Enterprise Strategy Group by TechTarget

经济白皮书

了解大型语言模型推理的总成本

与公有云或基于令牌的 API 相比，利用 Dell Technologies 本地解决方案具有更高的性价比，可将使用 RAG 进行 LLM 推理的经济效益提升 38% 到 88%

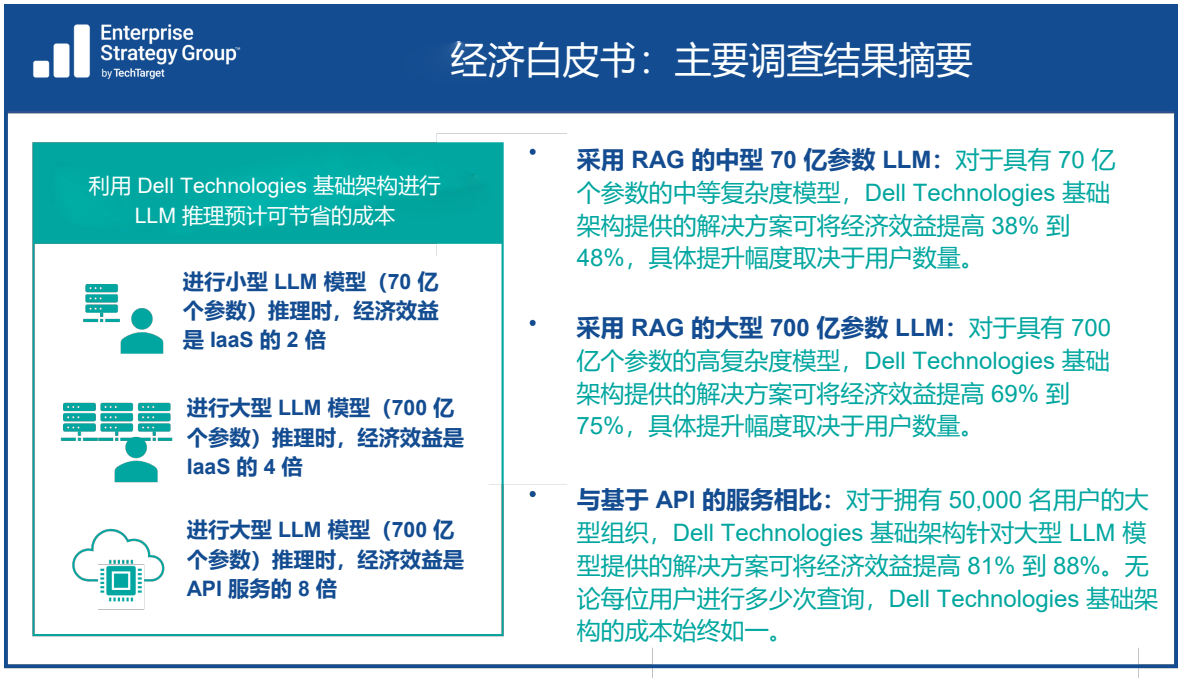
作者：Aviv Kaufmann，业务实践主管兼首席验证分析师
Enterprise Strategy Group

2024 年 4 月

本 Enterprise Strategy Group 经济验证报告受 Dell Technologies 委托撰写，经 TechTarget, Inc. 许可分发。

目录

简介 **3**
挑战 3
关于 LLM 推理的关键考虑因素 4
Enterprise Strategy Group 经济分析 **5**
Dell Technologies 本地基础架构与公有云 IaaS 的比较 5
规模较小的模型：具有 70 亿个参数的 Mistral 7B LLM 6
规模较大的模型：具有 700 亿个参数的 Llama 2 LLM 7
Dell Technologies 本地基础架构与基于 API 的 GenAI 服务的比较 8
要考虑的问题 **8**
Dell Technologies 的 LLM 推理解决方案 **9**
结论 **9**



