辰宇信息
WINMARKETRESEARCH

2026-2032全球与中国大模型压缩服务市场调研报告

【行业】:软件及商业服务 【报告编码】:179324119708463

【出版时间】:2026-10-29 【订购热线】:+86 180 2246 3983

【电子邮件】:market@winmarketresearch.com

【报告价格】： ¥18900.00 中文电子版
¥18900.00 英文电子版
¥37800.00 中文+英文电子版

内容摘要

根据统计及预测，2025年全球大模型压缩服务市场销售额达到了24.33亿美元，预计2032年将达到62.18亿美元，年复合增长率（CAGR）为 13.6%（2026-2032）。
本文研究全球及中国市场大模型压缩服务现状及未来发展趋势，侧重分析全球及中国市场的主要企业，同时对比北美、欧洲、中国、日本、东南亚和印度等地区的现状及未来发展趋势。
大模型压缩服务是指通过专业软件、平台化工具或项目制技术服务，在维持应用可接受的模型精度和生成质量的前提下，降低大语言模型、多模态基础模型及其他大型生成式人工智能模型的存储容量、内存占用、计算需求、推理时延或部署成本。该类服务通常覆盖模型评估、压缩策略设计、校准数据准备、训练后量化、量化感知训练、结构化或非结构化剪枝、稀疏化、知识蒸馏、低秩分解、算子与计算图优化、硬件感知编译、推理运行时适配、精度恢复及部署验证。根据客户需求，服务商可以交付压缩后的模型权重、更小的学生模型、面向特定硬件的推理引擎、优化容器、部署代码、性能测试报告或持续性模型优化支持。主要客户包括基础模型开发商、云计算与数据中心运营商、企业级AI解决方案商、软件企业、终端设备制造商、汽车及机器人企业、通信企业，以及需要在公有云、私有云、本地数据中心或边缘设备中部署大模型的公共机构。
核心总结
量化是当前商业化程度最高的大模型压缩服务类别
知识蒸馏和剪枝可实现更深层的模型与计算规模缩减
硬件感知优化逐步决定实际时延和成本改善效果
客户日益要求压缩模型与可部署推理引擎一并交付
市场参与者覆盖云平台芯片厂商开源生态和专业服务商
行业发展趋势
大模型压缩服务正在由单一模型量化向端到端压缩与部署优化发展。早期项目通常侧重于将FP32或FP16模型权重转换为INT8或INT4格式，目前服务内容则逐步融合混合精度量化、激活感知权重量化、剪枝、稀疏化、知识蒸馏、推测解码适配、计算图优化及硬件专用编译。NVIDIA Model Optimizer覆盖量化、剪枝、稀疏化、知识蒸馏和推测执行等技术，NeMo可执行FP8、INT8和INT4训练后量化及大模型知识蒸馏。AWS SageMaker AI则将量化、编译、推测解码和模型快速加载整合至托管式推理优化任务。上述发展表明，客户不再仅依据压缩文件大小评估服务效果，而是更加关注GPU显存占用、每秒生成Token数量、首Token时延、并发能力、功耗及模型质量保持程度等整体部署经济性。
第二项重要趋势是目标硬件适配的重要性持续上升。同一压缩模型在NVIDIA GPU、Intel CPU与Gaudi加速器、AWS Inferentia、Qualcomm NPU、移动处理器或其他边缘加速器上的表现可能存在较大差异，原因在于不同平台对INT4、FP8、稀疏计算、张量算子、内核库及存储架构的支持程度不同。Qualcomm AI Hub Workbench将量化、编译、真实设备性能测试和数值正确性验证结合；Microsoft Olive可针对指定硬件与执行后端自动组合优化流程；Intel Neural Compressor支持面向CPU、GPU及Gaudi加速器的量化、剪枝和知识蒸馏；Nota AI的NetsPresso也将硬件感知模型优化和部署适配作为核心能力。因此，大模型压缩服务正在由单纯面向模型算法的优化，转向模型架构、数值精度、推理运行时和目标硬件的协同优化。
第三项趋势是服务模式由工程师主导的定制项目逐步向自动化平台和优化智能体延伸。客户可以上传或引用训练完成的模型，设定目标硬件及性能约束，并获得自动生成的压缩模型资产和测试结果。AWS SageMaker AI、Qualcomm AI Hub Workbench和NVIDIA TensorRT Cloud代表托管式平台方向，Pruna AI与Nota AI则将专有压缩技术与付费软件或专家服务结合。不过，对于专有模型架构、多模态模型、混合专家模型，以及具有严格精度和安全要求的应用，全自动压缩仍存

在明显局限。高价值项目仍需要专业人员参与校准数据选择、敏感层分析、混合精度分配、学生模型设计及面向具体业务的效果评估。

市场动态

驱动因素

市场增长的首要驱动因素是大型生成式模型部署成本快速上升。模型参数、键值缓存、中间激活及并发服务共同形成较高的GPU或其他加速器显存、算力和能源需求。量化通过降低权重和激活值的数值精度，减少模型存储空间及存储带宽压力；剪枝和稀疏化则删除重要性较低的参数或连接；知识蒸馏可将大型教师模型的能力迁移至规模更小的学生模型，从模型架构层面降低部署资源需求。这些技术能够使客户在现有基础设施上支持更多并发请求，将模型迁移至成本更低的加速器，并推动私有化或端侧部署，降低对高显存GPU的依赖。

