽和限速，在不影响 ERP 核心系统性能的前提下最大化地提高了实时监控文件的传输效率，为实现其智能库存与预测的愿景奠定了基础。

场景三：AI赋能高效订单管理——将海量外部通用订单自动转为内部订单

延锋汽车每天收到整车厂和下游厂商的订单量巨大，之前需要通过人工根据经验把通用订单转为内部订单，每个工厂每天需要两名工作人员花 150 分钟进行手工分类。即使在这样的人工投入下，仍伴随 15%的分类错误，给延锋汽车带来成本和效率的双重挑战。

利用 IBM Watson Discovery 强大的自然语言学习能力，延锋汽车成功构建起 AI 模型， 从他们涵盖了 1.8 亿历史数据、200 多种排列组合、结构化数据和非结构化文本的混合数据中，学习通用订单对应的内部订单背后蕴藏的规则，变身智慧大脑，实现了全自动执行流程，无需人工操作。

场景四： AI赋能研发创新——ELM助力延锋电子优化研发流程，将效率转化为生产力

中国汽车的产业价值链在智能网联汽车发展趋势下，软件层面和智能化层面的价值逐渐被挖掘和放大，以高效研发为牵引，成为车企打赢价值战的致胜关键。与此同时，创新迭代和项目交付高速并行，质量管理难度也不断提升，客户对供应商的体量和质量追溯能力要求也越来越严苛。

IBM 汽车行业工程生命周期解决方案 ELM （Engineering Lifecycle Management)集合了研发效率管理、研发知识管理、研发能力构建和研发合规性四个主题，是市场领先的高效研发管理解决方案。助力延锋电子优化研发流程，将效率转化为生产力。

场景五： AI赋能设备管理创新——Maximo助力延锋提升设备管理绩效

随着延锋业务的快速发展，已经形成了在全球 20 多个国家拥有 240 多个生产基地的庞大规模。如何从集团的角度科学地持有并管好资产设备，对延锋提高生产运营水平、降低运营成本、以及实现智能制造都有重要意义。

IBM Maximo 是全球领先资产设备管理解决方案，近十年来一直与延锋在这一领域紧密合作，助力延锋在快速扩展的过程中实现资产设备的精细化管理，这包括：实现资产设备集团化管理，适应多语言、多时区、标准化管理的挑战；实现资产设备全生命周期管理；实现移动化应用，提升客户效率和用户体验；通过精细化的设备运维策略和执行，提升运维效率、保障生产执行并降低备件库存。

随着技术的不断发展，IBM 将 AI 和 IoT 技术不断地赋能到 Maximo 资产设备管理解决方案领域，在物联网监控、设备健康分析、预测性维护以及 AI 维修助手等领域都形成了领先的方案。随着延锋设备管理要求的不断提升、数据的不断积累和完善，IBM Maximo 将不断深化在延锋资产设备管理领域的应用与合作，助力延锋实现智能制造的战略目标。

6.1.4 苏州环球科技利用人工智能和自动化技术，成功构建企业智能业务流程管理平台

苏州环球科技股份有限公司（简称“环球科技”）始建于 1970 年，拥有 50 多年链条研发、制造经验，是国内链条行业的领头羊，也是集链条研发、制造、销售于一体的国家高新技术企业。近年来，数字化、智能化成为环球科技转型升级的重要手段，公司先后

开发部署了 MES、ERP、WMS、质检系统、供应商管理系统等多个 IT 系统，服务企业各个业务流程。

随着高新业务的快速发展和竞争加剧，环球科技急需实现从产品设计、生产、物流、销售等多环节多业务角色的紧密配合和上下游联动，构建统一的智能业务流程管理平台，提升效率、降低成本。环球科技要构建先进的智能业务流程管理平台，首先需要实现跨越不同业务流程间的多个系统的互联互通，串联不同流程上的工作角色使用的 IT 系统，实现各个业务环节之间的无缝连接，快速预警，全流程可视化、可追踪。

环球科技利用 IBM Cloud Pak for Integration 中的企业服务总线组件 App Connect 来提供可靠的应用集成解决方案,构建了统一的应用集成平台，解决了系统之间接口混乱的问题，实现了敏捷且轻量的应用集成，实现了多种应用接口联通、多种数据格式解析处理；还提供了信息同步、异步传输能力；同时具有高安全性、高稳定性和易扩展性；具有良好的统计、分析和监控能力。

实现现有信息系统的互联互通是走向智能自动化的第一步。第二步需要系统性梳理环球科技业务过程中的角色、流程、规则及当前 IT 系统架构。过去，环球科技的业务流程执行主要靠线下沟通，使得不同部门间存在信息断点，导致从订单到交付沟通成本高，经常难以按期兑现交付。同时，生产管理主要依靠经验判断，原料临时采购的情况经常发生，存在成本风险管理的问题，采购进度追踪也不及时。

环球科技采用 IBM Cloud Pak for Business Automation 中包含的业务自动化工作流 (BAW) 的能力，作为企业级 BAW 来整合业务系统与管理系统流程审批信息，实现对业务流程全面可视并综合管理。该产品为流程改进和 BAW 生命周期管理提供了共用软件平台，在流程管理和业务规则管理领域体现出优势，表现出关键任务解决方案所要求的强大

性能及稳健性，帮助环球科技建立他们所需要的工作流和可视化流程平台。IBM Cloud Pak for Business Automation 中除了 BAW，还在智能自动化领域提供客户所需的规则引擎、文档管理平台等能力，未来可以快速进行横向扩展，支持客户更多的智能化需求。

目前，环球科技利用 IBM Cloud Pak for Business Automation 建立了统一的业务自动化管理平台，这个平台的开发界面比较简单，公司的 IT 团队很容易就可以按照新的业务需求开发新的业务流程；而且开发速度快，调试方便快速，部署容易；开发出来的流程可以很容易地跟其他系统做整合集成；同时，流程自带内容和规则引擎，满足了环球科技对于简化规则引擎的需求。不仅如此，该平台从技术上实现了整体流程可视化，以便对内部复杂的流程进行管理，并且能做到流程的合规监控。基于混合云底座，方便客户切换不同的底层基础设施，简化了在混合 IT 环境下对其应用、数据以及业务流程的统一管理。

具体实施价值体现在：

- **数字化的生产流程超级自动化**：以前的零件生产计划需要人工手动用 Excel 计算完成，需要 5 天时间；现在全自动计算只需要几个小时，计划员工作量减少了 80%；
- **不良品的全数字化评审：**改进后的数字化流程将检验员和文员的工作效率大大提升，工作量减少了 60%；
- 数据驱动的采供销一体化流程将订单完成提高了 50%。

6.1.5 NASA 携手 IBM 发布 Hugging Face 平台最大开源地理空间 AI 基础模型

IBM 与开源 AI 平台 Hugging Face 共同宣布，基于美国宇航局（NASA）卫星数据构建的 IBM watsonx.ai 地理空间基础模型现已在 Hugging Face 发布。它将成为 Hugging Face 上至今最大的地理空间基础模型，也是首个与 NASA 合作构建的开源 AI 基础模型。

在环境条件几乎每天都在变化的气候科学领域，获取最新数据仍然是气候科学研究面临的主要挑战。尽管数据量不断增加（NASA 预估到 2024 年，其新任务将产生 25 万 TB 的数据），但科学家和研究人员们在分析这些大型数据集时仍面临障碍。作为与 NASA 签署的空间行动协议（Space Act Agreement）的一部分，IBM 在 2023 年初构建了一个用于处理地理空间数据的 AI 基础模型。现在，双方联手业内公认的开源领导者和 Transformer 模型库 Hugging Face，共同发布上述地理空间基础模型，以扩大气候和地球科学研究中对 AI 技术的访问和应用，从而加速创新。

该基础模型由 IBM 和 NASA 共同训练，使用了过去一年在美国大陆范围内的 Harmonized Landsat Sentinel-2（HLS）卫星数据，并基于洪水和焚烧区域的标记数据进行了调优。相比于目前的领先技术，该模型仅使用同等条件下一半的标记数据，便实现了 15%的效果改进。通过进一步的调优，该模型还可以应用于追踪森林砍伐、预测农作物产量、检测和监测温室气体等新任务。IBM 和 NASA 的研究人员还与克拉克大学合作，将该模型用于时间序列分割（time-series segmentation）和相似性研究等领域。

IBM 推出的人工智能和数据平台 watsonx，使企业能够利用可信数据扩展和加速最先进人工智能的影响。作为 IBM watsonx 的一部分，地理空间模型的商业版本也将通过 IBM 环境智能套件（EIS）推出。在开展 HLS 地理空间调频工作的同时，NASA 和 IBM 还在开发其他应用程序，以从地球观测中提取见解，包括基于地球科学文献的大型语言模

型。根据美国国家航空航天局的开放科学准则和原则，这项合作工作所产生的模型和产品将向整个科学界开放。

6.1.6 IBM 利用 watsonx 为温网锦标赛带来由基础模型与生成式 AI 赋能的数字体验

全英草地网球俱乐部是每年温布尔登网球锦标赛的主办方，保护这项世界上历史最悠久、最负盛名的网球赛事的丰富文化遗产对它来说至关重要。自 1877 年首次举办锦标赛以来，温布尔登网球赛已汇聚了各行各业的球迷，从英国皇室成员、企业主到业余体育迷，他们共同享受着世界上最好的网球赛事。

与此同时，社会正在发生日益巨大的变化，现有和潜在球迷的性质、深度和广度也是如此。数字技术处于变革的前沿，为球迷、媒体和球员提供了新的参与方式。温布尔登网球公开赛充分抓住了这些机遇，不仅加强了其已经非常强大的品牌，而且自身也成为了一家数字媒体公司，制作与网球相关的视频、网络和社交媒体内容，并在其数字平台上发布，包括 Wimbledon.com 和 Wimbledon 应用程序，以及通过新闻媒体发布。这种创新的基础是温布尔登与 IBM 的长期合作关系。两家公司于 1990 年携手合作，并在持续利用技术力量为温布尔登观众们提供全新、卓越的体验。

2023 年 6 月，IBM 和全英草地网球俱乐部公布了两项在 2023 年温布尔登网球锦标赛 (简称“温网”) 上推出的全新球迷数字体验新功能：

- 第一个新功能是利用 IBM watsonx 的生成式 AI 技术，为温网比赛期间所有视频集锦提供生成式 AI 网球评论解说的功能。
- 第二个新功能是叫做 IBM AI Draw Analysis 的应用，这是首个为网球比赛而打造的分析应用，可以提供一套全新的统计数据，以确定每个球员进阶决赛的潜力。

这两项新功能扩展了温网应用和 wimbledon.com 上的针对球迷的数字工具套件，是 IBM 和温网利用技术帮助球迷更深入地参与温网锦标赛的最新例子。

IBM AI 评论解说：新的 AI 评论解说功能将为观看比赛集锦视频的球迷提供关键时刻的音频评论解说及字幕，球迷可以打开或关闭字幕功能。该工具旨在为球迷提供更有见地的体验，让他们在温网应用和 wimbledon.com 上通过精彩视频来抓住比赛的关键时刻。为了开发此项新功能，IBM Consulting 的体验设计合作伙伴 IBM iX 的专家与全英俱乐部合作，利用 IBM 企业级 AI 和数据平台 watsonx 的基础模型，用网球的独特语言训练 AI。基于这些基础模型的生成 AI 被用来生成具有不同句子结构和特定词汇的旁白和解说，使剪辑的视频内容更具知识性，也更加引人入胜。

IBM AI Draw Analysis：引入温网球迷数字工具的另一个新功能是 IBM AI Draw Analysis，这是网球领域的首个此类统计数据，它使用 AI 来确定单打抽签中每个球员进入决赛的可能性。每个球员的进阶优势将通过评级的方式来呈现，基于包括球员与潜在未来对手的比赛以及球员在单打抽签中的位置与竞争对手的比较等因素。这一新的见解将帮助球迷发现单打抽签中的异常和潜在的惊喜，而这个仅通过查看球员的排名是无法察觉的.