简介

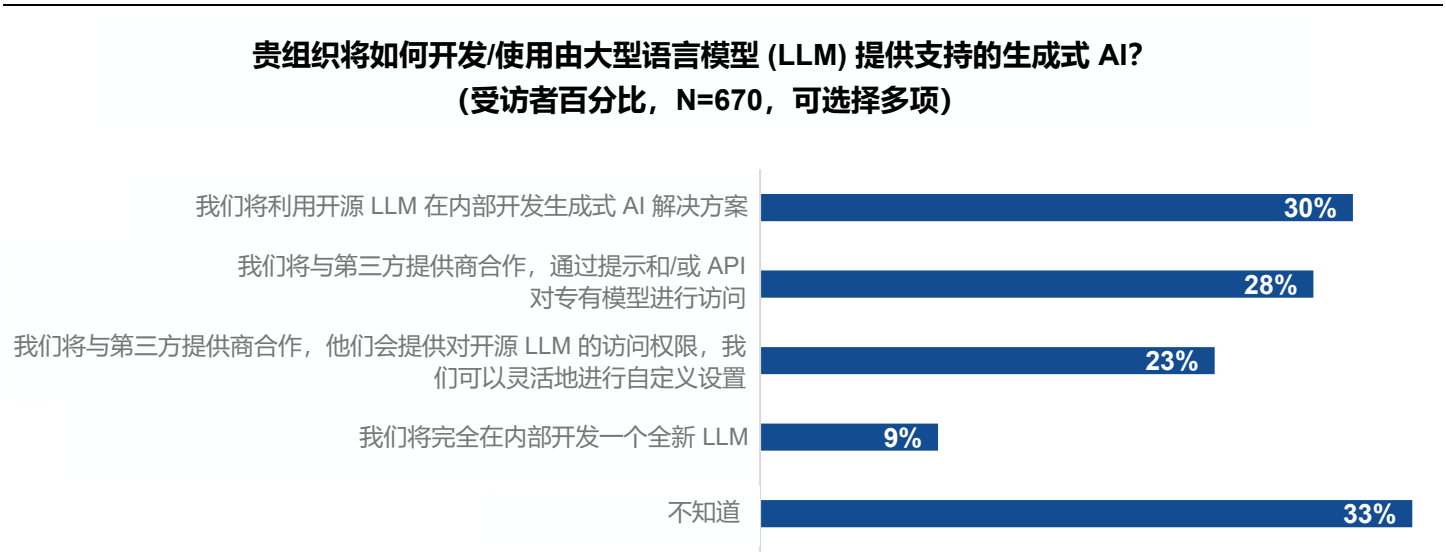
本经济白皮书介绍了在向组织提供基于文本的生成式 AI (GenAI) 功能时，一些可用的方案和考虑因素。TechTarget 旗下的 Enterprise Strategy Group 对利用检索增强生成 (RAG) 进行大型语言模型 (LLM) 推理的预期成本进行了建模，并比较了使用以下几种方案时的情形：使用本地 Dell Technologies 基础架构，以及使用原生公有云基础架构即服务 (IaaS) 或通过 API 使用 OpenAI GPT-4 Turbo LLM 模型服务。我们发现，Dell Technologies 在 LLM 推理方面具有更强的经济效益，最高是 IaaS 的 4 倍，是 GPT-4 Turbo API 的 8 倍。

挑战

各组织纷纷采用强大的 GenAI 和 LLM 技术，利用公司特定的数据和其他知识产权来自动生成内容、回答问题，并随时向决策者提供可用见解。组织使用 GenAI 将获益良多，一项 Enterprise Strategy Group 研究调查的受访者表示，在其组织中使用 GenAI 有几项主要好处，包括改进和/或自动执行流程和工作流、为数据分析和商业智能提供支持、提升员工的工作效率，以及提高运营效率。[1]

自行开发 LLM 不仅成本高昂，过程也相当复杂，但组织可以轻松地增强、微调和自定义现有开源 LLM 以满足自身需求。OpenAI GPT 等基于 API 的现成服务提供了一种更简单的解决方案，但在使用过程中推理（即查询）成本会迅速累积，大型组织和较为复杂的 LLM 尤甚。此外，组织也可以选择在支持 GPU 的强大企业服务器上，构建和控制自己的 LLM 推理解决方案，或使用支持 GPU 的同等云实例和运行开源 LLM 的机器学习平台（例如 NVIDIA 的 AI Enterprise）。不出所料，Enterprise Strategy Group 发现，在开发和使用由 LLM 提供支持的 GenAI 时，组织最热衷的战略是利用开源 LLM 在内部开发 GenAI 解决方案。[2]

