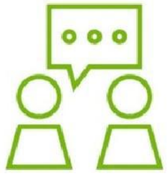
面向NVIDIA云合作伙伴的人工智能服务机会

通过生成A1提高企业价值

生成型人工智能对生产率的影响每年可能为全球经济增加4.4万亿美元。1

知识库副驾驶
人工智能助手来提升工作流



内容生成
图像、文本、三维模型

报告和数据分析
总结文本并生成可视化功能

语言翻译
多语言的实时交流

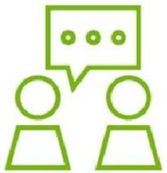
最重要的代理和聊天机器人
针对特定领域的专门聊天机器人

代码生成
高质量的规范建议

资料来源：生成型A1的经济潜力 生产力前沿出版社，麦肯锡出版社，2023年

流行的人工智能用例

GENERATIVE AI



LLM培训和推理

基于知识的副驾驶

代码/图像生成

文本摘要

翻译

DATA ANALYTICS

加速Apache火花

商业智能数据工程

AI/ML管道加速

SPEECH AI

加速语音Al

基于语音的代理助理

消费者应用程序

在线会议

NVIDIA AI企业

使ncp能够提供增值的人工智能服务

NCP人工智能服务

NV AI
企业启用

√较高的利润率
√成为值得信赖的顾问
，√开发粘性

NVIDIA云合作伙伴战略

参考资料

A1的架构

云

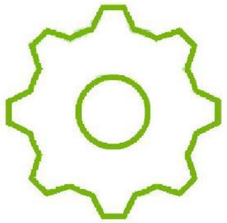
端到端计算，
网络和软件

优化A1工作负载

规模性能

+

NCP软件服务



NVIDIA A1企业+无软件

NVPS领导的集群部署

性能基准

确认

=

更快的上市时间，
卓越的客户
经验

性能和高可用性的基础设施

为A1服务做好准备

访问NVIDIA专家

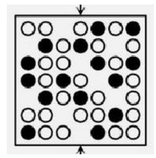
支持

将AI纳入您的数据的方法

云提供商
服务机会

价值

GenAI Propt
窗口用户



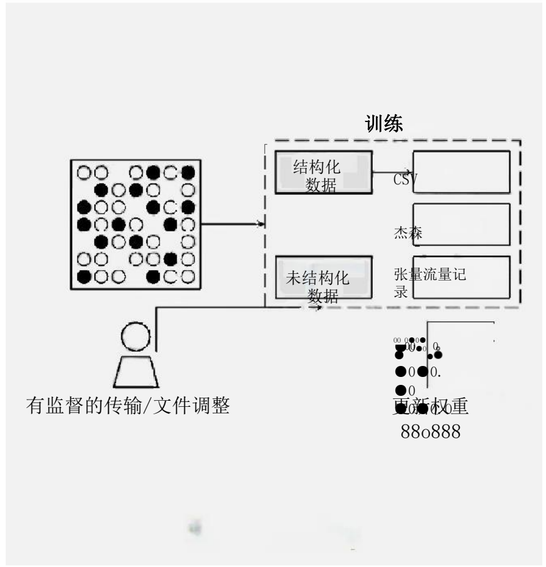
特定数
据集

Transformer模型OrgXDohain

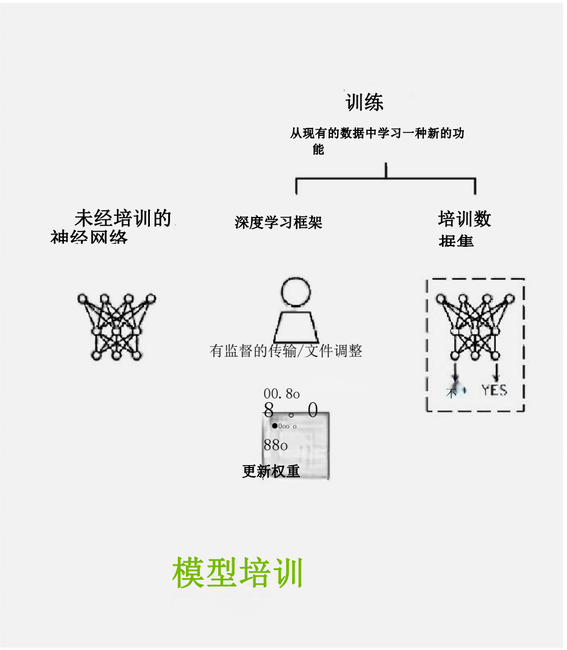
预先训练过的模型/引用模型增强

-努力

开发人员（要求）



和成本一



需要数据科学家(s)