新推出的两项新功能将添加到温网应用和 wimbledon.com 上为球迷提供的人工智能数字工具套件中。该套件里还包括 IBM Power Index 排行榜、IBM Match Insights 以及个性化推荐和精彩画面集锦等应用，这些数字功能使用来自温网比赛每次击球而得出的多达 100,000 多个数据点，由 IBM Cloud 上的 IBM Watson AI 技术进行分析，旨在让球迷更容易了解要关注哪些球员、这些球员与对手的比较，以及谁可能获胜等信息。球迷们在整个温网比赛期间都可以利用这些数字工具，持续关注他们喜爱的球员，不断更新和获取量身定制新见解。

6.1.7 美国最大的房车零售商 Camping World 通过 AI 驱动的虚拟助手重构客户体验

自 1966 年以来，Camping World 一直专门为车主和露营者提供产品和服务。现已发展成为美国最大的房车和房车相关产品和服务零售商，拥有 160 多个 Camping World SuperCenter。自 2009 年以来，该企业一直是 NASCAR 的官方房车和户外用品零售商。此外，Camping World 还与美国职业棒球大联盟建立了多年的合作伙伴关系。

Camping World 深知提供卓越的客户服务对于在竞争中保持领先地位至关重要。该企业在很大程度上依赖其呼叫中心来提供无与伦比的客户服务，但在新冠疫情之后，客户数量的激增暴露了其现有基础架构的一些问题。随着数量和流量的增加，客服代理管理和响应时间方面的缺口愈发突出。Camping World 为三类截然不同的喜欢房车生活方式的客户提供服务。第一是零售客户，第二是金融服务（如保险）或商品客户，第三是经销商客户。”Camping World 有一个规模不错的呼叫中心，但无法让一位客服代理来满足三个不同业务部门的需求。这给呼叫中心的人员配备带来了极大的复杂性。没有 24x7 全天候呼叫中心也是一个长期存在的问题。

Camping World 需要一个以人为本的解决方案，使其运营能够伸缩，并应对寻求快速帮助的客户数量的增加。在寻找最适合的呼叫中心现代化路线后，该零售商选择了 IBM 开发的认知 AI 方案。IBM 为 Camping World 提供了不同的场景，包括构建技术的路线图，最终使客户能够简化流程、提高客服代理效率，最重要的是极大地改善整体客户体验。

该解决方案由 IBM watsonx Assistant 提供支持，无缝集成了对话云平台 LivePerson，并在所有网络属性中进行了部署，增加了问题和电话功能的覆盖范围。它将 Camping World 客户与虚拟客服代理连接起来，使现场客服代理能够接管更复杂的对

话。虚拟客服代理名为 Arvee，通过动态路由和容量管理功能，确保更快、更高效的响应时间。Arvee 的潜在客户开发功能（尤其是在工作时间之后）是该团队以前没有的功能，可以让现场客服代理轻松跟踪并主动跟进客户询问。

IBM watsonx Assistant 可以识别客户的意图，并能将呼叫者转接至有空的现场客服代理以开展对话。实施后，客户参与率呈显著上升趋势，中断的对话数量有所减少。客户的等待时间越来越短，响应速度越来越快，客服代理的效率也得到了显著提高。借助客服代理桌面集成，以及 Arvee 在处理互联网和手机短信时主动收集客户数据的帮助，现场客服代理可以同时处理多个聊天，从而将整体效率提高 33%。截至 2022 年 3 月，客户参与度增加了 40%，等待时间降至 33 秒。

6.1.8 花旗银行采用 IBM 企业级 AI 解决方案实现业务数智化转型

作为一家全球领先银行，花旗银行为超过 2 亿客户提供服务，并一直在积极探索运用先进的企业级 AI 技术来增强企业运营。作为混合云和 AI 技术解决方案的领先供应商，IBM 为花旗银行提供了一套企业级 AI 解决方案和服务。

花旗银行拥有全球最大的公司审计部门之一，其中包括 2500 名审计员，他们需要处理大量的文档审查和风险评估工作。鉴于这一职能的规模和重要性，花旗银行深知替换现有审计平台必须谨慎考虑，并且必须对技术合作伙伴充满信任。在深入了解 IBM 的人工智能解决方案后，花旗银行选择了基于 IBM Watson Discovery、IBM Cloud Pak for Data 和 IBM OpenPages with Watson 的高级分析解决方案，从而协助 2500 名审计师协同在一个平台上工作。原来一个月可以完成 40 个审计项目，现在一个月可以完成上千个审计项目。同时，IBM 还为花旗银行创建了一个 AI 创新空间，以使他们能够继续在新审计平台上

应用 AI 来进行持续创新。通过引入 IBM 企业就绪的 AI 和数据平台 watsonx，花旗银行与 IBM 探索将 watsonx 和基础模型应用于内部管控，以进一步实现审计的智能化转型。花旗银行内部审计部门的 Mark Sabino 博士表示：“我们正在研究大型语言模型 (LLM) 的潜在用途，我认为有无限的可能。其中一个我在考虑内部使用的关键用例是，如何使用 LLM 来将您的管控与您的内部政策和法规联系起来。”

在部署 AI 工具时，如何在效率和创新之间找到平衡，离不开灵活、安全、可持续和可扩展的 IT 基础设施。花旗银行利用开源数据库 MongoDB 构建了全球最大的数据库平台，部署在多个全球数据中心，并选择在 LinuxONE 上托管 MongoDB。相较于传统的 IT 解决方案（如增加服务器），LinuxONE 提供了垂直扩展和对数据泄露与网络攻击的关键保护，从而优化了数据中心的运行效率，同时降低了碳足迹。与此同时，服务器节能比例达到 50%，性能提升 15%，安全性也得到了提升。此外，在 AI 应用领域，LinuxONE 也处于领先地位。去年发布的第四代 LinuxONE 搭载了业界首个集成的 AI 芯片，可以帮助客户在大规模交易等任务中实现实时 AI 推理能力。

花旗银行技术基础架构部常务董事 Martin Kennedy 表示："随着我们业务的增长和变得越来越‘数字为先’，采用传统的 IT 解决方案会增加更多的物理服务器，同时增加所需的楼层空间。而采用托管在 IBM LinuxONE 上的 MongoDB，则可以提供垂直扩展和针对数据泄露与网络攻击的关键保护，有助于优化数据中心，同时降低我们的整体碳足迹。"

6.1.9 Blendow Group 携手 IBM 获得基于 AI 的法律分析变革力量

立足于瑞典法律知识和情报传播的前沿，Blendow Group 已成为法律新闻、教育和专家分析的关键资源。Blendow Group 需要仔细分析、总结和评估无数法律文件，从法院裁

决到立法和判例法。由于这些分析基于大量的信息，在有限的员工资源下，Blendow Group 需要一个可扩展的解决方案。

为了应对这一挑战，Blendow Group 与 IBM 合作，应用基于 IBM AI 开发平台 watsonx.ai 以及 IBM 全栈的软硬件能力获得了法律分析中的变革力量。watsonx.ai，是一个专为今天与未来的业务而设计的 AI 开发平台。它结合了 IBM Watson Studio 的功能和利用基础模型能力的最新生成式 AI 的功能，使数据科学家、开发人员和数据分析师能够利用开放直观的用户界面来训练、测试、调整和部署由基础模型提供支持的传统机器学习和新的生成式 AI 功能，由此快速构建、运行和部署 AI。该人工智能解决方案擅长浏览大量法律文件，从详细的法院判决到广泛的法律文本和法规。它增强了研究、分析并简化了创建法律内容的过程，同时保持了敏感数据的最大机密性。

该解决方案不仅简化了内容准备过程，还大大提高了搜索和分析广泛法律文件的能力:

- 减少 70%发现和分析法律文件所需的时间
- 增加 80%各种法律文本的覆盖面
- 减少 90%总结和分析这些文档所需的时间

6.2 其他案例

以下由 COPU 提供的成员企业生成式人工智能相关应用案例。

6.2.1 大象声科（深圳）科技股份有限公司基于 Intel OpenVINO 平台构建智能语音增强和智能语音交互解决方案

【案例背景】

大象声科(深圳)科技有限公司(以下简称大象声科)成立于 2015 年，是全球领先的机器听觉人工智能公司。依托计算听觉场景分析理论(CASA)和深度学习技术，提供全球领先的智能语音增强和智能语音交互解决方案。随着人工智能的快速发展，语音交互成为新的交互形式，如何在复杂噪声环境下提供清晰的语音交互体验，是大象声科面临的挑战。

【业务需求】

在如今的快节奏生活中，人们需要在各种环境中进行语音交互，如地铁、商场、KTV 等噪声环境。在这些环境下，如何提供清晰的语音交互体验、实现语音增强和噪声抑制，为用户提供更佳的使用体验，是大象声科所面临的挑战。

【解决方案】

大象声科推出了智能语音增强和智能语音交互解决方案。通过将目标声音与噪声进行“理想二元掩模”处理，将声学信号处理转化为一个分类问题，基于深度学习和计算听觉场景分析理论，算法具有自适应能力，能够不断学习优化，实时分离人声和背景噪声，提取清晰人声。同时，借助 OpenVINO 集成在英特尔 GNA/VPU 平台上，大幅度提升了用户语音清晰度和语音交互体验。

【实施价值】

大象声科的智能语音增强和智能语音交互解决方案，能够有效提升各种环境下的语音交互体验，提供清晰、稳定的语音输出，大大提升了用户体验。通过自适应学习优化的算法，能够实时应对各种复杂的声音环境，满足了用户在各种环境下的语音交互需求，提升了其产品的竞争力和市场份额。

