统信软件技术有限公司
UnionTech Software Technology Co., Ltd.

统信AIOS白皮书

（2024）

中 国 操 作 系 统 领 创 者 · 给 世 界 更 好 的 选 择

目录

前言

信息技术的大潮浩浩荡荡，奔涌向前。自2022年底ChatGPT问世以来，全球范围内爆发了前所未有的AI浪潮，生成式人工智能成为了学术界与产业界皇冠上最璀璨的明珠。它创造了人类历史上最短时间获得亿级用户的世界纪录，把十年前市值不及Intel十分之一的显卡公司推向了世界市值第一。有人说，它将带来第四次产业革命，重塑世界格局，有人说，它将带来数字永生，还有人说，它将把人类社会彻底颠覆。

与以上观点相映射的是全球AI技术关键技术厂商市值、产业投资、市场应用各领域的数据变化：2024年6月5日，AI芯片厂商英伟达市值突破3万亿美元大关。6月18日，英伟达股价收盘上涨3.51%，市值达到3.33万亿美元，超越微软成为全球市值最高的公司，是英特尔的10倍以上。数据显示，2023年全年生成式AI投资达到252亿美元，较2022年的13.7美元增长17倍。

同时，更多行业研究数据预测和表明：随着生成式AI对各行各业的产业效率提升，到2032年，生成式AI有望在硬件、软件、服务、广告、游戏等众多领域创造1.3万亿美元收入，占科技领域总支出从目前的不足1%扩大到10%-12%，复合年增长率达到约42%。

一项针对全球37个国家和地区的公众调查显示，2023年，认为人工智能将在未来三到五年内极大影响他们生活的人比例从60%上升到66%。此外，52%的人表示对人工智能的产品和服务感到紧张，比2022年上升了13%，36%的人认为在未来5年内，AI将取代自己的工作。人们对AI充满了既爱又怕的矛盾，既认同AI能够改变世界，又担心AI取代人，人类变成机器的仆人。

身处时代大潮，惧怕毫无意义，回避亦是徒劳，唯有奋勇前进，方显英雄本色。统信软件2023年即推出了UOS AI 1.0，针对云侧模型与端侧模型、办公场景及生活场景打造了生成式人工智能的系统，树立了世界范围内第一个开源操作系统与人工智能融合的标杆。来到2024年，更在异构计算、云端一体、模型优化、交互融合等方面大展身手，与芯片、模型、应用等伙伴联合建设了UOS AI 2.0。我们借此机会特推出UOS AIOS白皮书，向业界各位伙伴分享经验与观

点，并诚挚邀请大家与我们同行。

在这一时代重大变革之际，让我们同心协力推进产品与技术创新，以开放的精神构筑产业繁荣，致力于让包括AI在内的各项技术帮助每个人获得更好的体验，让每个人都拥有实现自己梦想的能力，进而为整个世界创造出更光明的未来。

一、 产业趋势

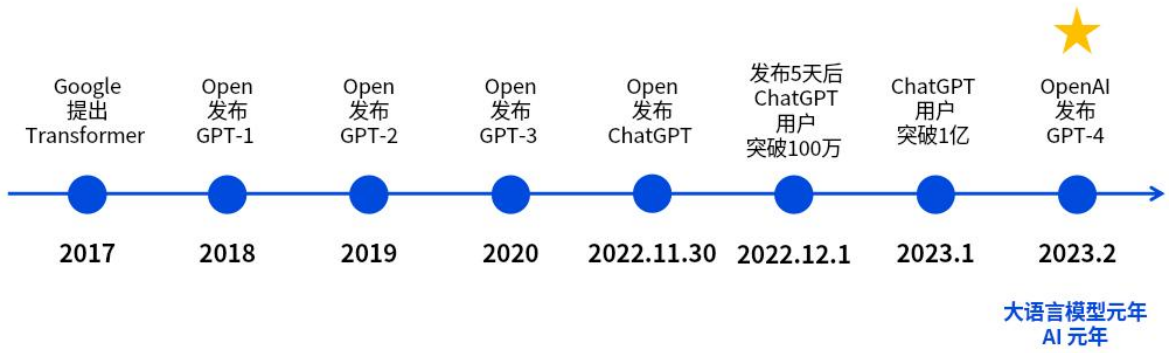
1.1 AI的发展趋势

1.1.1 AI技术快速进步

自2017年Google提出Transformer模型以来，NLP（自然语言处理）领域发生了革命性的变化。Transformer架构通过自注意力机制（self-attention）允许模型在处理文本时考虑整个句子，而不仅仅是上下文中的几个单词。这一创新使得模型可以更好地理解语言的语义和结构。

基于Transformer，非常快的出现了GPT模型Generative Pre-trained Transformer，使用Transformer架构通过大规模预训练学习语言结构和语义，然后进行生成性任务，如对话生成。GPT是LLM（Large Language Model）概念的一个实际兑现和落地。GPT引入了预训练概念，即先在大规模文本数据上进行无监督学习，然后通过微调来适应特定任务。这种训练策略使得模型能够在多个任务上表现出色，降低了对大量标注数据的依赖。GPT在自然语言处理领域实现了显著的性能提升，特别是在文本生成、对话系统、翻译和问答等任务上。它标志着深度学习技术在语言理解和生成方面的重要进步，使机器理解人成为可能和现实。

OpenAI 2018发布GPT-1，2019年发布GPT-2，2020年发布GPT-3。2022年11月30日，OpenAI推出ChatGPT应用，仅仅5天时间用户突破100万。2023年1月，GPT用户数突破1亿。2023年2月，OpenAI发布GPT-4，准确性大幅度提升，识图能力更高。可生成歌词、创意文本等风格多样。GPT开始影响和席卷全世界，2023年成为大语言模型和新的AI发展元年，AI自此迈入新的转折点和发展阶段。



为什么ChatGPT引发AI在全球的新发展和广泛的热度？AI努力的目标是让计算机像人类一样思考和行动。而GPT大语言模型将AI的性能和智能化进行了显著的提高，在自然语言理解和交互上，以及基于自然语言的内容生成和反馈上，AI已经接近于像人类一样思考。与ChatGPT的交互，已经看上去像与一个有智能的人在交互。

1.1.2 生成式AI的产业应用

新的生成式AI浪潮出现之后，在诸多行业和领域进行了应用，得到大量尝试和探索，包括和不限于：营销、办公、客服、人力资源、基础作业等领域，并且这种技术所带来的赋能与价值已经初步得到验证。有数据显示，33%企业在营销场景、31.9%的企业在在线客服领域、27.1%的企业在数字办公场景下、23.3%的企业在信息化与安全场景下迫切期望生成式AI的加强和支持。

- **营销场景：**营销场景是目前生成式AI渗透最快，也是应用最成熟的场景。生成式AI主要在营销动作中的内容生产、策略生成方面极大加强了数字营销的能力。例如市场认知阶段的核心价值是创意参考，可赋能环节包括：广告策略、品牌传播、市场分析、CEM、SEO、DSP、SSP，通过生成广告创意与投放优化参考，包括广告设计、广告内容、投放渠道策略和投放分析，从而提高广告效果和投放效率；

- **数字办公场景：**数字办公场景也是目前生成式AI渗透较快的场景之一，主要体现在对个体的办公效率提升。在文本内容生成、代码生成、流程设计和规范等方面表现出一定的提示和

优化。例如流程管理模块的核心价值是规范建议，可赋能环节包括：流程规范设计、流程路径设计、流程控制设计、流程优化，在一个新项目启动时，可以根据项目需求和历史经验自动生成流程规范建议，包括各阶段的任务分配、时间节点等；

- **在线客服场景：**在线客服是生成式AI音频生成最近距离的场景之一，声音合成、语义理解在智能化策略下，生成具有明确目的性的对话内容。例如全渠道接入模块的核心价值在于个性化模块，可赋能的环节：富文本沟通、自动主动对话、访客信息展现，生成个性化回复模板，更好地提供针对性服务，从而提升客户满意度；

- **人力资源：**生成式AI对人力资源服务的加成，是目前在企业经营管理体系中进展较快的领域。使人力资源管理体系的效率大幅提升的同时，在一定程度上也改变了传统人力三支柱的传统管理模型。例如招聘模块的核心价值在于简历推荐，可赋能的环节：筛选、面试筛选、笔试测评。以筛选简历阶段为例，可以分析各个候选人的简历，生成匹配结果报告，并根据公司需求智能推荐合适的候选人。大幅提高筛选准确性和效率，减少人力资源部门的工作负担；

- **基础作业：**生成式AI在基础作业场景中的表现十分突出，在设计、合同管理、法律服务等环节表现出很强的智能化以及可替代性；

1.1.3 生成式AI带来的新概念

生成式AI得到大量关注和尝试探索之后，也引发了许多新的概念，包括：

- **AIGC：**利用人工智能生成内容，包括文本、图像、视频、音乐、代码等。它代表了一种全新的内容生产方式，不再依赖于人类的手工创作，而是通过算法和模型自动生成。AIGC有可能完全改变知识和信息的传递方式，对信息传递、创意生产带来全新变化；

- **自然语言对话界面（LUI Language User Interface）：**传统的人机交互是基于GUI的，用户通过键盘、鼠标、触屏等设备进行人机交互。因为大语言模型的突破，对自然语言和人类意图的理解的进步。基于自然语言的对话和交互即LUI有可能成为替换GUI的全新的交互方式。用户可以通过语音或文字与计算机、智能设备等进行交流，就像与人类对话一样自然流畅，会大大提高了人机交互的效率和便捷性；

- **超级自动化：**结合人工智能、机器学习、自动化流程和机器人流程自动化等技术，实现业务流程的高度自动化。大语言模型可以理解和处理自然语言指令，与其他自动化工具协同工作，进一步提高业务流程的效率和准确性，减少人工干预；

- **智能助手与数字分身：**每个人都可以拥有自己的智能助手，帮助处理日常事务、管理日程安排、提供信息查询等服务。此外，数字分身的概念也逐渐兴起，即通过AI技术创建一个数字化的自己，可以代表自己进行一些活动，如参加虚拟会议、与他人交流等；

- **AI原生工作流：**由人工智能对工作流进行重新设计和优化，并且进一步由AI自动完成或者辅助人类完整；

- **渗透式人工智能：**将大模型应用从仅限于处理文本的任务拓展到与物理世界的交互，通过融合来自物理世界的真实数据，让AI直接与人类周围真实物理世界的感测数据进行交互、解析并作出反应；

这些概念，都对应一个全新的应用场景和应用领域，值得我们深度思考和探索，同时相信随AI技术的进一步发展，也会有更多的新概念和新场景产生。

1.1.4 生成式AI的未来展望

整体AI技术发展，可分为如下几个阶段：

- **第一阶段：计算，即机器对信息进行存储和计算。机器能够像人一样拥有记忆和计算的能力，主要表现为存储和处理海量数据**

- **第二阶段：感知智能，机器和计算机具备类似于人的部分感知能力，可以识别文本、图像、语音等的具体内容，并进行一定程度的基于规则和策略的处理**

- **第三阶段：认知智能，人工智能具备了独立思考和判断的能力，进行决策和行动**

当前的大模型和生成式AI的发展，正是由感知智能向认知智能的发展阶段。AI由算力、数据、算法三者共同决定和发展。感知智能发展了很多年，已经使得计算机和机器具备了不错的数据和内容感知能力，也出现了较多的AI应用场景，如图片识别、文字识别等等。而大模型对算法领域的突破，极大的提升了计算机的感知能力，由对数据和内容的感知进化到了对知识和意图的感知。感知是认知的基础，感知的进步促进了认知智能的实现可行性。

当前人智能的主要发展方向已经进入到了认知智能，表现形式为LUI、Agent等新概念的出现和应用，发展和增强独立规划、决策以及行动能力。对于AI在实际场景中进行应用和落地来说，人类需要的也是具备认知智能的AI。

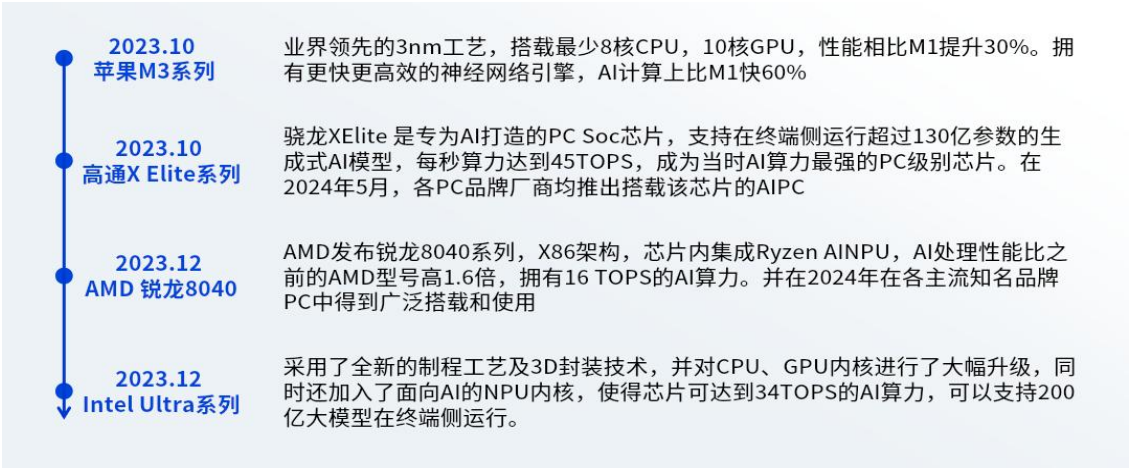
终极的具备意识的AI，能识别意义、价值等更加抽象的概念，是具备了“智慧”的AI，但同时也更加遥远。感知和认知层面的人工智能，在特定环境下能做出合理规划和判断，在某些方面替代人或者辅助人提供效率更具备实际意义，也是相当一段时间内人工智能的发展方向。