图 1.大多数组织计划在内部开发自己的生成式 AI 解决方案



来源：TechTarget, Inc. 旗下部门 Enterprise Strategy Group

[1]来源：Enterprise Strategy Group 研究报告，《*Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*》，2023 年 8 月。
[2] 同上。

© 2024 TechTarget, Inc. 保留所有权利。

关于 LLM 推理的关键考虑因素

基于文本的 LLM 主要用于学习、理解和生成基于文本的内容、回答、摘要和问题，这些内容可根据特定的行业、应用场景和组织量身定制。RAG 可使用从其他来源拉取的自定义数据增强 GenAI 模型的结果，有助于提高模型的准确性。此类 LLM 是企业部署最多的 LLM，不仅可用于聊天机器人、问答助手、实现流程改进和自动化，也可作为功能内置于自定义工具和应用程序中，此外还有许多其他应用场景。在交付 LLM 模型时，组织必须要考虑采用何种基础架构来进行训练（即执行必要的数据和计算密集型分析以构建有效模型）、推理（即在经过训练的模型上提供用户交互）和微调（即持续更新和优化模型）。本报告重点介绍执行推理工作负载应采用的基础架构。下面提供了几种适用于进行 LLM 推理的部署方法，其中包括：

- **传统基础架构。**购买或租赁的传统基础架构通常包括计算、内存、GPU 和存储资源，可以与商用或开源 AI 平台一起部署和管理，便于组织掌控各方面的部署事宜。在处理规模较大且可预测的工作负载时，此方法最为经济高效。
- **公有云 IaaS。**同样，组织可以一同部署具有 GPU 和存储的同等云计算实例，以及商用或开源 AI 平台。通过此方法也可以对平台进行类似控制，同时具有敏捷性，并且可与现有工具轻松集成。当部署规模较小以及具有不可预测或季节性需求时，此方法最为经济高效。
- **LLM API 服务。**OpenAI GPT 等成熟服务可用于快速提供相关功能，而无需管理基础架构或 AI 平台。在初步探索和开始使用阶段，部署规模较小，以及不需要进行很大程度的自定义或控制时，此方法是理想选择。

在决定采用哪种 LLM 平台之前，组织应该花时间了解自身的需求和能力，并探讨以下有关选择 LLM 推理平台的一些考虑因素，例如：

- **成本/ROI。**在落实和使用每项技术投资时，组织都应衡量该项投资的成本和收益。根据 Enterprise Strategy Group 的一项调查研究，很多组织表示，成本节约和 ROI 是用于衡量 AI 计划有效性的最常见指标。[3]
- **性能和可扩展性。**组织需要调整基础架构规模，确保处理器、GPU、内存和存储具有充足的资源，这点至关重要；这样一来，可以确保它能够在正常负载和峰值负载下，应对预期的推理并发性，并将平均推理延迟保持在极低水平，从而为用户营造积极的体验。在转移到推理平台之前，组织还应确定针对 LLM 的计算密集型训练是在同一平台上进行，还是使用性能更高的专用训练平台。
- **管理简易性。**在比较本地基础架构与云基础架构和服务时，组织必须考虑自身的内部能力，并了解运营相关基础架构和平台的成本（例如，管理成本、支持和维护成本以及电力/冷却成本）。借助托管选项，组织还可以享受在自己的数据中心进行托管的诸多好处，同时减轻运营基础架构和平台在资源和技能方面造成的负担。
- **预期用户工作负载。**选择解决方案时，组织需要考虑一项重要指标，那就是了解和预测每天有多少用户访问该工具以及他们提问的频率。如果需求量很小，使用 API 服务就能满足需要，但如果组织支持更多的用户和推理，构建专有平台将更具经济效益。组织必须考虑到随着时间的推移，采用率和使用频率的预期增长情况，以确保基础架构的规模适当，并可随业务需求的增长不断扩展。

