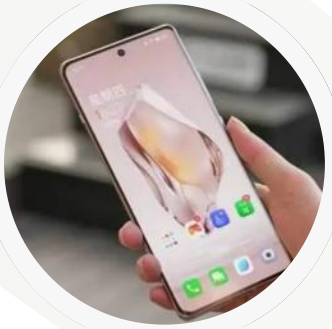
行业洞察

现状 | 挑战 | 趋势



行业洞察｜人工智能

能力越大，隐忧与机遇越大

2024 年，生成式 AI 迎来了从“ 能用 ”到“ 好用 ”的飞跃，彻底改变了人们对人工智能的认知。AI 技术通过不断进化的交互方式和强大的生成能力，正在深入改变人类的工作、生活以及科技生态。与此同时，生成式 AI 也面临着能耗压力、企业级落地挑战和模型“ 黑盒 ”难题等隐忧。在新技术的推动下，AI 不再是冷冰冰的工具，而是成为了生活中充满温度与智慧的助手。然而，这场技术革命带来的，不仅仅是人机交互的体验提升，更是一场从能力进化到责任担当的深刻变革。

文：魏德龄

现状 | AI从未如此好用

也许在去年还有人会怀疑所谓的生成式AI无非就是一个升级版的语音助手，其背后的原理仅仅是一个无比庞大的数据库而已。但当时间来到2024年，生成式AI通过自身的能力升级与实力展现，呈现了一个AI从未如此好用的新阶段。

现状1 体验升级：ChatGPT 4o引领新拐点

尽管有业内知情人士透露，ChatGPT 4o仅仅是OpenAI公司的一个后手，但5月13日的发布会绝对称得上是搅动整个AI行业的一个重要时间点。

作为OpenAI 推出的全新多模态模型，GPT-4o具备同时接受文本、音频和图像作为输入，并生成上述媒介输出的强大能力。这种进步让人机交互更加贴近人与人之间的自然对话，极大提升了语音交互体验。GPT-4o 的响应速度极快，音频输入的平均回应时间为320毫秒，与人类对话的反应时间相当，而在视觉和音频理解方面表现尤为卓越，能够生成多种音调并带有情感化表达。此外，该模型支持在线视频通话，为用户实时解答问题，并实现对话的动态打断与流畅衔接，优雅处理语音交互中的语调、背景噪声及多说话者情境，填补了传统语音助手延迟大、信息丢失严重的体验缺陷。与之前的语音助手（如Siri）的三阶段处理机制不同，GPT-4o 通过一个统一的神经网络直接实现音频、图像、文字和视频的实时转换，带来全新的跨越式体验。

GPT-4o 在性能上与 GPT-4 Turbo 不相上下，尤其在非英语文本处理、API响应速度和经济性方面表现优异，API价格较前代降低50%。这一模型适用于文本分析、数据可视化、图像解读等多场景应用，且免费用户即可体验 GPT-4o 的强大功能，包括通过 GPTs 和 GPT Store 访问更多工具、上传文件获取分析，以及利用记忆（Memory）构建个性化互动体验。

技术升级的最直观变化在于，通用人工智能可以低门槛的来学习用户所提供的专业内容资料，通过这些以往难以接触到的行业数据，来生成出更加符合使用者预期的内容，无论是文字、图片，GPT-4o的出现让人们见识到AI更加可用性的一面，而不再是经常出现“外行看着内行，内行看着外行”的奇怪创作表现。

现状2 用例大爆发：生成式AI融入多场景

生成式AI所带来的用例大爆发可谓是全方位的，AI视频生成同样是一个十分明显的案例。如今，人们已经偶尔会在网上看到通过AI生成的，并且内容生动有趣的视频内容。AI视频生成正从传统的检索生成和局部生成，逐步迈向依靠自然语言提示词的全量生成。这种技术趋势让生成内容更加灵活和丰富，显著拓宽了应用场景。检索生成主要基于现有素材，通过标签匹配和排列组合完成，具有一定效率但生成内容较受限。局部生成则能针对视频特定部分进行编辑，例如调整角色、背景、风格或添加特效，虽有创意性提升，但依然局限于预设元素。相比之下，提示词生成基于大规模模型，借助自然语言输入即可生成全新的视频内容，包括风格化场景、艺术效果或动画设计，极大扩展了创作空间和灵活性。这种新技术不仅提升了生成效率，还大幅降低了成本，为多领域应用提供了无限可能。