1.2 AI对操作系统的影响

1.2.1 新计算架构

AI的发展对硬件的计算架构也产生了巨大的影响，促进了全新的神经网络处理器（NPU）的诞生，并且推动了为AI专门设计的全新计算架构 —— 异构架构的形成和应用。新PC产品架构普遍结合NPU使用，并与合适处理器形成新的异构架构。异构计算可实现最佳应用性能、效能和电池续航，为生成式AI带来全新增强体验。

整体业界中，依托NPU和新的异构计算架构，硬件基础设施和硬件算力正在不断完善和提升，并增长显著：



在2023年，最高45TOPS算力的异构CPU得到推出和发布，并在开始在实际的终端产品中

得到应用。2024年，市面推出的新PC产品均搭载各不同厂商最新的AI异构CPU，并且用户侧终端算力平均值已突破10TOPS，预期在未来几年内，用户侧终端算力依然会显著增长。有观点认为，到2026年用户侧终端算力平均值将突破20TOPS，领先的终端算力将超过60TOPS。

可以看到，用户侧终端算力已经开始规模化的增长和普及。算力是AI应用的基础，终端算力的提升和普及将会极大程度的促进终端上AI应用的普适化，AI相关技术和应用将会更快速的在每一个终端用户手上得到普及。

终端计算架构变化、算力增长和AI技术的应用，也对操作系统产生影响提出新的要求，操作系统需要更好的支持异构架构和异构计算，并搭建出能更好支持AI应用的基础平台和服务。

1.2.2 新交互

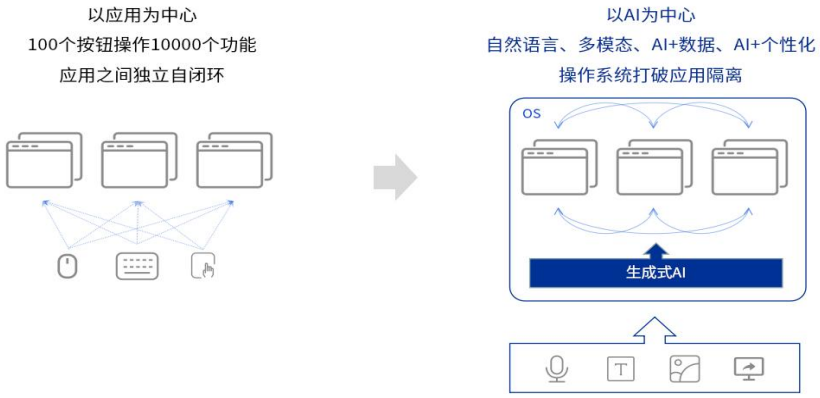
在生成式AI带来的新概念中，LUI（Language User Interface，语言用户界面）是一个重要的新概念，此概念对操作系统的影响是最大的，远超对其他应用、工具和系统的影响。操作系统是一切数字服务和信息服务的基础平台，为所有软硬件、应用等各种数字和信息化服务提供基础环境。用户所有的交互和操作都需要依托操作系统作为基础，提供交互界面、交互方式和交互手段。**操作系统的交互方式，支持和决定了用户面向任一数字应用和信息服务的交互方式。**

生成式AI为操作系统带来如下新的交互方式：

新LUI交互（AI + 自然语言）—— GUI图形化界面交互是以为功能为主体的，所有的操作都基于按钮进行操作，用户需要面对100个按钮操作10000个功能，而大部分的功能是不被用户直观可见的，这也导致了操作效率的低下和用户操作的困难。LUI则可能完全改变这一现状，生成式AI可通过对用户自然语言的理解完成对用户意图的掌握，从而自行调度应用、工具、功能完成以往需要用户主动来完成的任务，最终改变由用户和按钮构成的交互困局；

新任务交互（AI + 应用）—— LUI的交互的深入带来另一问题：应用、工具、系统的割裂性。

所有的应用都需要完成独立的功能和任务闭环，导致了应用之间的互相割裂。每个应用专注完成某个子领域的任务闭环，应用与应用之间则需要用户来进行操作和衔接来完成整体任务。所以，单个应用、工具、系统达成LUI是不够的，依然无法改变用户的交互困局。在操作系统层支持和整合LUI，能彻底解决应用独立的问题。**所有应用都依托和运行在操作系统上，操作系统可管理和调度不同应用，打通全量任务完成流程，从而最终实现完整的LUI交互；**



新个性化交互（AI + 个性化）：生成式AI在基于对用户的用户语言理解和意图理解之上，进一步分析用户的偏好和使用习惯，提供个性化的推荐和设置。并且可以基于用户的反馈，智能调整操作系统的界面布局和功能，即对传统GUI进行个性化的自动管理，从而更适应用户的需求；

新数据交互（AI + 数据）：操作系统中沉淀了大量的用户文件和用户数据，生成式AI可以分析和借读用户文件和数据，从而更智能的帮助用户管理数据。在用户的创作、创意领域提供智能的支持和辅助，以及为用户提供复杂问题的精准匹配回答和结果，形成全新的数据交互形式；

新多模态交互（AI + AI）：GPT模型出现之后，除在NPL自然语言领域演进和提升以外，也同步影响了多模态领域的发展，2023年OpenAI发布的GPT-4已开始具备多模态理解能力，直接导致了多模态交互的产生。用户可以通过不同的输入方式进行人机交互，包括文字、语音、图像甚至共享整个屏幕来与AI进行互动；

以上这些交互方式的变化，本质上是人机交互方式中心和主体的迁移和改变。以往所有的交互方式都是以人和应用为中心，设备、系统、AI等作为工具被人使用完成任务。而生成式AI

带来了中心和主体的迁移，以AI为中心，人指导AI，AI调用设备、系统完成任务。

1.2.3 新生态

生成式AI对操作系统计算架构影响、交互方式的影响以及产生的新概念最终的影响结果是对操作系统软硬件生态的重塑，生成式AI将会带来一个新的AI生态，并影响现有操作系统生态。

在硬件侧，新的硬件生态框架已经诞生 —— NPU的全新引入，异构计算的兴起，以及更多新硬件形态的出现如AIPC等，会促进和形成全新的AI硬件生态。

生成式AI的发展会影响和促进软件开发者使用AI相关技术提升应用的AI功能，越来越多的应用将会融入AI能力，实现功能的增强和创新。AI会成为应用创新与变革的基石，而最终所有的应用都是AI应用。**新的AI应用结合新的交互方式，则会诞生新的AI软件生态。**

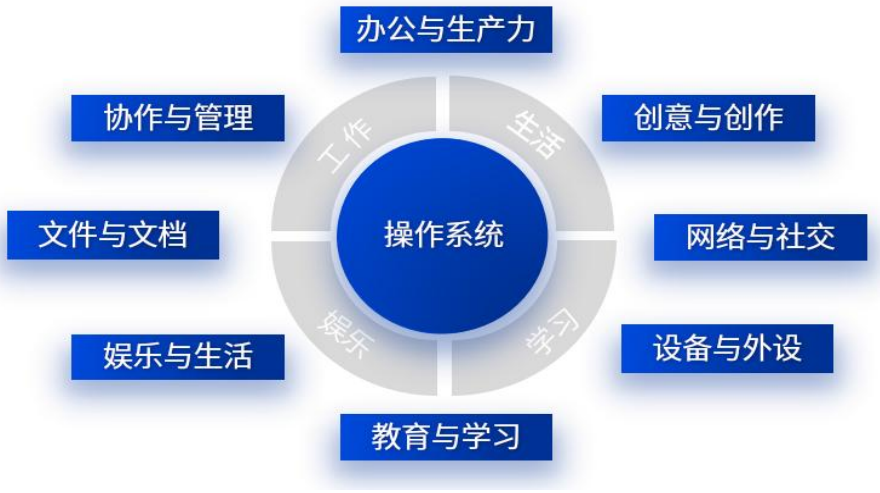
生成式AI的运行需要强大的算力和硬件资源的支撑，目前大部分的AI能力都是基于云测提供。同时在用户的使用和场景中，结合云侧服务已经成为主流的使用方式，办公、生产、创意、娱乐等各场景都是如此。**AI技术发展和应用会进一步加强云成为整个生态的组成部分，与AI硬件，AI软件一起协同构成整体新的AI生态。**

新AI生态

1.3 操作系统需要与AI深度融合

PC终端与操作系统贯穿在用户的工作、生活、娱乐和学习中，在诸多场景和环节为用户提

供赋能和服务：



生成式AI的创新和突破已赋能千行百业，但是PC终端系统的用户还没有从繁琐操作和复杂管理中解放出来。PC终端作为最主流最高频的生产力工具，承载着用户更深的期待：高效、智能、个性、便捷。在上述的操作系统的典型场景中，对于与AI的结合，用户也表达出了更多的要求，用户希望操作系统能够智能地识别、理解和适应他们的需求。比如，用户希望操作系统能够通过学习他们的使用习惯，自动优化系统设置、推荐应用、提供个性化的信息或建议。用户希望通过语音、更直接的命令/指令来操作PC和系统，而不是依赖传统的键盘和鼠标的一步一步的点击动作。在企业中，自动化和智能化客户的需求不断增加，企业和用户希望能够自动化进行不同业务系统之间的、不同应用、不同数据之间的智能联动和自动化操作。

生成式AI出现之后，主流的操作系统厂商，都已进行大量的探索并取得了一定的成果：

- 2024年5月19日，Windows发布Copilot+PC，通过Windows11 + Copilot + 全新的硬件，发布新的AIPC的概念和产品。将生成式AI的三大新影响一一兑现：
  - 新计算架构 —— 高通X Elite处理器，支持CPU中内置NPU进行AI计算调度
  - 新交互 —— Copilot 的全局交互中心和智能辅助
  - 新生态 —— AI Runtime集成诸多端侧模型，提供AIAPI和AISDK供应用使用
- 2024年6月11日，Apple发布Apple Intelligence，彻底革新Siri。集成端侧模型在协作、图片处理中提供强大的生成式AI的赋能，创新性的完成了跨应用的任务处理和联动，技术架构上提出了端+云的整体技术协同方式

操作系统承载着用户的大量场景、数据以及技能。**操作系统与AI的融合，有助于更深入地改变用户的工作和生活方式，这既是用户的需求，也是生成式人AI技术的要求。**生成式人工智能所带来的影响（包括新技术架构、新交互模式、新生态等）都需要操作系统承担起更多的责任，以助力生成式AI的发展与应用。

操作系统的未来也必然是更加AI化的，AI是操作系统发展的驱动力，也是操作系统发展的趋势。未来终极的操作系统本身即是AI，AI也即是操作系统。

1.4 操作系统与AI融合需要面临的挑战

生成式AI引发了诸多新的概念、应用、趋势，但也同时带来了诸多新的挑战：

安全与隐私问题 —— 当前的AI模型和应用大多运行在云侧，导致模型可能受到各种攻击，网络安全，单点故障等。同时在云侧的模型对用户来说，基本缺少控制手段和管理手段，完全依赖于服务商的安全措施。为更好的解决安全问题，必须引入端侧模型和端侧计算的解决方案。同时，模型训练和使用需要数据作为支持和基础，这些数据可能包含个人隐私信息。在数据的收集、存储和使用过程中，如果安全措施不到位，可能会导致个人隐私泄露。所以，操作系统作为用户数据的主要承载体，与AI结合时，如何做好安全和隐私保护是首要挑战；

计算资源的挑战 —— 大语言模型通常需要大量的计算资源进行支持，包括高性能的GPU、NPU等硬件以及大量的存储空间。端侧计算是很好的安全和隐私的解决方案，但对用户的终端也提出了算力要求。虽然用户侧算力持续增长，但是相对AI模型的算力要求和消耗增长，依然并不充分，二者依然存在一定的鸿沟。寻找均衡AI性能、AI体验与硬件资源的最优解，是当前操作系统与AI结合所不得不面对的挑战；

AI框架多样化导致软硬件生态复杂 —— 产业界各厂商在解决硬件算力时，同时会推出仅适合自身硬件架构的AI引擎和AI框架，典型的如NVIDIA的CUDA、Intel的OpenVino，不同的独立NPU厂商也会推出自身的引擎和框架。这样产生的直接结果是算力使能框架的多样化，任一大语言模型面对不同厂商的算力硬件，都需要单独适配该硬件对应的引擎，从而对整体硬件生态带来了极大的复杂度。操作系统必然面对各种不同的算力硬件以及配套软件工具链，硬件的生

态兼容建设也是操作系统重要挑战；

模型泛化能力不足 —— 大语言模型尽管在特定任务上表现出色，但模型在泛化到未见过的数据或任务时仍存在较大的局限性。当面对新的场景或数据分布发生变化时，模型的性能可能会下降，需要不断改进和优化模型以提高其泛化能力。同时在终端中用户场景又是极具个性化，模型的泛化提升，不免会滞后于用户场景的诉求变化。如何面向不同的终端用户进行个性化的模型泛化提升，使得模型能力能始终支持用户的体验和诉求存在一定挑战

商业应用挑战 —— 将AI技术应用到商业场景和具体的用户场景中，需要解决数据的整合、模型的定制化个性化、与现有业务系统的集成等问题。同时，如何确保AI体验符合商业需求和用户期望，以及如何实现商业价值的最大化，也是重要的挑战；