[3]来源：Enterprise Strategy Group 研究报告，《*Navigating the Evolving AI Infrastructure Landscape*》，2023 年 9 月。

- **数据监管。**训练和维护模型时，组织必须考虑所用数据源的位置和数据监管需求。如果数据驻留在本地或可在需要时轻松进行检索，那么使用混合云基础架构最为适用，而在某些情况下，使用公有云可以更轻松地收集和集中数据。使用本地实例，组织还能更好地控制安全性并确保敏感数据合规。维护最新、全面且客观公正的数据并使用此类数据进行训练，可改进 LLM 的效果并提高推理产生的见解的准确性。

Enterprise Strategy Group 经济分析

Enterprise Strategy Group 开展了一项经济分析，比较了利用 RAG 为几个复杂性各异的开源 LLM（包括 70 亿和 700 亿参数数量）和不同规模的组织（用户数量在 5 千到 5 万之间）提供推理的预期成本。我们假设该模型用于提供基于文本的内部问答，并且在数据所在位置进行推理，因此数据迁移成本不高。这项分析关注三年内与运行和推理模型相关的所有成本，包括提供和运行基础架构、管理系统以及根据需要支付云服务费用。

Dell Technologies 本地基础架构与公有云 IaaS 的比较

该模型首先比较了在以下两种平台上运行 LLM 推理的预期成本：传统基础架构（本地、托管环境、边缘位置等）以及 Amazon EC2 实例上具有类似配置的公有云 IaaS。根据推理基线测试的结果，我们针对每个工作负载调整了推理节点服务器和 NVIDIA H100 GPU 配置要求的规模，以确保它们能够应对常规负载和峰值负载（基于最大请求数和模型实例数）下的并发需求，并为每个预期工作负载提供适当的延迟和吞吐量。然后，我们对表 1 中所述的 Dell Technologies 基础架构和同等 EC2 配置的各项成本进行了建模。

表 1. 针对每项 LLM 推理工作负载需求构建的成本和假设模型

成本类别	Dell Technologies（本地）	公有云 IaaS (Amazon EC2)
初始购置成本（硬件和软件）	Dell Technologies 针对包含 ProDeploy 和 ProSupport 服务的戴尔 PowerEdge R760xa 和 R660 提供的价格	不适用
其他资本成本（利息）和折旧费用（收益）	已纳入模型中（8% WACC，6% 年折旧收益）	不适用
电力和冷却成本	根据系统规格计算得出（0.173 美元/千瓦时）	不适用
每月云支出	不适用	根据 3 年期预留折扣计算 p5.48xlarge EC2 实例成本
NVIDIA AI Enterprise 许可证/GPU	基于 5 年期许可证（按比例计算）	按每小时每实例计算，以每周 5 天、每天 16 小时为基础，从而限制成本
基础架构/实例管理	已建模（根据节点数量，系统管理员的比例为 10%-100%）	比本地模型低 66%
ML 模型和平台管理	已建模（根据模型实例的数量，ML 工程师的比例为 20%-100%）	与本地模型相同

来源：TechTarget, Inc. 旗下部门 Enterprise Strategy Group

规模较小的模型：具有 70 亿个参数的 Mistral 7B LLM

在第一次比较中，我们对提供一个包含约 70 亿个参数的小型模型的成本进行了建模，该模型类似于开源 Mistral 7B LLM。为了确定需求规模，我们根据测试结果使用了一种规模调整工具，该工具预测了服务器和 GPU 配置，结果表明，它可提供每个请求约 0.4 秒的平均延迟以及每秒 2.29 到 6.86 次推理的预计吞吐量。表 2 显示了有关实例和 GPU 数量的假设概况。