大模型部署环境日益分散，也在持续推动专业压缩需求。企业需要在公有云、私有数据中心、工作站、工业计算机、汽车、智能手机及嵌入式设备之间部署模型。由于不同硬件对INT4、FP8、稀疏矩阵、张量运算及推理内核的支持程度存在差异，在一种加速器上表现良好的模型不一定能够在另一平台取得相同收益。大模型压缩服务通过选择适当的压缩方式、适配目标运行时，并在代表性硬件上验证时延、内存和输出质量，连接研究阶段的模型权重与可投入生产的推理系统。

限制因素

市场发展的核心限制是模型质量下降风险。低比特量化可能影响事实准确性、推理稳定性、多语言能力、长上下文性能及对少数样本的识别能力；剪枝可能删除特定任务所需的重要参数；知识蒸馏也可能只能迁移教师模型的部分能力。压缩效果取决于校准和评估数据的代表性、数值格式、模型架构、目标任务及硬件实现。在通用公开测试中保持较高分数的压缩模型，仍可能在客户专有流程、监管型应用或低频边缘案例中出现不可接受的性能下降。

数据治理、模型许可和知识产权要求也会限制商业采用。压缩服务商可能需要接触专有模型权重、分词器文件、校准样本、评估数据及生产使用特征。金融、医疗、政府和国防客户可能禁止相关资产外传，从而增加对本地部署服务的需求，但同时提高项目实施复杂度。部分开源模型许可证对修改或再分发设置条件，而闭源模型提供商通常不向客户开放模型权重。此外，定制化压缩格式和专用推理引擎可能增加客户对特定硬件、运行时或服务商的依赖，提高后续迁移成本。

市场机遇

企业级开源权重大模型部署是最主要的市场机会。较多企业倾向于在私有基础设施中运行可控模型，但难以按照业务所需规模部署完整精度版本。压缩服务商可以提供由模型选择、量化、精度评估、运行时优化及部署集成组成的标准化服务包。金融、医疗、制造、法律服务、通信和政府等行业的需求尤其值得关注，因为这些客户普遍重视私有推理、确定性时延和运营成本控制。

端侧及设备侧生成式AI构成另一项重要增长机会。个人电脑、智能手机、汽车、机器人、工业网关及智能摄像头通常无法部署最大规模的全精度模型，但经过压缩的语言、视觉语言、语音及多模态模型能够支持本地助手、文档处理、语音交互、环境感知和设备控制。Qualcomm AI Hub支持端侧模型量化、编译和真实设备测试，Google AI Edge及Nota AI NetsPresso则强调硬件兼容与部署优化。能够面向多种NPU和嵌入式加速器适配大模型的服务商，可同时获取终端制造商和应用开发商的需求。

混合专家模型、推理模型和多模态基础模型正在形成更高技术壁垒的服务机会。这些架构通常不能仅通过统一权重量化取得最佳效果，服务商可能需要优化专家路由、删除冗余专家、为敏感层配置不同精度、分别压缩视觉编码器与语言解码器、降低键值缓存需求，或将复杂推理能力蒸馏至较小模型。Nota AI已开展针对大型混合专家模型的量化和专家剪枝工作，NVIDIA也在将量化感知训练或强化学习与知识蒸馏结合，以恢复低精度部署条件下的模型能力。此类项目技术要求更高，更适合形成高价值定制服务合同。

行业风险与挑战

行业面临的主要技术挑战是平衡压缩率、推理速度和模型质量。模型文件更小并不必然代表推理时延更低，实际性能还取决于计算内核、内存访问、批处理规模、输入序列长度、加速器利用率和运行时实现。当目标硬件缺乏原生低比特计算支持时，部分低精度格式只能减少显存占用，而难以获得明显速度提升。因此，服务商必须测试完整业务负载，而不能仅依赖理论压缩比例。

模型评估同样是重大挑战。大模型可以执行开放式生成、推理、检索、代码编写、翻译和多模态任务，难以通过单一精度指标判断压缩效果。服务商需要将公开基准与客户专有数据集、安全测试、幻觉评估、长上下文测试及人工审核相结合。模型行为还会受到提示词、语言和解码参数影响，因此验证压缩模型能否用于生产，远比比较总体基准分数更加复杂。

技术迭代速度过快也构成行业风险。新的模型架构、量化格式、推理引擎和加速器不断推出，服务商需要持续支持稠密Transformer、混合专家模型、状态空间模型和多模态架构，同时适配TensorRT-LLM、vLLM、ONNX Runtime、OpenVINO及各类硬件SDK。由于大量压缩算法已经开源，专业服务商必须依靠自动化能力、专有优化方法、可信测试结果、客户支持和生产部署经验建立差异化优势。

价值链分析

上游主要由基础模型开发商、开源模型平台、训练框架、评估数据集、半导体计算平台及推理运行时构成。关键技术投入包括模型权重、架构定义、分词器资产、校准数据、客户专有评估集、GPU或其他加速器资源，以及PyTorch、ONNX、vLLM、TensorRT-LLM和OpenVINO等软件框架。GPTQ、AWQ、SmoothQuant、稀疏化、剪枝和知识蒸馏等算法研究，为商业压缩服务提供方法基础。

中游参与者包括云端模型优化平台、芯片厂商工具链、开源优化框架、专业压缩企业和AI工程咨询服务商。核心活动涵盖模型分析、压缩技术选择、数据校准、量化、剪枝、知识蒸馏、计算图改写、内核适配、运行时编译、性能测试及精度恢复。服务交付模式包括自助式软件、云端优化任务、API平台、企业软件许可、专业咨询及与客户共同开发的定制优化项目。