6.2.2 深圳酷酷科技有限公司基于 Intel Realsense 技术架构构建 AI 可穿戴解决方案

【案例背景】

深圳酷酷科技有限公司(以下简称酷酷科技)成立于 2015 年，致力于 AR/MR 智能眼镜，AI 智能穿戴及新一代个人信息终端的研发和销售。在 AI 穿戴设备的发展过程中，酷酷科技面临着技术难关、市场竞争等各种挑战。

【业务需求】

在 AI 穿戴设备的高速发展中，如何实现技术的突破、满足市场的需求、提供优质的产品是酷酷科技所面临的挑战。酷酷科技在技术研发过程中，需要解决整机设计、微显示及光学设计、HCI 等方面的技术难题。

【解决方案】

酷酷科技从实际应用出发，结合科技发展的趋势，在 AI 穿戴设备、AR/MR 方向上努力攻克技术难关。目前已经在牙科的微创及显微手术上有所突破，基于英特尔 Realsense 等技术，在一些新产的手势识别，眼动交互等方面，进行研发和测试。同时，酷酷科技也在骨科、胸外科、脑外科、医美手术等方向进行技术及应用的突破。

【实施价值】

酷酷科技的 AI 穿戴设备解决方案，不仅实现了技术的突破，提供了优质的产品，也满足了市场的需求。通过攻克技术难关，酷酷科技提升了其产品的竞争力，增强了公司的市场地位。同时，酷酷科技的解决方案也为社会带来了实际的价值，如在医疗领域，酷酷科技的智能眼镜可以提升医生的手术效率，提高手术的精准度，为患者提供更好的医疗服务。

6.2.3 香港流形科技公司的基于 Intel OpenVINO 平台构建三维重建解决方案

【案例背景】

香港流形科技有限公司(以下简称流形科技)成立于 2015 年，专注于 3D 扫描、建模、机器人技术和算法开发，致力于构建虚拟与现实之间的桥梁。然而，如何提供实时、高精度的三维重建解决方案，降低作业成本，提升计算效率，是流形科技面临的挑战。

【业务需求】

在快速发展的 3D 扫描、建模行业中，提供实时、高精度的三维重建解决方案是业界的迫切需求。同时，如何将工期从以月为单位缩短到以分钟级，大幅度降低作业成本，提升计算效率，是流形科技所面临的挑战。

【解决方案】

流形科技以自研的高效多传感器融合算法，结合神经渲染技术，为三维重建行业提供实时、高精度的解决方案。通过自研的后处理技术，将传感器的原始精度提升 200%，精度媲美架站式扫描仪。流形科技的解决方案能够在分钟级时间内完成快速高精度三维重建，满足绝大部分下游 3D 行业应用，且流行机可以与多种机器人平台无缝对接，具有极强的通用性。

【实施价值】

流形科技的三维重建解决方案，实现了实时、高精度的三维重建，大幅度降低了作业成本，提升了计算效率，满足了绝大部分下游 3D 行业应用的需求。流形科技的解决方案，不仅提升了三维重建行业的效率，也推动了 3D 扫描、建模行业的发展，提升了流形科技的市场竞争力。

6.2.4 乘木科技（珠海）有限公司基于 Intel OpenVINO 平台构建的数字孪生解决方案

【案例背景】

乘木科技(珠海)有限公司(以下简称乘木科技)成立于 2015 年，是一家国家级高新技术企业，通过整合大数据、人工智能、物联网和数字孪生等前沿技术，提供智能、高效和创新的解决方案。然而，如何利用数字孪生技术，为用户提供更精准、高效的解决方案，是乘木科技面临的挑战。

【业务需求】

在高速发展的数字孪生技术中，如何将现实世界的实体、过程和系统映射到数字世界中，提供全面的数据分析和决策支持，帮助用户优化运营、提高效率和降低风险，是乘木科技所面临的挑战。

【解决方案】

乘木科技推出了数字孪生解决方案，应用了物联网、大数据、人工智能等前沿技术，具备实时监测、预测预警、仿真推演和智能决策等功能。乘木科技的数字孪生解决方案能够解决现实世界中的复杂问题，提供全面的数据分析和决策支持，帮助用户优化运营、提高效率和降低风险。

【实施价值】

乘木科技的数字孪生解决方案，实现了现实世界的实体、过程和系统的数字化映射，提供了全面的数据分析和决策支持。这一解决方案不仅帮助用户优化运营、提高效率和降低风险，也推动了数字孪生技术的发展，提升了乘木科技的市场竞争力。同时，乘木科技的数字孪生解决方案也为智慧城管、智慧工厂、智慧园区等领域的数字化转型提供了有力的技术支持。

6.2.5 深圳博通光电智能科技有限公司的基于 Intel OpenVINO 平台构建的电子纸智慧办公产品解决方案

【案例背景】

深圳博通光电智能科技有限公司(以下简称博通光电)，成立于 2011 年，是全球领先的物联网核心技术、产品和解决方案提供商。然而，在迅速发展的智慧办公行业中，如何利用电子纸，人工智能，大数据和物联网技术，提供高效的智慧办公解决方案，是博通光电面临的挑战。

【业务需求】

在智慧办公行业中，如何通过部署新一代智能办公终端，采集位置、呼叫通知等数据，智能管理投屏动作及显示内容，实现无纸化办公及数字化应用，高效管理办公信息，是博通光电所面临的挑战。

【解决方案】

博通光电推出了“电子纸”智慧办公产品解决方案。这一解决方案应用电子纸、人工智能、大数据和物联网技术搭建的智能办公系统，通过部署新一代智能办公终端，采集位置、呼叫通知等数据，智能管理投屏动作及显示内容，实现无纸化办公及数字化应用，高效管理办公信息。

【实施价值】

博通光电的“电子纸”智慧办公产品解决方案，实现了无纸化办公及数字化应用，高效管理办公信息，提升了办公精细化管理水平和企业办公效率，大幅减少了普通纸张的使用和降低了企业的人力成本。这一解决方案不仅优化了办公环境，也推动了智慧办公行业的发展，提升了博通光电的市场竞争力。

6.2.6 小惟科技（深圳）有限公司的基于 Intel OpenVINO 平台构建的 3D/XR 数字营销 SaaS 解决方案

【案例背景】

小惟科技(深圳)有限公司(以下简称小惟科技)成立于 2018 年，专注于 3D/XR 数字营销 SaaS 解决方案的研发和应用。然而，如何通过 AI1653D/XR 技术提供高效、经济、精准和智能的数字营销解决方案，降低制作成本，提高精准度，是小惟科技面临的挑战。

【业务需求】

在快速发展的数字营销行业中，如何通过 AI1653D/XR 技术实现快速、经济、精准和智能的数字营销，帮助电商企业在快速变化的市场中迅速生成高质量营销内容，降低制作成本，提高精准度，是小惟科技所面临的挑战。

【解决方案】

小惟科技推出了 3D/XR 数字营销 SaaS 解决方案。这一解决方案整合了硬件终端与在线 SaaS 平台，提供一体化的解决方案，帮助电商企业在快速变化的市场中迅速生成高质量营销内容，降低制作成本，提高精准度，同时通过智能广告创作、个性化推荐和数据驱动决策等工具，助力企业实现更为智能、创新的数字营销策略。

【实施价值】

小惟科技的 3D/XR 数字营销 SaaS 解决方案，实现了快速、经济、精准和智能的数字营销，大幅度降低了制作成本，提高了精准度，满足了电商企业的需求。这一解决方案不仅提升了数字营销行业的效率，也推动了 3D/XR 技术的发展，提升了小惟科技的市场竞争力。同时，小惟科技的解决方案也为企业提供了更为智能、创新的数字营销策略，助力企业实现 GMV 的快速增长。

七 企业级生成式人工智能的未来展望

近年来，人工智能相关技术持续演进，产业化和商业化进程不断提速，正在加快与千行百业深度融合。全球人工智能市场预计到 2024 年将超六千亿美元，复合增速 27%。世界各国纷纷布局人工智能，深化人工智能发展，将人工智能发展制定为国家未来数字化发展战略。

美国成立了国家人工智能倡议办公室、国家 AI 研究资源工作组等机构，各部门密集出台了系列政策，将人工智能提到“未来产业”和“未来技术”。2021 年 7 月，美国国家科学基金会联合多个部门和知名企业等，新成立 11 个国家人工智能研究机构，涵盖了人机交互、人工智能优化、动态系统、增强学习等方向，研究项目更是涵盖了建筑、医疗、生物、地质、电气、教育、能源等多个领域。

英国于 2021 年 9 月发布国家级人工智能新十年战略，这是继 2016 年后推出的又一重要战略，旨在重塑人工智能领域的影响力。英国支持人工智能产业化，启动人工智能办公室和英国研究与创新局联合计划等，确保人工智能惠及所有行业和地区，促进人工智能的广泛应用。

中国《中共中央关于制定国 民经济和社会发展第十四个五年规划和 2035 远景目标纲要的建议》指出，要瞄准人工智能等前沿领域，实施一批具有前瞻性、战略性重大科技项目，推动数字经济健康发展。十四五规划纲要明确大力发展人工智能产业，打造人工智能产业集群以及深入赋能传统行业成为重点。

日本继制定《科学技术创新综合战略 2020》之后，于 2021 年 6 月发布“AI 战略 2021”，致力于推动人工智能领域的创新创造计划，全面建设数字化政府。日本将基础设

施建设和人工智能应用作为重点，提出加快建设相关基础设施，重点强调了跨行业的数据传输平台以及人工智能相关标准等，全面推动人工智能在医疗、农业、交通物流、智慧城市、制造业等各个行业开展应用，并加大对中小企业的支援 [136]。

2022 年开始， OpenAI 发布的语言大模型 ChatGPT 由于能够通过自然语言交互完成多种任务，具备了多场景、多用途、跨学科的任务处理能力引发了社会的广泛关注。2023 年随着 GPT-4 的成功上市，语言大模型对于多模态领域也产生了重要影响，它从单调的文本交互，升级为可以接受文本与图像组合的多模态输入，相比传统的单模态大模型，多模态大模型更加符合人类的多渠道感认知方式，能够应对更加复杂丰富的环境、场景和任务。GPT-4 表明在多模态大模型中引入基于人类知识的自然语言能够带来模型在多模态理解、生成、交互能力上的大幅度提升。

以 ChatGPT 为代表的生成式人工智能，是人工智能技术的一种，它可以用来生成文本、图像、音频和视频，这些内容可以用来解决一些非常复杂的问题，也可以用来提高工作效率，有助于人类更好地理解世界，并创造出更多的价值。生成式人工智能之所以在最近几年可以飞速发展是由于深度学习的不断发展，社会数据的巨增，高效算力的普及化以及其它相关技术持续创新。基于基础大模型，人工智能领域正在经历从感知、理解、生成、创造的飞跃。

在未来的十年，人工智能未来会以可信安全人工智能为目标，以算法、算力，数据为核心，以开放公平框架为载体，以多元生态为驱动力，以法律制度为指导，全面赋能人类社会的智能化发展。生态参与方在都积极在算力可用性，数据准确性，模型泛化性，模型解释性，隐私安全性，模型适配性，模型可扩展性，模型高效性，社会伦理性，社会环保性这些领域进行发展和突破。

