重新思考关键 AI 基础设施

设备端基础设施如何成为企业 AI 开发与部署的新兴平台

调研机构

Omdia by informa techtarget

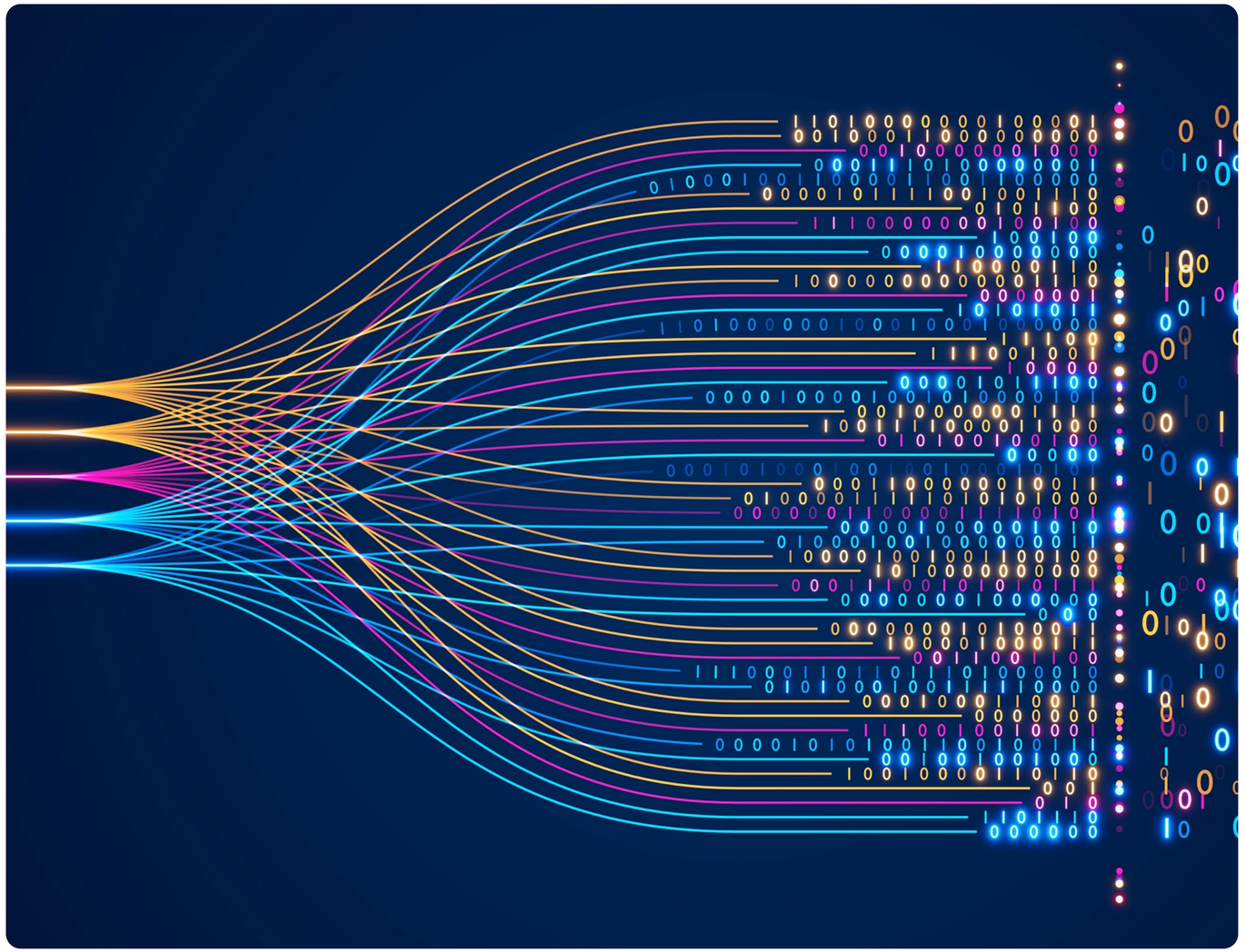
委托方

以设备端基础设施重塑 AI 发展思路

在人工智能 (AI) 技术飞速发展的当下，企业高度依赖云端和本地服务器基础设施来支撑 AI 的开发与部署。许多组织采用混合部署模式，为不同的业务任务匹配最适配的基础设施。但目前，多数企业尚未充分评估设备端 AI 在哪些场景下能够带来显著效益。

Omdia 针对 1500 余名企业技术负责人及从业者开展了独家调研，调研对象包括首席信息官、首席技术官、软件工程师及 AI 专业人员。调研结果显示，企业正愈发质疑现有基础设施是否能真正满足三大核心需求：安全合规要求、成本的不可预测性与持续性支出问题，以及基础设施与任务负载的匹配性需求。

设备端基础设施的核心优势正是围绕这三大核心需求构建，为现代企业面临的安全、经济效益与业务负载挑战提供了架构层面的解决方案。



安全

原生隐私架构保障替代流程性规避

四分之三的受访企业表示，担忧云端服务存在数据泄露风险；而设备端 AI 具备显著优势 — 数据全程无需传输，从根源上避免了传输环节发生泄露的可能性。鉴于 99% 的企业在 AI 工作流程中会处理专有数据，这一本地数据处理能力已成为满足安全需求的关键，让企业无需依赖流程管控即可彻底消除传输风险，同时保障 AI 辅助工作流程的安全。

经济效益

试验不受限且生产成本可控

设备端基础设施在完成初始投资后，边际成本几乎为零，企业可在无预算约束的前提下开展无限制的试验。在生产推理阶段，固定的硬件成本将取代云端随应用规模增长而不断攀升且难以预测的运营成本。此外，云端授权席位可能因用户采用率不如预期而沦为闲置资产；而设备端硬件无论 AI 使用模式如何变化，始终能保持完整的生产价值。

负载适配

精准匹配基础设施与模型需求

企业无需依赖超大规模基础模型即可挖掘 AI 价值 — 57% 的企业所使用的模型参数量均在 100 亿以下，这一规模完全可由 MacBook Air、入门级 MacBook Pro 等现代终端设备承载。出人意料的，大型企业向设备端迁移的趋势更为显著：主要采用设备端基础设施的企业，部署参数量超 1000 亿模型的概率近乎是其他企业的两倍。

企业AI 发展格局现状

AI 成熟型企业正通过深耕设备端应用，逐步转向混合部署模式

AI 应用发展成熟的企业，对混合部署模式的采纳率更高。这一现象表明，对于已实现或计划在多个业务板块规模化落地 AI 的企业而言，不同类型的基础设施具备各自独特的应用价值。

设备端基础设施主要承担开发流程、敏感数据处理、小模型跨团队分布式部署，以及需本地处理的生产推理工作。本地 GPU 集群则更适用于批量训练和多用户共享的使用场景，而云端基础设施的核心优势在于弹性扩展能力，且支持须具备高可用性要求的生产级 API 。

值得注意的是，企业不会单一选择某类基础设施，而是会根据战略需求进行组合搭配。56%使用 Mac 开展设备端 AI 负载部署的企业， 同时也会本地部署NVIDIA GPU ，这表明设备端平台是专业计算资源的补充，而非替代。