唯有克服这些挑战，操作系统方可实现与AI的深度融合，进而为用户提供更为智能、个性化且安全的使用体验。而要克服这些挑战，操作系统必须进行更为根本性的创新。这些变化与以往的发展路径截然不同，以往操作系统的发展是以自身为中心，对新的技术进行利用和集成。

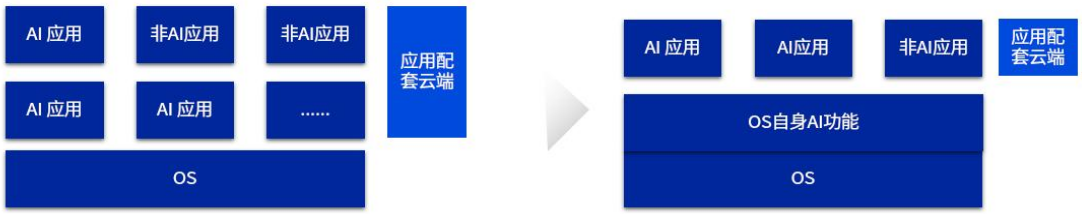
生成式AI带来的新变化赋予了更多新的可能，与AI的深度融合需要以AI为中心，从“OS + AI”向“AIOS”转变，即需要新一代的操作系统产品——AIOS。由AIOS承担起应对人工智能新挑战的责任，致力于创造和实现新概念与新场景。

二、 AIOS产品定义

操作系统的存在由来已久，有30年以上历史，经过多年的持续优化变得更加直观、友好。由于其特殊性，操作系统成为了连接软硬件的重要基础设施和载体。无论在何种形态的终端上，面向用户的应用、功能、体验和服务都需要以操作系统为基础来进行支撑和提供。

2.1 AI在操作系统中的应用历史回顾

伴随AI技术的整体发展历程，AI技术在操作系统中也有着长期的探索和应用历史。在过往整个AI的探索和应用过程中，操作系统充当了两方面的角色：

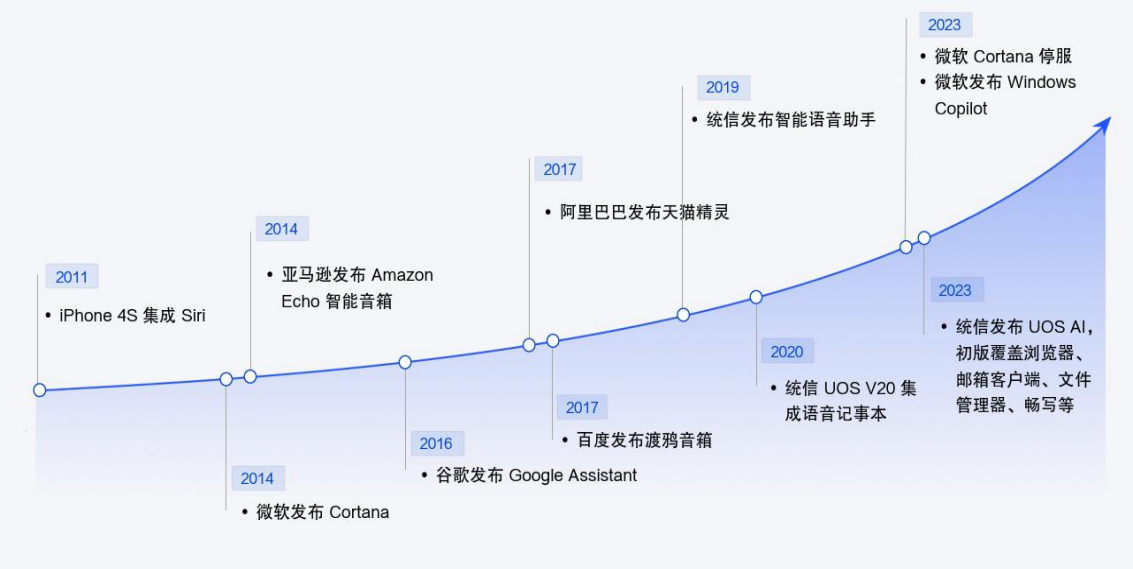


OS For AI　　AI For OS

OS For AI：AI应用作为众多应用之一被操作系统支撑。操作系统不直接参与AI技术的应用，由应用层自行集成AI，操作系统面向应用充当的原始应用支撑平台角色。如：图片处理应用、翻译应用，文字识别应用等；

AI For OS：操作系统自身应用AI技术提供AI功能和体验。操作系统在自身已有功能使用和体验中集成和应用AI技术，增强用户体验。如：操作系统支持语音识别、集成语音助手、使用面部识别技术来登录操作系统等；

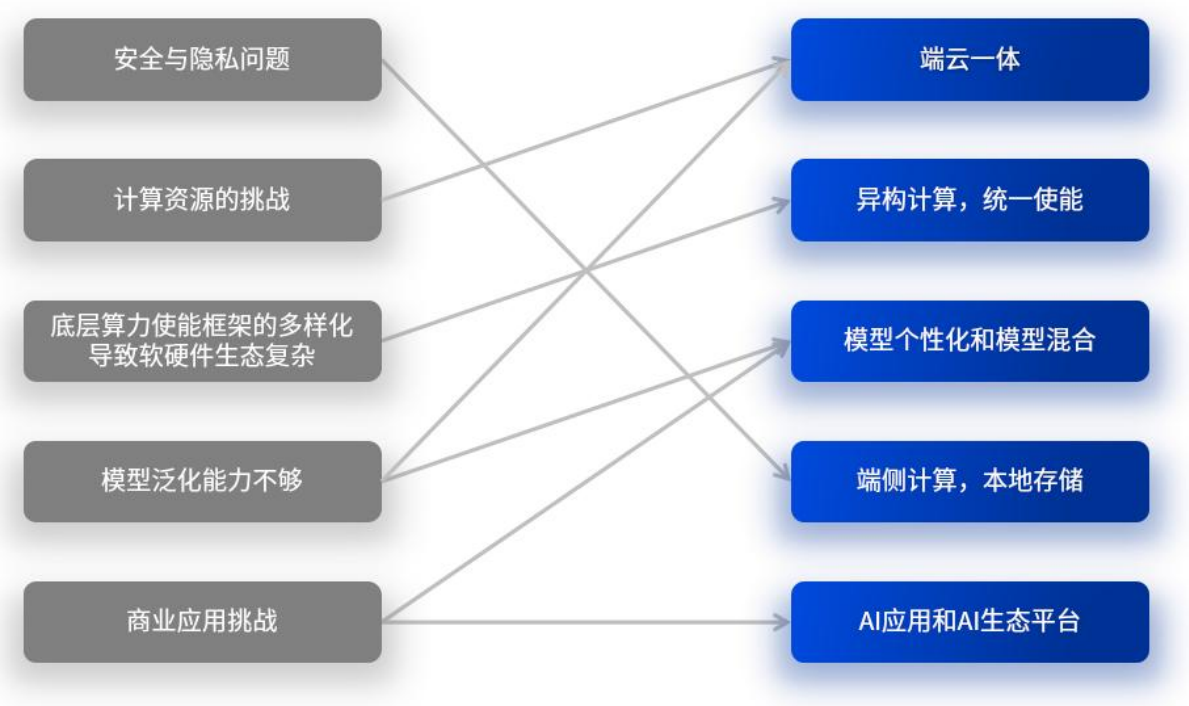
在 AI For OS中，统信操作系统也在进行过长期的探索和创新：



统信自2019年成立之初即开始探索AI在操作系统上的集成和应用，赋予用户直接的AI功能和体验，并在2023年推出UOSAI 1.0，成为国内首个AIOS概念提出和应用的厂商。

2.2 AI所带来新挑战的应对方案

针对操作系统与AI融合所需面临的挑战，首先需要有整体的应对和解决方案，统信针对挑战提出如下的应对方案：



端侧计算，本地存储 —— 端侧计算是针对安全和数据隐私问题的有效解决方案，在本地进行大语言模型的运行，可以相当程度上避免受到攻击。同时因为用户数据在本地进行存储和处理，减少数据在网络中的传输，极大的降低的数据泄露的风险。对于敏感数据，还可以进行更精细粒度的管理，始终保证数据在用户的管理和掌控中；

端云一体 —— 端云协同可以形成分布式负载均衡与弹性扩展，操作系统可以根据计算资源的可用性和任务复杂性，动态调整端和云的计算负担，实现更平衡的资源利用。当设备负载较低时，系统可以充分利用本地模型处理数据；在需要生成更复杂的内容时，则切换至云侧以获取更强的计算能力。通过端云一体化，对计算资源的不足和用户体验之间的均衡问题进行有效的优化和解决；

异构计算，统一使能 —— 通过构建统一的AI引擎技术栈，对不同NPU及算力厂商的AI引擎和推理框架予以统一，从而实现生态的归一化。针对新增NPU后衍生出的CPU、NPU、GPU两两组合或三者组合等不同算力组合，实施统一的异构技术调度与支持，以完成使能操作。对于必须使用独立引擎和框架的硬件设备，推出整体性的AI引擎框架层，对特殊的AI引擎框架进行封装，并向上层提供统一的模型接口，最终实现各种算力与各类AI引擎框架的兼容，达成对硬件的统一使能以及上层软件生态的归一；

模型个性化和模型混合 —— 操作系统可收集用户的输入数据（如文本、语音、图像等），对模型进行在线或间歇性微调，使模型逐渐适应用户的习惯，以解决模型泛化能力不够的问题。同时在此基础上，系统可集成和混合多个不同的模型，协同使用实现复合功能或更强的性能，根据任务的类型和复杂性选择合适的模型组合。最终形成多个个性化模型的混合结构，有效的解决模型泛化能力不够的问题；

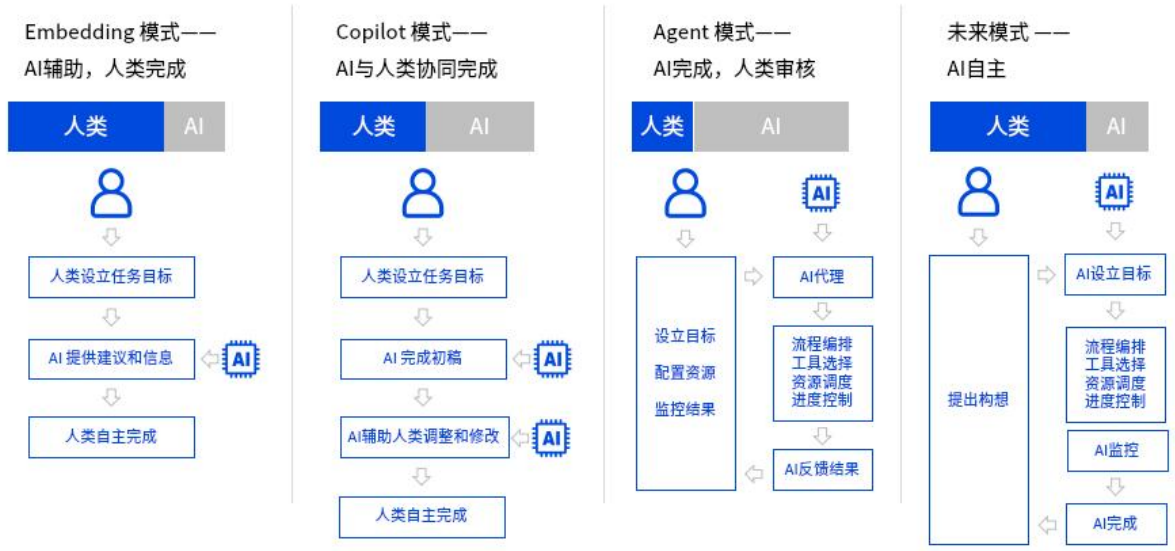
AI应用和AI生态平台 —— 具体的商业场景和具体的用户场景中，可通过模型的定制化个性化满足针对性的场景需要。同时也可针对具体的场景构建具体的AI应用和解决方案，满足用户的特定需求。AI生态平台则是支持AI应用开发、部署、和协同的综合性平台，汇集了模型、数

据、工具和开发资源。不仅为AI应用的开发提供基础架构，还通过整合多方资源和提供技术支持，建立一个有机发展的AI生态系统从而解决商业应用的挑战；

以上是统信针对AI所带来的新挑战提出的应对和解决方案，这些解决方案将组成统信AIOS产品的底层架构和核心理念。

2.3 AI与OS的融合模式选择

为了达成新一代AIOS构想，首先需要对AI与人的协同有更清晰和直观的认识。人与AI的协同模式当下总共有三种分类。分别是Embedding模式、Copilot模式、Agents模式。未来可能还会有一种超越Agents模式的新模式产生。



- Embedding模式在生活中非常常见，例如传统的AI辅助工具，车牌识别、图片分类、语音朗读等工具应用，这些应用已经伴随人类生活多年，成为生活中不可或缺的一部分
- Copilot模式则相较于Embedding模式更加智能，如今我们使用AI应用，大部分倾向于Copilot模式。在这个模式下，人类只需要做调整和修改就可以达成目标
- Agents模式则是当下最热门的AI应用范式，它的宗旨是让AI充当人类的代理，帮助人类完成既定任务，可以自己进行流程编排、工具选择、资源调度和进度控制的工作