加速大模型及技术创新

提升模型泛化性、解释性和适配性，同时注重隐私安全和环保性，是人工智能未来发展的关键。我们需要加速大模型及技术创新，以突破模型规模、复杂性、跨模态学习、自学习能力、长期记忆和时间建模、适应性和鲁棒性等方面的限制。此外，我们还需要从算法、算力和数据各领域解决社会伦理性、社会环保性、隐私安全性等问题，确保人工智能大模型的可持续发展和社会价值。这将为各个领域的应用带来更高的性能、更广泛的应用场景和更好的用户体验。

赋能人工智能产业应用

生成式人工智能已经在零售，教育、医疗、汽车、金融、公共服务等行业有许多应用的实践。

零售行业：通过提供有针对性的信息、促销和建议，可以提高客户参与度和忠诚度，及时吸引消费者。同时，通过缺陷跟踪在潜在缺陷发生之前进行检测，以提高产品质量并减少浪费，从而提升与客户互动以改善购物体验的策略。

教育行业：个性化学习已成为教育数字化颠覆的首要任务。首席信息官正在寻求将生成性人工智能纳入其机构的业务方面向领导提供建议。此外，学生支持服务和咨询问题也是教育行业的重要应用。

医药行业：研发团队正在评估药物发现和开发的加速，人工智能正在接受研究、临床试验和数据培训，以识别关系和潜在的新药。

医疗行业：评估聊天机器人和应用程序以提供医疗信息和治疗建议的简单语言描述，通过分析患者数据（如病史和生活方式因素）来个性化医疗，帮助确定每个患者最有效的

治疗方案，包括个性化药物剂量和靶向疗法。这些都是人工智能在各个领域的应用实践，将为行业带来巨大的价值

科技行业：利用自动化错误检测、代码生成和测试等任务，科技行业致力于简化软件开发过程。人工智能算法可以从过去的软件开发项目中学习，并利用这些知识来改善未来的开发周期。此外，实时检测和响应网络威胁也正在改善网络安全。通过分析大量数据，如网络流量日志，以识别可能表明安全漏洞的行为。

汽车行业：汽车行业正在利用大量来自传感器和其他来源的数据来开发和部署自动驾驶车辆。通过实时分析这些数据，汽车制造商可以做出决策，例如检测和避免障碍或预测交通模式。此外，分析需求、库存水平和供应商绩效数据也被用于优化汽车供应链。购买体验也得到了增强，通过虚拟和语音功能。呼叫中心、召回和索赔处理也得到了协助。

金融行业：通过分析大量实时交易数据来提高欺诈检测和预防能力。银行正在考虑通过分析客户数据（如交易历史记录、支出模式和通信偏好）来提供更个性化的客户服务。银行通过分析市场趋势和财务数据来提高其投资管理能力。

尽管目前只有 10%的企业已应用生成式人工智能技术，但随着时间的推移，人工智能将与各行各业深度结合，创造出巨大的经济社会价值。为了应对这一趋势，我们需要加快人工智能基础设施的建设，并建立人工智能的生态。

加快人工智能基础设施

我们需要布局人工智能相关的基础设施，包括算法、算力和数据。业界普遍认为，算法、算力和数据是人工智能发展的三大支柱，也是推动人工智能技术创新始终的主旋律。因此，各地政府可以开始规划如何整合各方资源，建设以“算法框架、算力资源，数据资源”为核心能力，以“开放可信平台”为主要赋能载体的公共普惠的智能化服务的基础设

施。这将促进政府、事业单位以及企业在人工智能领域的使用，同时带动人工智能生态上游基础设施、人工智能大模型、框架的发展。

建立人工智能的生态

此外，我们还需要建立人工智能的生态，包括培养人才、推动研究、促进合作等。这将有助于推动人工智能技术的持续创新和应用，为社会带来更多的福祉。人工智能生态是一个涵盖基础设施层、服务组件层、生态应用层的复杂系统，其中的相关参与方在其中发挥着重要的作用。基础设施层，以预训练模型为基础搭建实现框架的基础核心功能，包括编程开发、编译优化以及硬件三个子层，位于产业上游，具有较高的进入门槛，主要由头部科技企业、科研机构等进入。服务组件层，提供 AI 模型生命周期的可配置高阶功能组件，实现细分领域性能的优化提升，支持开发者以更灵活的姿态支持人工智能模型训练、应用适配。随着兼具大模型和多模态模型的 AIGC 模型加速成为新的技术平台，新型的商业模式模型即服务已经开始推出市场。生态应用层，用以支持基于 AI 框架开发的各种人工智能模型的应用、维护和改进，为个体用户就细分场景使用人工智能提供更好的产品和服务，位于产业下游。政府在建立人工智能产业生态中扮演着重要的角色，可以积极与顶尖高校、科技企业、各行各业企业、开源社区等合作，储备强大的人才资源，建立创新科研实验室，优化模型算法，开发 AI 框架，推广人工智能的商业化应用。从基础层的模型到服务层的模型工具，再到应用层的产品和服务，促进生态合作，循环式带动产业发展，为未来提供源源不断的动力。

制定人工智能的政策和治理

制定人工智能的政策和治理也是至关重要的。人工智能相关的治理工作将影响人工智能的持续健康发展，关乎未来社会如何更安全地使用人工智能。因此，需要建立完善的政

策体系，规范人工智能的发展，确保其安全、可靠和可持续性。通过加强监管、制定标准、鼓励创新和合作等方式，可以推动人工智能的健康和可持续发展，为社会带来更多的利益和价值。各国和地区由于文化背景、经济发展和生态环境的差异，需要结合自身国情和技术发展情况，制定符合人工智能的合规标准、知识产权与数据权益保护规则、大模型的研发、训练和部署指南、安全要求、最佳实践、风险管理、伦理审查评估以及大模型能力评估方法。这有助于行业组织、研究机构和企业等在人工智能治理方面先行先试，将安全可信的理念贯穿于人工智能的全生命周期。这也有助于行业组织、研究机构和企业建立健全的治理机制和风险管理体系，推动更多实践范式的创新发展。

虽然近年生成人工智能展现出颠覆性特质，有彻底改变现有的经济和社会框架的潜力，但是目前依然处于早期发展阶段。全社会都面临的共同问题是如何善用人工智能来服务人类，支持人类社会发展。政府，行业机构，企业，以及个体对于这个问题会有不同的视角和关注点。全球各界人员根据各自关注点共同参与到其发展，提高生成样本的质量，多样性，实用性，生成式人工智能与其他深度学习模型结合性。生成式人工智能的可解释性问题也会有改善，以提高模型的可靠性和可应用性。面向各个特定领域的应用的生成式人工智能也将在更加符合伦理和法律的基础上可持续快速发展以满足不同领域的需求。

八 参考文献

[1] 赵竹青, “AI 观察 | 政府工作报告首提“人工智能+”有何深意？,” 人民网, 9 3 2024. [联机]. Available: http://finance.people.com.cn/n1/2024/0309/c1004-40192366.html.

[2] M. G. China, “生成式 AI 在中国：2 万亿美元的经济价值,” [联机]. Available: https://www.mckinsey.com.cn/生成式 ai 在中国：2 万亿美元的经济价值.

[3] IBM, “《2022 年全球 AI 采用指数》,” 13 6 2022. [联机]. Available: https://china.newsroom.ibm.com/2022-06-13-IBM-2022-AI-AI-.

[4] 百度微信公众号, “《6000 万、4500 万和 10 亿，百度世界 2023 重磅发布都在这里了！》,” 18 10 2023. [联机]. Available: https://mp.weixin.qq.com/s/Y7BELEKaSQaeuXcItgxrnA.

[5] “US federal AI governance: Laws, policies, and strategies,” 6 2023. [联机]. Available: https://iapp.org/resources/article/us-federal-ai-governance/.

[6] “Europe’s world-first AI rules get final approval from lawmakers. Here’s what happens next,” 3 2024. [联机]. Available: https://apnews.com/article/ai-act-european-union-chatbots-155157e2be2e42d0f1acca33983d8c82.

[7] “Artificial intelligence act: MEP adopt landmark law,” 3 2024. [联机]. Available: https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law.

[8] “Unleashing the AI Imagination: A Global Overview of Generative AI Regulations,” 11 8 2023. [联机]. Available: https://www.pillsburylaw.com/en/news-and-insights/ai-regulations-us-eu-uk-china.html.

[9] IBM, “《IBM 董事长兼首席执行官观点：如何推进可信的人工智能》,” 27 9 2023. [联机]. Available: https://china.newsroom.ibm.com/2023-09-27-IBM.

[10] IBM 商业价值研究院, “《数据故事：生成式 AI 市场现状》,” 7 2023. [联机]. Available: https://www.ibm.com/downloads/cas/7KX4Q06G.

[11] IBM 商业价值研究院, “《数据故事：生成式 AI 解析企业优先事项》,” 7 2023. [联机]. Available: https://www.ibm.com/downloads/cas/DJAN8YO7.

[12] IBM 商业价值研究院, “《AI 时代的 CEO 决策力》,” 8 2023. [联机]. Available: https://www.ibm.com/downloads/cas/1YOKEERG.

[13] 第一财经, “《全球算力指数评估：生成式 AI 市场潜力大，制造业增幅大》,” 13 7 2023. [联机]. Available: https://new.qq.com/rain/a/20230713A05IC000.

[14] “2023 IBM Institute for Business Value generative AI integrity and compliance pulse survey. 200 US CxOs,” 9. [联机]. Available: 2023.

[15] IBM 商业价值研究院, “《企业生成式 AI 市场现状》,” 7 2023. [联机]. Available: https://www.ibm.com/downloads/cas/ZJLX7LP6.

[16] “2023 IBM Institute for Business Value generative AI impact on hybrid cloud pulse survey. 414 US CxOs,” 2023. [联机].

[17] I. I. f. B. Value, “AI ethics in action: An enterprise guide to progressing trustworthy AI,” 4 2022. [联机].

[18] Cisco, “CISCO 2022 Consumer Privacy Survey.,” 9 2 2023. [联机]. Available: https://www.cisco.com/c/dam/en_us/about/doing_business/trust-center/ docs/cisco-consumer-privacy-survey-2022.pdf.

[19] J. C. F. C.-G.-W. C. N. G. O. a. S. P. I. I. f. B. V. Cheung, “Balancing sustainability and profitability: How businesses can protect people, planet, and the bottom line,” 4 2022. [联机]. Available: https://ibm.co/2022-sustainability-consumer-research.

[20] I. I. f. B. Value, “CEO decision-making in the age of AI: Act with intention,” 6 2023. [联机]. Available: https://ibm.co/c-suite-study-ceo.

[21] I. I. f. B. Value, “Generative AI impact on labor pulse survey. 300 US CxOs,” 2023. [联机].