这样的模式反映了任务负载的成本效益逻辑：安全敏感型开发工作在受管控的工作站开展；大规模训练任务在需要时调用专用集群；生产部署则根据规模需求匹配对应的基础设施。

企业会将基础设施部署在能为特定任务实现绩效、成本、管控三者最优结合的场景中。以金融服务企业为例，可在工作站完成反欺诈检测模型的原型开发；当业务规模要求提升时，在 GPU 集群上训练生产级模型；将高吞吐量的推理任务部署至云端；同时在本地系统运行分支机构级别的模型，让每种架构都在其价值最优的场景中发挥作用。

平台多元性带来战略灵活性。企业可根据需求匹配基础设施，而非将所有业务负载强行部署在单一架构上。随着 AI 能力的持续演进和任务负载模式的变化，多元化的基础设施组合比单一平台投入更具灵活适应性。

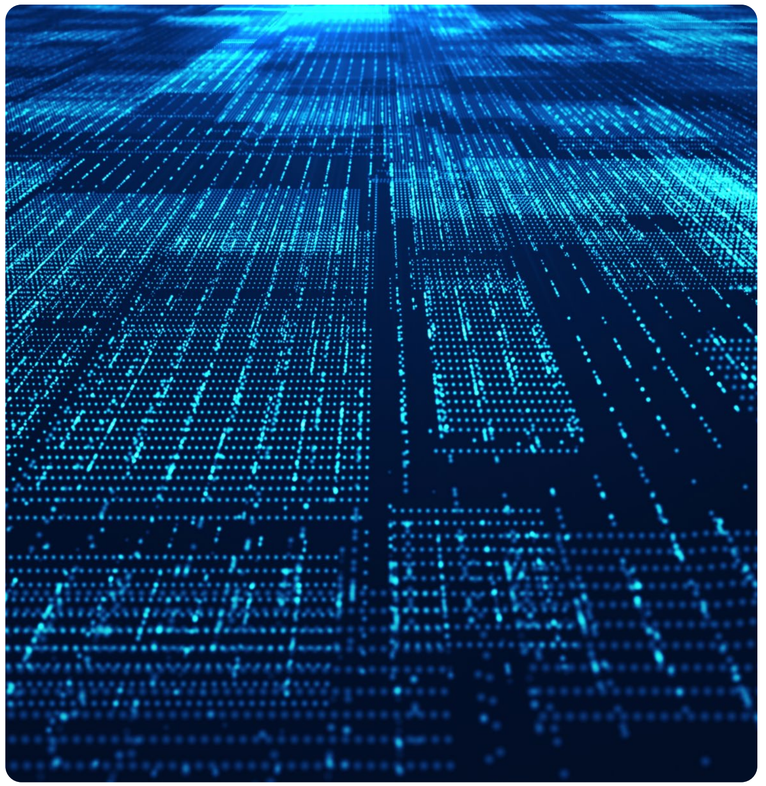


图 1: 41% 的企业已在业务流程中积极应用 AI

贵公司当前的 AI 应用成熟度如何?

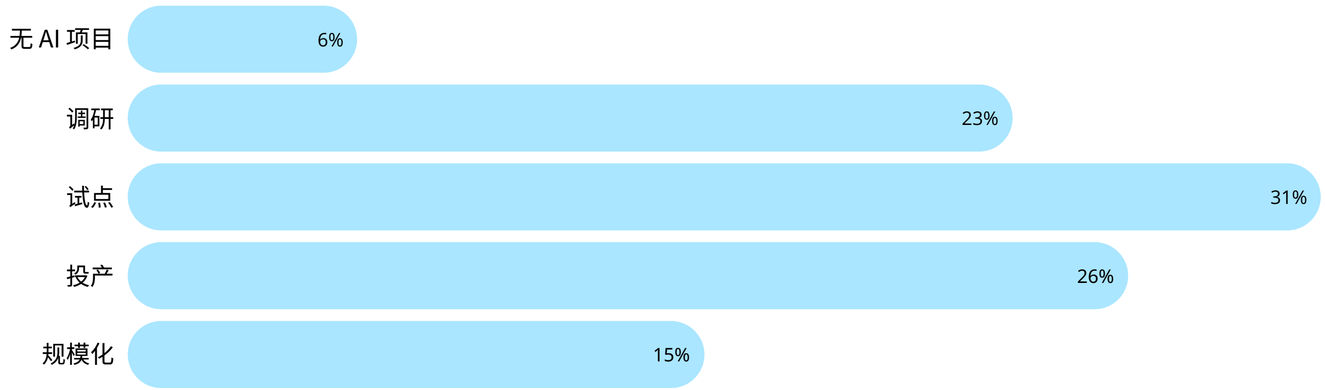
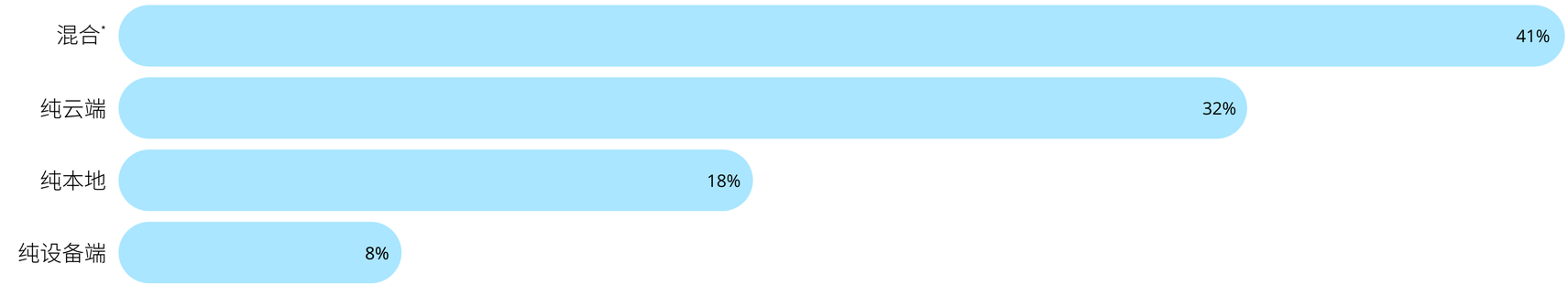
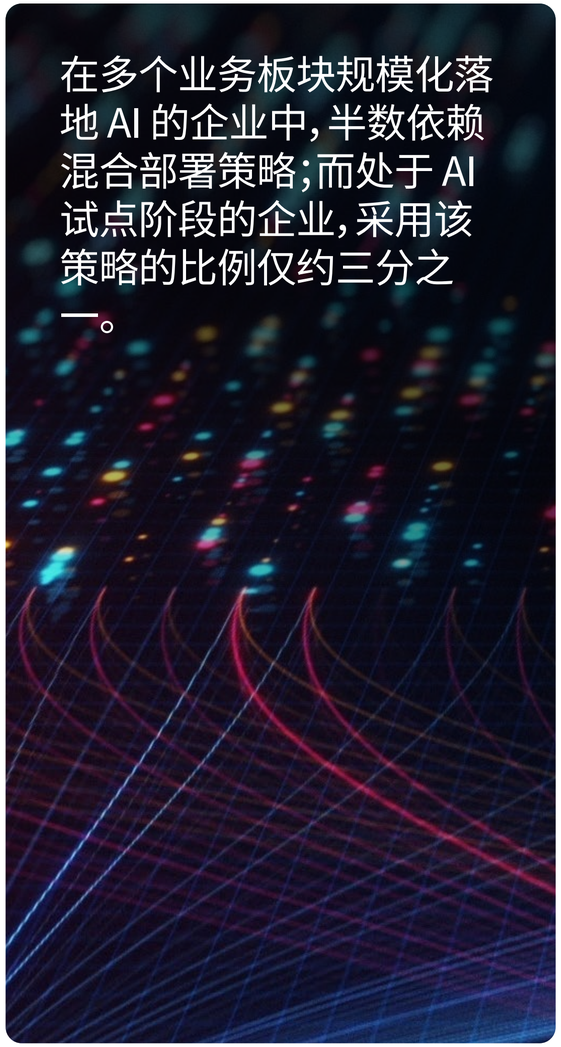