下游包括基础模型企业、AI应用公司、云服务商、部署私有AI的企业客户、设备制造商及嵌入式系统开发企业。客户通过减少基础设施支出、提高推理吞吐、缩短响应时间、改善设备兼容性，并支持原本无法承载大模型的应用场景实现价值。

领先服务商的业务范围正由一次性模型压缩进一步延伸至部署架构、推理软件、模型监控、版本更新和持续优化服务。
细分市场分析
按照服务技术划分，大模型压缩服务可分为量化服务、剪枝与稀疏化服务、知识蒸馏服务、低秩及结构压缩服务，以及多技术融合优化服务。量化是标准化程度较高的类别，具体包括训练后量化、量化感知训练、仅权重量化、权重与激活联合量化及混合精度优化，常见部署格式包括FP8、INT8和INT4，但具体支持能力取决于模型、运行时和目标硬件。剪枝与稀疏化服务可删除权重、通道、网络层、注意力头或模型专家，知识蒸馏则利用大型教师模型的输出或中间表示训练更小的学生模型。
多技术融合服务的重要性正在上升，因为单一压缩方式通常难以同时实现最优的质量和效率。服务商可以先删除部分网络层或降低模型深度，再通过知识蒸馏恢复能力，之后实施低比特量化，并将模型编译至特定推理引擎。NVIDIA已展示结合结构化剪枝和知识蒸馏生成紧凑型语言模型的方法，Red Hat和Neural
Magic则提供融合量化与稀疏化、面向vLLM部署的LLM
Compressor工作流。此类项目需要更多工程投入，但可能比标准量化获得更明显的基础设施成本改善。
按照交付模式划分，市场可分为项目制定制压缩、自助式优化平台、托管式云端优化、企业软件许可，以及压缩加推理一体化服务。定制项目适用于专有模型、特殊硬件及严格质量要求；自助式平台适合拥有内部工程团队和重复性优化任务的客户；托管云服务可以降低基础设施和工具链复杂度；压缩加推理服务商则同时通过模型优化和后续推理托管实现商业变现。
按照部署目标划分，市场可覆盖云端GPU优化、CPU与私有数据中心优化、云端定制加速器优化、工作站和个人电脑部署、移动及汽车NPU部署，以及嵌入式边缘部署。云端项目更加关注吞吐量、服务并发和单Token成本，端侧项目则强调模型内存、功耗、启动速度和硬件兼容性，不同场景需要采用差异化压缩策略和评估标准。
下游市场机会
云计算和数据中心运营商是重要客户群体，因为模型压缩可以提升单台加速器承载的模型数量或并发会话数量。AI软件企业和企业级解决方案商利用压缩技术控制智能助手、AI智能体、检索增强生成、文档分析、代码辅助及智能客服等应用的运行成本。使用私有基础设施部署模型的企业还需要根据现有GPU或CPU集群实施优化，为硬件中立型咨询和平台服务创造需求。
终端设备制造商构成另一条增长路径。智能手机、个人电脑、汽车、机器人和工业设备企业正寻求无需持续连接云端即可运行的本地生成式AI功能。这些客户需要模型适配特定NPU、内存容量、电池预算和操作系统，相关压缩服务可能由芯片厂商、设备制造商和应用开发商共同采购。
基础模型开发企业本身也是潜在客户。模型开发商不再仅发布全精度模型，而是逐步提供多个参数规模和量化版本，以扩大可部署范围。压缩服务商可以协助生成官方低比特版本、小型蒸馏模型和面向特定硬件的软件包，并随着推理框架和加速器更新持续开展测试及维护。
地区分析
北美是大模型压缩技术和商业平台较为集中的地区，拥有主要云服务平台、GPU及处理器供应商、基础模型开发商和开源推理社区。NVIDIA提供Model Optimizer、NeMo及TensorRT相关优化能力，AWS通过SageMaker
AI提供托管式推理优化任务，Microsoft开发Olive与ONNX Runtime优化工具，Intel提供Neural
Compressor，Qualcomm运营AI Hub Workbench，Red Hat则整合了Neural
Magic在大模型压缩与vLLM方面的能力。该地区已形成覆盖算法、软件框架、托管服务和企业部署的完整生态。
欧洲拥有一批研究驱动型专业优化企业。Pruna AI提供模型压缩软件、付费高级优化功能和优化模型推理服务。欧洲企业对自主可控AI、私有化部署及开源权重大模型低成本运行的需求，将推动硬件中立、能源效率、数据保护及本地云部署相关服务发展。
亚太地区在端侧及设备导向的大模型压缩领域具有较强需求。韩国Nota AI通过NetsPresso提供模型压缩、硬件感知优化和边缘部署服务，并逐步扩展至大型语言和混合专家模型。中国、日本、韩国和中国台湾地区拥有较强的半导体、消费电子、汽车和工业设备产业基础，对大模型适配本地NPU及嵌入式处理器形成持续需求。模型压缩服务由此成为基础模型与商业化边缘硬件之间的重要连接环节。
竞争格局
大模型压缩服务市场由多类参与者共同竞争，尚未形成由单一类型企业主导的市场结构。AWS等云服务商提供与模型部署基础设施相结合的托管式优化流程；NVIDIA、Intel和Qualcomm等半导体及计算平台供应商提供面向自身或所支持硬件的模型优化工具链；Microsoft和Hugging Face分别通过Olive、ONNX Runtime及Optimum构建通用优化软件生态，尽管相关工具不一定以独立定制压缩服务形式销售，但对市场技术体系和客户自研能力具有较大影响。
专业服务商主要通过专有压缩算法、自动化能力和工程支持建立竞争优势。Red Hat与Neural
Magic重点布局企业级开源AI、LLM Compressor、稀疏化、量化及vLLM部署；Nota
AI依托NetsPresso提供硬件感知模型压缩和边缘部署；Pruna AI则将压缩软件、优化模型和推理服务相结合。当客户需要跨硬件平台优化、复杂模型适配或项目制定制支持时，此类企业通常比单一硬件平台厂商具有更高灵活性。
未来市场竞争将更多围绕可验证的生产环境结果，而不是单纯比较支持多少种压缩算法。压缩比例、任务质量保持率、推理吞吐量、时延、显存降低幅度、模型架构覆盖能力、目标硬件范围、自动化程度、数据安全方案和生产推理系统集成能力，均将成为关键竞争指标。能够提供可重复评估、本地化实施、多硬件支持及持续模型维护的服务商，将比仅提供单一量化工具的企业拥有更强的企业客户竞争力。
报告范围
本文重点分析在全球及中国有重要角色的企业，分析这些企业大模型压缩服务产品的市场规模、市场份额、市场定位、产品类型以及发展规划等。
主要企业包括：