[22] “Pytorch,” [联机]. Available: https://pytorch.org/.

[23] “Tensorflow,” [联机]. Available: https://www.tensorflow.org/.

[24] “Keras,” [联机]. Available: https://keras.io/.

[25] “Transformers,” [联机]. Available: https://huggingface.co/docs/transformers/en/index.

[26] “Ray Docs,” [联机]. Available: https://docs.ray.io/en/latest.

[27] “Colossal-AI,” [联机]. Available: https://colossalai.org/.

[28] “DeepSpeed: Extreme Speed and Scale for DL Training and Inference,” [联机]. Available: https://www.microsoft.com/en-us/research/project/deepspeed/.

[29] “Kubeflow: The Machine Learning Toolkit for Kubernetes,” [联机]. Available: https://www.kubeflow.org/.

[30] “Caikit: an AI toolkit that enables users to manage models through a set of developer friendly APIs.,” [联机]. Available: https://github.com/caikit/caikit.

[31] “NVIDIA Triton Inference Server,” [联机]. Available: https://www.nvidia.com/en-us/ai-data-science/products/triton-inference-server/.

[32] “NVIDIA TensorRT,” [联机]. Available: https://docs.nvidia.com/tensorrt/index.html.

[33] “Vector database,” [联机]. Available: https://en.wikipedia.org/wiki/Vector_database.

[34] K. Martineau, “What is retrieval-augmented generation?,” [联机]. Available: https://research.ibm.com/blog/retrieval-augmented-generation-RAG.

[35] R. Kundu, “F1 Score in Machine Learning: Intro & Calculation,” [联机]. Available: https://www.v7labs.com/blog/f1-score-guide.

[36] Z. a. J. R. Guo, “Evaluating large language models: A comprehensive survey,” *arXiv preprint arXiv:2310.19736,* 2023.

[37] J. H. N. L. J. B. Alon Talmor, “CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge,” [联机]. Available: https://aclanthology.org/N19-1421/.

[38] H. R. D. C. R. L. B. Y. C. Maarten Sap, “Social IQA,” [联机]. Available: https://allenai.org/data/socialiqa.

[39] W. a. J. Z. Yu, “Reclor: A reading comprehension dataset requiring logical reasoning,” *arXiv preprint arXiv:2002.04326,* 2020.

[40] J. a. C. L. Liu, “Logiqa: A challenge dataset for machine reading comprehension with logical reasoning,” *arXiv preprint arXiv:2007.08124,* 2020.

[41] “LSAT dataset,” [联机]. Available: https://search.r-project.org/CRAN/refmans/testforDEP/html/LSAT.html.

[42] W. a. Z. H. Chen, “Hybridqa: A dataset of multi-hop question answering over tabular and textual data,” *arXiv preprint arXiv:2004.07347,* 2020.

[43] “MTEB Leaderboard - a Hugging Face Space by mteb,” *https://huggingface.co/spaces/mteb/leaderboard.*

[44] “AI 大模型的训练数据来源详解,” [联机]. Available: https://www.sohu.com/a/747117709_828277.

[45] “什么是数据增强,” [联机]. Available: https://aws.amazon.com/cn/what-is/data-augmentation/.

[46] Anthropic, “HH-RLHF,” [联机]. Available: https://huggingface.co/datasets/Anthropic/hh-rlhf.

[47] W. X. a. Z. K. Zhao, “A survey of large language models,” *arXiv preprint arXiv:2303.18223,* 2023.

[48] X. a. L. J. Zhu, “A survey on model compression for large language models,” *arXiv preprint arXiv:2308.07633,* 2023.

[49] “大语言模型量化简介,” [联机]. Available: https://www.bilibili.com/video/BV1zm4y1u72W/ .

[50] J. a. T. Y. Wei, “Emergent abilities of large language models,” *arXiv preprint arXiv:2206.07682,* 2022.

[51] “GPU Performance Background User's Guide,” *https://docs.nvidia.com/deeplearning/performance/dl-performance-gpu-background/index.html#understand-perf.*

[52] G.-I. a. J. J. S. Yu, “Orca: A distributed serving system for Transformer-Based generative models,” 出处 *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, 2022, pp. 521--538.

[53] “Comparing LLM serving frameworks — LLMOps,” [联机]. Available: https://medium.com/@plthiyagu/comparing-llm-serving-frameworks-llmops-f02505864754..

[54] J. a. W. X. Wei, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems,* 卷 35, pp. 24824--24837, 2022.

[55] D. a. S. N. Zhou, “Least-to-most prompting enables complex reasoning in large language models,” *arXiv preprint arXiv:2205.10625,* 2022.

[56] “什么是 LangChain？,” [联机]. Available: https://www.ibm.com/cn-zh/topics/langchain.

[57] “Mini-Chain: A tiny library for large language models.,” [联机]. Available: https://srush.github.io/MiniChain/#why-mini-chain.

[58] “什么是多模态机器学习？,” [联机]. Available: https://blog.csdn.net/electech6/article/details/85142769.

[59] S. a. F. C. Yin, “A survey on multimodal large language models,” *arXiv preprint arXiv:2306.13549,* 2023.

[60] “Our next-generation enterprise studio for AI builders,” [联机]. Available: https://www.ibm.com/products/watsonx-ai.

[61] “Sample foundation model prompts for common tasks,” [联机]. Available: https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/fm-prompt-samples.html?context=wx&audience=wdp.

[62] “Guiding Llama 2 with prompt engineering by developing system and instruction prompts,” [联机]. Available: https://developer.ibm.com/tutorials/awb-prompt-engineering-llama-2/.

[63] “Tuning Studio,” [联机]. Available: https://www.ibm.com/docs/en/watsonx/w-and-w/1.1.x?topic=solutions-tuning-studio.

[64] AWS, “数据块、对象和文件存储有什么区别,” [联机]. Available: https://aws.amazon.com/cn/compare/the-difference-between-block-file-object-storage/.

[65] IBM, “Apache Avro,” [联机]. Available: https://www.ibm.com/cn-zh/topics/avro.

[66] Apache, “Apache orc,” [联机]. Available: https://orc.apache.org/docs/.

[67] AWS, “AWS Big Data Blog,” [联机]. Available: https://aws.amazon.com/blogs/big-data/choosing-an-open-table-format-for-your-transactional-data-lake-on-aws.

[68] Apache, “Apache Hive,” [联机]. Available: https://hive.apache.org/.

[69] IBM Torsten Steinbach, “IBM’s Metastore aaS,” [联机]. Available: https://www.ibm.com/blog/ibms-metastore-aas-there-is-no-lake-without-metadata/.

[70] F. Baradari, “Data Virtualization and the U.S. Federal Data Strategy,” [联机]. Available: https://www.datamanagementblog.com/data-virtualization-us-federal-data-strategy/.

[71] Alluxio, “AN INTRODUCTION TO THE PRESTO ARCHITECTURE,” [联机]. Available: https://www.alluxio.io/learn/presto/architecture/.

[72] Apache, “Apache spark cluster overview,” [联机]. Available: [3] https://spark.apache.org/docs/latest/cluster-overview.html.

[73] Apache Drill, “Apache Drill,” [联机]. Available: https://drill.apache.org/.

[74] IBM Kim Martineau, “What is retrieval-augmented generation,” [联机]. Available: https://research.ibm.com/blog/retrieval-augmented-generation-RAG.

[75] Y. Wu, “Why You Shouldn’t Invest In Vector Databases,” [联机]. Available: https://blog.det.life/why-you-shouldnt-invest-in-vector-databases-c0cd3f59d23c.

[76] Chroma, “Chroma v0.4,” [联机]. Available: https://www.trychroma.com/blog/chroma_0.4.0.

[77] “kubernetes.io,” [联机]. Available: https://kubernetes.io/.

[78] Red Hat, “《OpenShift Container Platform 架构概述》,” Red Hat, [联机]. Available: https://access.redhat.com/documentation/zh-cn/openshift_container_platform/4.3/html/architecture/architecture.

[79] 周立旸, “IBM AI 存储：算力稀缺时代的"破局者",” IBM, [联机]. Available: https://china.newsroom.ibm.com/2023-11-21-IBM-AI.

[80] P. N. C. C. D. R. Talia Gershon, “A cloud-native, open-source stack for accelerating foundation model innovation,” IBM, [联机]. Available: https://research.ibm.com/blog/openshift-foundation-model-stack.

[81] “multi-nic-cni,” [联机]. Available: https://github.com/foundation-model-stack/multi-nic-cni.

[82] “NVIDIA GPU Operator,” [联机]. Available: https://github.com/NVIDIA/gpu-operator/tree/master.

[83] “Multi-Cluster App Dispatcher,” [联机]. Available: https://github.com/project-codeflare/multi-cluster-app-dispatcher.

[84] “instascale,” [联机]. Available: https://github.com/project-codeflare/instascale.

[85] S. S. J. J. E. G. D. T. Talia Gershon, “Why we built an AI supercomputer in the cloud,” IBM, [联机]. Available: https://research.ibm.com/blog/AI-supercomputer-Vela-GPU-cluster.

[86] I. Research, “codeflare.dev,” IBM, [联机]. Available: https://codeflare.dev/.

[87] “智能时代的生产力变革：AIGC 产业应用实践,” [联机]. Available: https://www.sohu.com/a/766527769_121924299.

[88] “IBM 陈旭东在腾讯产业合作伙伴大会的发言：携手共创 AI 新生态,” [联机]. Available: https://www.kjnews.org/238.html?rkey=20240118ZH15160&filter=21254.

[89] IBM 商业价值研究院, “《CEO 生成式 AI 行动指南———开放创新和生态系统》,” 11 2023. [联机]. Available: https://www.ibm.com/downloads/cas/PRQ1DPEV.

[90] IBM 商业价值研究院, “《专家洞察：扩展 AI 的公认概念》,” 9 2020. [联机]. Available: https://www.ibm.com/downloads/cas/M4GLJV1B.

[91] “IBM 和 Meta 与 50 多个创始成员及协作者成立 AI 联盟,” [联机]. Available: https://china.newsroom.ibm.com/2023-12-07-IBM-Meta-50-AI.

[92] “AI 联盟网站,” [联机]. Available: https://thealliance.ai/.

[93] “中国信通院启动工业和信息化企业合规典型案例征集,” [联机]. Available: http://www.caict.ac.cn/xwdt/ynxw/202308/t20230828_460565.htm.

[94] 世界互联网大会人工智能工作组, “《2023 年发展负责任的生成式人工智能研究报告及共识文件》”.

[95] IBM 商业价值研究院, “《专家洞察：借助 AI 驱动的工作流程，建立供应链弹性》,” 11 2020. [联机]. Available: https://www.ibm.com/downloads/cas/JRARGNBD.

[96] IBM 商业价值研究院, “《CEO 生成式 AI 行动指南———供应链》,” 11 2023. [联机]. Available: https://www.ibm.com/downloads/cas/4VQ1B3L5.