*基于戴尔公司的数据。nvida。

架构解决方案

NVIDIA AI企业

AI生产级软件

人工智能解决方案
生产最快路径

NVIDIA NIM AI微服务

常见的应用程序和
优化模型
便于部署

砌块
对AI
发展

为生产部署而构建的企业级基础容

企业级
安全性、稳定性、可管理性
&support

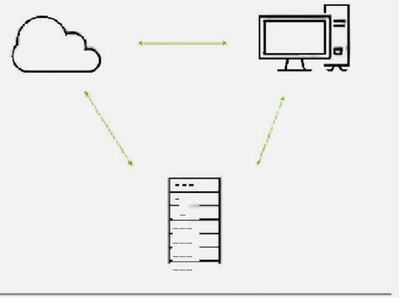
CVE修补程序

API稳定性

端到端可管
理性

sla与
NVIDIA支持

云本地和认证
到处跑



云|数据中心|工作站|Edge

英伟达
认证

NVIDIA AI企业

云、数据中心和工作站的边缘

NVIDIA AI企业

NVIDIA AI企业

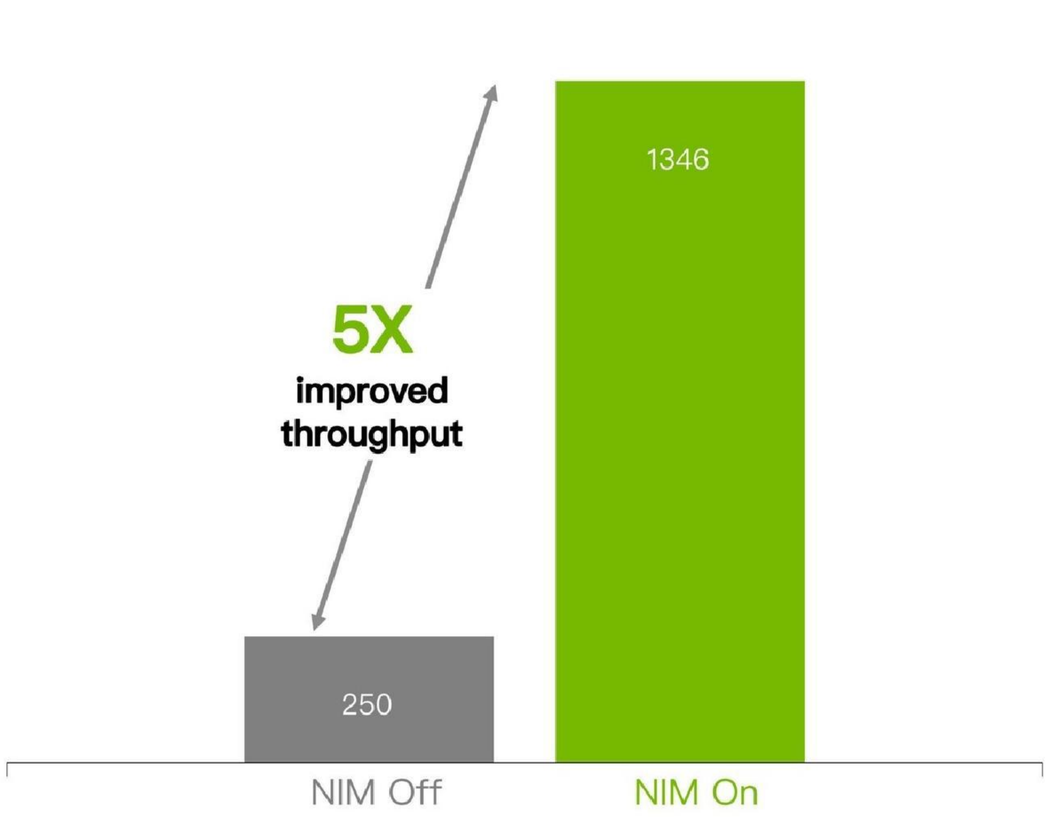
云、数据中心和工作站的边缘

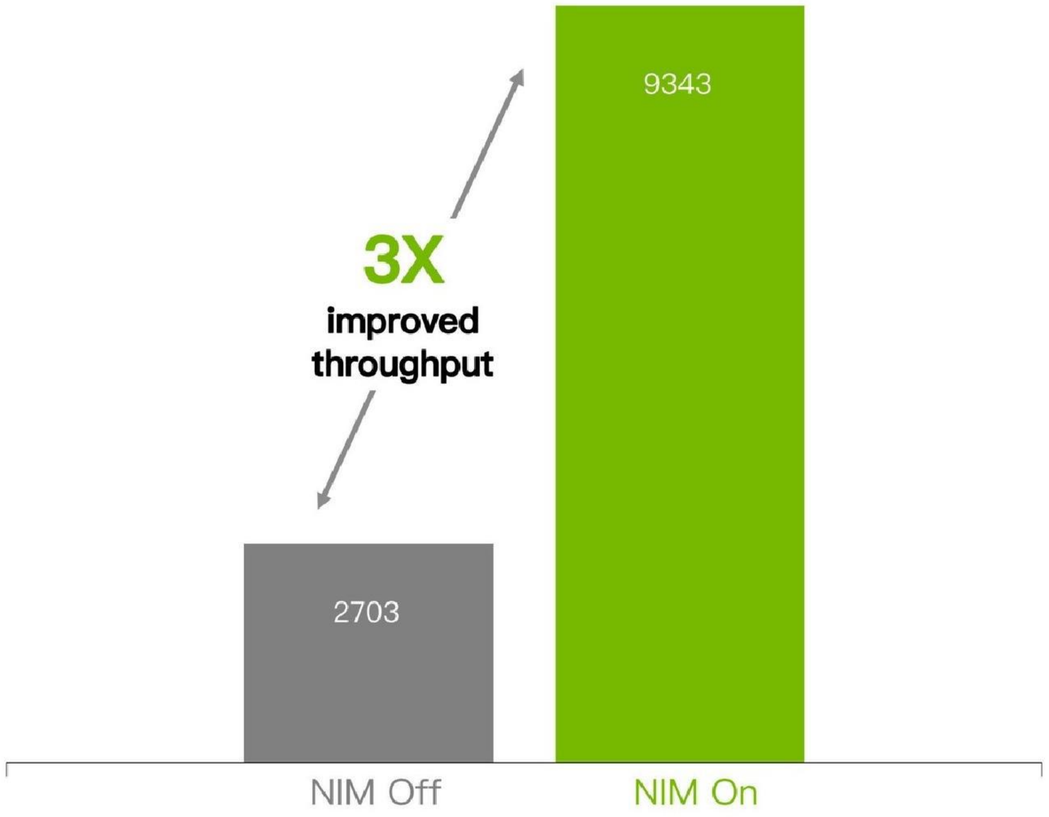
NVIDIA AI企业效益

最多可节省5倍的成本

提高的效率降低了解决方案的总体成本

3-70B在4xH100 SXM-3-8B在1xH100





喇嘛3-70，长度：700年：100借给eguests：100.4100 sMLink。NMOf: FP6,TF:~120s,TL:~180ms.NMOn: FP8.TF:~4.5,TL:-70ms