图 2: 73% 的企业采用云端或混合基础设施开展部署

贵公司采用何种策略部署 AI 业务负载?



来源: OMDIA
注释: N=1584, *“混合”策略指 AI 部署融合云端、本地服务器及设备端的计算资源。主流形式为云端与本地服务器相结合。



企业正逐步将部分 AI 任务负载迁移至设备端

超三分之一的企业计划在未来 12 个月内，将更多 AI 业务负载迁移至设备端。这些企业已明确识别出设备端基础设施能够弥补现有部署方案短板的特定任务负载场景。

对于安全敏感型负载，流程管控会产生持续的管理成本。每一个涉及专有信息、客户数据或受监管内容的项目，都需要完成文件备案、审计周期管理和传输风险评估等工作。设备端处理通过将数据保留在本地，大幅降低了这类管理成本。不传输的数据就无需进行传输安全审查，也无需申请第三方访问授权。

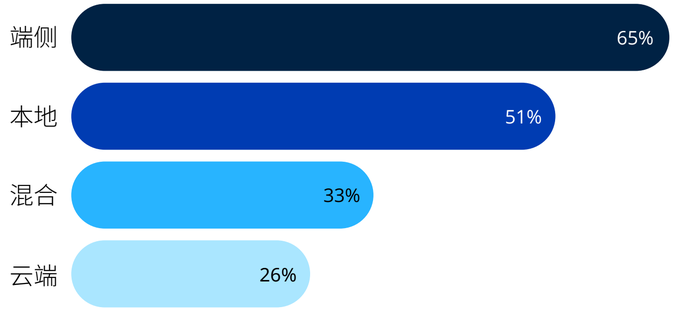
在开发流程中，迭代速度直接决定团队优化模型和应用的效率。团队需要在不等待云端资源分配或不监控使用配额的前提下，完成假设验证、参数优化和方案有效性验证。设备端基础设施彻底打破了这些限制：开发团队可直接基于本地数据集开展工作，无传输延迟、无 token 数量限制、且无迭代边际成本。

企业并非放弃云端基础设施，而是遵循最初推动其采用混合策略的核心逻辑 — 为合适的任务匹配合适的工具。各类架构均有其专属的应用场景。AI 成熟型企业的发展模式是战略多元化，而非对单一平台的依赖。

在使用本地服务器基础设施的企业中，51% 计划提升设备端能力。已采用设备端基础设施的企业中，65% 计划将更多业务负载迁移至设备端。

图 3: 计划向设备端迁移更多 AI 业务负载的企业占比（按当前部署策略划分）

未来 12 个月，预计贵公司设备端与本地服务器及云端的 AI 业务负载占比将如何变化?*



来源: OMDIA
注释: N=1568, *仅展示各当前主流基础设施下，企业“增加设备端负载”的反馈结果

设备端基础设施填补 AI 服务商短板

仅 9% 的企业表示，与其战略合作的 AI 服务商无明显能力短板

构建战略 AI 伙伴关系能赋予企业难以独立建立的能力，如预训练模型、部署基础架构和专有化工具。但伙伴关系也会带来一些局限，例如要求开放数据访问权限、定价复杂化、增加集成相关性等问题。通常，不同企业感知到这类问题的节点有所不同：注重安全的企业，会在部署敏感数据相关业务时遇到此类问题；而高度关注成本的企业，则往往是在收到账单时才真正意识到问题。

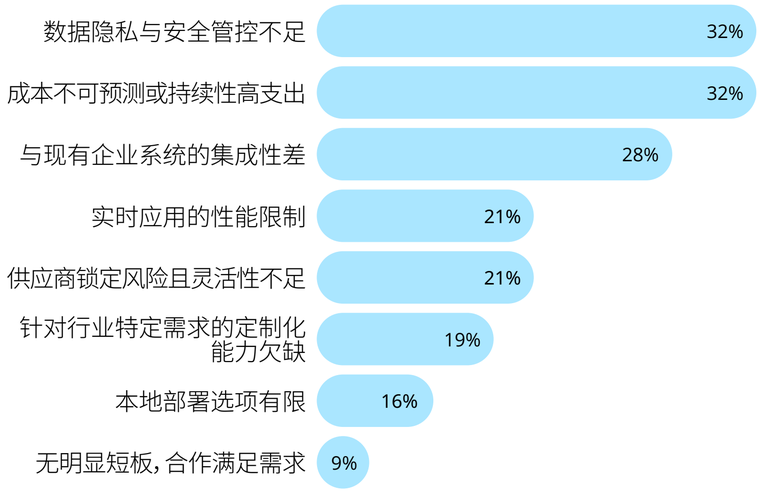
Omdia 的调研结果显示，企业对此类合作的不满情绪普遍存在。无论合作对象是主流云服务商、传统企业软件厂商还是原生 AI 企业，数据隐私管控不足、成本不可预测、与现有系统集成性差，都是企业反复遇到的核心痛点。仅有 9% 的企业表示，合作方能够完全满足其需求。

这种不满情绪为混合部署模式创造了发展机遇。企业并非拒绝 AI 合作，而是寻求更灵活的方式，解决纯云端策略无法充分应对的特定运营挑战。通过在可控的设备端基础设施上开展本地开发，企业能够为敏感数据处理和实时任务打造高速运行环境，既实现成本可预测，又能直接获取数据访问权限。

以设备端计算补充基础设施能力，能够填补纯云端伙伴模式固有的功能短板，通过推动 AI 全流程落地简化运营。这意味着，各个环节都能实现 AI 赋能，即使是此前因隐私控制要求或数据出站成本而受阻的敏感数据标注与清洗也能覆盖。以设备端基础设施作为发展基础确保了企业可将核心云端资源优先用于大规模分发和集中化协调工作，而非让其被更适合设备端处理的任务占用。

设备端基础设施为企业打造了安全的核心基底，助力企业管理整个 AI 生态系统

图 4: 安全、成本、集成性差是 AI 合作中持续存在的三大痛点
当前 AI 伙伴关系中面临的两大主要局限是什么？(最多选择两项)