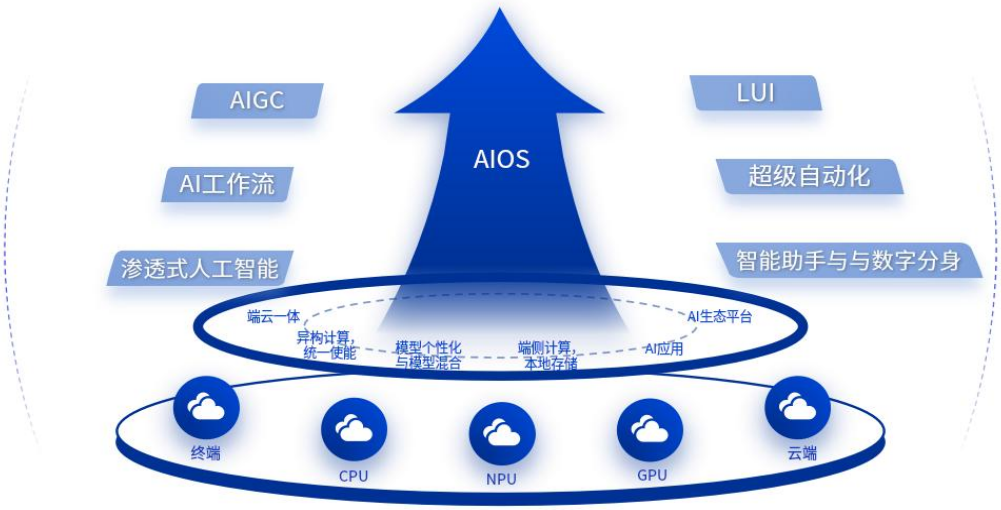
当前统信AIOS的目标为：在操作系统层面达到Agents模式的效果。Agent模式可以充分

发挥AIOS的优势，作为平台OS可以建立全局的感知Agent，打破应用间自闭环，达成AI感知、AI理解、AI完成，人类审核的高级协同模式。

未来可能产生专家模式、超智能模式，人类参与程度可能越来越少，甚至越来越多的工作都能由AI自主完成，只需要人类提出构想或参与授权即可。随着现在技术的飞速发展，相信不远的将来就能看到AI领域新的技术突破。统信AIOS也将不断适应AI发展变革，提出新的技术架构以支撑未来庞大的AI需求。

2.4 AIOS产品定义

基于以上历史分析、AIOS所需面对的挑战和应对方案以及AI与OS协同模式的选择，统信提出如下AIOS的定义:



AIOS 全面融合硬件的多种 AI 算力，通过端云结合的方式为用户提供系统级、全场景智能化体验，并提供丰富的智能体生态和原生 AI 能力，是智能时代的新型计算平台。

AIOS 需具备如下特点和能力：

- **用户层，通过 AI 重构系统体验** —— AIOS 通过 AI 重构用户与设备，用户与数据，用户与应用的交互体验，构建简单高效、安全可靠、智能的服务获取和主动服务能力
- **算力层，高效管理硬件资源** —— AIOS 提供异构计算的能力和解决方案，具备统一的 AI 计算调度使能能力，融合 CPU、GPU、NPU 多种组合的 AI 计算能力，高效管理硬件资源；
- **模型层，算力解耦端云协同** —— AIOS 中原生集成多种端侧 AI 模型与能力，并支持接入云侧，进行端侧模型与云侧模型的协同，同时提供全生命周期的多模态模型管理，兼顾数据安全与隐私，成本与体验；
- **生态层，重构全新应用生态** —— AIOS 通过自身的 AI 能力和服务，提供 AISDK、AIAPI 等开发接口，降低 AI 应用门槛，提升 AI 应用效率，构建全新的 AI 应用生态

2.5 AIOS目标用户

面向个人用户：统信AIOS为个人用户提供生产力的全面智能化支持。包括：全局的AI感知和系统级Agent，基于用户个性化的知识库和个人数据的AIGC内容创作和辅助，图片、PPT等创意场景的智能支持，日常系统管理的智能推荐和智能控制，个人日常事务、备忘、日程的智能管理等，显著提升生产力和工作效率。

面向企业用户：在企业环境中，统信AIOS通过结合企业自有业务和资源，在信息流、资源流、工作流三个维度提供智能化服务和体验，提升企业效率。包括：私有化调优和训练的模型；结合企业级数据和知识库为员工提供统一的企业公有AI信息服务打通信息差；企业级统

一Agent和个人Agent混合多数字助理；结合企业业务系统和应用的自然语言AI工作流；邮件、会议、日程等各种企业场景的AI内容、AI管理、AI安排等服务；在管理场景中提供数据的AI收集、AI分析、AI策略等。

面向政府用户：在政府场景中，行政效率和安全隐私是最重要最核心的场景和需求。AIOS可通过智能化的任务自动化、流程优化和文档管理功能，帮助政府部门减少人工干预，提高行政效率。如：公文智能协助、本地的垂直领域知识库和数据库、自然语言AI工作等。同时AIOS可通过本地计算，AI数据加密技术、智能访问控制机制、智能身份认证等，能够为政府用户提供高标准的数据安全保护。

面向开发者：AIOS为开发者提供开放的AI开发平台，支持开发者创建和集成自己的AI应用或AI能力。通过丰富的API和开发者工具，AIOS能够与各种应用无缝集成，支持开发者基于系统本地模型，AI基础设施扩展个性化的AI能力，构建自定义应用或工作流。此外，AIOS的架构还兼容AI模型训练、调试和测试的需求，支持开发者在本地进行轻量化的AI模型训练和推理操作。并可以在最终实际场景中，下载和运行自有模型。

2.6 AIOS主要场景和功能

统信AIOS定位为信息技术提供新质生产力，围绕生产力场景提供丰富AI功能，赋能用户和企业生产力生产效率提升。

统信AIOS核心两大场景：

- **生产力场景** —— 功能上围绕生产力提升和生产效率提升。此场景下，统信AIOS提供：全局AI助手、AI文件处理、AI图片处理、AI数字员工、AI自然语言搜索、AI会议助手、AI个人知识库等功能；

- **AI生态场景** —— 面向底层硬件构建AIOS底座和硬件生态，面向上层，构建AI应用生态，统一服务，统一赋能。此场景下，统信AIOS提供：异构计算、用户和应用自主模型管理、AI开发者支持等功能；

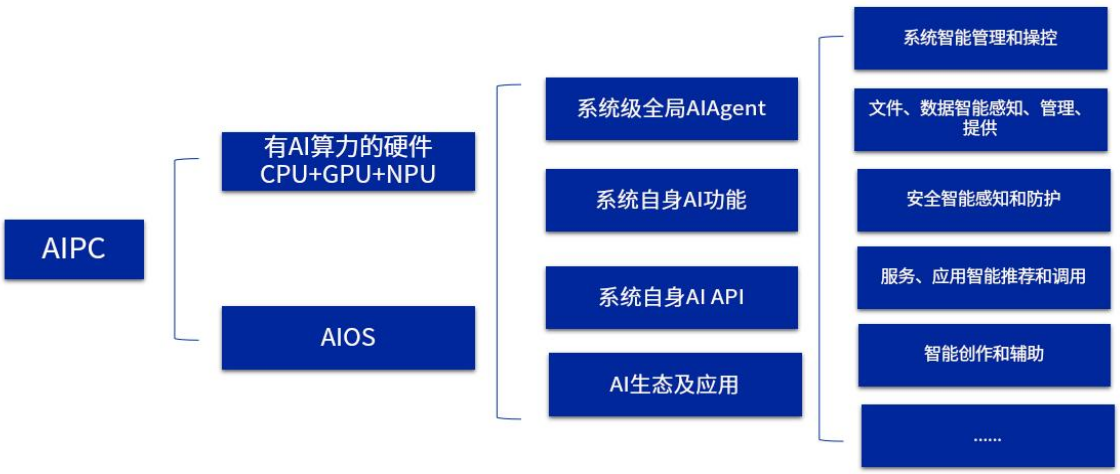
在此两主要场景之外，在企业级管理、运维、安全等场景中，统信AIOS也会结合AI和生态软件，探索和提供更多的AI功能。

2.7 AIOS配合的AIPC解决方案

OS与PC的关系密不可分，OS需要基于PC运行，PC需要通过OS才能为用户提供服务。在AIOS和AIPC之间，AIOS是操作系统层面的创新，而AIPC则是硬件层面的创新。AIOS与AIPC

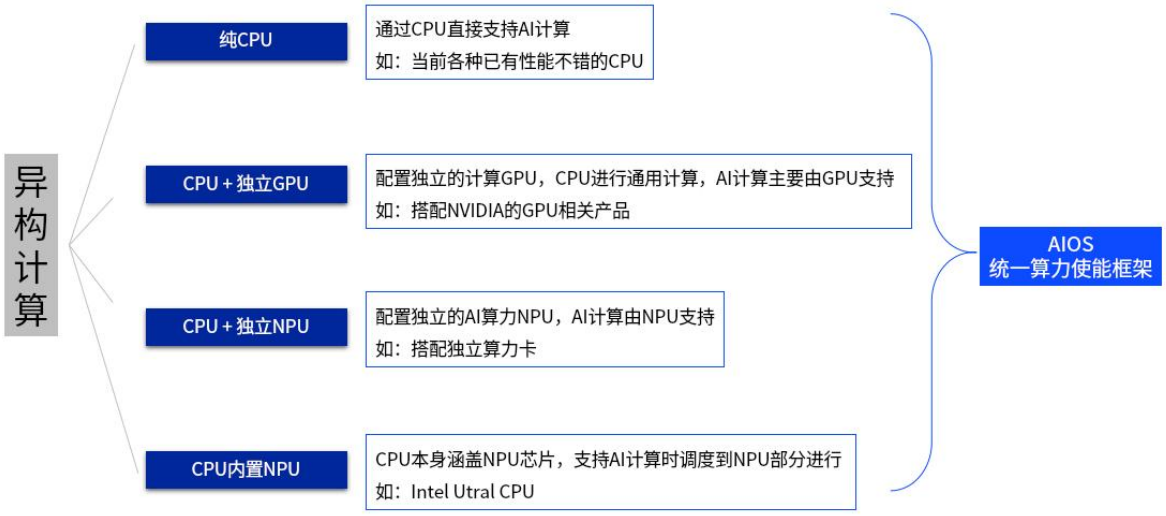
之间是软硬件密切协同的关系，通过二者的紧密结合，AIOS能够在AIPC的支持下，充分发挥AI的潜力，为用户带来更加智能、高效和个性化的使用体验。

在此，统信提出AIPC的基本定义：



- **AIPC = 有 AI 算力的硬件 + AIOS**
- AIPC 整体提供如下能力和服务：
  - 具备本地 AI 计算硬件和能力
  - 具备系统级全局 Agent，面向用户提供全局的 AI 服务和体验
  - 软件+硬件整体协同提供了丰富的 AI 功能
  - 提供了自身的 AIAPI，支持 AI 能力调用
  - 具有丰富的 AI 生态和应用
- AIPC 在 系统智能管理和操控、文件数据管理、安全保护、服务与应用、生产力与创作等场景为用户提供丰富的 AI 功能和体验

根据以上AIPC定义，统信通过AIOS实现丰富的功能和体验搭载在AIPC硬件产品上，支持有算力AI硬件的使能。**针对AIPC硬件层的不同算力架构组合，AIOS通过统一使能框架，实现了全量的支持和覆盖**：



AIOS统一算力使能框架对异构计算的支持，可以兼容各种算力解决方案的组合，AIPC可以在各种算力解决方案中进行合适的选择，最终都可通过AIOS完成整体软硬件协同，并通过

AIOS实现丰富的AI体验，为用户提供高价值的AIPC产品。

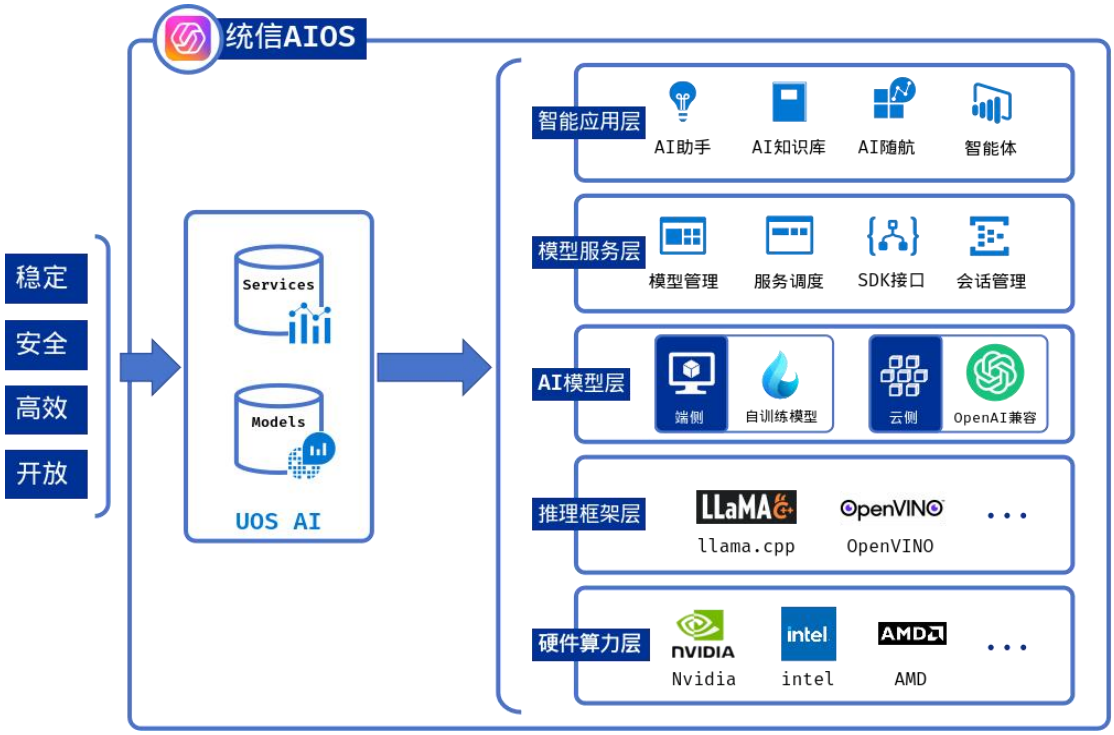
三、 AIOS技术架构

3.1 AIOS技术思想

人工智能的持续发展离不开底层服务支撑和软件平台优化。有越来越多的行业用户关注到框架、平台产品中的大模型能力，但在通往应用和大规模落地的过程中，还需要不断面对算力、数据、效果、成本、安全等多维度带来的挑战。