为所有所有部署提供灵活的软件分支

安全，API稳定，和安心为您所有的投资

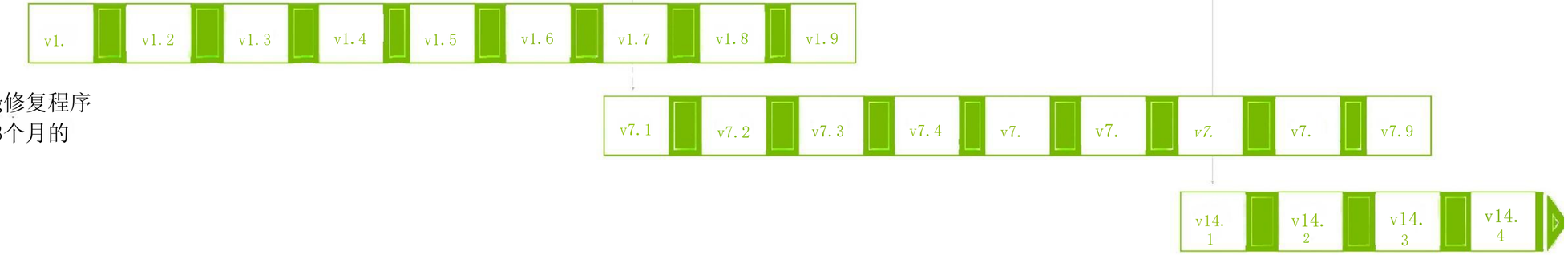
英伟达人工智能企业

功能分支

树顶软件优化
每月发布节奏
CVE补丁和bug修复

生产分公司

API稳定性
每月发布一次的CVE补丁程序和bug修复程序
2个分支机构/年，9个月的寿命，3个月的
重叠



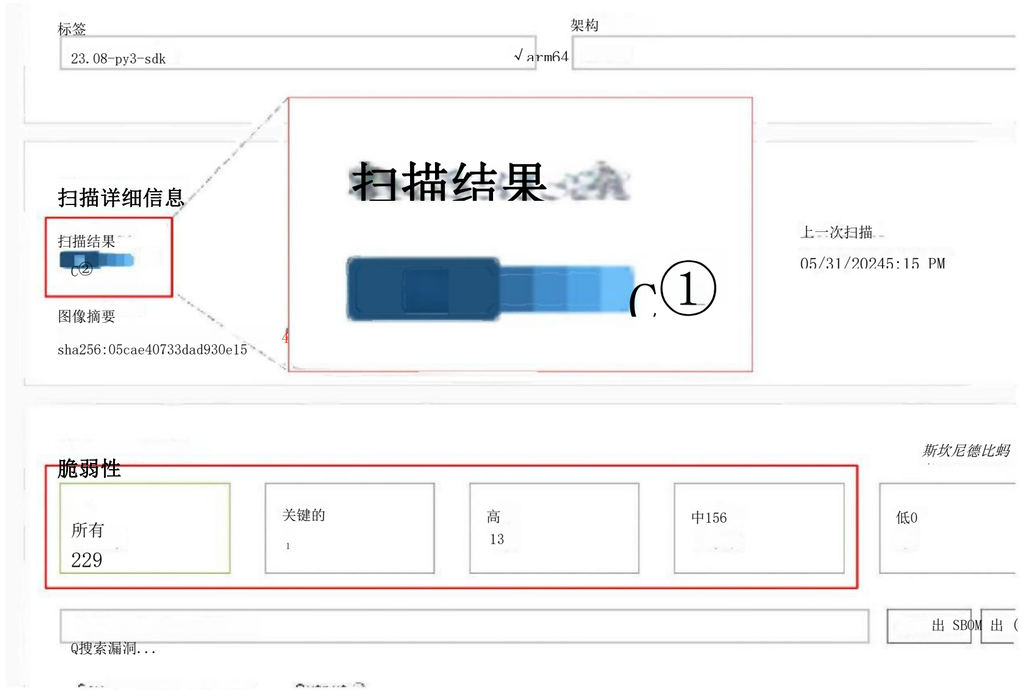
长期支持部门

对于高度监管的行业，季度CVE补
丁/bug
固定
最多3年，支持6个月的重叠
期

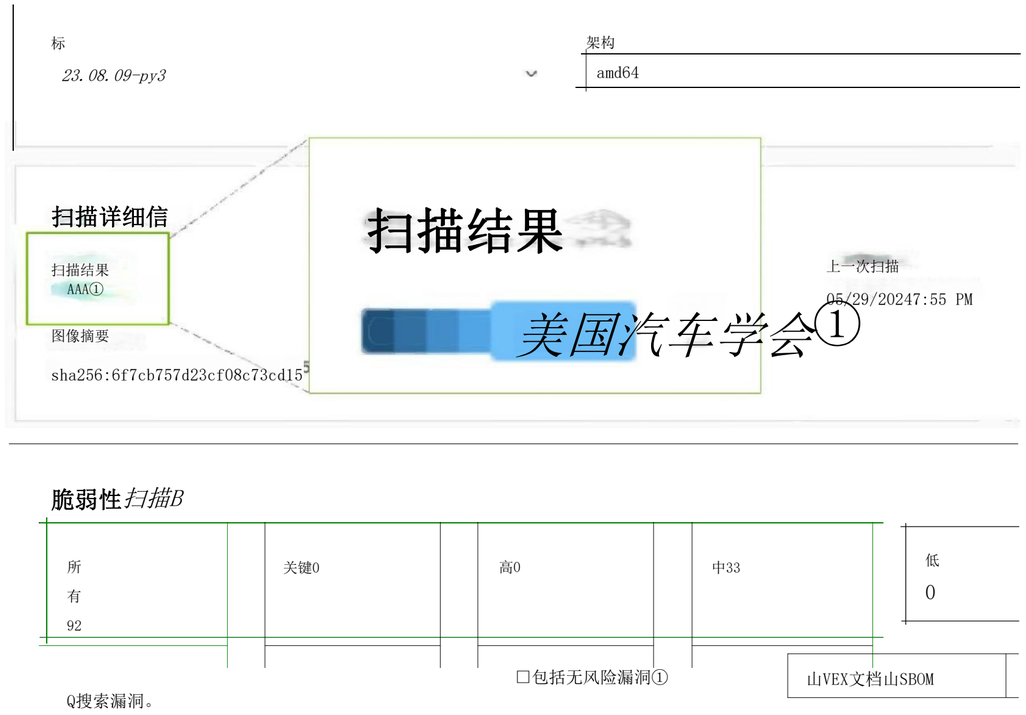
主动安全评估

每晚扫描和报告每个NIM容器

NVIDIA AI企业关闭



NVIDIA AI企业成立



易用性最新报告

- 非安全专家使用的容器　明白
- 按每个容器内的包装进行分级

每30天更新一次详细报告

- 快速识别旧版本上的安全问题

NVIDIA企业服务

提供客户成功

支持服务

包含的支持

业务标准支持

√Standard SLA

√访问NVIDIA AI专家，√功能分支（FB），
生产分公司（PB）和长期分公司支持
（LTS）

√优先级通知

增值支持

业务关键支持

√24/7支持

√1小时响应SLA

技术客户经理（TAM）

√客户冠军，√指定的支持

√绩效审查和支持计划

专业服务

人工智能工作负载登录

√指导和支持

√特定用例

自定义约定

教育服务

√NVIDIA AI企业管理公共训练营

√在数据中心介绍AI

√加速计算、数据科学、深度学习和图形学(自定节奏&结构导向型)