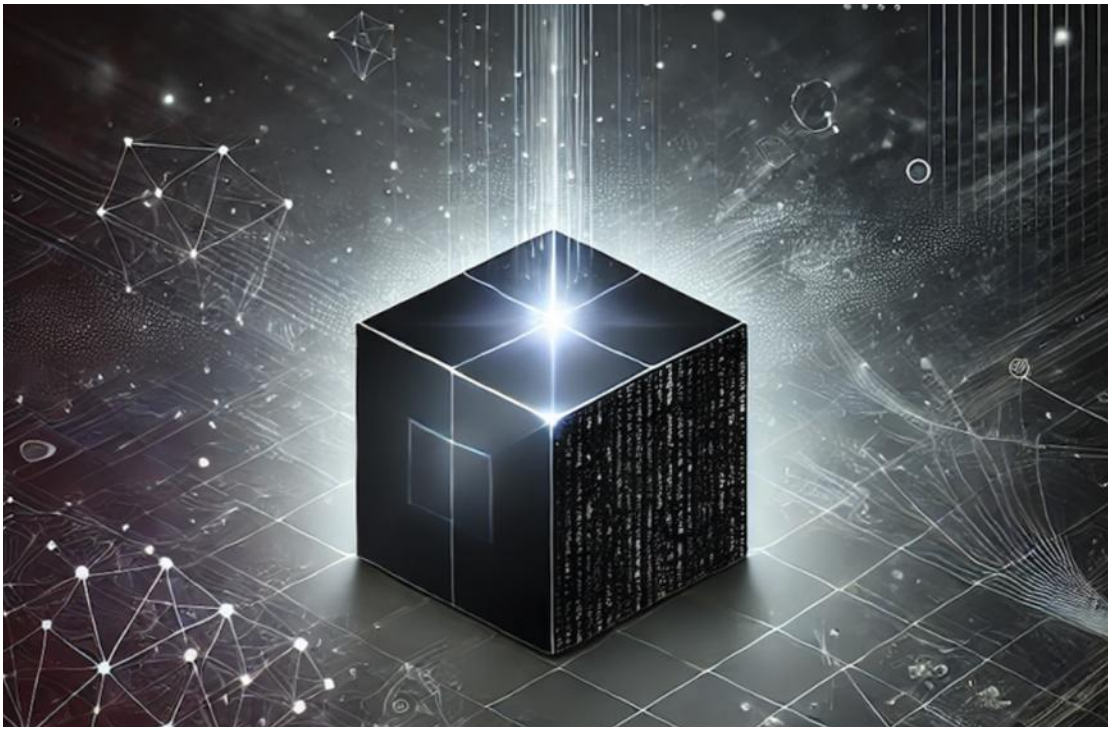
国内的生成式AI产品同样能够看到从能用到好用的趋势，科大讯飞发布的讯飞星火大模型4.0 Turbo在数学能力和代码能力上取得了重大突破。根据行业实用数学任务构建的测试集CAppliedMath-1.0，讯飞星火4.0 Turbo在计算、财务、金融、度量等多个维度的任务中均超过GPT-4o水平，已完成超长思维链、树搜索和自我反

思评价等算法验证；根据代码生成HumanEval测试集上的效果对比，讯飞星火4.0 Turbo在Python、Java、JavaScript等任务上和GPT-4o的差距微弱，在C++能力上超过GPT-4o，同时推出星火代码7B版本，满足代码生成、代码补全等极速响应型任务，效果业界最优。

2024年，生成式AI技术正快速从“能用”迈向“好用”和“实用”，以GPT-4o和讯飞星火大模型4.0 Turbo为代表的新一代多模态模型，显著提升了跨媒介交互体验、生成效率及准确性，广泛应用于文本分析、数据可视化、代码生成等领域，推动了AI更贴近人类需求的全面发展。然而，在技术持续突破的同时，AI的发展也面临诸多挑战，例如高能耗带来的环境负担、模型思维过程的“黑盒”特性导致的透明性不足，以及如何在企业级场景中实现更高价值的落地。

挑战 | 能力越大，隐忧越大

在超级英雄电影中，有这样一句脍炙人口的台词：“能力越大，责任越大”。不过，对于AI来说，随着能力的增强，所对应的责任一面，也同样代表着隐忧。下面的这一年中观察到的问题，也同样是业界在反复热议的话题，AI的隐忧主要表现在三个问题上：