NVIDIA
Red Hat
Intel
Nota AI
Qualcomm
Hugging Face
Multiverse Computing
Picovoice
Fireworks AI
Baseten
Predibase
Edge Impulse
SiMa.ai
Anyscale
阿里云计算
华为云
面壁智能
硅基流动
无问芯穹
潞晨科技
清程极智

按照不同产品类型，包括如下几个类别：

量化服务
知识蒸馏服务
剪枝与稀疏化服务
推理引擎与算子优化服务
硬件适配与部署优化服务
其他

按照不同优化目标，包括如下几个类别：

模型体积压缩
显存占用降低
推理延迟降低
推理吞吐提升
其他

按照不同应用，主要包括如下几个方面：

云端大模型优化
数据中心推理优化
私有化部署优化
边缘端模型优化
其他

重点关注如下几个地区

北美
欧洲
中国
日本
东南亚
印度

章节概要

第1章：报告统计范围、产品细分及全球总体规模及增长率等数据
第2章：全球不同应用大模型压缩服务市场规模及份额等
第3章：全球大模型压缩服务主要地区市场规模及份额等
第4章：全球范围内大模型压缩服务主要企业竞争分析，主要包括大模型压缩服务收入、市场份额及行业集中度分析
第5章：中国市场大模型压缩服务主要企业竞争分析，主要包括大模型压缩服务收入、市场份额及行业集中度分析
第6章：全球主要企业基本情况介绍，包括公司简介、大模型压缩服务产品、收入及最新动态等
第7章：行业发展机遇和风险分析
第8章：报告结论

报告目录

1 大模型压缩服务市场概述
1.1 大模型压缩服务市场概述
1.2 不同产品类型大模型压缩服务分析
1.2.1 量化服务
1.2.2 知识蒸馏服务
1.2.3 剪枝与稀疏化服务
1.2.4 推理引擎与算子优化服务
1.2.5 硬件适配与部署优化服务
1.2.6 其他
1.2.7 全球市场不同产品类型大模型压缩服务销售额对比（2021 VS 2025 VS 2032）
1.2.8 全球不同产品类型大模型压缩服务销售额及预测（2021-2032）
1.2.8.1 全球不同产品类型大模型压缩服务销售额及市场份额（2021-2026）
1.2.8.2 全球不同产品类型大模型压缩服务销售额预测（2027-2032）
1.2.9 中国不同产品类型大模型压缩服务销售额及预测（2021-2032）
1.2.9.1 中国不同产品类型大模型压缩服务销售额及市场份额（2021-2026）
1.2.9.2 中国不同产品类型大模型压缩服务销售额预测（2027-2032）
1.3 不同优化目标大模型压缩服务分析
1.3.1 模型体积压缩
1.3.2 显存占用降低
1.3.3 推理延迟降低
1.3.4 推理吞吐提升
1.3.5 其他
1.3.6 全球市场不同优化目标大模型压缩服务销售额对比（2021 VS 2025 VS 2032）
1.3.7 全球不同优化目标大模型压缩服务销售额及预测（2021-2032）
1.3.7.1 全球不同优化目标大模型压缩服务销售额及市场份额（2021-2026）
1.3.7.2 全球不同优化目标大模型压缩服务销售额预测（2027-2032）
1.3.8 中国不同优化目标大模型压缩服务销售额及预测（2021-2032）
1.3.8.1 中国不同优化目标大模型压缩服务销售额及市场份额（2021-2026）
1.3.8.2 中国不同优化目标大模型压缩服务销售额预测（2027-2032）
2 不同应用分析
2.1 从不同应用，大模型压缩服务主要包括如下几个方面
2.1.1 云端大模型优化
2.1.2 数据中心推理优化
2.1.3 私有化部署优化
2.1.4 边缘端模型优化
2.1.5 其他
2.2 全球市场不同应用大模型压缩服务销售额对比（2021 VS 2025 VS 2032）
2.3 全球不同应用大模型压缩服务销售额及预测（2021-2032）
2.3.1 全球不同应用大模型压缩服务销售额及市场份额（2021-2026）
2.3.2 全球不同应用大模型压缩服务销售额预测（2027-2032）
2.4 中国不同应用大模型压缩服务销售额及预测（2021-2032）
2.4.1 中国不同应用大模型压缩服务销售额及市场份额（2021-2026）
2.4.2 中国不同应用大模型压缩服务销售额预测（2027-2032）
3 全球大模型压缩服务主要地区分析
3.1 全球主要地区大模型压缩服务市场规模分析：2021 VS 2025 VS 2032
3.1.1 全球主要地区大模型压缩服务销售额及份额（2021-2026）
3.1.2 全球主要地区大模型压缩服务销售额及份额预测（2027-2032）
3.2 北美大模型压缩服务销售额及预测（2021-2032）
3.3 欧洲大模型压缩服务销售额及预测（2021-2032）
3.4 中国大模型压缩服务销售额及预测（2021-2032）
3.5 日本大模型压缩服务销售额及预测（2021-2032）
3.6 东南亚大模型压缩服务销售额及预测（2021-2032）
3.7 印度大模型压缩服务销售额及预测（2021-2032）
4 全球主要企业市场占有率
4.1 全球主要企业大模型压缩服务销售额及市场份额