业务连续性加速达到了价值

投资回报

英伟达NIM

企业生成铝的NIM

生成型A1的企业运行时

英伟达NIM

推理微服务

加快进入市场的时间，
持续维护的微量服务

通过从专有数据源调整自定义模型来提高准确性

部署在任何地方的安全和控制的铝的应用程序和数据

授权开发人员使用最新的A1模型、标准apl和企业工具

在生产环境中部署、具有API稳定性、安全补丁和企业支持

优化了最大的吞吐量

最大限度地提高总体

云

数据中心

原

工作站

NVIDIA NIM推理微服务

在一个容器中作为微服务交付的一个完整的AI推理堆栈

英伟达NIM

企业基础容器

社区或自定义模型

域特定代码

库伯内特

优化的推理引擎

行业标准应用程序

掌舵机图和K8s操作符

GPU加速系统

快速和方便的AI模型部署

NVIDIA NIM上的加速基础设施

推理视觉设计检索语音生物学

不公正的人工智

谷歌

一个部门

shutterstsck

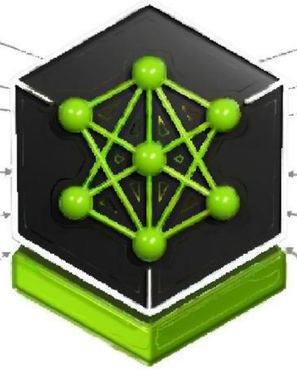
英伟达

元

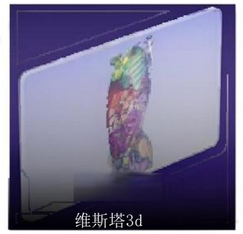
英伟达

getty图像

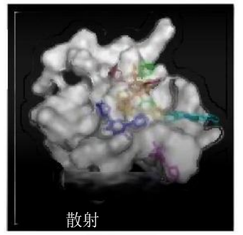
微软公司



英伟达
企业容器



英伟达



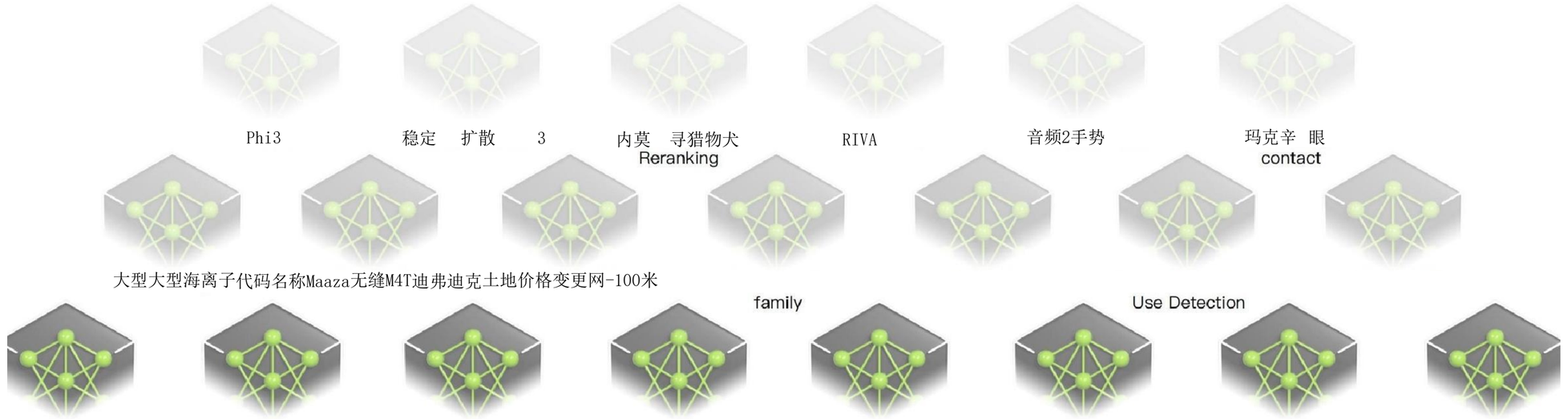
MIT

英伟达

NVIDIA模型、社区模型、自定义模型和合作伙伴模型