■ 挑战1 算力与能耗：失效的摩尔定律

无论是云端 AI 还是端侧 AI，都正在让摩尔定律失效，尽管 AI 的性能拥有可见性的飙升，但不妨重温下该定律的全部描述：“ 半导体芯片上集成的晶体管数量每隔 18 到 24 个月翻一番，性能提升一倍，价格下降一半的现象。”如今 AI 性能提升的背后，并不意味着价格或是成本会相对进行下降，代工制程升级的成本水涨船高，云端 AI 的提升方式也更多依赖于更多的 GPU 数量，并对应了更大的能耗。

人工智能的掣肘之处已经凸显，那就是能耗问题。本质上来说，ChatGPT 的强大表现表源自于“大力出奇迹”。根据估算，GPT-4 可能使用了约 10,000 至 25,000 张 A100 显卡完成训练，而 Stability AI 则使用了约 5,000 张 A100，Falcon-40B 仅需 384 张 A100 即可完成训练。相比之下，Inflection 通过 3,500 张 H100 显卡训练出了与 GPT-3.5 能力相当的模型。据业内人士透露，GPT-5 的训练可能需要 30,000 至 50,000 张H100显卡，这一数字远超现有模型的资源需求，进一步凸显了先进AI模型对计算力的极高依赖。

算力增长所对应的便是能耗。预估GPT-6的耗电将达700万度。相比大型AI系统的百万瓦级功耗和海量数据需求，而人类大脑则能以很小样本和30瓦功耗实超高计算效率和识别。

这就意味着，AI算力背后所依附的数据中心正在面临巨大的能耗压力。有数据统计显示，中国的数据中心正在面临巨大的能耗问题，在2022年已经接近2700亿度的用电，预计到2025年会翻倍，达到4000亿度电。这就意味着，到2025年，中国数据中心的能耗约等于4个三峡或葛洲坝的发电总量。

如果找不到解决途径，算力的尽头将会是能源。

■ 挑战2 难有大作为的企业级领域

如今 GenAI 作为一种新的产品卖点，在消费电子领域确实风生水起，产品逻辑多为通过生成式能力带来如系统交互、图片处理、文字信息汇总等方面的升级。然而，当类似的逻辑应用于企业级领域的时候，GenAI 技术本身目前的种种不足之处，却会被放大，从而成为了落地过程中的掣肘。

最大问题便是所谓的致幻率问题，“ 一本正经胡说八道 ”的情况在消费电子领域或许可以被用户一笑了之，但在 IT 运营管理的过程中，却可铸成大错，当 ToB 领域对于安全性和准确性的要求变高，以及对高可靠性的要求，就难以有过多的容错性。从而导致 GenAI 的方案可能难以被厂商最终采用。

准确性问题显然与训练数据的专业性与量级存在强关联，但企业往往并不愿意对外分享数据，如何在构建便利 AI 条件的情况下来平衡安全性和隐私性成为了比较大的挑战。在使用相关 GenAI 来实现产出的时候，知识产权问题也应运而生，生成的图像、归纳的总结、构建的代码的知识产权到底属于谁，企业对于此类的担忧同样一直与 GenAI 的发展而相生相伴。

GenAI 的出现也在打破企业内部的一些边界，对于员工而言很容易自然而然地把如会议纪要、产品资料等内容上传在云端 AI 来快速获取会议总结。企业难以遏制这种员工简化工作流程的渴望，但对于合规与安全性又提出了更大挑战。

这无疑影响了企业对于部署相关落地方案的决心与判断。而从很多企业在今年所对外提供的 AI 解决方案也不难发现，在产品功能上多聚焦于通过自然语言来优化操作流程，并一般会避免让 A 涉及到相关决策的环节，此举无疑也映衬了厂商对于自身产品信心的不足，显示出企业级应用仍有较长的探索与完善之路。

挑战3 神秘的思维黑盒

随着生成式GenAI和深度学习模型的广泛应用，其强大的能力在自然语言处理、医疗诊断、自动驾驶等领域展现出巨大潜力。然而，这些技术的核心问题之一——思维“黑盒”特性——正在引发越来越多的关注。所谓“黑盒”，是指这些模型的推理过程高度复杂、难以解释，对其内部决策逻辑的透明度存在重大欠缺。这种特性不仅引发了学术界对AI可解释性的讨论，也对其在关键行业中的应用构成了显著障碍。