随着大模型从基础研发走向应用落地，AIOS技术架构的设计和选型成为人工智能软件基础设施建设的重要一环，AIOS技术架构的重要性和价值进一步凸显。大模型预训练完成了“从0到1”的技术统一，而大模型在通往“从1到100”的应用和大规模落地过程中，标准化的全栈基础软件和工作流是支撑大模型基础研发和应用落地的核心环节。

统信AIOS技术架构思想秉持了稳定、安全、高效、开放的原则，整体上从硬件算力层，到推理框架层、AI模型层、模型服务层、智能应用层五个方面进行了总体性设计。



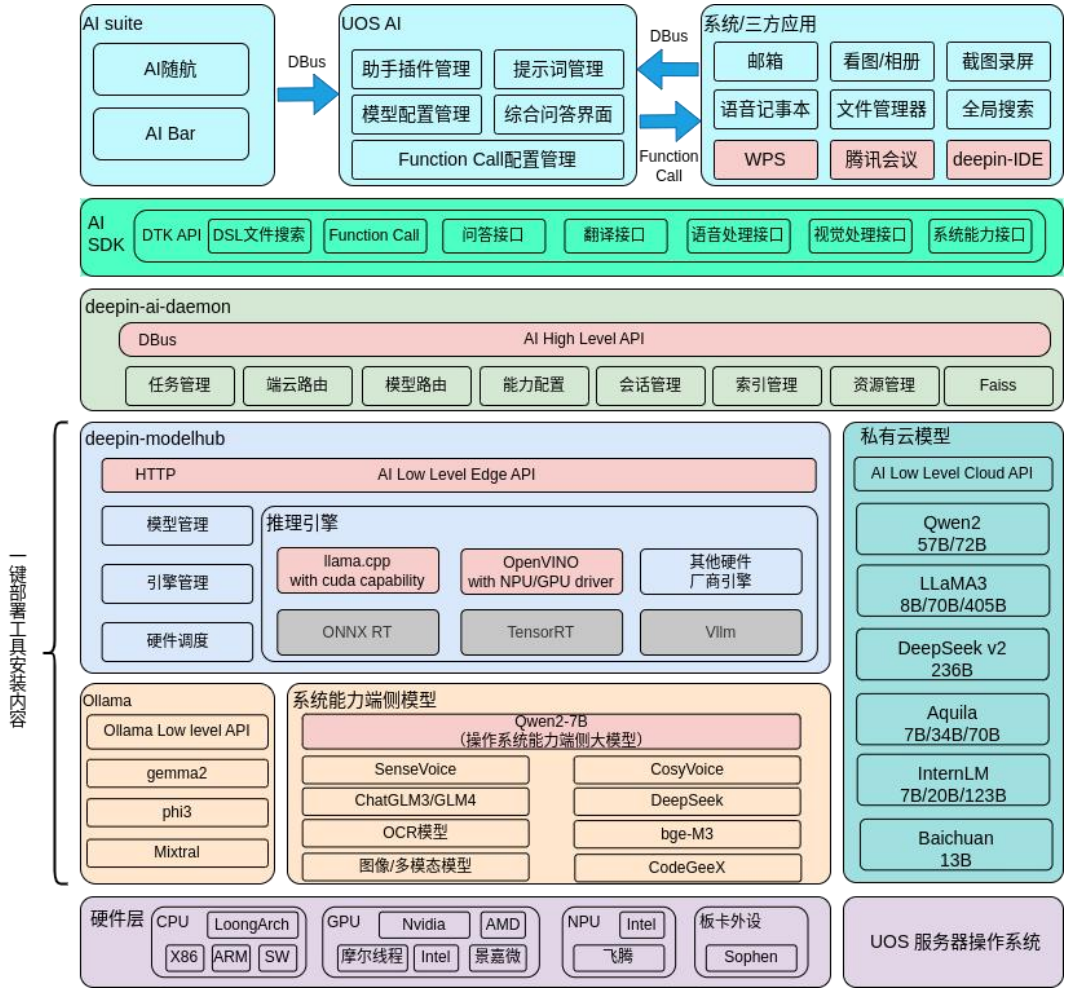
从技术上来看，一个成熟的AIOS技术架构在设计上必须考虑以下几点：

- **支持不同的硬件架构进行AI计算，实现异构计算和统一使能功能。**
- **支持常见的推理框架，具备端侧模型推理功能，例如llama.cpp、OpenVINO。**
- **在AI模型层，能提供端云协同管理的方式，对端侧模型和云侧模型进行统一管理。**
- **具备完全离线的端侧模型计算能力，保护用户的安全与隐私。**
- **提供统一的AI能力接口服务，用于应用层的AI赋能和三方生态应用接入。**

3.2 AIOS总体技术架构

统信AIOS在人工智能领域投入了深入的研究与探索，致力于将业界领先的AI模型、高效的推理框架、尖端的硬件设备与UOS操作系统的核心优势相结合，打造了一个创新的、开放的、多层次AI技术架构。这一架构旨在为用户提供一个智能且易用的工作环境，同时为开发者提供一个便捷、高效的智能平台，以促进创新和提升生产力。

统信AIOS总体技术架构如图所示：



AIOS架构总体架构分为5个方面：分别是异构计算与使能、端侧计算模型、模型管理、AI能力服务与AI SDK

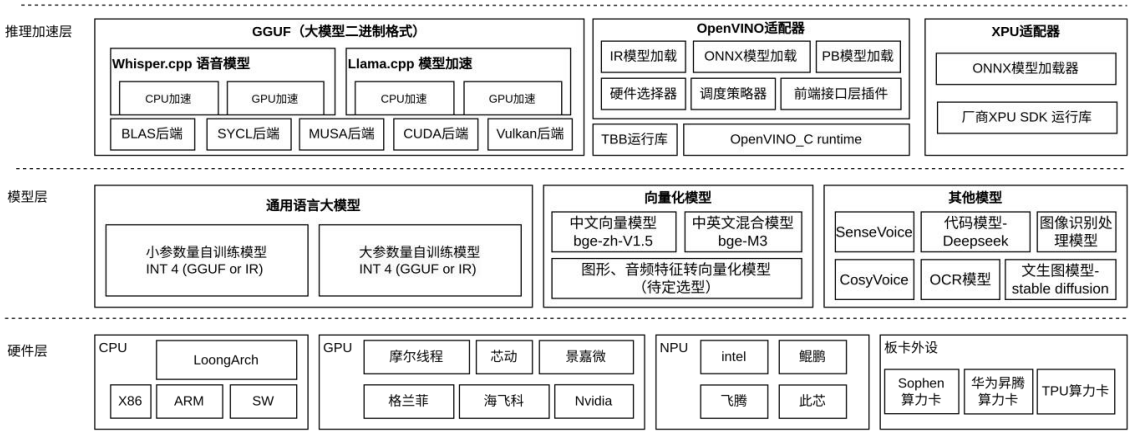
3.3 异构计算与使能

为支持不同的硬件架构都能进行AI计算，统信AIOS开发和设计了一套完整的硬件和调度使能框架：

- 针对Intel，完整支持OpenVINO，支持针对Intel CPU和NPU的AI计算调度
- 针对NVIDIA，完整支持CUDA，支持任意CPU与NVIDIA GPU组合下的AI计算调度
- 针对其他的架构（如x86、ARM、LoongArch等）CPU 和AI加速器（如GPU、NPU）的异构组合，通过llama.cpp作为端侧部署的主推计算调度方案进行支持

在对Intel、NVIDIA的兼容之外，在所有流行的端侧模型推理工具中，从资源占用、部署

难度、维护难度、扩展性、社区支持程度方面综合考虑，以及与各生态伙伴充分讨论，**统信选用llama.cpp作为主推和公共技术方案，进行深度开发和支持**。并将以上三大使能方案，整合成为统信AIOS的硬件使能层，对AI硬件和算力的各种方案和架构进行统一支持。



llama.cpp对于端侧推理和计算有显著优势：支持通用CPU本地推理加速，同样也支持CUDA和OpenCL推理加速，具备CPU+GPU+NPU的混合调度能力。在计算精度上具备FP16和FP32的混合精度，在模型量化方面，支持4bit、8bit模型量化。llama.cpp运行期占用内存更小，推理速度也更快。

3.4 端侧模型推理

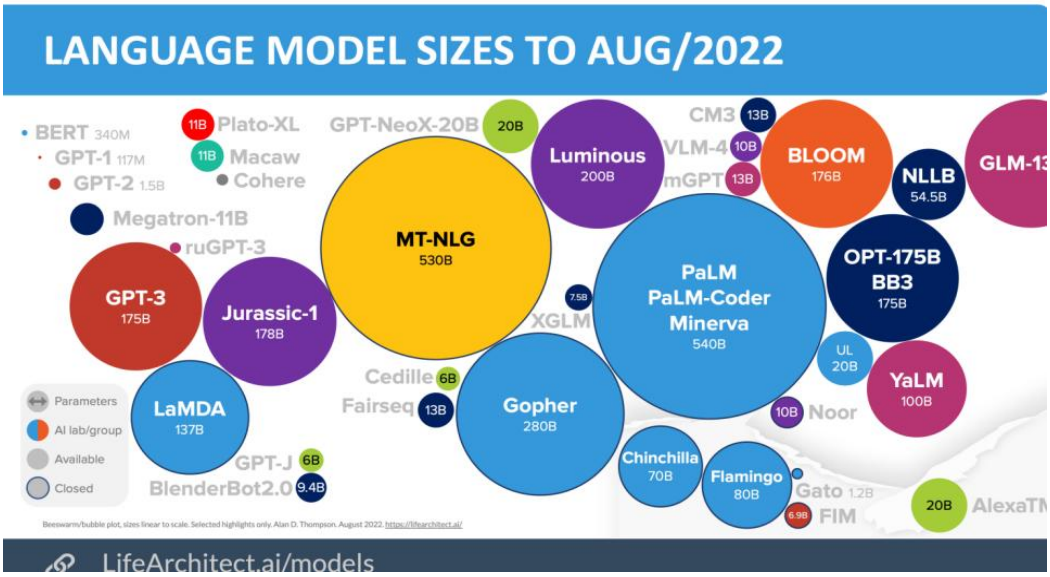
要在端侧完成大模型推理，目前存在不可能三角：性能、参数量和内存及功耗占用，这也是各家厂商最需要解决和考虑的问题：

要想模型的效果优异，就要参数量大；而参数量大，就意味着内存占用大，功耗也会大；功耗过大又可能会影响性能。

对于三者如何平衡，统信UOS的观点是：**针对模型进行量化，有效减少模型能耗。**一方面，量化可以降低模型的内存或者显存占用。另一方面，量化也可以大大减少系统的算力开销。

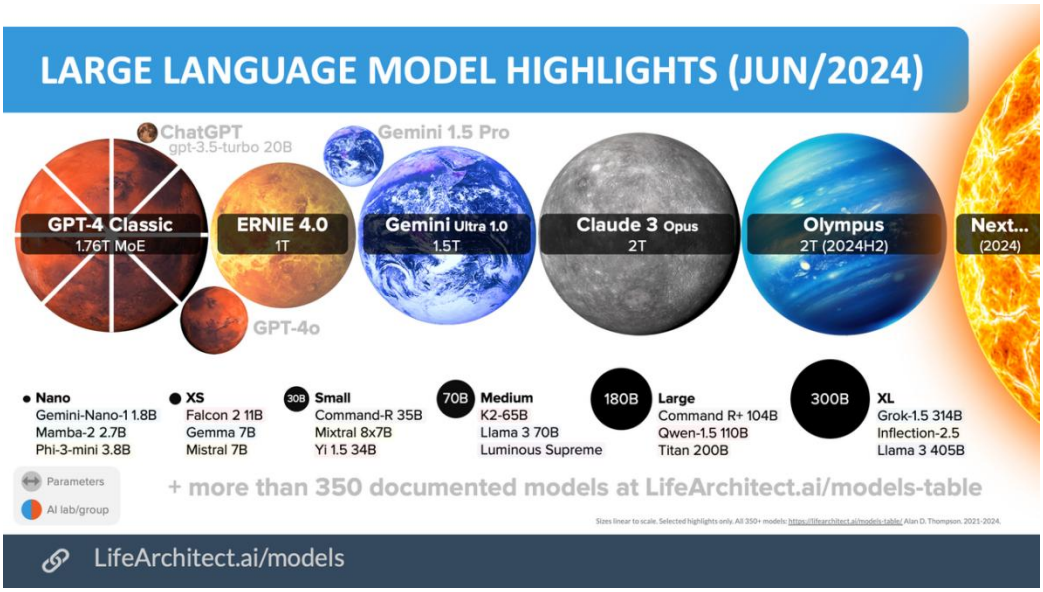
统信AIOS从模型参数量级、硬件平衡策略两个方面，针对端侧模型进行了系统级优化：