nvida。

NIM对许多领域的微服务



文字区域-视觉

优化的文本

RAG语音

杜吉坦

医疗保健

计算机仿真

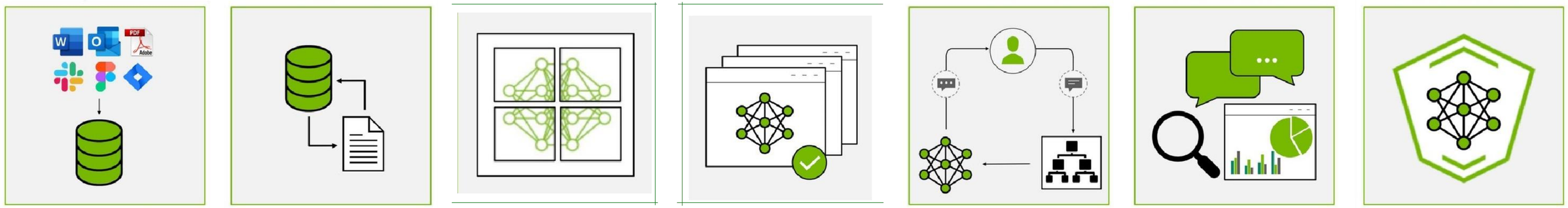
视觉

NVIDIA内莫

使用NVIDIA NeMo构建企业生成型所有应用程序

定制、增强和部署生成式A1模型

数据数据，数据模型信息
收购持续时间预训练自定义检索推断护栏

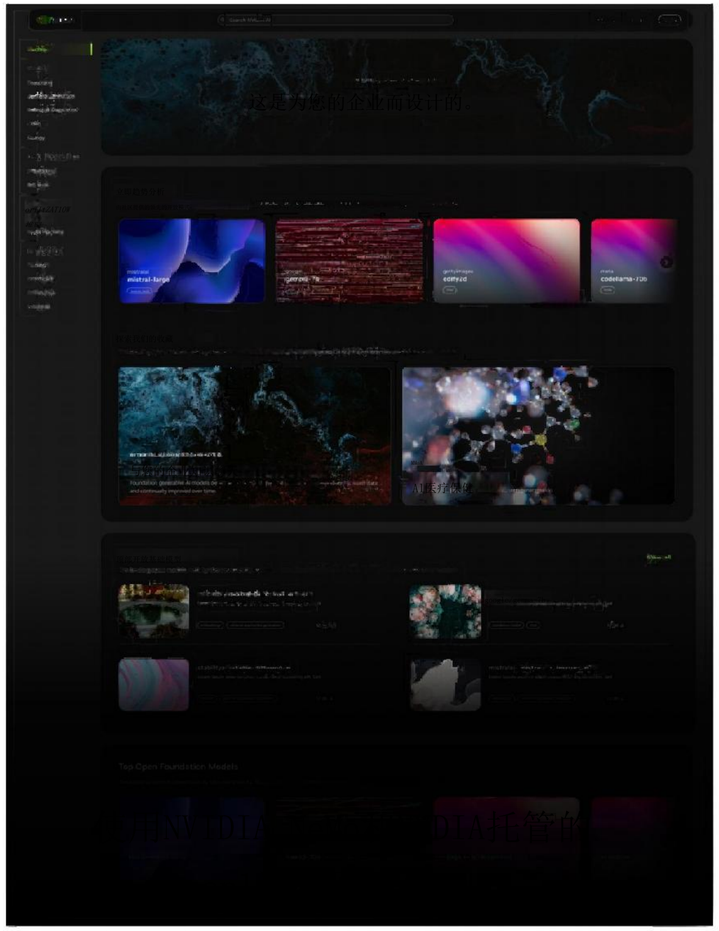


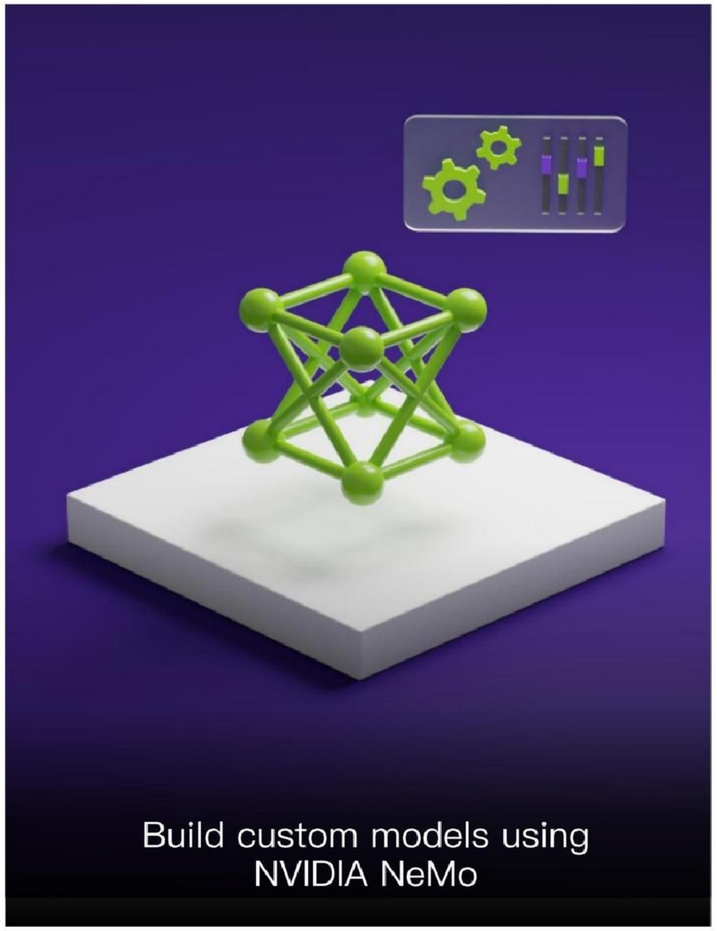
数据准备、微调和自定义部署

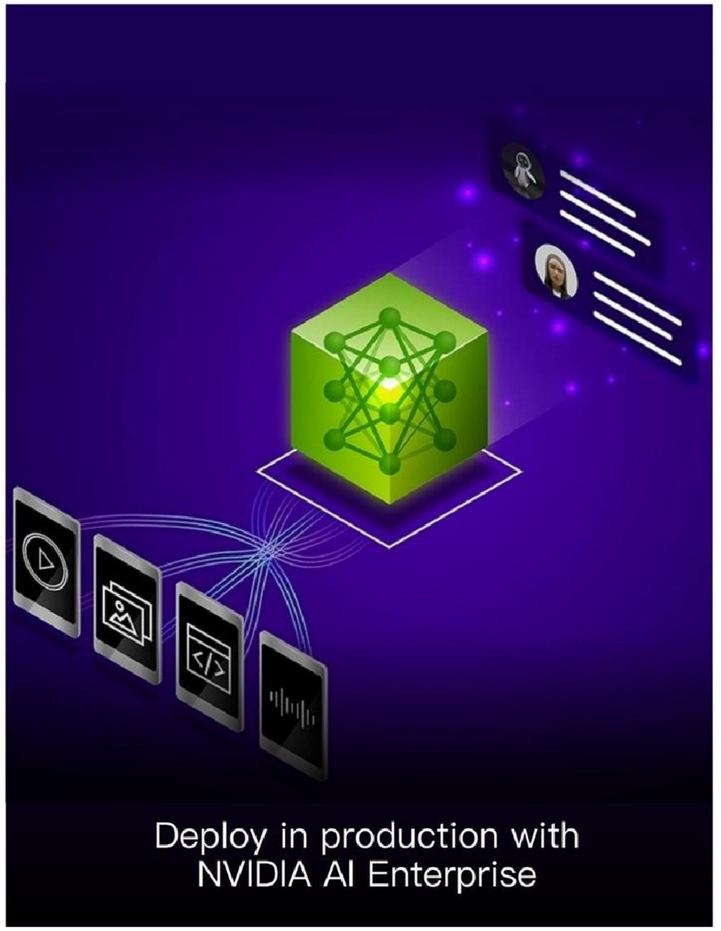
提高数据的质量和广度调整到使用域规模最终用户实施

开始使用NVIDIA NeMo

经验、原型和部署最新的A1模型。英伟达。com





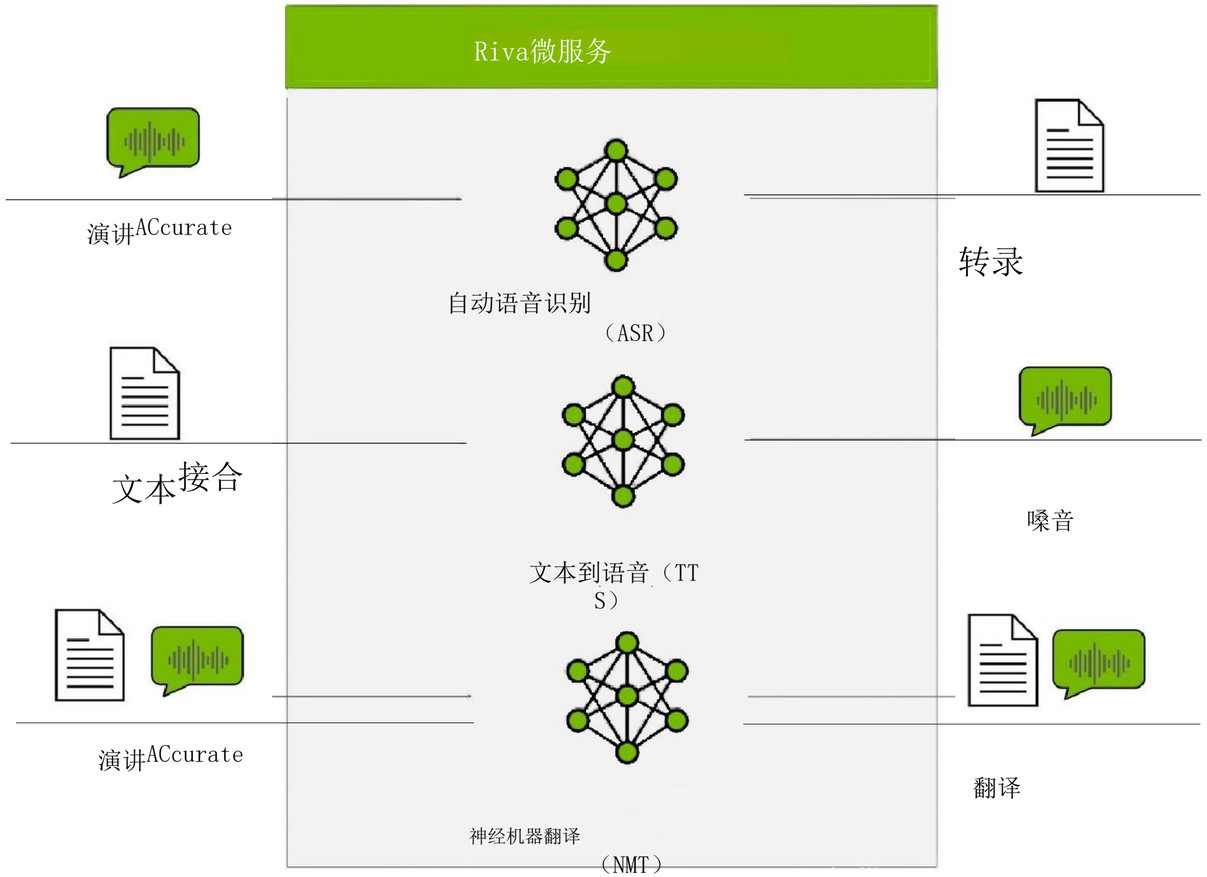


NVIDIA里瓦

NVIDIA里瓦

面向会话应用程序的世界级多语言语音接口：

聊天机器人，智能虚拟助手，数字化身



注：Riva的NIM正在进行早期访问

- Models:
  - 实时性，完全GPU加速
  - ●的开箱即用的最先进的状态
  - 高度准确
  - 完全可根据具体需要进行定制
  - 多语言：
    - 用5种语言编写的TTS
    - NMT适用于英语<→31语言的语言 • 扩展到数十万