来源: OMDIA
注释: N=1352

企业关注显性成本，却忽视了核心因素

超三分之二的企业会追踪云端成本，但仅有不到半数的企业会统计安全相关成本

云端计算、网络安全和软件授权费用，是企业在 AI 基础设施总拥有成本 (TCO) 中最关注的三大项。但关注并不等同于能够实现精准的追踪和归因核算。

在将云端计算列为总拥有成本关注项的企业中，72% 会进行系统化追踪，将其作为 AI 基础设施预算中的独立科目。这一现象符合行业逻辑，因为云服务商会定期提供账单，并清晰拆分各服务的费用明细。

而网络安全的追踪则呈现出完全不同的现状。尽管是第二大总拥有成本关注项，但仅有不到半数的企业会将其作为 AI 基础设施总拥有成本的独立组成部分进行统计。网络安全相关支出是真实存在的，包括合规管理成本、审计周期投入、安全事件响应成本和专业工具费用等，但这些支出分散在企业现有各项预算和各业务部门中，难以进行清晰的归因核算。

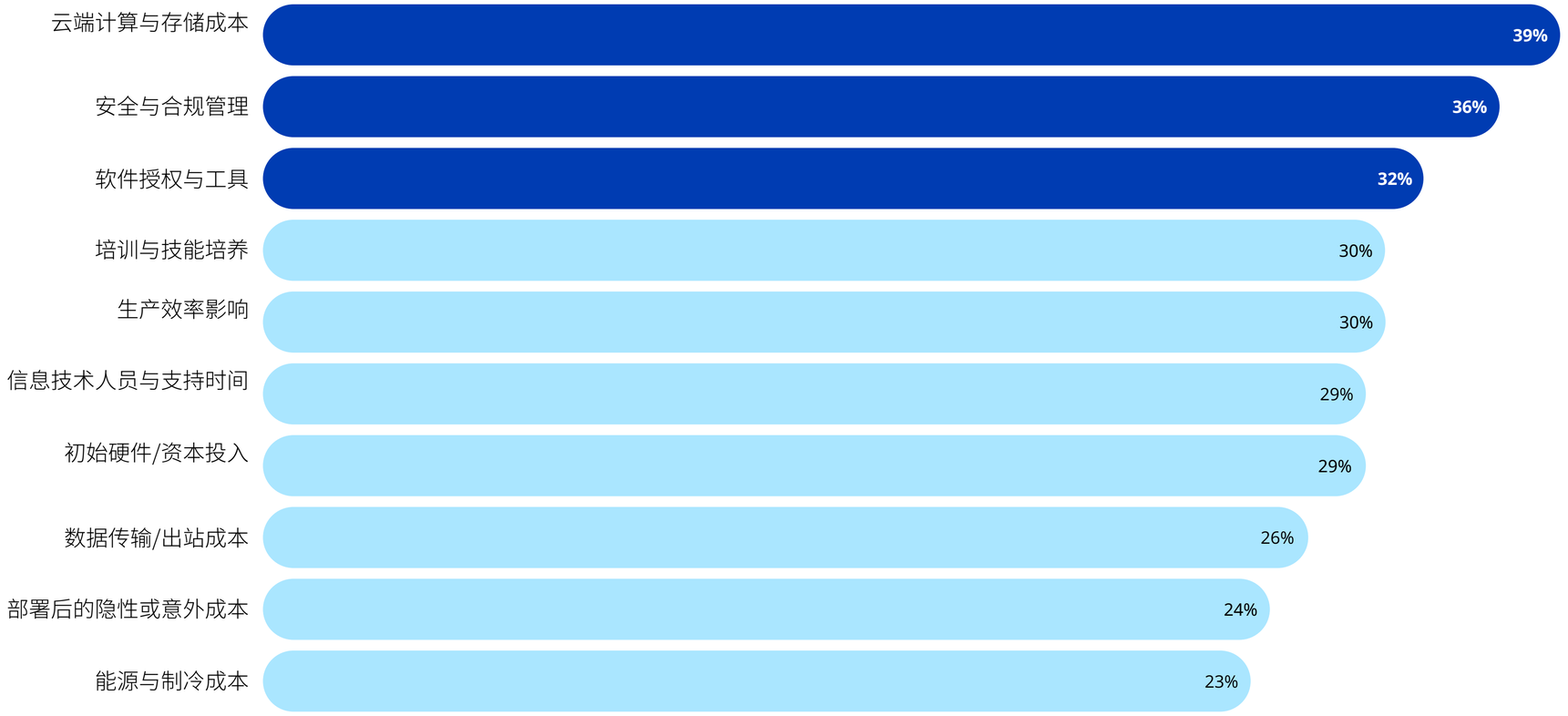
这种可视化缺口造成了企业的战略盲区。企业只会对可量化的指标加以优化。当云端成本以明细账单形式呈现，而网络安全投入却分散在采购、人力和运营预算中时，决策者自然会将注意力放在显性的数字上。但这些隐性成本的规模可能更大，且对企业的长期发展更为关键。

这一缺口也使得企业难以发现非纯云端部署策略的成本优势。由于潜在的成本节约效果无法即时体现，而前期投入会在成本评估中占据主导地位、掩盖长期价值，企业往往会系统性低估混合部署模式的价值。这种成本可视化的不对称性，可能导致企业的决策偏向云端中心化策略，即便混合部署模式的总拥有成本更具优势。

成本归因核算的缺口和预算体系的特点，可能掩盖混合部署模式在总拥有成本上的优势。

图 5: 云端、安全、软件授权是 AI 基础设施总拥有成本中最主要的三大关注项

贵公司对 AI 基础设施总拥有成本的核心关注项是什么?选择三项。



来源: OMDIA
注释: N=1584

安全模型，正当其理

安全即阻力 vs 安全即架构

从设计上尽量减少数据暴露

当前的云端中心化策略存在一个固有矛盾：要利用云端 AI 能力，数据就必须离开企业内部，传输至第三方基础设施。76% 的企业认为，这种数据传输存在暴露风险。企业的担忧并非仅源于恶意攻击，更来自于更为常见的非故意传输失误或配置错误。

当企业使用客户数据、内部知识库、专有研究成果等核心价值信息微调或训练模型时，这种担忧会进一步加剧。87% 的企业认为，将敏感数据保留在本地服务器或设备端设备至关重要，但云端优先的架构设计只能降低风险，无法从根源上消除风险。

这一问题的核心矛盾在于：技术创新需要开放数据访问权限，而数据保护则要求限制数据传输。开发团队需要敏感的专有数据来构建高效的模型。而每一个新项目的启动，都需要开展合规审查，产生大量的规划管理成本，限制了新应用场景下的试验速度。

在受监管的行业中，这一矛盾尤为突出。医疗企业使用患者健康信息训练模型时，需遵守强制性管控要求；金融服务企业使用交易记录和客户金融数据时，必须遵循数据属地化存储规定。专业服务企业处理客户机密数据时，会受到第三方数据访问的合同限制。