4.2 全球大模型压缩服务主要企业竞争态势
4.2.1 大模型压缩服务行业集中度分析：2025年全球Top 5厂商市场份额
4.2.2 全球大模型压缩服务第一梯队、第二梯队和第三梯队企业及市场份额
4.3 2025年全球主要厂商大模型压缩服务收入排名
4.4 全球主要厂商大模型压缩服务总部及市场区域分布
4.5 全球主要厂商大模型压缩服务产品类型及应用
4.6 全球主要厂商大模型压缩服务商业化日期
4.7 新增投资及市场并购活动
4.8 大模型压缩服务全球领先企业SWOT分析
5 中国市场大模型压缩服务主要企业分析
5.1 中国大模型压缩服务销售额及市场份额（2021-2026）
5.2 中国大模型压缩服务Top 3和Top 5企业市场份额
6 主要企业简介
6.1 NVIDIA
6.1.1 NVIDIA公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.1.2 NVIDIA 大模型压缩服务产品及服务介绍
6.1.3 NVIDIA 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.1.4 NVIDIA公司简介及主要业务
6.1.5 NVIDIA企业最新动态
6.2 Red Hat
6.2.1 Red Hat公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.2.2 Red Hat 大模型压缩服务产品及服务介绍
6.2.3 Red Hat 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.2.4 Red Hat公司简介及主要业务
6.2.5 Red Hat企业最新动态
6.3 Intel
6.3.1 Intel公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.3.2 Intel 大模型压缩服务产品及服务介绍
6.3.3 Intel 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.3.4 Intel公司简介及主要业务
6.3.5 Intel企业最新动态
6.4 Nota AI
6.4.1 Nota AI公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.4.2 Nota AI 大模型压缩服务产品及服务介绍
6.4.3 Nota AI 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.4.4 Nota AI公司简介及主要业务
6.5 Qualcomm
6.5.1 Qualcomm公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.5.2 Qualcomm 大模型压缩服务产品及服务介绍
6.5.3 Qualcomm 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.5.4 Qualcomm公司简介及主要业务
6.5.5 Qualcomm企业最新动态
6.6 Hugging Face
6.6.1 Hugging Face公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.6.2 Hugging Face 大模型压缩服务产品及服务介绍
6.6.3 Hugging Face 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.6.4 Hugging Face公司简介及主要业务
6.6.5 Hugging Face企业最新动态
6.7 Multiverse Computing
6.7.1 Multiverse Computing公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.7.2 Multiverse Computing 大模型压缩服务产品及服务介绍
6.7.3 Multiverse Computing 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.7.4 Multiverse Computing公司简介及主要业务
6.7.5 Multiverse Computing企业最新动态
6.8 Picovoice
6.8.1 Picovoice公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.8.2 Picovoice 大模型压缩服务产品及服务介绍
6.8.3 Picovoice 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.8.4 Picovoice公司简介及主要业务
6.8.5 Picovoice企业最新动态