用户

- 灵活部署：本地部署、所有云部署、边缘部署、嵌入式部署

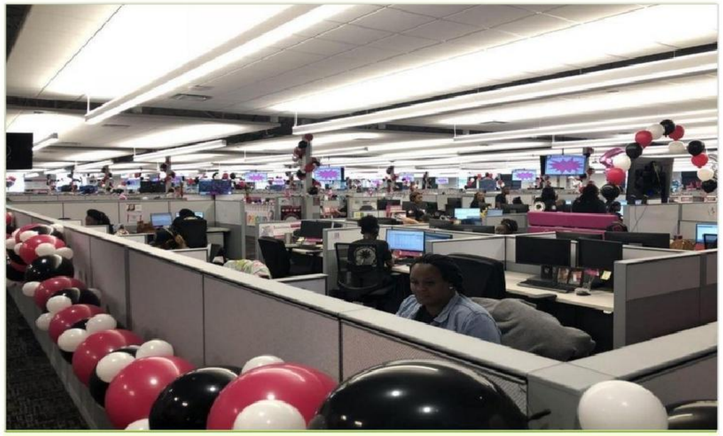
• NVIDIA AI的一部分 公司

可用性：API 目录，NGC 目录，GitHuk

• 尝试一下： API，自定节奏 训练，动手操作 实验室，90天 试验

英伟达RIVA ASR客户获胜

电信|UCaaS|医疗保健|消费者应用程序



分子分子

~3倍精度和10倍速度提高

客户支持ASR

代理协助

实时分析

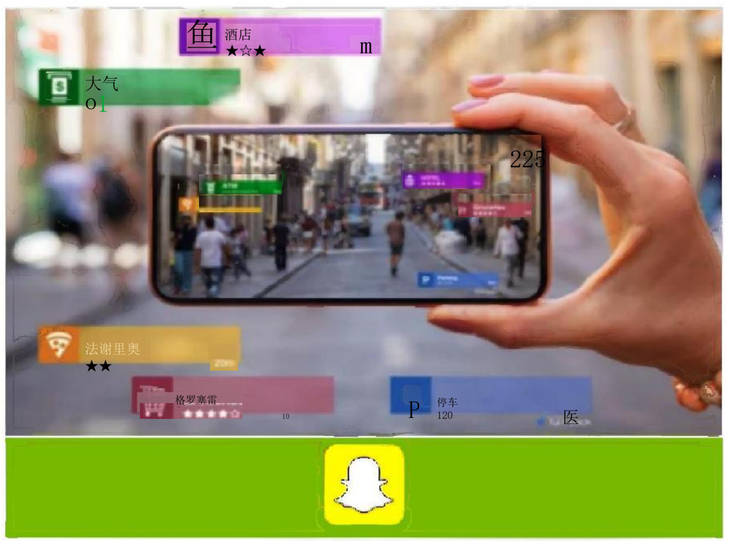
环中心

2倍更好的准确性和
无口音和噪音大的环境问题

视频通话记录的ASR

实时会议记录

总结



SOTA WER用于嘈杂的环境和表达性的合成声音

ASR和TTS为消费者应用程序提供1亿个
月活跃用户

• 集成在语音支持的开发工具中

△I²L△3sAVAyA botpress计算机中心

数据怪物

德洛伊特

浮子机器

惠普企业

InstaDeep kore.aDle兴istems .