安全需求是企业采用设备端基础设施的首要原因之一。正是出于这一担忧，企业往往更愿意选择从架构层面解决问题，而非以流程管控。

担忧

76%

的企业担忧或高度担忧云端服务存在数据泄露风险

偏好

87%

的企业认为，将敏感数据保留在本地服务器或设备端设备至关重要

设备端 AI 开发强化安全防护

无需流程管控，依托架构层面的隐私保障，显著降低安全风险

设备端开发从架构层面解决了创新与安全之间的矛盾。数据全程不离开安全环境，彻底省去了各类安全审查和合规管理成本。在数据主权要求的背景下，这一优势更为突出 — 企业在多个司法管辖区开展业务时，需遵守各国的属地化数据存储规定。云端架构会增加合规难度，而设备端处理则能让合规工作更易落地。

设备端基础设施的安全优势十分明确：不传输的数据就不会发生传输环节的泄露。开发团队可直接基于本地数据集开展工作，无需申请第三方基础设施的访问权限。模型在数据存储的本地完成训练，无需管理数据传输成本、验证传输加密效果、审计云服务商的管控措施，也无需准备跨境数据传输的相关文件。

设备端基础设施创造了以往无法实现的流程优势：利用 AI 技术为 AI 工作做数据准备。部署在云端的企业，必须在数据传输前完成敏感数据的清洗和预处理，剔除个人身份信息或编辑机密内容。若缺乏设备端 AI 能力，这一数据准备工作只能通过传统工具完成，形成严重的流程瓶颈。

而借助设备端 AI，开发团队可在本地通过 AI 工具加速数据清洗和预处理流程 — 利用模型识别敏感字段、实现编辑工作自动化、验证数据质量、完成数据集转换，所有操作均在数据离开安全环境前完成。

企业可以基于敏感数据全速开展创新，而不必在其他方面取舍。以往需要数天完成的安全审查工作，如今周期大幅压缩。工程团队重获高效的开发速度，安全团队则通过架构层面的保障而非流程管控，建立起对数据安全的信心。

这并非意味着设备端模式可免除所有安全需求。设备物理安全、访问权限管控和审计日志记录仍至关重要。但它能从设计上规避云端架构固有的传输相关安全风险。设备端基础设施消除了架构风险，同时也消除了流程性的管控摩擦。


99%
的企业在训练或微调 AI 模型时会使用专有数据，使得设备端 AI 成为企业工作流程的核心优势，有助于消除传输风险、降低流程管理成本，并推动 AI 辅助的数据准备流程落地

经济效益，正当其道

设备端开发消除额外试验成本

AI 开发的核心在于迭代。团队只有拥有自由试验的空间，才能诞生最优解决方案。这意味着团队可在无约束的前提下，验证假设、优化模型、探索创新方案。对于仍在调研或试点 AI 解决方案的 54% 的企业而言，尽早实现基础设施的平衡配置，能在试验工作深入、任务负载向生产阶段规模化推进时，直接转化为成本节约。

设备端基础设施彻底改变了试验探索的成本逻辑。一次性的硬件投资，即可解锁无限制的迭代可能。硬件部署完成后，每一次迭代、测试和验证均不会产生额外的成本。

而云端中心化策略下，每一次探索都会产生相应成本，计算时间、数据传输、存储服务的费用会不断累积。即便一个有潜力的方案在 20 次迭代后宣告失败，这 20 次迭代产生的费用仍会体现在账单中。从失败方案中学习的成本，与成功方案的试错成本完全相同。

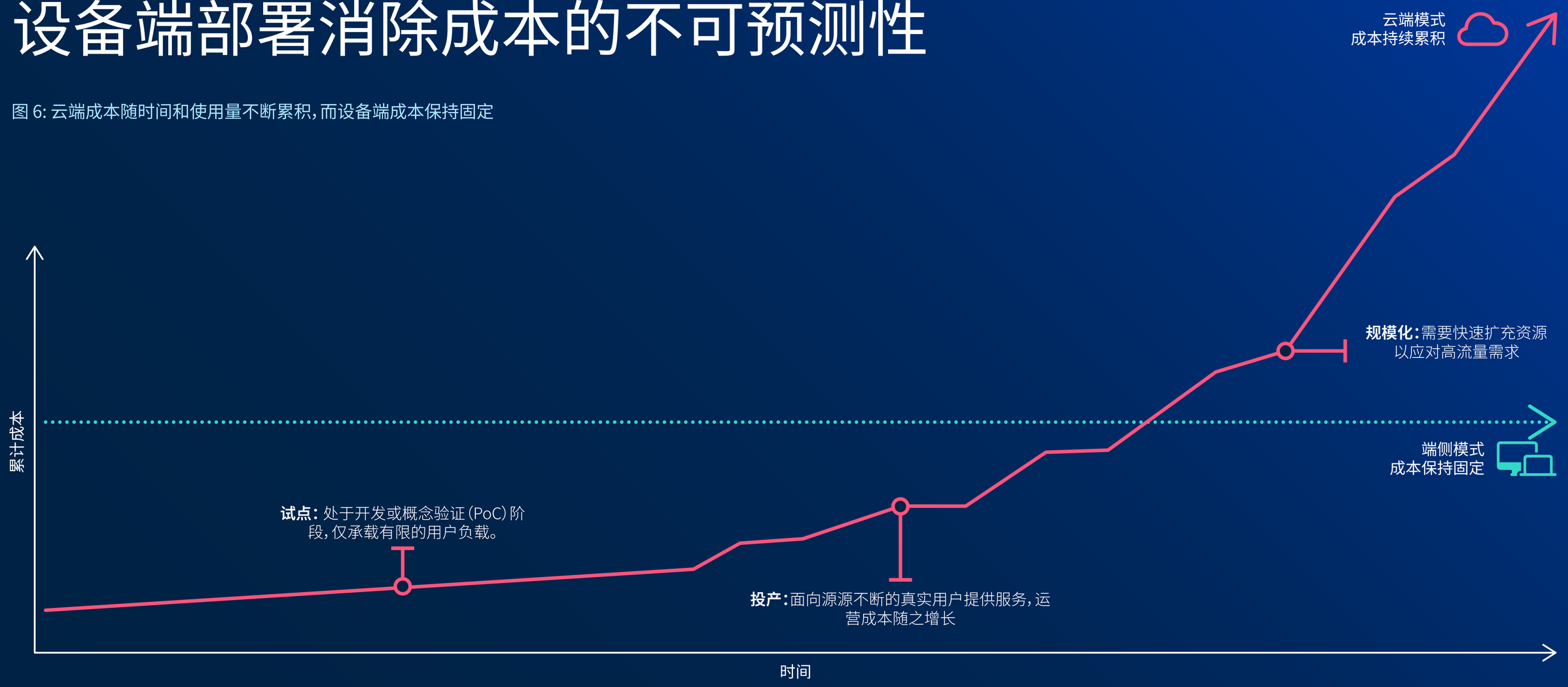
对于周期较长的项目，云端和设备端模式的总拥有成本存在显著差异。云端费用会随模型版本更新不断叠加，而能支持全流程设备端 AI 运行的硬件，在完成初始投资后，可实现零边际成本的无限制试验。此外，同一硬件投资在其生命周期内，可支撑多个项目的开展。