模型参数量级一直是大模型军备竞赛的红海区域，当今各路友商们都认为端侧模型趋势已经是箭在弦上，不得不发，但又在端侧模型庞大的参数量级面前犯了难。我们可以回顾看看2022年统计的大模型训练参数量级情况：



在2022年，当时不可一世的GPT-3的训练参数量级已经来到了175B，但在参数量这个赛道上，它依旧不是王者。从图上我们还可以看到有Luminous的200B、Gopher的280B、PaLM的540B，越来越大的参数量，带来的计算压力和存储压力也是巨大的。在2022年的当时，几乎没有个人PC能够承载大模型的运算参数。可以说当时的大模型仅属于硬件俱乐部的宠儿，个人想使用大模型，除非搭建Nvidia计算卡集群，如此苛刻的条件下，全世界大模型厂商几乎是要靠不断砸钱才能维持运转。

如今发展到2024年，大模型世界发生了一些新的变化。



从图上可以看出，大热门GPT-4的总参数量达到了惊人的1.76T，另一个横空出世的Claude3 Opus 的参数量达到了更加惊人的2T，同样Olympus的参数量也达到了2T，在参数量上，各个大模型厂商依旧不遑多让。但另外一方面，我们也观察到一些大模型开始把自己越做越小，比如微软的Phi-3-mini只有3.8B，Gemma也有1.5B和7B版本，这些模型在通用任务和一些特定任务上的表现依旧非常出色。

参数量小的模型，依旧保持了很好的泛化能力。这是由于Transformer不仅是另一个神经网络，而是一个极具通用性的“差分计算机”。它通过前向和后向传播进行自我调整，能够高效处理复杂任务。Transformer的扩展性是AI领域的重大突破，使得大规模模型成为可能，但同时也可以通过蒸馏、剪枝等技术，能够将大模型的能力压缩到更小的模型中，实现更高效的认知处理，甚至1～10亿参数的小模型就能完成复杂任务。

UOS LM模型把参数量定义在1.5B和7B两个区间，使得模型既能保持原始模型优秀的泛化能力，同时小参数加上量化技术后，可以把模型压缩至1G以下，尽最大可能节约用户空间，使用户流畅体验端侧模型能力成为可能。

统信AI技术架构在端侧模型的部署和运行中，采用了一种创新的方法，即基于预训练模型来增强操作系统的智能应用场景。这种方法涉及到从操作系统的日常使用中提取数据，构建专门的训练数据集，这些数据集反映了用户的实际需求和操作习惯。通过这些数据集，统信AI对模型进行进一步的训练和微调，赋予模型更加丰富的功能和更高的适应性。

3.5 端云协同管理

统信AIOS系统赋予了用户加载个性化AI模型的能力，统信AIOS设计了一套完善的模型管理框架deepin-modelhub端侧模型管理组件，提供本地模型的管理、下载、安装、启动停止、模型参数配置、端口配置、自动添加UOS AI对接配置等功能。核心是硬件策略配置、模型会话管理、模型资源管理内容。并且当前支持统信AIOS自训练模型、Ollama生态兼容模型、厂商

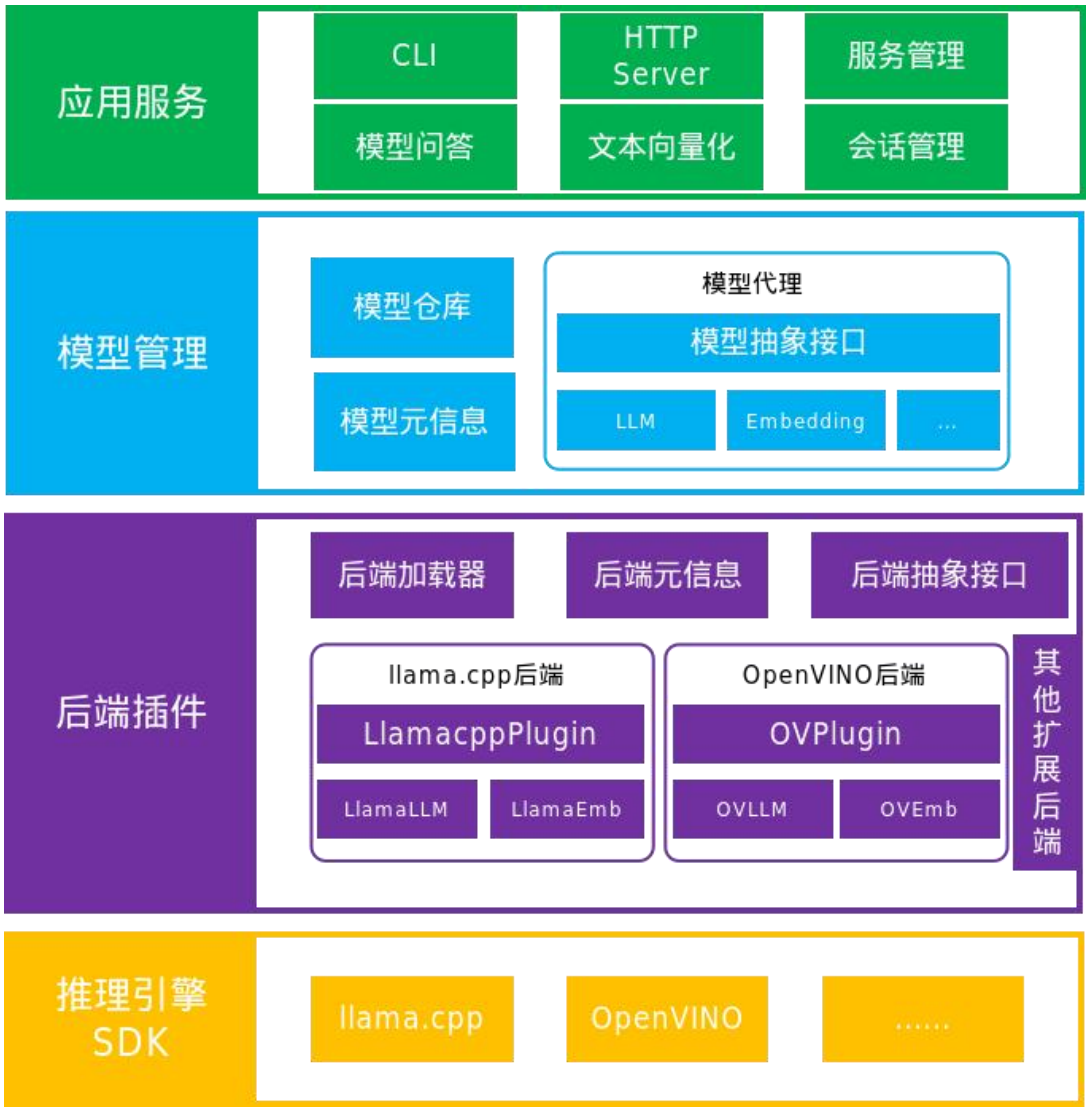
的适配模型等。

模型管理组件（deepin-modelhub）是模型管理与推理的核心，它专门设计来应对AI模型在多样化硬件环境中部署和运行的挑战。该组件通过提供标准化的接口和流程，极大地简化了模型的集成、部署和管理过程。它支持多种硬件平台，能够智能地根据不同的硬件环境自动加载和执行推理任务，确保了模型在各种设备上的高效运行。此外，组件提供了统一的HTTP访问接口，这些接口按照OpenAI接口规范实现，提供模型原始的能力。

组件还实现了模型的全生命周期管理，包括模型的启动、运行到停止，以及状态查询等操作，为模型的持续运行和维护提供了强有力的支持。它支持多种推理框架，如llama.cpp和OpenVINO，能够根据系统和硬件环境自动选择最优的推理框架，以确保模型性能的最大化。支持接入硬件厂商的模型推理框架，以兼容、适配各大AIPC硬件。组件的易用性和可扩展性通过命令行界面（CLI）和HTTP服务得到体现，这使得模型访问和交互能够适应不同的应用场景。最后，组件采用统一的模型组织方式，这不仅便于模型的管理和维护，也提高了整个AI系统的可操作性和稳定性。

AI模型管理组件基于对现有模型管理解决方案、本地部署策略及推理框架的调研，结合UOS AI的特定业务需求，旨在实现以下关键目标。

- 高效开发：采用 C/C++与 Qt 框架进行开发，以确保高性能和跨平台兼容性。
- 模型管理：实现模型的全生命周期管理，包括启动、停止、状态查询等基础操作。
- 框架兼容性：支持包括 open vino、llama.app 在内的多种主流推理框架，实现模型推理的封装，并提供标准化的模型运行时接口。
- 访问接口统一化：支持通过控制台和 HTTP 协议进行访问，确保用户可以通过统一的接口与模型交互。
- 现有模型适配：对 UOS AI 当前使用的模型进行适配和封装，以实现无缝集成。



这项技术可以让用户加载属于自己的模型，技术路线上兼容了llama.cpp、OpenVINO、

NVIDIA的生态，用户除了可以在deepin-modelhub里运行自训练大语言模型之外，还可以自己添加llama3.1、phi-3、gemma2、qwen2等模型。这些模型可不是只能躺在命令行里提供指令，而是可以直接链接UOS AI助手，用户可以自己手动训练一个专属模型，直接享受端侧模型带来的便利，最大效能发挥机器硬件算力带来的魅力。

3.6 安全与隐私

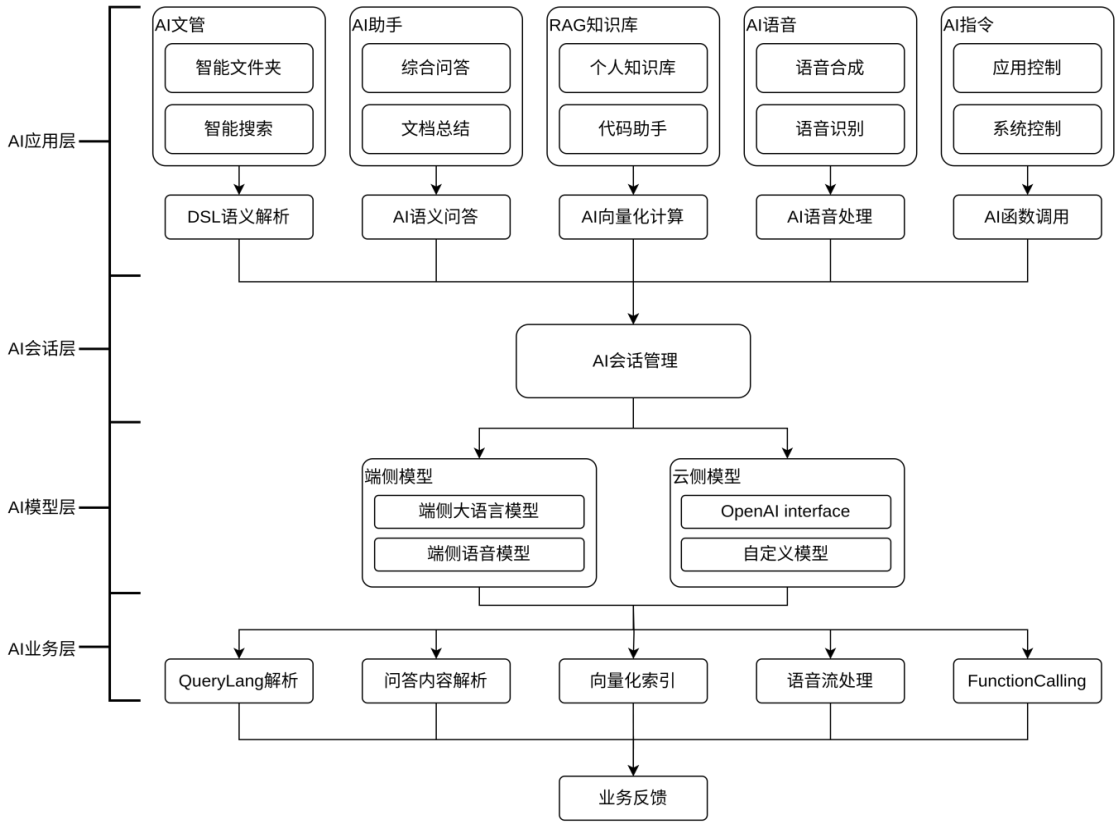
当前AI领域的安全问题主要来自于两个方面：

- 云侧推理时，用户数据的采集、传输、存储过程中的安全漏洞导致的安全问题。
- 端侧推理时，AI模型本身的数据污染、设计缺陷导致的错误决策问题。

在统信AIOS设计的技术架构中，端侧模型能力和云侧模型能力都能统一管理并提供一致性的AI能力服务，但当前大多数AI安全问题都来自于云侧模型推理过程中导致的数据泄漏问题。这些问题主要集中在以下三个方面：

- 在云侧AI平台大模型的运行过程中，需要收集大量的用户数据，如个人信息、操作习惯、地理位置等。如果这些数据在采集过程中未得到妥善处理，可能会导致用户隐私泄露。
- 在云侧AI平台收集到的数据需要进行存储，若存储系统存在安全漏洞，可能会被黑客攻击，导致数据泄露、篡改或丢失。
- 与云侧AI平台通信过程中，数据在网络上进行传输时，如未采用加密技术，可能会被窃取或篡改，影响数据的安全性和完整性。