人R

读

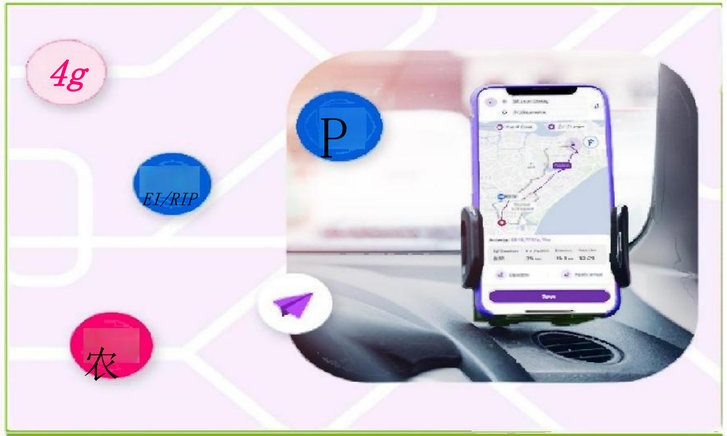
intefact

uOiiipll

英伟达

英伟达里瓦TTS客户获胜

消费者应用程序|医疗保健|UCaaS|电信



不允许的

本地英语-新加坡语音驾驶应用程序

- 发音、音调和口音正确的TTS语音
- 1000个每月活跃用户中的10秒
- 1000个并发用户

  K8可扩展云部署
- 可移植到任何场所或云上

十 西北部 荔物 人为的

减少50%的患者等待时间

- TTS用于签到亭和病人公告

  根据《美国新闻与世界报道》
- 协调1K+患者的护理
- 改善手术和患者体验

SOTA WER用于嘈杂的环境和表达性的合成声音

- ASR &TTS

  每月1亿活跃用户
- 集成在具有语音支持功能的开发工具中
- 16个定制的交流声音

botpress

plaboo

Computacenter

quantiphi

数据怪物

r复制者软冰
淇淋V

马拉莫特

英伟达

矢量风险投资

基本命令管理器要点

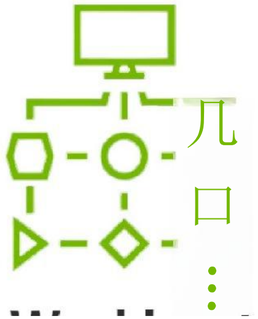
基本命令管理器要点

专为企业人工智能数据中心管理而打造

基础设施

资源调配

维护一个最安全、可靠新的，和基 的铝础设施



经营

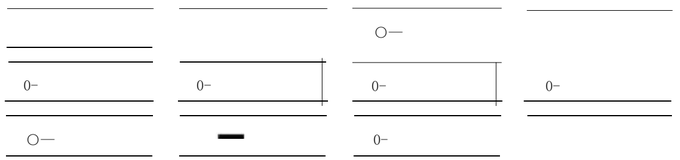
轻松地为数据科学家提供他们所需要的所有工具和资源

资源

监控

获得详细的见解决策

共享基础设施

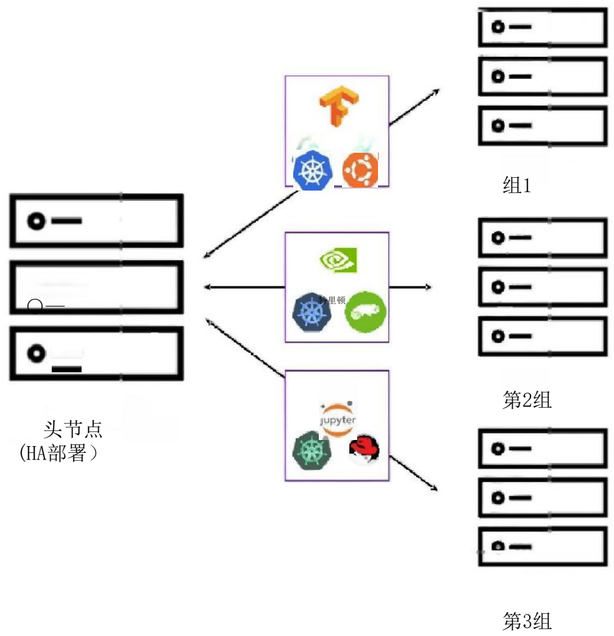


nvida。

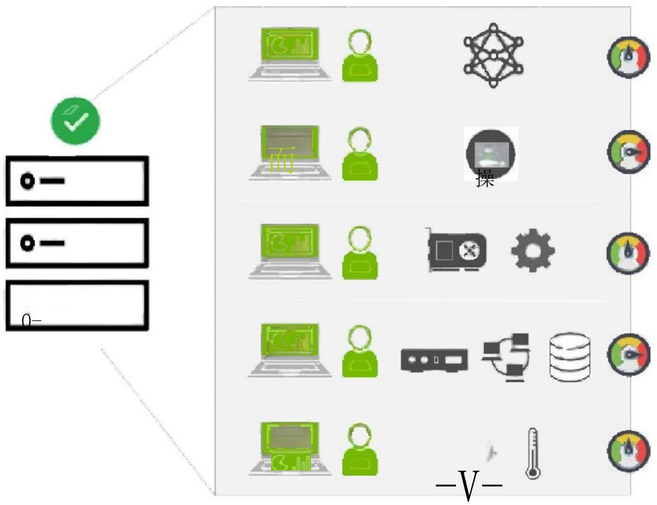
基本命令管理器要点

映像管理、利用率和自动缩放

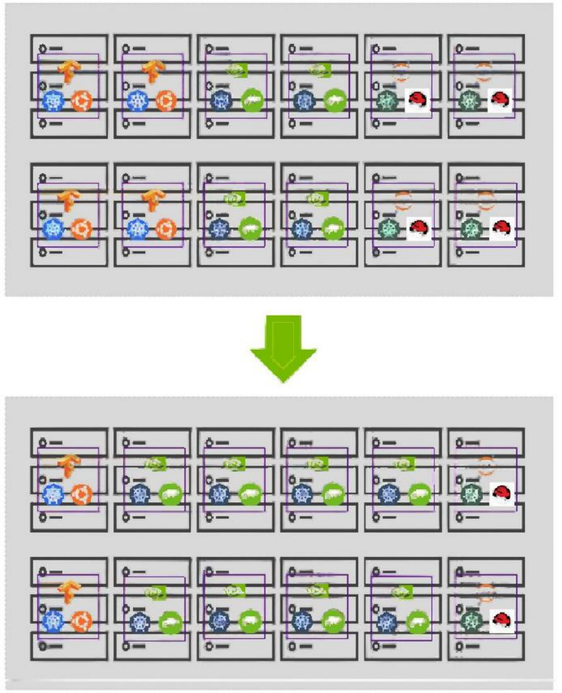
映像管理和配置



利用率指标



自动放大器

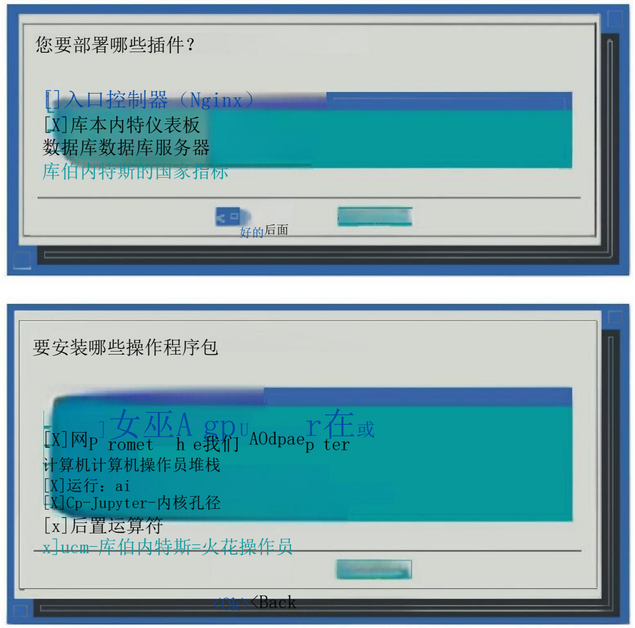


英伟达。

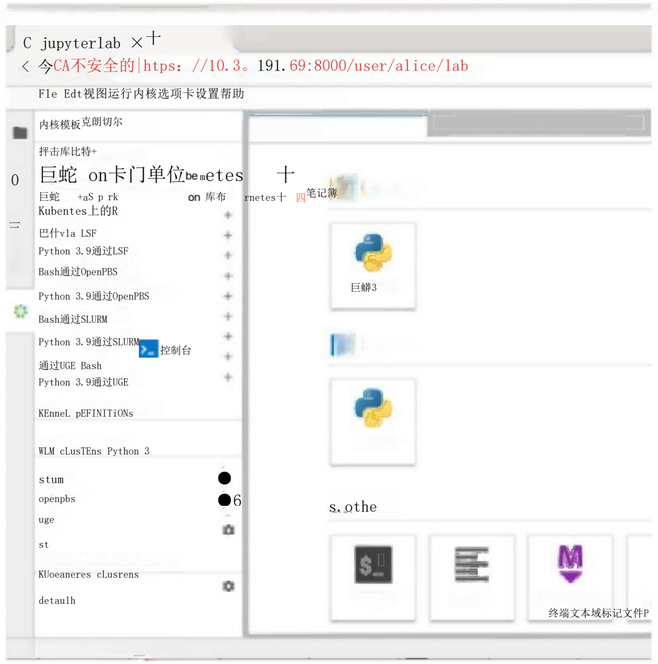