其战略优势会随时间不断放大。开发团队可开展充分的验证工作、探索各类可能性、实现快速迭代，无需为预算问题担忧。此时，项目的限制因素从基础设施成本转变为人力投入和工程时间。这就彻底消除了因财务压力导致的试验受限、验证缺失、创新方案探索不足等问题。

设备端基础设施不仅能降低成本，更能支撑孕育优质 AI 成果的快速迭代周期。团队能够开展更多测试、从更多失败中学习，最终交付更优质的产品和解决方案。

设备端部署消除成本的不可预测性

图 6: 云端成本随时间和使用量不断累积，而设备端成本保持固定



设备端 AI 部署能帮助企业规避推理成本失控和云端授权闲置的问题

除开发阶段外，基础设施的选择对生产推理阶段的成本也具有显著影响。云端 AI 主要采用两种定价模式：按 API 调用次数计费，费用随请求量累积；或按席位订阅计费，费用固定且与使用强度无关。两种模式均会产生成本叠加的经济挑战。

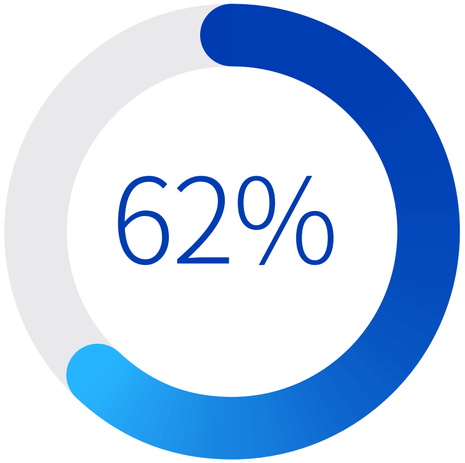
设备端基础设施则从根本上改变了 AI 推理的经济逻辑。完成初始硬件投资后，每一次推理请求均不会产生额外的成本。在设备端生成 1000 个或 10 万个 token，其总拥有成本保持不变。这种成本可预测性，对于在全企业范围内规模化落地 AI 的企业而言价值尤为突出 — 企业内部各部门的独立运营，往往导致云端成本难以在全企业范围内实现精准、实时的追踪。

按席位授权的定价模式则带来另一类挑战。企业采购云端订阅服务时，往往预期会有广泛的用户采用，但实际使用量常未达预期。32% 的企业表示，推动终端用户使用 AI 工具的效果未达预期，另有 30% 的企业表示效果一般。这些未被使用的授权，成为了无生产价值回报的持续性成本。

设备端基础设施则能彻底规避这一风险。无论用于 AI 业务负载还是标准计算任务，硬件在其生命周期内始终能持续创造价值。企业可消除闲置订阅带来的资源浪费，同时在需求变化时保持灵活适应性。

图 7：按席位计费的云端授权无论使用强度如何均需照常收费

假设某企业以每月每用户 30 美元的价格，采购 1000 个云端 AI 授权，若仅有 10% 的授权未被使用，企业每月将产生 3000 美元的无效成本。



的企业表示，推动终端用户使用 AI 工具的效果一般或未达预期。

负载适配，正当其用

为特定目标优化的企业 AI 模型

参数量达万亿级的预训练模型频繁出现在公众视野，让人们形成了“所有有价值的 AI 工作都需要超大规模数据中心支撑”的认知。但这并未能反映企业部署的真实现状。

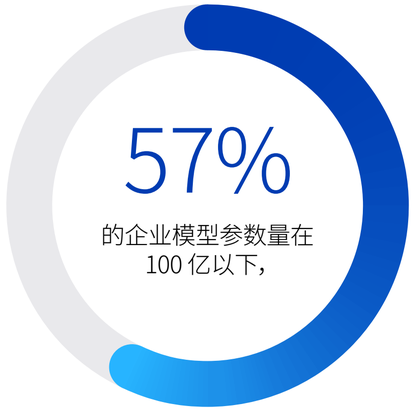
95% 的企业并非从零构建模型，而是在现有模型的基础上开展微调和推理工作。这些任务对计算能力的要求明显更低。

此外，57% 的企业日常部署的模型，参数量均在 100 亿以下。即便对于参数量达数千亿的超大规模模型，企业也可通过桌面设备集群，实现本地开发与部署。换言之，企业 AI 开发的基础设施需求，往往与现代个人电脑的能力相匹配。

采用设备端基础设施的企业，部署参数量超 1000 亿模型的概率近乎其他企业的两倍

图 8: 95% 的企业开展模型微调与推理工作

贵公司日常部署的 AI 模型参数量规模如何?贵公司开展 AI 推理、模型微调，还是从零构建模型?



大多数企业模型的参数量在 100 亿以下，完全可由 MacBook Air 或入门级 MacBook Pro 等现代终端设备承载

来源: OMDIA
注释: N=1529

经量化处理后，参数量 100 亿的模型仅需约 12GB 内存。24GB 内存的设备即可轻松承载，且能为同时运行的其他应用和工作流程预留充足的内存空间。即便基础配置的 16GB 内存设备，也能支撑参数量 10 亿以下的小模型运行。

统一内存架构使得超大规模模型也得以实现设备端部署

行业内对 AI 基础设施的一种普遍认知是：模型规模越大，就越需要部署在云端。但 Omdia 的分析发现，企业部署的模型规模与基础设施选择偏好之间，并无显著关联。

部署参数量超 1000 亿模型的企业，相比还是部署 100 亿参数量模型的企业，对云端基础设施的偏好程度并无增加。真正决定基础设施部署选择的因素是安全需求、成本考量和开发流程的实际需要。

这一规律背后的技术逻辑是：限制模型部署的并非规模本身，而是可用的内存容量和架构设计。核心决定因素是基础设施能否为业务任务负载提供充足的内存，而非模型是否突破某一随意设定的规模阈值。

对于部署参数量超 1000 亿模型的 20% 的企业而言，内存可用性成为核心考量因素。无论是具备大型统一内存池的单系统配置，还是多设备集群部署，只要能满足 100GB 以上的内存需求，就能支撑这类超大规模模型的本地处理。