[97] IBM 商业价值研究院, “《研究洞察：把握 AI 和自动化的机遇》,” 8 2023. [联机].

[98] IBM 商业价值研究院, “《CEO 生成式 AI 行动指南———营销》,” 11 2023. [联机]. Available: https://www.ibm.com/downloads/cas/OZN8PNVA.

[99] IBM 商业价值研究院, “《CEO 生成式 AI 行动指南———利用生成式 AI 推动变革》,” 12 2023. [联机].

[100] IBM, “AI Ethics,” IBM, [联机]. Available: https://www.ibm.com/impact/ai-ethics.

[101] “Ethics by Design and the AI Lifecycle,” IBM, [联机]. Available: https://www.ibm.com/docs/en/cpod?topic=ebd-ethics-by-design-ai-lifecycle.

[102] J. J. Thomas, “Foundations of trustworthy AI: Operationalizing trustworthy AI,” IBM, [联机]. Available: https://www.ibm.com/blog/operationalizing-trustworthy-ai/.

[103] “《可解释机器学习》人工智能汇集,” COPU, [联机]. Available: https://gitcode.net/COPU/copu/-/blob/master/static/images/《可解释机器学习》人工智能汇集（电子版）.pdf.

[104] IBM, “AIX360,” IBM, [联机]. Available: https://aix360.res.ibm.com/resources#guidance.

[105] IBM, “AIX360 算法选择树,” [联机]. Available: https://github.com/Trusted-AI/AIX360/blob/master/aix360/algorithms/README.md.

[106] R. M. M. ,. I. J. ,. K. K. ,. F. R. ,. a. A. K. A. Chatzimparmpas1, “The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations,” *STAR – State of The Art Report,* 2020.

[107] “ART360,” IBM, [联机]. Available: https://art360.res.ibm.com/resources#overview.

[108] IBM, “AI Fairness 360,” IBM Research, [联机]. Available: https://aif360.res.ibm.com/resources#guidance.

[109] IBM, “AI Privacy 360,” IBM, [联机]. Available: https://aip360.res.ibm.com/resources.

[110] J. Holdsworth, “What is model drift?,” IBM, [联机]. Available: https://www.ibm.com/topics/model-drift.

[111] IBM, “AIX360 项目资源,” IBM, [联机]. Available: https://aix360.res.ibm.com/resources#guidance.

[112] 大数据安全标准特别工作组, “人工智能安全标准化白皮书,” 全国信息安全标准化技术委员会, 2019. [联机]. Available: https://www.tc260.org.cn/file/rgznaqbz.pdf.

[113] IBM Research, “AI FactSheets 360,” IBM, [联机]. Available: https://aifs360.res.ibm.com/.

[114] “InstructLab,” [联机]. Available: https://github.com/instructlab/taxonomy.

[115] A. B. A. P. K. X. D. D. C. A. S. Shivchander Sudalairaj, “LAB: Large-Scale Alignment for ChatBots,” [联机]. Available: https://arxiv.org/abs/2403.01081.

[116] K. Martineau, “A faster, systematic way to train large language models for enterprise,” IBM, [联机]. Available: https://research.ibm.com/blog/LLM-generated-data.

[117] huggingface, “Evaluate,” [联机]. Available: https://huggingface.co/docs/evaluate.

[118] IBM, “Evaluating AI models,” [联机]. Available: https://dataplatform.cloud.ibm.com/docs/content/wsj/model/getting-started.html?context=cpdaas&audience=wdp&locale=en.

[119] IBM, “OpenPages on Cloud Pak for Data,” [联机]. Available: https://www.ibm.com/docs/en/cloud-paks/cp-data/4.8.x?topic=services-openpages.

[120] 《主数据管理实践白皮书》，中国信通院, 2018/12.

[121] 国家标准化管理委员会, GB/T 36073-2018 数据管理能力成熟度评估模型，中华人民共和国国家质量监督检验检疫总局、中国国家标准化管理委员会, 2018-03-15.

[122] 数据管理协会（DAMA 国际）, DAMA 数据管理知识体系指南[M]，北京：机械工业出版社, 2020.

[123] DAMA China 成于念，“企业数据质量管理考核评分实践,” [联机]. Available: http://www.dama.org.cn/wordpress/2023/04/06/成于念企业数据质量管理考核评分实践/.

[124] IBM, “什么是数据生命周期管理,” [联机]. Available: https://www.ibm.com/cn-zh/topics/data-lifecycle-management.

[125] Amundsen Project, “amundsen,” [联机]. Available: https://www.amundsen.io/amundsen/.

[126] datahubproject.io, [联机]. Available: https://datahubproject.io/docs/features.

[127] OpenMetadata, “github,” [联机]. Available: https://github.com/open-metadata/OpenMetadata.

[128] Apache Atlas, [联机]. Available: https://atlas.apache.org/#/.

[129] D. S. G. L. ,. J. L. S. J. ,. C. N. Ajuma Bella Salifu, “Observe GenAI with IBM Instana Observability,” IBM, [联机]. Available: https://www.ibm.com/blog/announcement/genal-llm-observability.

[130] IEA, “Electricity 2024,” [联机]. Available: https://www.iea.org/reports/electricity-2024.

[131] kepler, [联机]. Available: https://github.com/sustainable-computing-io/kepler.

[132] CNCF, “TAG Environmental Sustainability,” [联机]. Available: https://github.com/cncf/tag-env-sustainability.

[133] IBM，中国信通院, “可持续计算蓝皮报告（2022 年）,” [联机]. Available: http://www.caict.ac.cn/kxyj/qwfb/ztbg/202301/t20230105_413693.htm.

[134] IBM 商业价值研究院, “《研究洞察：把握 AI 和自动化的机遇》,” 2023.

[135] 红杉中国, “2023 企业数字化年度报告,” 2023.

[136] 中国信通院, “人工智能白皮书 2022,” [联机]. Available: http://www.caict.ac.cn/kxyj/qwfb/bps/202204/t20220412_399752.htm.

[137] “Foundation models in watsonx.ai,” [联机]. Available: https://www.ibm.com/products/watsonx-ai/foundation-models.

附录一 watsonx.ai 基础模型库

模型名称	提供商	用例	上下文长度
granite-7b-lab	IBM	支持问题解答 (Q&A)、摘要、分类、生成、提取和 RAG 任务	8128
granite-13b-chat	IBM	支持问题解答 (Q&A)、摘要、分类、生成、提取和 RAG 任务	8192
granite-13b-instruct	IBM	支持问题解答 (Q&A)、摘要、分类、生成、提取和 RAG 任务	8192
granite-20b-multilingual	IBM	支持法语、德语、葡萄牙语、西班牙语和英语的问题解答 (Q&A)、摘要、分类、生成、提取、翻译和 RAG 任务	8190
granite-8b-japanese	IBM	支持日语的问题解答 (Q&A)、摘要、分类、生成、提取、翻译和 RAG 任务	4096
llama-3-8b-instruct	Meta	支持摘要、分类、生成、提取和 翻译任务	8192
llama-3-70b-instruct	Meta	支持 RAG、生成、摘要、分类、问题解答（Q&A）、提取、翻译和代码生成任务	8192
llama-2-70b-chat	Meta	支持问题解答 (Q&A)、摘要、分类、生成、提取和 RAG 任务	4096
llama-2-13b-chat	Meta	支持问题解答 (Q&A)、摘要、分类、生成、提取和 RAG 任务。可用于提示调优。	4096
llama2-13b-dpo-v7 (Korean)	MindsAndCompany	支持韩语的问题解答 (Q&A)、摘要、分类、生成、提取和 RAG 任务	4096
codellama-34b-instruct	Meta	从自然语言提示生成和翻译代码的特定任务模型	16384

mixtral-8x7b-instruct	Mistral AI	支持问题解答 (Q&A)、摘要、分类、生成、提取、RAG 和代码生成任务。	32768
merlinite-7b	ibm-mistralai	支持问题解答 (Q&A)、摘要、分类、生成、提取、RAG 和代码生成任务。	32768
jais-13b-chat (Arabic)	core42	支持阿拉伯语的问题解答 (Q&A)、摘要、分类、生成、提取和翻译任务	2048
flan-t5-xl-3b	Google	支持问题解答 (Q&A)、摘要、分类、生成、提取和 RAG 任务。可用于提示调优。	4096
flan-t5-xxl-11b	Google	支持问题解答 (Q&A)、摘要、分类、生成、提取和 RAG 任务。	4096
flan-ul2-20b	Google	支持问题解答 (Q&A)、摘要、分类、生成、提取和 RAG 任务。	4096
elyza-japanese-llama-2-7b-instruct	ELYZA	支持问题解答 (Q&A)、摘要、RAG、分类、生成、提取和翻译任务	4096
mt0-xxl-13b	BigScience	支持问题解答 (Q&A)、摘要、分类、生成任务	4096
starcoder-15.5b	BigCode	从自然语言提示生成和翻译代码的特定任务模型	8192

* 出处： [137]

* 截止 2024-05-08

附录二 人工智能指标

可信AI特性	指标名	指标含义	传统AI可用	LLM可用	LLM任务	指标来源
可解释性	ProtoDash	数据解释类。通过选择一组代表性的样本或“原型（prototypes）”来提高模型的可解释性。算法的目标是最小化所选原型和数据集之间的距离。Protodash可以用来识别对模型预测影响最大的样本，还可以应用于特征选择。	Y	Y	RAG 相关任务（在 Retriever 检索到相关的文档后，Protodash 算法可以选择出代表性的“原型”文档，并计算文档权重。帮助理解 RAG 决策过程）	AIX360 ProtoDash for RAG
	Disentangled Inferred Prior VAE (DIP-VAE)	数据解释类。变分自编码器（VAE）的扩展，它用于学习解耦（disentangled）的隐含特征表示。DIP-VAE 学习到的隐变量之间相互独立，每个隐变量对应于输入数据的一个特定的、可解释的特征。	Y			AIX360
	Contrastive Explanations Method (CEM)	模型解释类。一种用于生成对抗性和对比性解释的算法。借助保留模型决策不变的最小特征集合（PP）和不改变输入太多的情况下，能够改变模型决策的最小特征集合（PN），帮助理解哪些特征对模型决策最为重要。	Y			AIX360
	Contrastive Explanations Method with Monotonic Attribute Functions (CEM-MAP)	模型解释类。在 CEM 的基础上，进一步考虑输入特征的单调性。可用于解释模型的预测结果是否会随着某些特定属性的增加或减少而单调增加或减少。	Y			AIX360
	LIME	模型解释类。围绕目标预测生成一个新的、局部的数据集，然后在这个局部数据集上训练一个简单的模型（如线性回归或决策树），用这	Y			AIX360