大模型的“黑盒”特性源于其设计与运行方式。首先，模型通过多层神经网络捕捉数据中的复杂模式。这些多层抽象形成的高层次内部表示往往不具备直观的语义信息，难以被人类理解。其次，大模型采用分布式表示，信息以神经元激活模式的形式存储，任何单一神经元都无法直接对应具体的特征或概念。此外，非线性激活函数引入的非线性变换，使得模型在面对输入数据微小变化时可能产生难以预测的输出。最后，端到端学习方式虽然省去了人工设计特征的步骤，却将特征提取与决策过程紧密集成，进一步加剧了模型的不透明性。

黑盒特性在某些关键领域可能会引发一系列问题。例如在自动驾驶领域，黑盒模型可能在突发情况下做出难以预测的决策，例如在面对未知路况或标志时，模型的错误反应可能直接导致安全事故。或是在目前正在大量尝试融入AI能力的金融行业，黑盒模型如果被用于信用评估或风险管理，可能无法满足监管机构的合规性要求，原因在于一旦模型拒绝了某一贷款申请，银行却无法提供拒绝理由。

尽管黑盒问题尚未彻底解决，学界和业界正在积极探索可能的解决方案。一些研究者尝试通过可视化技术和模型简化来揭示模型的内部结构，另一些人则采用基于知识的解释方式，为模型的决策提供更加直观的解释。

在相关技术实现之前，AI的黑盒特性仍是限制其在高风险领域大规模应用的重要因素。

趋势 | 洞察

假如能力的另一面是隐忧的话，隐忧所对应的则是需求与机遇。面向即将到来的2025年，AI的未来将会继续引发出无限的可能性。在此，根据市场风向，可以预测以下三大趋势：

趋势1 端侧AI继续牙膏爆挤

上文中已经提到了云端AI所带来的在数据中心侧的压力，与此同时当AI开始与众多行业产生深度融合，对于时延性的要求也在提升，如果是像使用云端AI助手时的转圈圈般的响应表现，甚至可能会引发安全隐患。

例如在通信领域，将AI融于AI系统设计之初几乎已经成为业界的普遍共识。但在对于AI与通信融合的思考中，接入网的实时性要求，也对AI在处理海量数据时的响应速度提出了很大挑战。如今以智能手机处理器在端侧AI上的成果无疑提供了对应的解题思路。有预测表示，未来的6G终端将利用端侧AI能力，能够在本地处理大量数据，而不需要跟云端做过多的互通操作，这样既可以保护隐私，又可以提高响应速度。

端侧AI的算力也在显著提升，并且没有依赖于更高的能耗。以骁龙8至尊版为例，搭载的全新架构Hexagon NPU性能提升了45%，能效提升45%，基础大语言模型上的token生成速率提升了高达100%。快速响应方面，在目前业界流行的一些大语言模型上，骁龙8至尊版的处理速度达到超过70 tokens/s。在MLPerf BenchMarks测试中，相比骁龙8 Gen3，性能提升达到了104%。

受益于端侧AI能力的不止于智能手机。在汽车领域，骁龙座舱至尊版集成的最新NPU，其性能相比8125提升至最高12倍，能够处理高达几十亿参数的大语言模型，通过搭配检索增强生产技术，以及基础模型，能够实现车辆维修助手、故障分析、问题上报等功能。在PC领域，骁龙X Elite 45TOPS的NPU算力和异构计算架构，为开启终端侧生成式AI体验提供了优势，让骁龙X系列成为支持首批Windows 11 AI PC的平台，让个人用户体验更加智能和个性化。高通还在投资日期间透露了第三代Oyon CPU架构的相关信息，预计明年在AI性能上还将带来进一步的提升。

趋势2 功能从设想到现实

关于AI的设想，业界已经开始试图利用这项技术跳脱出以往思维的窠臼。变革传统的交互方式就是一项正在从设想走向现实的案例进行时。其背后的技术根基在于AI已经具备了看得懂、听得懂、能理解的基本功，使其能够实现以往语音助手所不能达到的高度。

2024骁龙峰会上，高通总裁安蒙抛出了这样一个观点，他认为随着AI将在终端设备上所带来的体验维度升级，所谓的“杀手级应用”概念将不复存在，它只是一个过去式的思考问题的角度。未来，每个应用都将借助AI实现融合与互通，具备“杀手级应用”的潜力。2023年，他还曾就这一设想表示：“AI引擎在终端运行与云端交互，你可以在终端本地运行一个应用，或者终端按照你的需求去云端交互。至此，大家看到了5G和AI是如何把一切都连接到一起。尽管我们有一个以应用为中心的终端，但不一定需要所有应用，它和云端整合就知道你的需求，你可以在终端或者云端上挑选应用。”