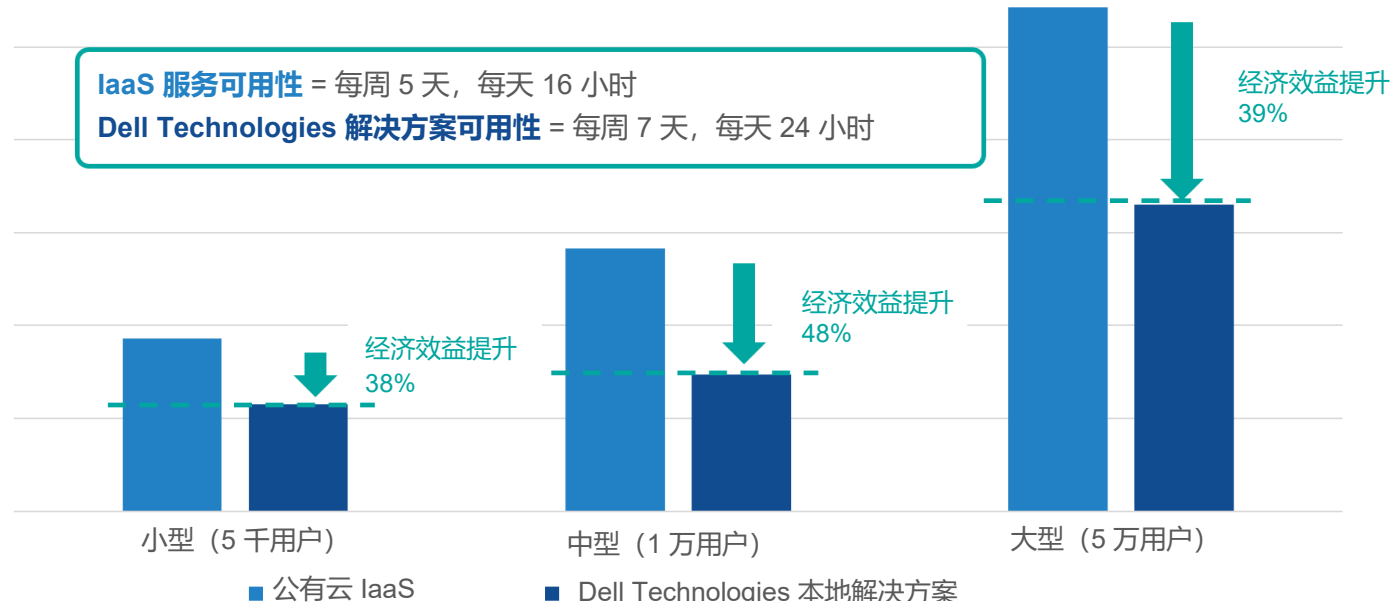
表 2. Mistral 70 亿参数模型推理的配置假设

LLM 模型（参数数量）	用户数量	推理节点/实例数量	H100 GPU 数量
Mistral（70 亿）	5,000	1	1
	10,000	1	2
	50,000	1	4

来源：TechTarget, Inc. 旗下部门 Enterprise Strategy Group

然后，我们根据表 1 中汇总的所有成本对每种配置进行了建模。如图 3 所示，就为组织提供推理而言，Dell Technologies 基础架构的经济效益高出 1.6 倍到 1.9 倍（38% 到 48%），并且全天候随时可供组织使用。

图 2.使用 RAG 为 70 亿参数 Mistral LLM 提供推理的预期成本



来源：TechTarget, Inc. 旗下部门 Enterprise Strategy Group

规模较大的模型：具有 700 亿个参数的 Llama 2 LLM

我们随后对提供具有 700 亿个参数的大型模型的预期成本进行了建模，该模型类似于开源 Llama 2 70B LLM。我们再次使用该规模调整工具确定了需求规模，以预测服务器和 GPU 配置，结果表明，它可提供约 1.8 秒的每请求平均延迟，比之前略高，并且可提供每秒 2.29 到 22.86 次推理的预计吞吐量。表 3 显示了有关实例和 GPU 数量的假设概况。