6.9 Fireworks AI
6.9.1 Fireworks AI公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.9.2 Fireworks AI 大模型压缩服务产品及服务介绍
6.9.3 Fireworks AI 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.9.4 Fireworks AI公司简介及主要业务
6.9.5 Fireworks AI企业最新动态
6.10 Baseten
6.10.1 Baseten公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.10.2 Baseten 大模型压缩服务产品及服务介绍
6.10.3 Baseten 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.10.4 Baseten公司简介及主要业务
6.10.5 Baseten企业最新动态
6.11 Predibase
6.11.1 Predibase公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.11.2 Predibase 大模型压缩服务产品及服务介绍
6.11.3 Predibase 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.11.4 Predibase公司简介及主要业务
6.11.5 Predibase企业最新动态
6.12 Edge Impulse
6.12.1 Edge Impulse公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.12.2 Edge Impulse 大模型压缩服务产品及服务介绍
6.12.3 Edge Impulse 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.12.4 Edge Impulse公司简介及主要业务
6.12.5 Edge Impulse企业最新动态
6.13 SiMa.ai
6.13.1 SiMa.ai公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.13.2 SiMa.ai 大模型压缩服务产品及服务介绍
6.13.3 SiMa.ai 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.13.4 SiMa.ai公司简介及主要业务
6.13.5 SiMa.ai企业最新动态
6.14 Anyscale
6.14.1 Anyscale公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.14.2 Anyscale 大模型压缩服务产品及服务介绍
6.14.3 Anyscale 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.14.4 Anyscale公司简介及主要业务
6.14.5 Anyscale企业最新动态
6.15 阿里云计算
6.15.1 阿里云计算公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.15.2 阿里云计算 大模型压缩服务产品及服务介绍
6.15.3 阿里云计算 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.15.4 阿里云计算公司简介及主要业务
6.15.5 阿里云计算企业最新动态
6.16 华为云
6.16.1 华为云公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.16.2 华为云 大模型压缩服务产品及服务介绍
6.16.3 华为云 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.16.4 华为云公司简介及主要业务
6.16.5 华为云企业最新动态
6.17 面壁智能
6.17.1 面壁智能公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.17.2 面壁智能 大模型压缩服务产品及服务介绍
6.17.3 面壁智能 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.17.4 面壁智能公司简介及主要业务
6.17.5 面壁智能企业最新动态
6.18 硅基流动
6.18.1 硅基流动公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.18.2 硅基流动 大模型压缩服务产品及服务介绍
6.18.3 硅基流动 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.18.4 硅基流动公司简介及主要业务
6.18.5 硅基流动企业最新动态

6.19 无问芯穹
6.19.1 无问芯穹公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.19.2 无问芯穹 大模型压缩服务产品及服务介绍
6.19.3 无问芯穹 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.19.4 无问芯穹公司简介及主要业务
6.19.5 无问芯穹企业最新动态
6.20 潞晨科技
6.20.1 潞晨科技公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.20.2 潞晨科技 大模型压缩服务产品及服务介绍
6.20.3 潞晨科技 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.20.4 潞晨科技公司简介及主要业务
6.20.5 潞晨科技企业最新动态
6.21 清程极智
6.21.1 清程极智公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
6.21.2 清程极智 大模型压缩服务产品及服务介绍
6.21.3 清程极智 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
6.21.4 清程极智公司简介及主要业务
6.21.5 清程极智企业最新动态
7 行业发展机遇和风险分析
7.1 大模型压缩服务行业发展机遇及主要驱动因素
7.2 大模型压缩服务行业发展面临的风险
7.3 大模型压缩服务行业政策分析
8 研究结果
9 研究方法与数据来源
9.1 研究方法
9.2 数据来源
9.2.1 二手信息来源
9.2.2 一手信息来源
9.3 数据交互验证
9.4 免责声明

报告图表

表格目录

表 1： 量化服务主要企业列表
表 2： 知识蒸馏服务主要企业列表
表 3： 剪枝与稀疏化服务主要企业列表
表 4： 推理引擎与算子优化服务主要企业列表
表 5： 硬件适配与部署优化服务主要企业列表
表 6： 其他主要企业列表
表 7： 全球市场不同产品类型大模型压缩服务销售额及增长率对比（2021 VS 2025 VS 2032）&（百万美元）
表 8： 全球不同产品类型大模型压缩服务销售额列表（2021-2026）&（百万美元）
表 9： 全球不同产品类型大模型压缩服务销售额市场份额列表（2021-2026）
表 10： 全球不同产品类型大模型压缩服务销售额预测（2027-2032）&（百万美元）
表 11： 全球不同产品类型大模型压缩服务销售额市场份额预测（2027-2032）
表 12： 中国不同产品类型大模型压缩服务销售额列表（2021-2026）&（百万美元）
表 13： 中国不同产品类型大模型压缩服务销售额市场份额列表（2021-2026）
表 14： 中国不同产品类型大模型压缩服务销售额预测（2027-2032）&（百万美元）
表 15： 中国不同产品类型大模型压缩服务销售额市场份额预测（2027-2032）
表 16： 模型体积压缩主要企业列表
表 17： 显存占用降低主要企业列表
表 18： 推理延迟降低主要企业列表
表 19： 推理吞吐提升主要企业列表
表 20： 其他主要企业列表
表 21： 全球市场不同优化目标大模型压缩服务销售额及增长率对比（2021 VS 2025 VS 2032）&（百万美元）