在这一设想的落地方面，荣耀已经成为最具代表的产品。其手机产品中的AI智能体，带来了“一句话关闭自动续费”“一句话点饮品”“一句话旅行规划与订票”等颠覆性端侧AI体验，甚至在其中还能选择出用户最喜欢的产品类别，比如是美式还是拿铁。Copilot+PC也正在焕发出新的潜力，用户可以仅仅通过一张儿童画般的草图生成出海边的风景，任意搜索全部文档中的信息内容，无论是文字、图片，或是仅仅是一种关于物品的形容。以及在离线状态下，也能即刻生成出相关美食必吃榜推荐的AI助手。

这种交互方式的变革已开始在企业级应用中出现，同样是通过自然语言的方式来简化运维过程中的操作。例如元景2.0中通过采用自适应的表格拆分和整合，自动补齐了表头和标题等信息，使表格问答的准确率提升了20个百分点；针对车牌号、故障码等字符串查询“找不对”的问题，元景2.0采用多路检索融合的方式，使回答准确率提升近20个百分点。

随着AI技术的快速发展，传统的交互方式正在被重新定义，从设想到现实的转变已然开始。在终端设备和云端深度融合的驱动下，AI不仅提升了用户体验的高度和广度，也拓展了技术的应用边界。无论是在消费级市场上实现“一句话解决”的便捷操作，还是在企业级场景中优化复杂任务处理流程，AI都展现出了强大的变革潜力。可以预见，未来的技术生态将以更加智能、个性化和高效的方式重塑人类与设备、服务的关系，真正实现“所想即所得”的数字化生活与工作体验。

■ 趋势3 企业级用例静待花开

尽管企业级AI解决方案面临种种技术本身的制约，但这一市场无疑在近年来成为了聚焦点。原因在于企业对于通用AI平台难以建立信任，独立定制的解决方案，由于采用了相对隔离式且专业度更高的数据库，提升了隐忧之下的信心。

很多商业AI的底气在于数据，这意味着AI和一个企业的业务流程、运营管理深度融合，充分挖掘利用企业内部和行业的数据，释放数据的价值和潜能，让企业的决策运营更卓越、更智能，让商业社会更高效。部分公司对于致幻问题的解决方式在于用企业核心的业务数据进行训练，而且是一个真实的、实时的、准确的业务数据来训练这个模型。

针对不同行业需求的定制化设定也是企业级AI们所集中表现出的特色，以思特奇九思大模型为例，为企业提供开发态、训练态、运行态、运营态的全生命周期能力支持。针对特定行业和企业数据进行模型训练，思特奇构建1套智算基础设施、1套大模型通用平台、N个AI技术能力、X个应用场景的思特奇人工智能体系总体架构。

在安全领域，以AI对抗AI的概念同样成为了新的共识性路径。未来将成为AI对抗AI的时代，不可能光靠人力去进行事件响应，必须用AI来协助。网络安全企业的产品框架中通过专为实现卓越安全分析与威胁防护而构建的即开即用AI模型，安全团队能够创建自己的ML模型并将其集成到架构体系中，从而实现欺诈检测、安全研究、复杂数据可视化等独一无二独特的用例。

未来，企业级AI产品的发展将更加注重深度融合和个性化定制，以满足不同行业和场景的复杂需求。AI不仅将成为企业生产力提升的核心驱动力，还将在业务创新中扮演重要角色。从定制化模型到实时数据驱动的智能决策平台，企业级AI有望打破传统工具的局限，成为企业管理和运营的全方位助手。

随着AI技术的成熟，企业将更倾向于构建专属的私有化模型和独立的智算基础设施。这种模式能够在保护数据隐私的前提下，充分挖掘数据价值，实现更高效的资源调配和风险控制。此外，随着边缘计算和混合云技术的普及，AI在企业级应用中的分布式部署能力将进一步增强，为跨部门、跨区域的智能化协同提供技术保障。

结语

生成式AI的迅速发展为各行业带来了无限可能，也提出了新的挑战。从多模态模型的跨越式提升到企业级场景中的深度融合，AI正在以更贴近人类需求的方式展开全面赋能。尽管面对能耗、隐私、可信性等诸多难题，这些技术正在推动全行业的数字化转型与创新。可以预见，在接下来的发展中，生成式AI将通过更智能、更高效的应用场景，将“技术想象”变为“现实可能”，为未来社会构建一个更加智能、便捷和可持续的世界。