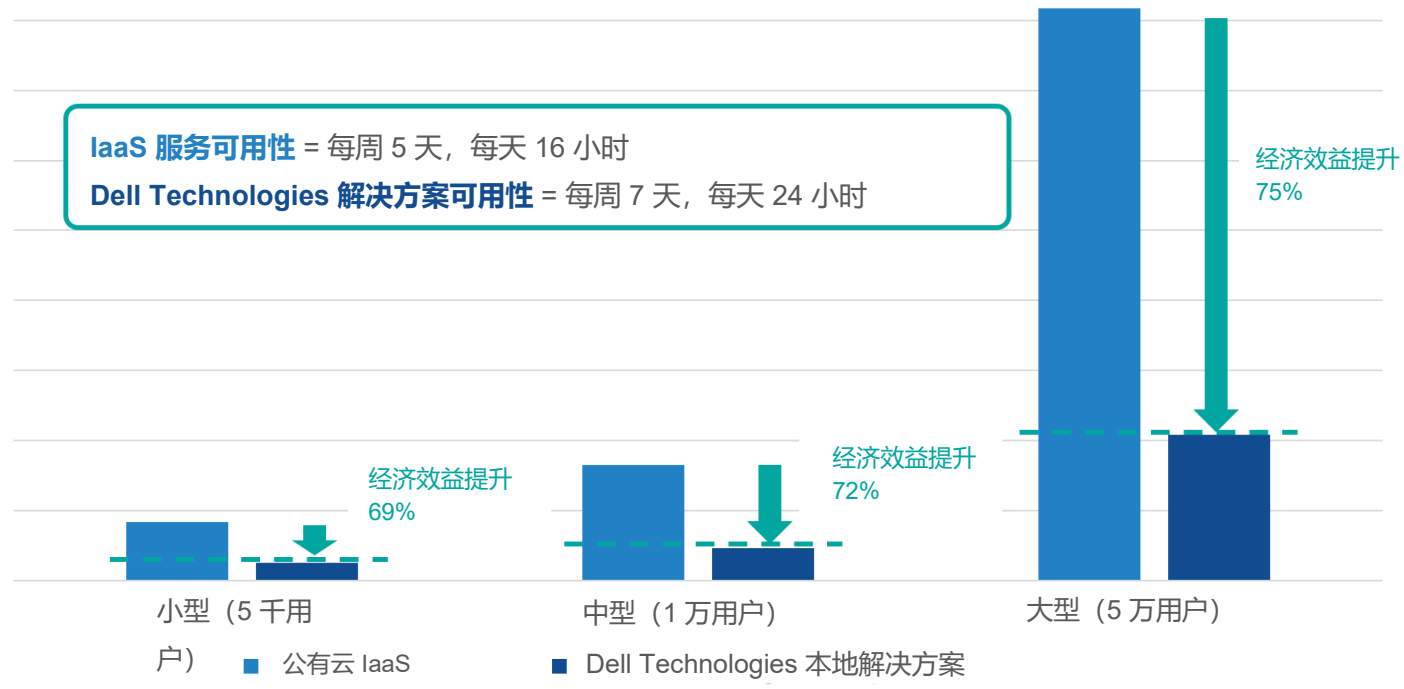
表 3. Llama 2 700 亿参数模型推理的配置假设

LLM 模型（参数数量）	用户数量	推理节点/实例数量	H100 GPU 数量
Llama 2（700 亿）	5,000	2	8
	10,000	4	16
	50,000	20	80

来源：TechTarget, Inc. 旗下部门 Enterprise Strategy Group

再次根据表 1 中汇总的所有成本对上述每种配置进行建模后，我们发现，就为组织提供推理而言，Dell Technologies 基础架构的经济效益高出 3.3 倍到 4 倍（69% 到 75%），并且全天候随时可供组织使用。