表 22： 全球不同优化目标大模型压缩服务销售额列表（2021-2026）&（百万美元）
表 23： 全球不同优化目标大模型压缩服务销售额市场份额列表（2021-2026）
表 24： 全球不同优化目标大模型压缩服务销售额预测（2027-2032）&（百万美元）
表 25： 全球不同优化目标大模型压缩服务销售额市场份额预测（2027-2032）
表 26： 中国不同优化目标大模型压缩服务销售额列表（2021-2026）&（百万美元）
表 27： 中国不同优化目标大模型压缩服务销售额市场份额列表（2021-2026）
表 28： 中国不同优化目标大模型压缩服务销售额预测（2027-2032）&（百万美元）
表 29： 中国不同优化目标大模型压缩服务销售额市场份额预测（2027-2032）
表 30： 全球市场不同应用大模型压缩服务销售额及增长率对比（2021 VS 2025 VS 2032）&（百万美元）
表 31： 全球不同应用大模型压缩服务销售额列表（2021-2026）&（百万美元）
表 32： 全球不同应用大模型压缩服务销售额市场份额列表（2021-2026）
表 33： 全球不同应用大模型压缩服务销售额预测（2027-2032）&（百万美元）
表 34： 全球不同应用大模型压缩服务市场份额预测（2027-2032）
表 35： 中国不同应用大模型压缩服务销售额列表（2021-2026）&（百万美元）
表 36： 中国不同应用大模型压缩服务销售额市场份额列表（2021-2026）
表 37： 中国不同应用大模型压缩服务销售额预测（2027-2032）&（百万美元）
表 38： 中国不同应用大模型压缩服务销售额市场份额预测（2027-2032）
表 39： 全球主要地区大模型压缩服务销售额：（2021 VS 2025 VS 2032）&（百万美元）
表 40： 全球主要地区大模型压缩服务销售额列表（2021-2026）&（百万美元）
表 41： 全球主要地区大模型压缩服务销售额及份额列表（2021-2026）
表 42： 全球主要地区大模型压缩服务销售额列表预测（2027-2032）&（百万美元）
表 43： 全球主要地区大模型压缩服务销售额及份额列表预测（2027-2032）
表 44： 全球主要企业大模型压缩服务销售额（2021-2026）&（百万美元）
表 45： 全球主要企业大模型压缩服务销售额份额对比（2021-2026）
表 46： 2025年全球大模型压缩服务主要厂商市场地位（第一梯队、第二梯队和第三梯队）
表 47： 2025年全球主要厂商大模型压缩服务收入排名（百万美元）
表 48： 全球主要厂商大模型压缩服务总部及市场区域分布
表 49： 全球主要厂商大模型压缩服务产品类型及应用
表 50： 全球主要厂商大模型压缩服务商业化日期
表 51： 全球大模型压缩服务市场投资、并购等现状分析
表 52： 中国主要企业大模型压缩服务销售额列表（2021-2026）&（百万美元）
表 53： 中国主要企业大模型压缩服务销售额份额对比（2021-2026）
表 54： NVIDIA公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 55： NVIDIA 大模型压缩服务产品及服务介绍
表 56： NVIDIA 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 57： NVIDIA公司简介及主要业务
表 58： NVIDIA企业最新动态
表 59： Red Hat公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 60： Red Hat 大模型压缩服务产品及服务介绍
表 61： Red Hat 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 62： Red Hat公司简介及主要业务
表 63： Red Hat企业最新动态
表 64： Intel公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 65： Intel 大模型压缩服务产品及服务介绍
表 66： Intel 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 67： Intel公司简介及主要业务
表 68： Intel企业最新动态
表 69： Nota AI公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 70： Nota AI 大模型压缩服务产品及服务介绍
表 71： Nota AI 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 72： Nota AI公司简介及主要业务
表 73： Qualcomm公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 74： Qualcomm 大模型压缩服务产品及服务介绍
表 75： Qualcomm 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 76： Qualcomm公司简介及主要业务
表 77： Qualcomm企业最新动态
表 78： Hugging Face公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 79： Hugging Face 大模型压缩服务产品及服务介绍
表 80： Hugging Face 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 81： Hugging Face公司简介及主要业务

表 82： Hugging Face企业最新动态
表 83： Multiverse Computing公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 84： Multiverse Computing 大模型压缩服务产品及服务介绍
表 85： Multiverse Computing 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 86： Multiverse Computing公司简介及主要业务
表 87： Multiverse Computing企业最新动态
表 88： Picovoice公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 89： Picovoice 大模型压缩服务产品及服务介绍
表 90： Picovoice 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 91： Picovoice公司简介及主要业务
表 92： Picovoice企业最新动态
表 93： Fireworks AI公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 94： Fireworks AI 大模型压缩服务产品及服务介绍
表 95： Fireworks AI 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 96： Fireworks AI公司简介及主要业务
表 97： Fireworks AI企业最新动态
表 98： Baseten公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 99： Baseten 大模型压缩服务产品及服务介绍
表 100： Baseten 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 101： Baseten公司简介及主要业务
表 102： Baseten企业最新动态
表 103： Predibase公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 104： Predibase 大模型压缩服务产品及服务介绍
表 105： Predibase 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 106： Predibase公司简介及主要业务
表 107： Predibase企业最新动态
表 108： Edge Impulse公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 109： Edge Impulse 大模型压缩服务产品及服务介绍
表 110： Edge Impulse 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 111： Edge Impulse公司简介及主要业务
表 112： Edge Impulse企业最新动态
表 113： SiMa.ai公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 114： SiMa.ai 大模型压缩服务产品及服务介绍
表 115： SiMa.ai 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 116： SiMa.ai公司简介及主要业务
表 117： SiMa.ai企业最新动态
表 118： Anyscale公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 119： Anyscale 大模型压缩服务产品及服务介绍
表 120： Anyscale 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 121： Anyscale公司简介及主要业务
表 122： Anyscale企业最新动态
表 123： 阿里云计算公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 124： 阿里云计算 大模型压缩服务产品及服务介绍
表 125： 阿里云计算 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 126： 阿里云计算公司简介及主要业务
表 127： 阿里云计算企业最新动态
表 128： 华为云公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 129： 华为云 大模型压缩服务产品及服务介绍
表 130： 华为云 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 131： 华为云公司简介及主要业务
表 132： 华为云企业最新动态
表 133： 面壁智能公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 134： 面壁智能 大模型压缩服务产品及服务介绍
表 135： 面壁智能 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 136： 面壁智能公司简介及主要业务
表 137： 面壁智能企业最新动态
表 138： 硅基流动公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 139： 硅基流动 大模型压缩服务产品及服务介绍
表 140： 硅基流动 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 141： 硅基流动公司简介及主要业务