		个简单模型能够模拟原始复杂模型在局部的行为，以帮助理解原模型。				
	SHAP	模型解释类。基于博弈论中的 Shapley 值来解释各个特征对模型预测的贡献。	Y			AIX360
	Teaching AI to Explain its Decisions (TED)	模型解释类。提供了一个训练框架，使模型不仅输出决策，还输出用户能够理解且相关的解释。	Y			AIX360
	ProfWeight	模型解释类。根据给定的可解释模型和高性能复杂神经网络，对训练集进行重新加权学习。	Y			AIX360
	Generalized Linear Rule Models (GLRM)	模型解释类。在 GLRM 中，模型由一系列的逻辑规则组成，这些规则描述了输入特征与目标变量之间的关系。每个规则都与一个线性系数相关联，这些系数一起确定了输入如何影响输出。	Y			AIX360
	Boolean Decision Rules via Column Generation (BRCG)	模型解释类。该算法基于线性规划中常用的列生成方法来优化决策规则。算法迭代地添加最有用的布尔特征（或规则），以提高模型的预测准确性同时保持模型的可解释性。	Y			AIX360
	Faithfulness（忠实度）	描述解释或特征能够正确反映模型行为的程度。一个忠实的解释应当能够准确地反映出模型对特定输入作出预测决定的真实原因。具体实现方式：解释是否与模型的输出紧密相关；改变解释认为重要的特征时，模型输出是否会相应变化。	Y	Y		AIX360
	Monotonicity （单调性）	模型输出与一个或多个输入特征之间保持一致的单调关系。如果模型是单调的，那么当一个输入特征的值增加（或减少）时，输出也会相应地增加（或减少）。	Y			AIX360
公平性	Selection rates and error rates including rich subgroup	衡量不同群体（如不同种族、性别等）获得正面预测（如获得贷款批准）的比率是否相等。以及检测假阳性率和假阴性率等指标需要在不同群体中保持一致，以避	Y			AIF360

	fairness 选择率，错误率和子组公平性	免对某些群体产生不利影响。针对可能存在交叉的身份（如种族与性别的交叉）也要保证比例一致。				
	Sample distortion metrics	评估训练数据是否存在扭曲或偏差。确定数据集中是否存在代表性不足的群体或过度代表的群体，从而导致模型在应用时对某些群体产生偏见。	Y			AIF360
	Generalized Entropy Index	通过比较个体结果与平均结果的差异来描述数据分布的不平等程度。	Y			AIF360
	Differential Fairness and Bias Amplification	确保对于所有可能的敏感属性组合（如种族与性别的交叉），模型的行为在统计上是一致的，来减少对任何子群体的不公平对待。	Y			AIF360
	Bias Scan with Multi-Dimensional Subset Scan	通过高效地扫描多个维度（如年龄、性别、种族等），寻找在特定子群体中表现出显著统计偏差的模式。	Y			AIF360
健壮性	CLEVER	评估神经网络模型健壮性的指标。通过观察输入数据的微小变化可能导致的模型最大损失变化来预测模型对抗攻击的敏感性。	Y			ART
	Loss sensitivity	描述模型输出对其输入的微小扰动或变化的敏感程度。通过测量输入的微小变动可能导致损失函数显著变化的程度来衡量模型对抗攻击的潜在脆弱性。	Y			ART
	Empirical robustness	衡量模型在真实世界条件下的表现，包括其处理噪声、损坏或应对对抗性输入数据的能力。该指标评估了模型在面对扰动时与理想状态相比的性能下降。	Y			ART
	Randomized Smoothing	一种增强深度学习模型健壮性的技术。通过对输入数据添加随机噪声，然后对噪声数据进行平滑或集成预测来增加模型的健壮性。	Y			ART
	Clique Method Robustness Verification	一种用于验证神经网络模型健壮性的方法。通过构建模型输入空间的“clique”（一组紧密连接的节点），	Y			ART

		来评估模型在面对小范围输入扰动时的输出稳定性。				
隐私性/安全性	PII	评估模型的输入数据，以及模型生成的输出中是否含有个人敏感信息	Y	Y	文本摘要，内容生成，问题解答	watsonx.governance
	HAP	评估模型的输入数据，以及模型生成的输出中是否含有任何有毒信息	Y	Y	文本摘要，内容生成，问题解答	watsonx.governance
模型质量	accuracy	准确率是指在已处理的全部预测中，预测正确的占比。	Y	Y		Hugging Face
	bertscore	利用 BERT 预训练上下文嵌入，通过余弦相似度计算候选句子和参考句子中的词语间距离。		Y		Hugging Face
	bleu	通过将翻译片段与一组参考译文进行比较，评估机器翻译文本从一种自然语言翻译成另一种自然语言时的质量。		Y	文本摘要，内容生成，问题解答	Hugging Face
	bleurt	用于自然语言生成的评估指标，通过多阶段的迁移学习构建。基于预训练 BERT 模型进一步训练和构建而成。		Y		Hugging Face
	brier_score	分类任务的一种评估指标。衡量两个概率分布之间误差。	Y	Y	分类任务	Hugging Face
	cer	字符错误率 (CER) 是衡量自动语音识别系统性能的常用指标。CER 类似于文字错误率 (WER)，但针对的是字符而不是文字。		Y	语音识别	Hugging Face
	character	评估机器翻译准确性的字符级指标。通过计算将机器翻译输出转换为参考翻译所需的最少字符级编辑次数（插入、删除、替换）来进行评估。		Y	机器翻译	Hugging Face
	charcut_mt	可将机器翻译的输出结果与参考译文进行比较。该匹配算法基于对最长公共子串的迭代搜索，并结合基于长度的阈值来限制短字符和噪声字符的匹配。		Y	机器翻译	Hugging Face
	chrf	使用字符 n-gram 来计算精确度和召回率，然后计算这些的 F-score 用于衡量翻译的准确性和完整性。		Y	机器翻译	Hugging Face

	code_eval	计算了在一组参考文献中预测结果的好坏。		Y		Hugging Face
	comet	使用一个预训练的跨语言句子嵌入模型（如 XLM-R）作为其基础，该模型能够生成不同语言之间的高质量句子表示。		Y	机器翻译	Hugging Face
	competition_math	该指标用于评估启发式数学能力测试（MATH）数据集的性能。它首先对输入进行规范化处理，然后计算准确率。		Y	MATH 数据集评估	Hugging Face
	confusion_matrix	评估分类的准确性。混淆矩阵中的每一行代表一个真实类别，每一列代表一个预测类别中的实例。	Y	Y		Hugging Face
	coval	是针对 CoNLL 和 ARRAU 数据集的核心参照评估工具，它实现了常用的评估指标。		Y	CoNLL 和 ARRAU 数据集	Hugging Face
	cuad	本指标对合同理解 Atticus 数据集 (CUAD) 第 1 版的官方评分脚本进行了包装。合同理解 Atticus 数据集 (CUAD) v1 是一个语料库，包含 510 份商业法律合同中的 13,000 多个标签，这些标签经过人工标注，以识别律师在审查与公司交易相关的合同时所关注的 41 类重要条款。		Y	合同理解 Atticus 数据集	Hugging Face
	exact_match	返回输入的预测字符串与其引用完全匹配的比率。	Y	Y		Hugging Face
	f1	F1 分数是精确度和召回率的调和平均值。	Y	Y		Hugging Face
	frugalscore	基于蒸馏方法，使用摘要、反向翻译和去噪模型构建的合成数据集上继续对小型模型进行预训练而获得。		Y		Hugging Face
	glue	通用语言理解评估基准是用于训练、评估和分析自然语言理解系统的资源集合。		Y	自然语言理解	Hugging Face
	google_bleu	BLEU 在用于单句时有一些不理想的特性，Google BLEU 对此做了优化。		Y	机器翻译	Hugging Face
	indic_glue	印度语言的自然语言理解任务基准。		Y	自然语言理解（印度语）	Hugging Face

	mae	平均绝对误差（MAE）是预测值与实际数值之间差值的平均值。	Y			Hugging Face
	mahalanobis	mahalanobis 距离度量了点与分布之间的距离。它常用于多元异常检测、高度不平衡数据集的分类和单类分类。		Y		Hugging Face
	mape	平均绝对误差（MAPE）是预测值与实际数值之间的百分比误差的平均值。	Y			Hugging Face
	mase	平均绝对缩放误差（MASE）是预测值的平均绝对误差除以样本内简单基准预测的平均绝对误差。	Y			Hugging Face
	matthews_correlation	matthews 相关系数在机器学习中被用来衡量二元分类和多类分类的质量。它考虑到了真假阳性和阴性，通常被认为是一种平衡的测量方法，即使分类的规模相差很大，也可以使用。	Y			Hugging Face
	mauve	MAUVE 是衡量生成文本与人类文本之间差距的指标。它使用大型语言模型量化嵌入空间中文本两个分布之间的 KL 散度来计算。		Y	内容生成	Hugging Face
	mean_iou	IoU 是预测的分割结果与真实结果之间的重叠面积除以预测的分割结果与真实结果之间的结合面积。		Y	图形分割，图像识别等	Hugging Face
	meteor	根据精确度和召回率的调和平均值计算，召回率的权重高于精确度。		Y	文本摘要，内容生成，机器翻译	Hugging Face
	mse	平均平方误差（MSE）表示误差平方的平均值，即估计值与实际值之间的平均平方差。	Y			Hugging Face
	nist_mt	DARPA 委托 NIST 开发基于 BLEU 分数的机器翻译任务评估工具。		Y	机器翻译	Hugging Face
	pearsonr	Pearson 相关系数可用于衡量两个数据集之间的线性关系。	Y			Hugging Face
	perplexity	在给定一个模型和一个输入文本序列的情况下，复杂度衡量的是模型生成输入文本序列的可能性。它被定义为		Y		Hugging Face

		一个序列的指数平均负对数似然值				
	poseval	poseval 指标可用于评估 POS 标记。它将数据集中的每个标记视为独立的观察结果，并计算精确度、召回率和 F1 分数，而与句子无关。		Y		Hugging Face
	precision	精确度是指在所有正例中，被正确标记为阳性的示例所占的比例。	Y			Hugging Face
	r_squared	R 方值为 1 表示模型完全解释了因变量的方差。R 方值为 0 表示模型不能解释任何方差。介于 0 和 1 之间的值表示模型解释因变量方差的程度。	Y			Hugging Face
	recall	召回率是指被模型正确标注为正例的样本占所有正例的百分比。	Y			Hugging Face
	rl_reliability	一套用于衡量强化学习（RL）算法可靠性的度量标准				Hugging Face
	roc_auc	ROC 曲线下面积。返回值表示所使用的模型根据输入数据预测正确类别的程度。得分为 0.5 意味着模型的预测完全是偶然的，即模型预测的正确率与掷硬币的正确率相同。得分高于 0.5 表示模型的表现好于随机掷硬币的概率，低于 0.5 则表示模型的表现低于随机掷硬币的概率。	Y			Hugging Face
	rouge	一套用于评估自动文本摘要和机器翻译软件的指标和软件包。将自动生成的摘要或翻译与参考文献（人工生成的）摘要或翻译进行比较。		Y	文本摘要，机器翻译，问题解答，命名实体识别等	Hugging Face
	sacrebleu	提供计算可共享、可比较和可复制的 BLEU 分数。		Y	文本摘要，内容生成，问题解答	Hugging Face
	sari	SARI 将预测的简化句子与参考句子和源句子进行比较，并明确衡量系统添加、删除和保留词语的好坏。		Y	文本摘要	Hugging Face
	seqeval	seqeval 是一个用于序列标记评估的 Python 框架。它可以评估命名实体识别、		Y	命名实体识别等	Hugging Face