内存架构决定了适配场景：统一内存系统可通过单设备实现 100GB 以上的内存配置，无需依赖多 GPU 基础设施。

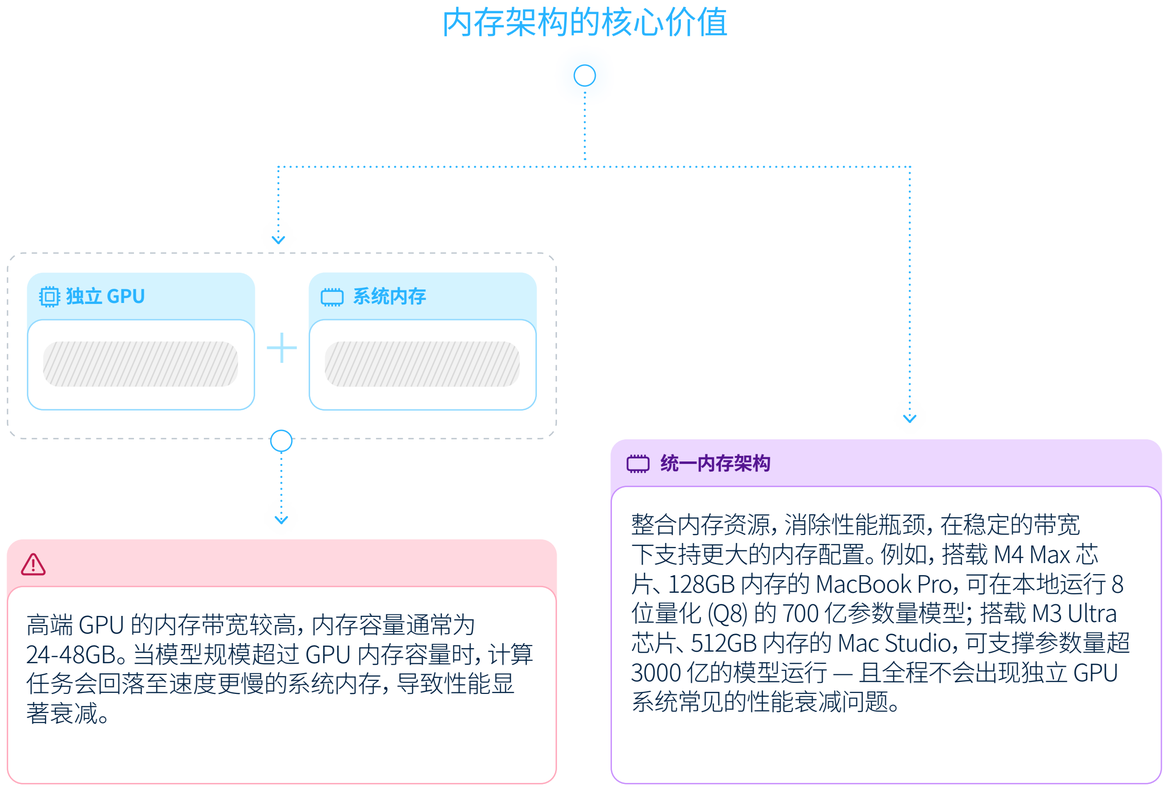


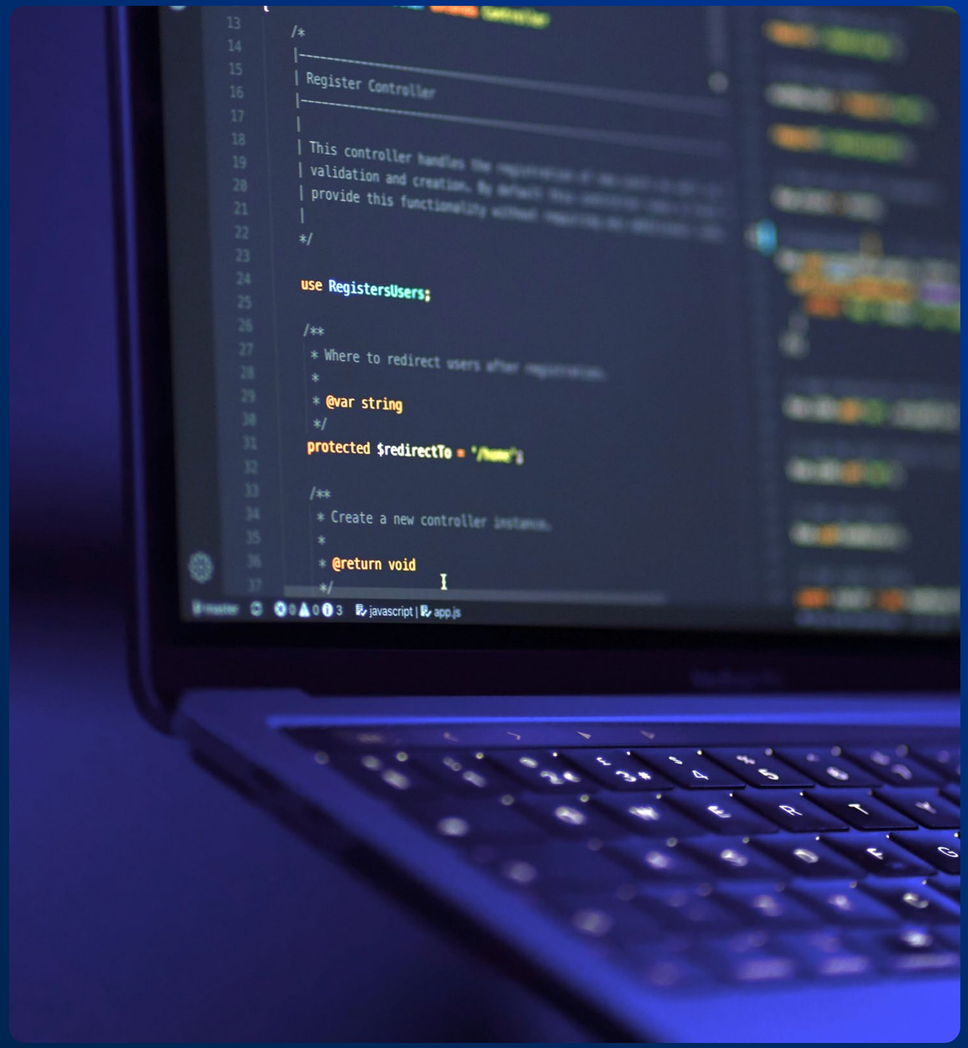
图 9: 不同精度下，模型推理的内存需求
（含模型权重 + 20% 推理开销）

参数量	Q8	FP16	FP32
10 亿: 文本分类、情感分析、简单信息抽取	1.2GB	2.4GB	4.8GB
100 亿: :对话式 AI、问答、摘要生成、基础代码辅助	12GB	24GB	48GB
150 亿: 文档分析、翻译、结构化内容生成	18GB	36GB	72GB
700 亿: 高级代码生成、复杂任务处理	84GB	168GB	336GB
1000 亿: 多模态分析、研究综述与资料整合	120GB	240GB	480GB
3000 亿: 前沿级别、复杂推理、自主智能体任务编排	360GB	720GB	1440GB

不同规模模型的能力参考示意 (实际性能取决于具体架构及微调方式)。

Q8 (8位量化)、FP16 (16位浮点) 和 FP32 (32位浮点) 是指存储模型权重时所采用的数值精度。较低的精度能在显著降低内存占用的同时，确保输出质量不受明显影响——在大多数推理任务中，8位量化被公认为具有 '近乎无损' 的表现。在本文展示的方案中，Q8 是显存效率最高的选项；而在实际应用中，针对内存资源更加受限的设备端部署，4位量化 (4-BIT QUANTIZATION) 也得到了广泛采用。

来源: OMDIA



大力投入 AI 业务负载的企业普遍选择 Mac

AI 战略与平台选择之间，呈现出显著的关联规律。自主研发 AI 解决方案的企业，采用 Mac 开展设备端 AI 业务负载的比例，近乎是采购商业解决方案企业的两倍。