为规避云侧AI平台带来的安全隐患，统信AIOS提供了完整的端侧模型运行全套链路，在离线状态下，统信AIOS也可以提供完整的AI能力服务。



从整体结构上，统信AIOS凭借自训练端侧模型的能力，即使在离线状态下也可以提供一系列AI应用业务功能。极大削减了云侧模型带来的安全隐患问题。

统信AIOS的端侧模型在训练过程中，也采用了高安全保障方案进行设计和训练。

AIOS自训练端侧模型的训练过程一目了然，通过微调提升大模型在UOS理解、自我认知和多个端侧任务的处理能力。数据集包括预训练、微调数据，来源于UOS公开信息爬取、人工生成，在数据来源层面保障了安全。

在模型优化方面，使用了多种量化技术为不同推理后端生成模型。确保模型的整体性能处于高水平。持续优化模型表现。在优化技术上保障了模型的安全性。

训练阶段合并多个任务的微调数据集，提升模型在多个下游任务上的能力，通过对齐微调，提升模型任务的精度，确保模型输出与预期一致。避免模型产生混乱的情况，保障了输出的安全性。

评估阶段统信构建了完整的自动化多维度评估系统，及时监控训练后模型以及量化后模

型的指标和精度变化。在模型评估和输入输出前后一致的层面上保障了安全性

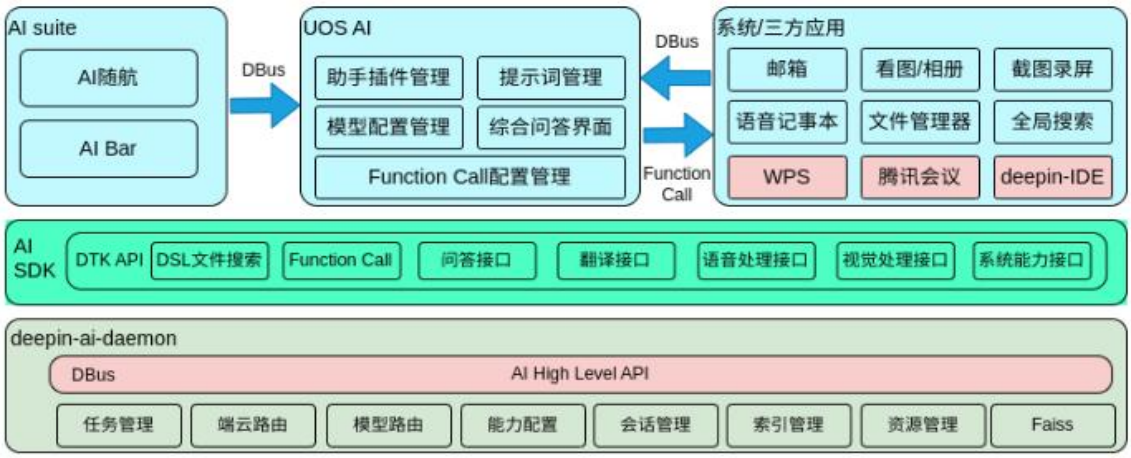
通过上述四个阶段的迭代改进和多阶段优化，确保模型在端侧多任务处理能力上的提升，实现自训练模型在统信AIOS中的高效、准确和安全。

3.7 AI能力服务与AISDK

AI服务（deepin-ai-daemon）是构建在大模型上的服务组件，旨在将大模型的原始能力转化为实际的上层产品功能。这一服务组件的设计考虑了灵活性和多样性，支持本地模型与云侧模型的无缝接入，使得用户可以根据具体需求和场景选择使用端侧模型或云侧模型。

为了进一步扩展AI服务的能力，统信AI正在积极探索和开发新的应用服务。使用向量化模型，将文档和文件转换为高维空间中的向量标识，这些向量能够捕捉内容的语义信息。结合Faiss数据库的高效搜索能力，可以为用户提供快速、准确的文档和文件向量搜索服务。

AI服务组件的能力也将通过DBus接口的形式开放出高层次的API，允许更广泛的系统组件和第三方应用程序能够无缝地集成和利用AI服务的能力。



在AI业务能力层面，统信AIOS系统可以从通用语言问答、文档总结、Function calling、文件搜索、语音识别、语音合成、图像识别等方面，把模型能力装在一个具体的业务场景中。开发者可以通过AI业务能力接口来调用到自己想要使用的大模型能力，用户在使用文件管理器、全局搜索、看图、相册、UOS AI助手这些应用的过程中，也在无感地使用到大模型提供

的业务能力。将智能体Agent概念融化到系统和应用中，是操作系统AI能力跟用户之间的桥梁，也是模型能力的最直观体现方式。

统信AIOS在AI应用层面，AI服务（deepin-ai-daemon）提供了不同的应用接口场景设计，技术点如下：

- AI SDK：提供high level能力接口，面向应用层
- UOS AI：对标Copilot能力，可接入云侧和端侧模型接口，主要功能场景为综合对话、系统功能调用、智能体嵌入、提示词框架等。主打AI综合能力
- AI 随航：提供划词后针对文字的扩展体验，可提供文字的翻译、扩写、润色等实时性体验。主打WPS、微信生态体系
- 自然语言引擎：利用自然语言在大模型的泛化特性，在系统中实现类似中文编程的效果，利于用户使用系统和传播。主打系统体验
- 智能IDE：提供类似cursor的智能编程体验。需要结合CodeGeeX端侧模型。主打开发者
- 全局智能搜索：主要利用模型输出结构化数据能力，比如DSL、Partial Mode提炼关键词，做文件搜索场景。主打文件体验

在统信AIOS技术架构中，AI服务（deepin-ai-daemon）不仅仅是模型管理和推理的平台，它还致力于将AI技术应用于更广泛的服务领域，以提供更加丰富和个性化的用户体验。为了进一步扩展AI服务的能力，我们正在积极探索和开发新的应用服务，这些服务将利用先进的向量化模型和Faiss数据库等技术，构建和优化个人知识库。

随着AI服务在应用服务方面的不断深化和扩展，统信AI技术架构将能够为用户提供更加全面和深入的智能体验。从基础的模型管理和推理到高级的知识管理和搜索，AI服务（deepin-ai-daemon）将成为连接用户与智能应用的桥梁，推动个人和企业在信息时代中的创新和发展。

配套AI服务，统信在AI SDK整体上提供三个层面的接口能力：

- AI 业务插件（AI - Daemon）：灵活的 AI 业务插件能力，具备系统级和应用级调用配置，三方应用可注册自己功能接口进入 AI 推理逻辑；
- AI 接口能力（DTK AI）：面向编程的 DTK AI 接口能力，提供文件搜索、问答、总结、润色、语音等业务能力，减少应用开发工作；
- AI 模型管理（Model hub）：原生模型接口能力，融合模型原生输入输出接口能力，方便用户调试自定义提示词或内容解析。

Deepin Tool Kit（DTK）是UOS操作系统提供的应用开发框架，专门用于操作系统及其相关应用的开发。DTK的设计目标是为开发者提供一套简单、高效、统一的API，从而提升应用程序的开发效率，并确保在操作系统上实现一致的用户体验。在面向AI场景下，统信AI在DTK中加入了AI开发接口DTK AI。DTK AI模块通过对AI服务接口的封装，简化了AI功能的调用流程，使得开发者可以以更加简单、统一的方式接入统信AI提供的智能服务。

DTK AI模块对复杂的AI服务接口进行了二次封装，提供了易于使用的高层次API，开发者无需深入理解底层的AI计算和模型推理过程，即可快速调用AI功能，如语音识别、图像处理、自然语言处理等。DTK AI接口是简洁的，开发者可以通过少量代码实现复杂的AI功能。这种方式大大降低了AI技术的应用门槛，让更多的开发者能够快速上手，将AI能力集成到应用程序中。DTK AI不仅封装了常用的AI服务接口，还提供了与操作系统深度集成的API。开发者可以通过这些接口直接调用操作系统中的AI能力，借助DTK AI实现更加智能的操作系统交互。

AI接口能力（DTK AI）一览表：

接口能力	接口描述
翻译接口	提供中英文翻译功能，输入原文本，输出译文
TTS	中英文字转语音功能，输入文本，输出音频的字节流数据
ASR	中英文语音转文字功能，输入音频的字节流数据，输出文本

文档总结	总结文本功能。输入大片文本，输出总结后的精简文本
通用问答	聊天对话功能。输入上下文，输出模型回答
系统调用	提供系统接口的能力调用和注册机制
文件搜索	传入自然语言，解析关键词后进行本地文件搜索能力

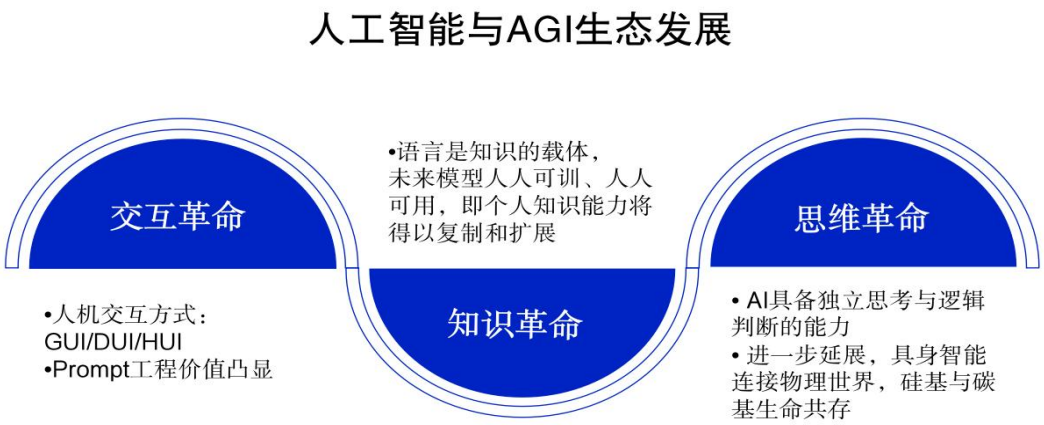
AI插件能力（AI-Plugins）一览表：

业务能力	业务描述
模型路由	查询模型运行状态，提供模型的Low API服务地址
模型调度	调度modelhub运行模型、关闭模型、管理模型功能
索引管理	管理文件的特征索引、向量索引
AI业务插件	提供复杂的AI能力插件。如向量化的索引创建、更新、搜索接口、问答，DSL，Function Calling，语音文字转换，翻译等

AI模型服务层（Modelhub）一览表

业务能力	业务描述
推理引擎	基于推理框架提供的API二次开发，封装llama.cpp、OpenVINO、ONNX等推理引擎
模型推理	根据提供的本地模型格式选择对应的推理框架，执行模型推理
硬件策略	根据硬件环境（CPU指令、GPU、intel NPU等）、运行时（如CUDA、Open Vino Runtime）和模型格式选择最优的推理框架和推理设备
模型接口	提供模型的HTTP访问接口，遵循Open AI接口规范。根据所运行模型，提供对应的接口。如语言模型提供chat接口、向量化模型提供embedding接口
推理框架	执行模型推理的框架，提供llama.app或其他模型加速框架插件能力

四、 统信AIOS生态愿景



近年来，人工智能的创新能力正在推动着操作系统行业走向新的发展阶段，通过AI技术的引领，操作系统从传统的功能提供者转变为智能化赋能者，为用户提供更加智能、高效的操作体验。因此，发展全新的AI软硬件生态不仅激发了操作系统行业的创新动力，还带来了转型效应，推动着各行各业的信息化技术升级与变革。统信软件持续发展AI技术在操作系统上的不断演进和融合，携手生态合作伙伴共同创造出更加智能、便捷的数字化生态，持续探索AI在操作系统领域的潜力，不仅着眼于技术层面上的突破，更希冀在商业模式、服务范畴等方面能够带来全新的市场机遇。

4.1 软件生态愿景——智慧融合，AIOS定义数字未来

在软件生态发展层面，随着AIGC技术的爆发，相关技术公司迅速崛起，大模型及后续调优及适配技术能力强，各企业纷纷借助大语言模型甚至多模态模型不断地切入市场刚需。与此同时，垂类企业也依托自身的领域积累了大量的专业数据，当下也直接借助第三方或者自建AI技术，构建垂直的大模型或应用，极大地推动AI智能应用的发展。随着AI大模型持续更新迭代，从模型架构、训练数据、训练规模、指令微调等维度持续提高整体核心性能，也将不断推动AI应用场景的拓展和功能体验的创新。

统信软件在AI软件生态构建上进行了前瞻布局，2024年统信正式推出中国首款操作系统级端侧模型——UOS LM，实现端云高效协同，并努力打造下一代新AIOS操作系统。与此同时，统信软件可以提供专门针对AI应用开发、部署及优化的AI操作系统能力，结合高效的原生开发工具和AI框架支持，为国内外主流的AI大模型和AI应用企业提供全面的系统平台能力和技术服务。目前，80%以上的国内移动互联网热榜AI软件已经上架统信UOS应用商店，累计下载超过50万次，更是有10余款与第三方ISV开展合作的AIGC和Agent产品方案正式投入商用使用，例如公文写作、智能白板、图像、会议助手等等场景。通过深入合作和不断积累，在供给侧，统信AIOS的软件生态发展已经形成“OS+通用大模型+领域大模型+行业大模型+企业/个人小

模型”这一基础AI软件业态，通过模型部署、开发工具、算法优化、定制集成以及联合研发等合作维度，助力不同行业的用户实现高性能、高效率的AI软件解决方案：

（1）模型部署和推理：拥有良好的文件系统支持和管理、开放的工具实现自动化操作和稳定的多任务管理、安全的用户管理和权限控制等能力，可以满足互联网、垂直行业、企业定制的AI大模型基于统信AIOS进行快速部署和高效推理。