图 3.使用 RAG 为 700 亿参数 Llama 2 LLM 提供推理的预期成本

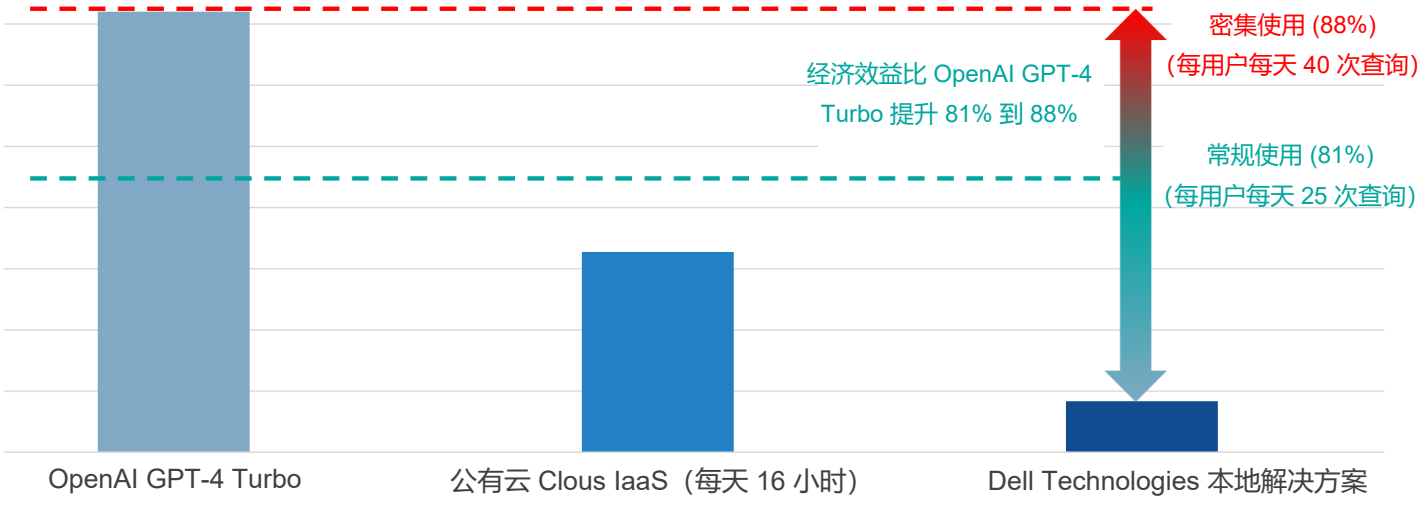


来源：TechTarget, Inc. 旗下部门 Enterprise Strategy Group

Dell Technologies 本地基础架构与基于 API 的 GenAI 服务的比较

然后，我们比较了大型组织使用基于 OpenAI API 的成熟 GenAI 服务 GPT-4 Turbo，为其 50,000 名用户提供同等 700 亿参数模型的预期成本。该服务按输入和输出“令牌”进行定价，具有出色的成本效益。基于文本的问答要求每个查询的令牌强度适中，峰值负载期间没有太大差异，因此所需的输入和输出令牌数量相对平衡。对于本地解决方案和基于 API 的解决方案，我们假定每个查询总共有 1,500 个（输入和输出）令牌，每个用户平均每天进行大约 25 次查询。我们对公开声明进行的研究表明，该假设设定的每用户查询数量适中，成熟度较低的组织中每个用户进行的查询次数偏少，而在成熟度较高的组织中，平均每个用户每天最多会进行 40 次查询。我们对 GPT-4 Turbo 进行的计算预测其成本约为每用户每月 12.50 美元，这一结果优于基于套件的 AI 辅助工具，后者的成本约为每用户每月 30 美元。根据这些假设，我们发现，与使用基于 API 的服务相比，Dell Technologies 本地基础架构提供推理的经济效益高出 5.4 倍到 8.6 倍（81% 到 88%），提供 GenAI 功能的成本仅为每用户每月 2.31 美元。

图 4.针对 5 万名用户为 700 亿参数 Llama 2 LLM 提供推理的预期成本



来源：TechTarget, Inc. 旗下部门 Enterprise Strategy Group

要考虑的问题

虽然 Enterprise Strategy Group 的模型以诚信为本，建立在保守、可信和经过验证的假设之上，但任何一种建模场景都无法充分代表每个潜在环境。客户所能节省的费用取决于各自具体的应用场景、数据性质、专业知识水平，及其模型和基础架构需求。Enterprise Strategy Group 建议您对可用产品进行自行分析，并咨询 Dell Technologies 的意见，以了解和讨论您通过概念验证测试证实的解决方案之间的差异。