		语音部分标记、语义角色标记等。				
	smape	对称平均绝对误差 (sMAPE) 是预测值与实际数值之间差异百分比误差的对称平均值。	Y			Hugging Face
	spearmanr	类似 Pearson 相关系数，可用于衡量两个数据集之间的线性关系。与 Pearson 相关性不同，Spearman 相关性并不要求两个数据集都是正态分布。	Y			Hugging Face
	squad	本指标封装了斯坦福问答数据集（SQuAD）第一版的官方评分脚本。		Y	SQuAD 数据集	Hugging Face
	squad_v2	本指标封装了斯坦福问答数据集（SQuAD）第 2 版的官方评分脚本。		Y	SQuAD 数据集	Hugging Face
	super_glue	SuperGLUE 是以 GLUE 为蓝本设计的新基准。该指标用于计算与 SuperGLUE 每个子集相关的评估指标。		Y	自然语言理解	Hugging Face
	ter	TER (Translation Edit Rate)，也称为 Translation Error Rate，用来评估机器翻译质量的指标。首先计算将机器翻译的输出修改为一个参考翻译所需的最少编辑操作数。然后将这个编辑数除以参考翻译的总词数，从而计算出错误率。		Y	机器翻译	Hugging Face
	trec_eval	TREC Eval 指标结合了许多信息检索指标，如精确度和归一化累积增益（nDCG）。它用于根据参考值对检索文档的排名进行评分。	Y			Hugging Face
	wer	词错误率（Word Error Rate，WER）是一种常用于评估语音识别系统或机器翻译系统性能的指标。它通过系统输出与参考文本进行比较，从而确定系统产生的错误的程度。WER 的计算方式是将编辑距离（Edit Distance）除以参考文本中的总词数，得到一个百分比值。		Y	语音识别，机器翻译	Hugging Face

	wiki_split	WikiSplit 是三个指标的组合： SARI、exact match 和 SacreBLEU。		Y		Hugging Face
	xnli	XNLI 指标可以评估模型在 XNLI 数据集上的得分，该数据集是从 MNLI 数据集中的几千个例子中挑选出来的一个子集。		Y	预测文本引申义（XNLI 数据集）	Hugging Face
	xtreme_s	XTREME-S 指标旨在评估模型在跨语言多语言编码器评估（XTREME-S）基准上的性能。		Y		Hugging Face
模型健康/模型性能	Number of scoring records (total / min/ max / median / avg)	记录模型部署期间接收到的预测请求数量。（总量、最大、最小、平均、中位数）	Y	Y		watsonx.governance
	Input token count (total / min/ max/ median/avg)	记录输入数据的 token 数量。（总量、最大、最小、平均、中位数）	Y	Y		watsonx.governance
	Output token count (total / min/ max/ median/avg)	记录输出内容的 token 数量。（总量、最大、最小、平均、中位数）	Y	Y		watsonx.governance
	API Latency (median /avg/min /max)	记录模型部署期间，API 调用模型的响应时间。（最大、 最小、平均、中位数）	Y	Y		watsonx.governance
	API Throughput (median /avg/min /max)	记录模型部署期间，每秒处理的 API 预测调用请求数量。（最大、 最小、平均、中位数）	Y	Y		watsonx.governance
	Record Latency (median /avg/min /max)	记录模型部署期间，处理 1 条数据的响应时间。（最大、 最小、平均、中位数）	Y	Y		watsonx.governance
	Record Throughput (median /avg/min /max)	记录模型部署期间，每秒处理的数据行数。（最大、 最小、平均、中位数）	Y	Y		watsonx.governance

	User count	用户数量	Y	Y		watsonx.governance

附录三　名词解释

- 2-dimensional：二维
- Actuators：执行器
- adapter tuning：适配器微调
- AGI：通用人工智能
- AI Alliance : AI 联盟
- AI Ladder：人工智能阶梯
- AIGC: Artificial Intelligence Generative Content，人工智能生成内容
- AIIA：中国人工智能产业发展联盟
- Alignment：对齐
- API: Application Programming Interface，应用程序编程接口
- Autonomous AI Agents：自主 AI Agents
- BAW：业务自动化工作流
- BI：商业智能
- Caikit：开源的人工智能工具包
- CCSA TC601：中国通信标准化协会大数据技术标准推进委员会
- CEO：首席执行官
- chain-of-thought prompting：思维链提示
- CMO：首席营销官
- CNN: convolutional neural networks，卷积神经网络
- Co-learning：协同学习
- Colossal-AI：分布式深度学习框架
- Continuous Batching：连续批处理
- CoT: Chain-of-thought，思维链
- CPU: Central Processing Unit，中央处理器
- Data Augmentation：数据增强
- DCMM：Data Management Capability Maturity Assessment Model，数据管理能力成熟度评估模型
- Decision-making mechanism：决策机制
- DeepSpeed：分布式深度学习优化库

- Discriminative AI：判别式 AI
- Emergent abilities：涌现能力
- Environment：环境
- GANs: generative adversarial networks，生成式对抗网络
- GDPR：通用数据保护条例
- Generative AI：生成式人工智能
- GPU: Graphics processing unit，图形处理器
- HDFS：Hadoop 分布式文件系统
- IBM IBV: IBM 商业价值研究院
- ICL: in-context learning，上下文学习
- IDC: Internet Data Center，互联网数据中心
- JAX：是一个面向加速器的数组计算和程序转换的 Python 库，专为高性能数值计算和大规模机器学习而设计
- Keras：基于 Python 的深度学习库
- KPI: Key Performance Indicators，关键绩效指标
- Kubeflow: Google 主导的一个开源项目
- LAVR：构建任务解决系统的通用框架
- least-to-most prompting：由少到多提示
- LLMs: large language models，大语言模型
- LoRA：低秩适配：
- M-ICL: Multimodal In-Context Learning，多模态上下文学习
- M-IT: Multimodal Instruction Tuning，多模态指令调优
- MAE: Mean absolute error，平均绝对误差
- MCoT: Multimodal Chain of Thought ，多模态思维链
- ML: Machine Learning，机器学习
- MLOpts: Machine Learning Operations，机器学习运维
- MS: mean-square error，均方误差
- Multimodal Fusion：多模态融
- Multimodal Representation：多模态表示学习
- NASA：美国国家航空航天局
- NLP: Natural Language Processing，自然语言处理

- offline stage：离线阶段
- online stage：在线阶段
- ONNX：Open Neural Network Exchange，开放神经网络交换（ONNX）是一个开放的生态系统，为人工智能模型（深度学习和传统机器学习）提供开源格式
- parameter-efficient fine-tuning：参数高效微调
- pipeline parallelism：流水线并行
- PPO: Proximal Policy Optimization，近端策略优化
- prefix tunin：前缀微调
- prompt tuning：提示微调
- Prompt-tune：提示微调
- PTQ: Post-training quantization，后训练量化
- QAT: Quantization-aware training，量化感知训练
- RAG: Retrieval Augmented Generation，检索增强生成
- ResNet: residual network，残差网络
- RL: Reinforcement learning，强化学习
- RNN: recurrent neural networks，循环神经网络
- Schema evolution：模式演变
- Sensors：传感器
- Static Batching：静态批处理
- tensor parallelism：张量并行
- TensorFlow：开源深度学习框架
- TensorRT：高性能深度学习推理的平台
- TorchScript ：是一种从 PyTorch 代码创建可序列化和可优化模型的方法
- Translation：模态转化

致谢

《可信赖的企业级生成式人工智能白皮书》由中国开源软件推进联盟以及 IBM 的专家和志愿者共同编写完成，编写过程中得到许多开源人士、企业单位、社区、高校的大力支持，在此表示感谢！

编委会

顾问

陆首群 中国开源软件推进联盟名誉主席

策划组

谢东 IBM 中国研发中心总经理，大中华区首席技术官
程海旭 IBM 全球兼大中华区首席技术官（标准及开源）
刘澎 中国开源软件推进联盟副主席兼秘书长
梁志辉 中国开源软件推进联盟常务副秘书长
孟繁晶 IBM 中国系统开发中心首席技术官

主编组

程海旭 IBM 全球兼大中华区首席技术官（标准及开源）
刘泽宇 IBM 中国开发中心高级架构师
石延霞 IBM 商业价值研究院高级咨询经理
罗东文 IBM 科技事业部市场经理
张颖 IBM 咨询大中华区数字化转型区块链和可持续发展负责人
刘晓金 IBM 全球技术服务部高级咨询经理
孟迎霞 中国开源软件推进联盟副秘书长，CSDN 副总裁
鞠东颖 中国开源软件推进联盟执行副秘书长

工作组

COPU 工作组成员

隆云滔 中国科学院科技战略咨询研究院副研究员
田忠 COPU 专家委员会副主任委员
李博文 北京国家金融科技认证中心实验中心负责人
张侃 COPU 专家委员会委员
荆琦 中国开源软件推进联盟副秘书长，北京大学副教授
陈伟 COPU 专家委员会副主任委员

IBM 工作组成员

白默涵 程文杰 初德高 董琳 樊斐 冯媛 葛巍 韩艳艳 姜朋慧 李青 廖文静 刘佳怡 刘默驰 徐斌 徐孝天 杨军 杨悦 元中方 袁怿 原雪洲 臧倩 张玉明 赵则名 朱茉 庄雪吟

贡献者

曹岚 陈栋 丁伟 都娟 何蕾 李变 李玲 刘俊 刘胜利 倪栋 聂锦程 庞文峥 沈海军 孙盛艳 王彩彩 王积杰 王君 吴敏达 杨继辉 姚勇 张家驹 赵登科 赵蓉 郑维珺

中国开源软件推进联盟秘书处
电话: +86 010-88558999
联盟公共邮箱: office@copu.org.cn
联盟官网: http://www.copu.org.cn
地址：北京市海淀区紫竹院路66号赛迪大厦18层

COPU开源联盟
微信公众号

如您对本册感兴趣，欢迎拨打免费电话
IBM 专家顾问团队竭诚为您服务：

数据与AI: 400-810-1818 转 6216
IT及业务自动化: 400-810-1818 转 6210
安全软件: 400-810-1818 转 6212
可持续发展软件: 400-810-1818 转 2392
基础架构（存储及服务器）: 400-810-1818 转 6255
IBM 咨询: 400-810-1818 转 6218

IBM 中国
官方微信

IBM 中国
官方视频号