表 142： 硅基流动企业最新动态
表 143： 无问芯穹公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 144： 无问芯穹 大模型压缩服务产品及服务介绍
表 145： 无问芯穹 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 146： 无问芯穹公司简介及主要业务
表 147： 无问芯穹企业最新动态
表 148： 潞晨科技公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 149： 潞晨科技 大模型压缩服务产品及服务介绍
表 150： 潞晨科技 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 151： 潞晨科技公司简介及主要业务
表 152： 潞晨科技企业最新动态
表 153： 清程极智公司信息、总部、大模型压缩服务市场地位以及主要的竞争对手
表 154： 清程极智 大模型压缩服务产品及服务介绍
表 155： 清程极智 大模型压缩服务收入及毛利率（2021-2026）&（百万美元）
表 156： 清程极智公司简介及主要业务
表 157： 清程极智企业最新动态
表 158： 大模型压缩服务行业发展机遇及主要驱动因素
表 159： 大模型压缩服务行业发展面临的风险
表 160： 大模型压缩服务行业政策分析
表 161： 研究范围
表 162： 本文分析师列表

图表目录

图 1： 大模型压缩服务产品图片
图 2： 全球市场大模型压缩服务市场规模（销售额）, 2021 VS 2025 VS 2032（百万美元）
图 3： 全球大模型压缩服务市场销售额预测:（百万美元）&（2021-2032）
图 4： 中国市场大模型压缩服务销售额及未来趋势（2021-2032）&（百万美元）
图 5： 量化服务 产品图片
图 6： 全球量化服务规模及增长率（2021-2032）&（百万美元）
图 7： 知识蒸馏服务产品图片
图 8： 全球知识蒸馏服务规模及增长率（2021-2032）&（百万美元）
图 9： 剪枝与稀疏化服务产品图片
图 10： 全球剪枝与稀疏化服务规模及增长率（2021-2032）&（百万美元）
图 11： 推理引擎与算子优化服务产品图片
图 12： 全球推理引擎与算子优化服务规模及增长率（2021-2032）&（百万美元）
图 13： 硬件适配与部署优化服务产品图片
图 14： 全球硬件适配与部署优化服务规模及增长率（2021-2032）&（百万美元）
图 15： 其他产品图片
图 16： 全球其他规模及增长率（2021-2032）&（百万美元）
图 17： 全球不同产品类型大模型压缩服务市场份额2025 & 2032
图 18： 全球不同产品类型大模型压缩服务市场份额2021 & 2025
图 19： 全球不同产品类型大模型压缩服务市场份额预测2026 & 2032
图 20： 中国不同产品类型大模型压缩服务市场份额2021 & 2025
图 21： 中国不同产品类型大模型压缩服务市场份额预测2026 & 2032
图 22： 模型体积压缩 产品图片
图 23： 全球模型体积压缩规模及增长率（2021-2032）&（百万美元）
图 24： 显存占用降低产品图片
图 25： 全球显存占用降低规模及增长率（2021-2032）&（百万美元）
图 26： 推理延迟降低产品图片
图 27： 全球推理延迟降低规模及增长率（2021-2032）&（百万美元）
图 28： 推理吞吐提升产品图片
图 29： 全球推理吞吐提升规模及增长率（2021-2032）&（百万美元）
图 30： 其他产品图片
图 31： 全球其他规模及增长率（2021-2032）&（百万美元）
图 32： 全球不同优化目标大模型压缩服务市场份额2025 & 2032
图 33： 全球不同优化目标大模型压缩服务市场份额2021 & 2025
图 34： 全球不同优化目标大模型压缩服务市场份额预测2026 & 2032
图 35： 中国不同优化目标大模型压缩服务市场份额2021 & 2025
图 36： 中国不同优化目标大模型压缩服务市场份额预测2026 & 2032

图 37： 云端大模型优化
图 38： 数据中心推理优化
图 39： 私有化部署优化
图 40： 边缘端模型优化
图 41： 其他
图 42： 全球不同应用大模型压缩服务市场份额2025 VS 2032
图 43： 全球不同应用大模型压缩服务市场份额2021 & 2025
图 44： 全球主要地区大模型压缩服务销售额市场份额（2021 VS 2025）
图 45： 北美大模型压缩服务销售额及预测（2021-2032）&（百万美元）
图 46： 欧洲大模型压缩服务销售额及预测（2021-2032）&（百万美元）
图 47： 中国大模型压缩服务销售额及预测（2021-2032）&（百万美元）
图 48： 日本大模型压缩服务销售额及预测（2021-2032）&（百万美元）
图 49： 东南亚大模型压缩服务销售额及预测（2021-2032）&（百万美元）
图 50： 印度大模型压缩服务销售额及预测（2021-2032）&（百万美元）
图 51： 2025年全球前五大厂商大模型压缩服务市场份额
图 52： 2025年全球大模型压缩服务第一梯队、第二梯队和第三梯队厂商及市场份额
图 53： 大模型压缩服务全球领先企业SWOT分析
图 54： 2025年中国排名前三和前五大模型压缩服务企业市场份额
图 55： 关键采访目标
图 56： 自下而上及自上而下验证
图 57： 资料三角测定