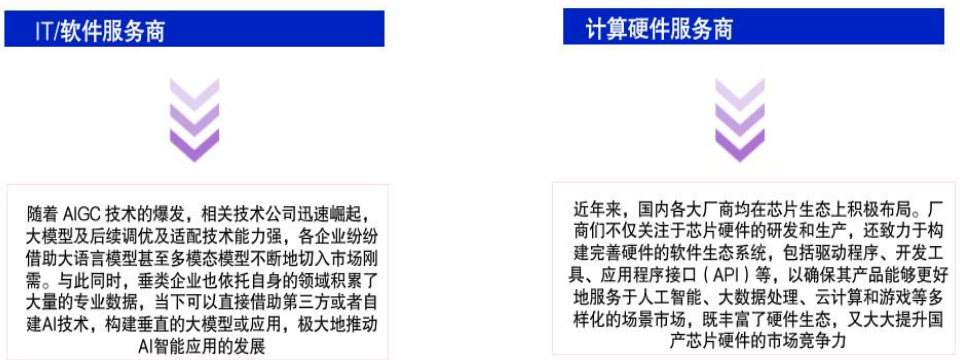
（2）开发工具和库支持：提供丰富的AI开发工具和库，包括深度学习框架、数据处理工具等，帮助企业及个人开发者快速构建和优化AI应用。

（3）算法优化和加速：针对AI大模型和复杂AI应用，提供算法优化和加速技术，提高计算效率和性能。

（4）定制化集成和应用：通过广泛的生态合作，将不同业务场景的AI大模型和应用集成到统信AIOS产品中，并提供个性化、定制化的功能和服务，满足用户不同需求和场景的要求。

（5）联合研发与实验局：与产业链上下游企业合作，以及倡议合作伙伴共同探索AI模型和应用在跨行业场景的应用与创新，实现算法能力的整合和交叉共享，实现数据和功能的互通互联，提升产品的智能化水平。

（6）开展技术交流和培训活动：帮助ISV了解最新AI技术趋势和开发工具，提高其研发能力和竞争力，协助ISV解决技术难题，优化应用功能。



4.2 硬件生态愿景——智能驱动、创新领航，AIOS加速算力变革

近年，政府出台一系列政策优化算力基础设施布局，在政策的支持下国产硬件数量剧增，极大的提升了智能算力比例。同时，国内各大厂商均在芯片生态上积极布局，大大提升国产芯片硬件的市场竞争力。统信软件通过与CPU、GPU、整机以及各类具备AI加速能力的硬件算力、方案厂商的广泛合作，以联合进行统一的接口标准制定、产品与技术融合、行业场景用户生态共建等方式，共同打造基于中国操作系统底座的卓越AI产品和解决方案，目前已经实现的有：

（1）硬件兼容性和优化：统信AIOS支持全架构CPU平台、国内外主流GPU卡，并联合头部OEM品牌，提供针对不同硬件组合的优化和适配能力。

（2）统一标准与开放合作：统信AIOS与硬件厂商紧密合作，提供开放的驱动程序、开发工具、应用程序接口（API）等，支持第三方进行AI能力的集成和优化，并通过行业场景的生态合作，加速AI硬件生态的拓展和支持，为用户提供更多选择和定制化服务。

同时，统信AIOS仍将不遗余力探索与硬件能力的深度融合：

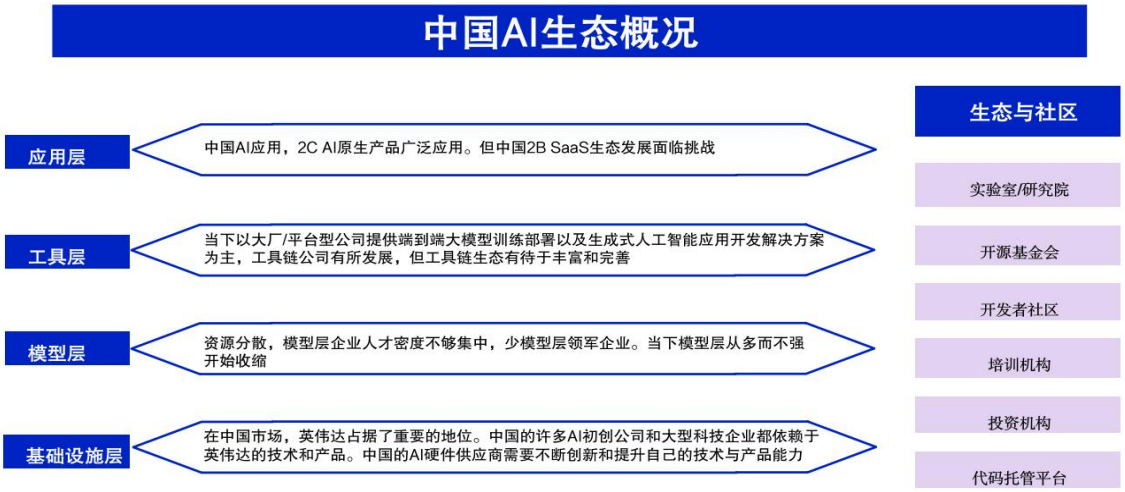
（1）性能优化和加速：针对AI模型和应用需求，持续优化系统内核、调度策略和资源分配，提供AI任务加速和性能提升的支持。

（2）弹性扩展和灵活部署：支持异构AI算力工作、自动识别算力资源并实现灵活部署，提高系统的弹性和可扩展性。

（3）数据保护和网络安全：支持跨不同终端之间的数据传输的硬件加解密能力，支持私有化部署，确保用户数据在整个AI产品端到端的安全传输和处理。

随着政策支持下国产硬件数量的增加，华为等国内各大厂商均在硬件生态上积极布局。例如，华为推出了 CANN 以及对应的生态体系，力求突破技术瓶颈。此外，像景嘉微、摩尔线程等公司也在自主研发硬件产品，并努力打造与之兼容的软件环境，以减少对外部供应商

的依赖，提升国产硬件的市场竞争力。统信软件将持续与合作伙伴一起构建一个完备的AI操作系统硬件生态，以支持各类整机和AI算力卡的优化和适配为目标，并最终为用户提供高效、安全的AIOS环境在各行业的广泛应用，促进数字化转型和创新发展。



AI时代大模型带来生产力水平的显著提升，推动软件应用生态进入新的发展阶段，也必将诞生新的超级用户入口。各类厂商也纷纷提前布局，将结合自身资源禀赋和技术迭代趋势布局最具创新价值的AI应用类型，其次构建AI开发层能力，支撑AI应用生态落地，厂商也加速布局AI Agent，掌握AI时代的用户入口。AI时代C端+B端应用价值的实现将带来全产业的生态重塑，从底层基础设施到AI开发到解决方案，不同环节的生态参与方迎来新的布局机会。统信将携手软件伙伴在AIOS软件产品能力的聚合、集成与创新过程中，一起完善AI软件生态系统。从软件生态伙伴的产品维度，AIOS将为AI大模型和应用伙伴提供全面的产品支持和技术服务，助力他们在AI领域取得创新突破。从广大的软件开发者维度，则更着重提供便捷的开发工具和资源能力，以辅助他们快速构建、部署和管理复杂的AI应用，促进AI技术的广泛应用和产业化。最终，统信软件通过推动AI软件生态技术在各行业的落地和价值体现，将为用户提供更多元化、智能化的产品和服务，推动整个产业向前发展，实现数字化转型的新高度。

4.3 AIOS与软硬件生态伙伴的相互赋能

软硬件协同涉及要求AIOS高效管理多类型资源，实现算例的弹性扩展、跨平台部署、多

场景兼容等特性，例如，可以不断优化深度学习编译技术，提升算子库性能、开放性和易用性，尽可能屏蔽底层处理器差异，向上兼容更多AI框架。在算力中心大型基础设施方面，通过AIOS平衡算例提升和能耗降低，综合管理IT设备提升算力利用效率。

软硬结合提供高效的计算能力和优化的算法，是统信AIOS和广大软硬件生态合作伙伴共同推动人工智能技术的发展和应用的关键因此，而这也离不开彼此在市场拓展、协同发展上的相互赋能：

（1）市场拓展：

- 联合推出AI解决方案，适配不同硬件设备，覆盖多样化的市场需求，提供全面的解决方案；
- 开展技术培训和合作项目，推广AI技术在操作系统和硬件上的应用，培养更多的开发者和用户；
- 不断优化产品和服务，满足市场需求，加强品牌合作，提升软硬件整体市场占有率；

（2）协同发展:

伙伴联盟：建立开放的生态合作伙伴关系—UAIP智能联盟，共同发展AI生态系统，促进生态闭环的形成；统信软件提供开放的平台和共享资源，吸引更多第三方开发者和合作伙伴加入合作，并共享技术与数据资源。

原生开发：基于统信AIOS原生环境开发AI应用，合作推进需求分析、设计，并适用统信UOS的原生开发工具如DTK、Code、玲珑等进行编码、测试以及维护等开发周期工作。

产业孵化：双方共同培育AI产业创新项目，包括向产业下游伙伴提供AI技术人力、市场拓展，以及商业建议等资源，联合将产品推向市场，并提供持续的支持。

成果展示：通过动态和静态的展示内容来宣传、推广AI产品技术的联合适配工作成果及行业客户成功案例。

人才培养：基于AI技术规划一套系统化的人才培养，包括联合编制教材与课程，开展和院

校及科研机构合作，共同培养下一代行业AI技术人才，开展相关技术人员技术认证。

五、 联盟与合作

5.1 联盟定义

“UAIP智能联盟（UnionTech Artificial Intelligence Plan）”项目是由统信软件发起并运营的一项活动，该项目吸引了从事操作系统相关信息技术领域的企业单位、科研院所、社会组织及专家学者等共同参与，旨在共同构建以统信自研操作系统等基础软硬件为核心的智能创新生态。

5.2 联盟愿景

统信软件将携手合作伙伴，共同打造一个具有强大影响力和创新能力的智能产业联盟。通过与产业链上下游的企业、科研机构、高校等多方合作，深入探索智能技术的应用与发展，为联盟成员创造更多的价值和机遇。

5.3 合作厂商清单

UAIP智能联盟汇聚了众多行业领先的科技企业，目前联盟已拥有30多家重量级伙伴，包括但不限于讯飞、360、智谱、百度云思必驰、万兴科技、饼干科技、方寸无忧、实在智能、博思云创、亦心科技、九科、星环科技、蜜度、美图、文修智能等，以及持续深化合作的浪潮计算机、中科可控、航天706、联想开天、软通计算、紫光计算机、视睿、海光、龙芯、兆芯、飞腾、Intel、此芯、百敖、昆仑太科等知名企业。此外，联盟还积极寻求与软硬件AI制造商等更多领域的合作伙伴建立联系。

5.4 联盟共建

（1）明确智能联盟战略目标

统信AIOS目前已覆盖业内90%以上的主流开源大模型与框架，持续优化接入方式和接口，以提升模型的调用效率与稳定性。同时，不断接入国内外主流大模型，并支持本地模型的接入与融合。联盟秉持积极开放的态度，全面拥抱AI技术，推动其深度融入智能产业链上下游业务中，助力客户提高生产效率、改善客户体验、创新商业模式，从而帮助客户选择合适的AI技术路径，最终增强客户的核心竞争力，实现可持续发展。

（2）明确智能需求和场景，联合打造AI应用解决方案

当前，AI技术的应用场景正持续拓宽，众多应用的关注点正从表面问题转向更深层次的应用领域。联盟携手头部智能伙伴，针对不同垂直行业的特定需求与场景特点，集成先进的AI算法和模型等资源，联合打造一体化的AI应用解决方案。

（3）加强产业链上下游协同合作

联盟致力于加强智能产业链上下游企业、科研机构等各方之间的协同合作。通过共同制定人工智能操作系统相关的标准和规范、重视人才培养与知识分享、密切关注国内外政策动态和法规要求以及争取政策与资金支持等措施，为联盟成员创造良好的发展环境，推动整个智能产业链的协同发展。

（3）着眼长远发展，共建市场品牌

UAIP智能联盟积极与各行业龙头智能企业建立战略合作关系，共同拓展市场份额。通过联合宣传、案例分享等形式向市场展示联合方案的优势与创新成果。同时与行业媒体、科技媒体等建立良好合作关系以加强对联盟伙伴的宣传报道力度进而提升多方在市场上的品牌知名度与影响力。

近年来，随着大语言模型、硬件算力平台等技术的全面革新，为AI产业的发展注入了巨大

的活力。同时，AI软件企业也以此为契机，百花齐放、创新突破，推动着行业应用的蓬勃发展。在这样的背景下，UAIP智能联盟的诞生，是中国AI产业发展历程的重要一幕。操作系统作为基础软件平台，承载着AI软硬件生态的繁荣兴起，深知建立伙伴联盟对于推动AI产业持续发展的重要助力。

统信软件作为国产操作系统领军企业，倡议并邀请更多AI产业领域的伙伴企业加入UAIP智能联盟，共同引领AI技术在技术创新和应用落地上迈出更加坚实的步伐。

统信软件愿与全体UAIP智能联盟的合作伙伴一道，汇集各方优势的智慧与力量，合力开发以统信UOS为底座的智能算力平台，激发大模型训练的潜能，为AI应用企业提供稳健的基础架构支持，携手并肩负起推动中国AI产业发展加速发展和应用的重大使命，共同引领人工智能产业的创新浪潮。

5.5 如何加入

统信软件诚邀AI产业链伙伴踊跃加入UAIP智能联盟，填写入会申请表后请盖章并邮寄至指定地址（具体邮寄地址请向联盟专员咨询）。对于本项目有任何咨询或建议，欢迎随时联系我们的联盟项目专员，联系方式：18810869628。期待与您携手共进，智启时代新篇章。