这一现象清晰体现了企业的实际偏好。自主研发团队的技术需求最为深度。他们需要无边际成本的迭代灵活性、专有信息的绝对数据安全，以及可直接观察、管控和优化的性能表现。

这些需求与设备端基础设施的优势高度契合，而开展高强度本地开发的研发人员，也最倾向于选择 Mac。

对于将基础设施支撑开发流程视为核心需求的企业而言，迭代速度、本地数据访问、稳定的性能表现是关键，这类企业往往具备成熟的 AI 能力，也是 Mac 的核心用户群体。明确自身需求的技术团队，最终都会选择 Mac。

当技术从业者掌握平台决策权时，他们会优先选择那些能提升生产力的工具。Mac 在内部开发人员中的采用优势表明，该平台具备某些其他替代方案所不具备的特性，能够满足他们的工作流程需求。

企业之所以选择 Mac 开展自主研发，核心原因在于其出色的性能表现，以及开发人员对该平台的高度熟悉

…其比例几乎是商业方案买家的两倍。当技术从业者掌握平台决策权时，他们会优先选择那些能提升生产力的工具—在企业内部 AI 开发中，Mac 的采用率高达 1.8 倍。

核心结论

构建均衡的 AI 基础设施体系

本次调研揭示了企业现有基础设施方案与 AI 需求之间的差距。企业面临着增加管理成本的安全壁垒、掩盖运营支出的成本结构，以及降低开发效率的流程约束。

企业的未来发展方向，是通过设备端能力完善基础设施体系，实现业务负载需求与基础设施的精准匹配，而非采用以云端为中心的策略来部署所有 AI 项目。能够建立这种灵活适配能力的企业，将在 AI 技术持续演进、业务负载模式不断变化的过程中，持续积累竞争优势。

让 AI 基础设施与任务负载精准匹配

根据模型规模、数据敏感度、使用模式和用户需求，确保任务负载与基础设施的适配。运行 100 亿参数量以下模型的多数企业，与研发前沿模型的 AI 研究机构，他们所需的最优基础设施存在本质差异。开发阶段的任务负载，与生产服务阶段的需求也并不相同。企业应摒弃“一刀切”的基础设施默认方案，避免出现需求与资源错配的问题。

让不可预测成本变得可预测

将总拥有成本的核算范围，从云端计算账单、硬件采购等易追踪的科目，拓展至更多维度。量化安全审查的时间损耗、合规管理的成本、资源约束对生产效率的影响，以及按使用量定价模式下试验探索的边际成本。企业只会优化可量化的指标，不完整的成本可视化会导致基础设施决策的非最优化。

尽可能从架构层面解决安全问题

对于涉及专有或敏感信息的项目，评估能从架构层面消除传输风险的基础设施方案，而非仅通过流程管控进行风险治理。设备端处理并非唯一方案，但对于开发流程和小模型推理任务而言，该模式能彻底省去各类合规管理成本，消除安全审查的时间延迟。

赋能内部技术团队打造混合解决方案

赋予内部开发团队自主整合各类基础设施的能力，让每种基础设施都能在最适配的流程场景中发挥价值。开展复杂 AI 研发工作的团队，早已在打造混合基础设施体系，实现业务负载与基础设施的精准匹配。企业应建立相应的采购框架和技术政策，助力而非限制这种战略多元化发展。

Mac 的 AI 优势是设计使然，而非偶然

企业对 Mac 的偏好并非随机选择，而是源于其架构设计能直接解决云端中心化或本地服务器中心化策略的痛点与短板。

Mac搭载的 Apple 芯片的统一内存架构，既能为企业日常的 AI 业务负载提供充足的内存容量，又能在特定价格区间内实现出色的内存规模扩展，支持超大规模模型的设备端运行。这一设计解决了业务负载与基础设施的错配问题，在无需持续投入云端成本、无需采购昂贵本地服务器的前提下，为企业提供充足的性能支撑。

数据安全

设备端处理从设计上消除了传输风险。对于绝大多数需要将敏感数据保留在本地的企业而言，Mac 提供的是架构原生的隐私保护，而非流程性的规避方案。

可预测的成本模型

固定的硬件投资取代了随时间不断增长的云端按使用量计费成本。开发团队可开展无限制的试验探索，且不会产生边际成本；生产部署则能规避随使用量增长的成本支出。

适配 AI 推理的统一内存

20% 的企业部署了参数量超 1000 亿的模型 — 这类模型的规模或超出了独立 GPU 显存的承载能力，或使用成本过高，但却非常适配 Apple 芯片的统一内存架构：MacBook Pro 的统一内存容量最高可达 128GB，Mac Studio 更是高达 512GB。

开发人员的偏好

开展 AI 自主研发的企业，选择 Mac 的比例是采购商业解决方案企业的 1.8 倍。当技术团队掌握平台决策的主导权时，他们会基于性能表现、安全需求和平台熟悉度进行最优选择。

附录

关于

Omdia

用心聆听，洞见未来；专业建议，引领发展。

Omdia 是全球领先的科技研究机构，由 Informa TechTarget 的研究部门（Ovum、Heavy Reading、Tractica 和 Canalys）与收购的 IHS Markit 科技研究业务整合而成。我们汇聚了 300 余名分析师的专业力量，研究领域覆盖全科技产业链，涉及 200 多个市场。每年发布超 3000 份研究报告，深度分析数千家科技、媒体和电信企业。凭借全面的行业情报和深厚的技术专业能力，我们为客户挖掘具有实际行动价值的洞察，助力客户在瞬息万变的科技环境中理清发展脉络，实现企业当下与未来的业务增长。

关于本次调研 — 研究方法与样本详情

本次关键 AI 基础设施调研由 Omdia 在 2025 年第四季度开展并完成分析，调研对象为以下全球五大市场的企业信息技术决策者：美国、英国、德国、中国和印度。

调研共收集 1584 份有效问卷，调研对象为员工规模 500 人以上的医疗、金融服务和专业服务企业。所有受访者均拥有技术采购的直接预算审批权，或对企业 AI 战略具有重要影响力，包括但不限于首席信息官、首席技术官、软件工程师、AI 专业人员等。

本次调研围绕企业对 AI 基础设施决策的态度、安全与隐私需求、总拥有成本考量，以及 AI 业务负载的平台偏好等核心问题展开。

企业所在国家/地区

国家	
美国	20%
英国	20%
德国	20%
中国	20%
印度	20%

企业所属核心行业

行业	
医疗保健	33%
金融服务	34%
专业服务	33%

企业全球员工规模

企业年营收规模

注：由于四舍五入，百分比总和可能不等于100%。

Omdia
by informa techtarget

The Omdia team of 400+ analysts and consultants are located across the globe

Americas

Argentina
Brazil
Canada
United States

Asia-Pacific

Australia
China
India
Japan
Malaysia
Singapore
South Korea
Taiwan

Europe, Middle East, Africa

Denmark
France
Germany
Italy
Kenya
Netherlands
South Africa
Spain
Sweden
United Arab Emirates
United Kingdom

Omdia

E insights@omdia.com
E consulting@omdia.com
W omdia.tech.informa.com

OmdiaHQ
Omdia

Citation Policy

Request external citation and usage of Omdia research and data via citations@omdia.com