Dell Technologies 的 LLM 推理解决方案

无论数据位于何处，Dell Technologies 均可帮助组织轻松地将 AI 应用于其数据。为此，戴尔提供从桌面到数据中心再到云的广泛 AI 服务组合，让组织能够合理调整投资规模并利用数据构建 AI 工厂，同时以可持续的方式高效、安全地实现各种 AI 应用场景。为实现此目标，除提供全面的服务组合外，戴尔还提供了一个广泛、开放的合作伙伴生态系统，无论组织是制定 AI 战略，还是加速和扩展其 GenAI 投资，戴尔都能为客户提供鼎立支持，满足他们在 AI 之旅不同阶段面临的需求。

如果组织面临数据安全威胁、合规性问题、数据孤岛和未经验证的数据集，借助面向生成式 AI 的 Dell Professional Services，组织可帮助业务和 IT 主管就优先应用场景达成共识，并制定可行的路线图以实现目标、为 LLM 集成准备企业数据、提高网络安全成熟度，以及建立符合具体业务需求的 AI 平台。此外，借助 Dell APEX，组织可以订阅 AI 解决方案，并针对多云应用场景对其进行优化。

要详细了解戴尔提供的解决方案，请访问戴尔 AI 网页。

结论

GenAI 的应用范围不断扩大，几乎涵盖每个业务领域，这是确保改善运营和助力企业未来取得成功的关键因素。Enterprise Strategy Group 的研究表明，当今组织应用 GenAI 的主要领域包括研究、市场营销、软件开发、产品开发和 IT 运营，预计该技术在各个领域的普及热度将只增不减。[4]组织可以对自己的自定义 LLM 版本进行训练和推理，以获得更具影响力和更有意义的结果。

进行 LLM 推理可以利用几种部署方法，每种方法都有一些独特优势，分别适用于特定的应用场景和需求。对于拥有数千名用户的大型组织来说，他们可以随时利用自定义 LLM 提供的功能，在这种情况下，Dell Technologies 基础架构可提供高性能 LLM 推理，其经济效益是 IaaS 的 4 倍，是 OpenAI GPT-4 Turbo 的 8 倍。Enterprise Strategy Group 强烈建议广大企业在实施 LLM 为其组织提供支持时，考虑利用 Dell Technologies 提供的经济高效的技术和知识渊博的服务，以确保取得卓越的成效、加快落实其 GenAI 计划，并在更短时间内实现预期可节省的成本。

[4]来源：Enterprise Strategy Group 研究报告，《*Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*》，2023 年 8 月。

©TechTarget, Inc. 或其子公司。保留所有权利。TechTarget 和 TechTarget 标识是 TechTarget, Inc. 的商标或注册商标，已在全球各司法辖区注册。其他产品和服务名称及标识，包括 BrightTALK、Xtelligent 和 Enterprise Strategy Group 可能是 TechTarget 或其子公司的商标。所有其他商标、标识和品牌名称均为其各自所有者的资产。

本出版物中包含的信息来自 TechTarget 认为具有可靠性的来源，但 TechTarget 对此不作担保。本出版物可能包含 TechTarget 的观点，这些观点可能随时发生改变。本出版物可能包括预测、推测和其他预测性陈述，代表 TechTarget 根据当前可用信息做出的假设和预期。这些预测基于行业趋势，包含变数和不确定性。因此，TechTarget 不保证本出版物所包含的特定预测、推测或预测性陈述的准确性。

未经 TechTarget 明确同意，任何以硬拷贝形式、电子形式或其他形式将本出版物的全部或部分内容复制或再分发给无权接收的人员的行为，均属违反美国版权法，将承担民事损害赔偿责任并受到刑事诉讼（如适用）。如有任何疑问，请联系客户关系部 cr@esg-global.com。

Enterprise Strategy Group 简介

TechTarget 的 Enterprise Strategy Group 提供有针对性且切实可行的市场情报、需求方研究、分析师咨询服务、GTM 战略指导、解决方案验证以及为企业技术的采购和销售提供支持的定制内容。

contact@esg-global.com

www.esg-global.com