工作负载管理

基本命令管理器要点

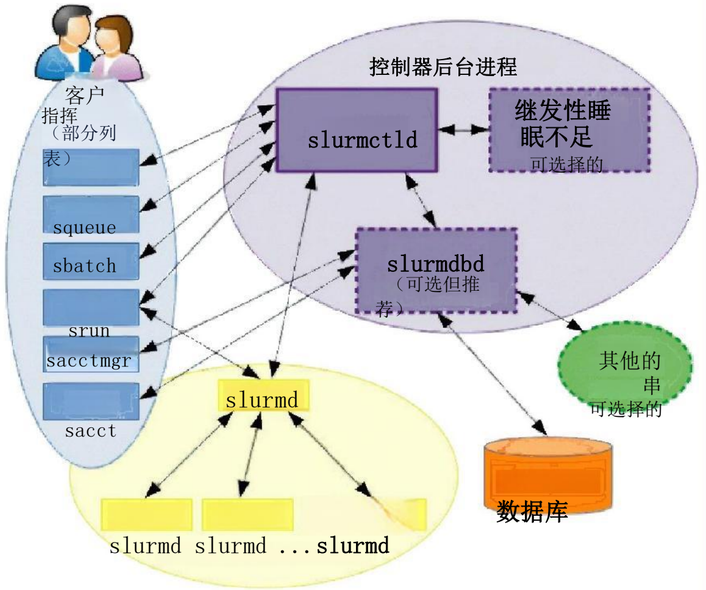
库伯内特



朱皮特



斯拉姆

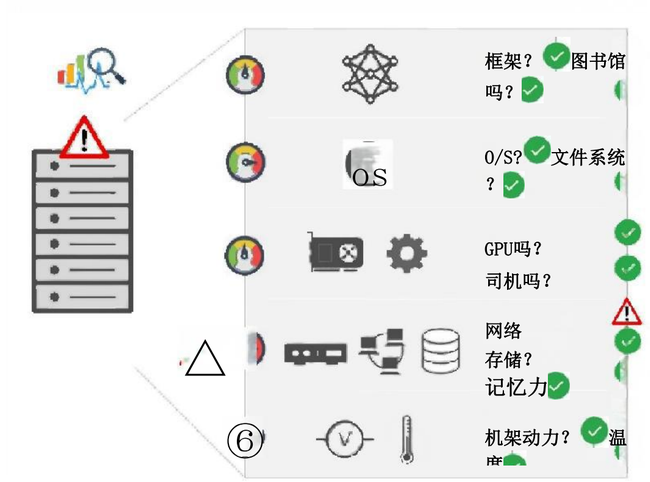


英伟达。

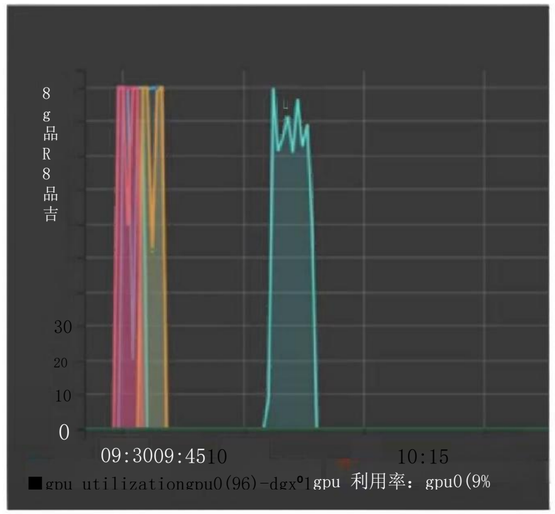
资源监控

基本命令管理器要点

健康监测和补救



GPU利用率



报告和退款

用户	Jbbs	运行时（s	CPU（$）
bob	1	1804	0.3608
charline	1	1804	0.7216
丹尼斯	2	3606	1.0818

英伟达。

云本机管理和编排

简化库伯特基础设施管理

GPU操作员

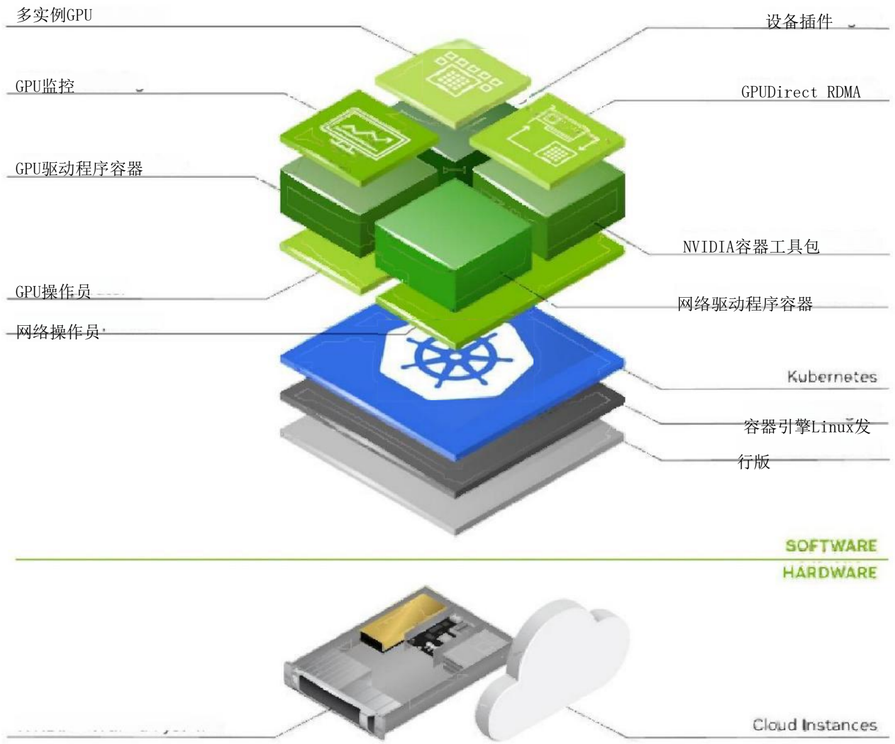
- 自动化本系统的生命周期管理
  在库伯特上公开gpu所需的软件。
- 启用高级功能，包括更好的GPU
  性能、利用率和遥测技术。

●已认证和验证的兼容性与
行业领先的库伯特解决方案

网络操作员

●促进了RDMA和GPUDirect RDMA的执行
库伯内特斯集群中的工作负载。

- NVIDIA网络驱动程序，以启用高级功能

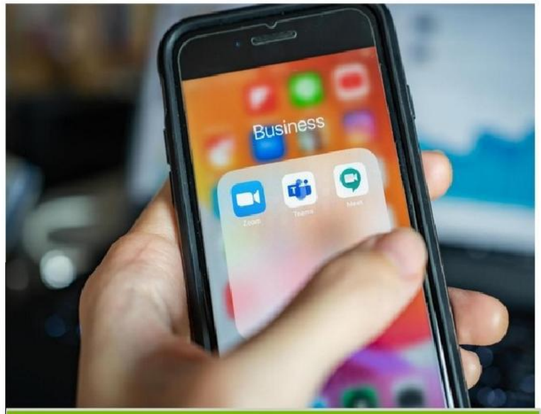


nvida。

Apache火花用例

阿帕奇火花在现代企业中广泛存在

60%的财富500强用户使用ApacheSpark3. x[1]



CONSUMER INTERNET

推荐系统

广告分析

观众细分

金融服务

欺诈检测

风险分析

投资组合管理

TELECOM

网络质量

IOT分析

过渡到5G

FEDERAL & SLED

欺诈、虐待、浪费

威胁检测

情报分析

1 https://spark.apache.org

NVIDIA急流加速器为阿帕奇的火花

改进您现有的数据处理工作流

5x

更快的执行时间，降低成本，提供全面的企业支持

更快地将数据进出数据湖
。充分利用速度更快的分析方法
加速铝管路

4x

节省云计算的使用成本，降低电力消耗和碳足迹

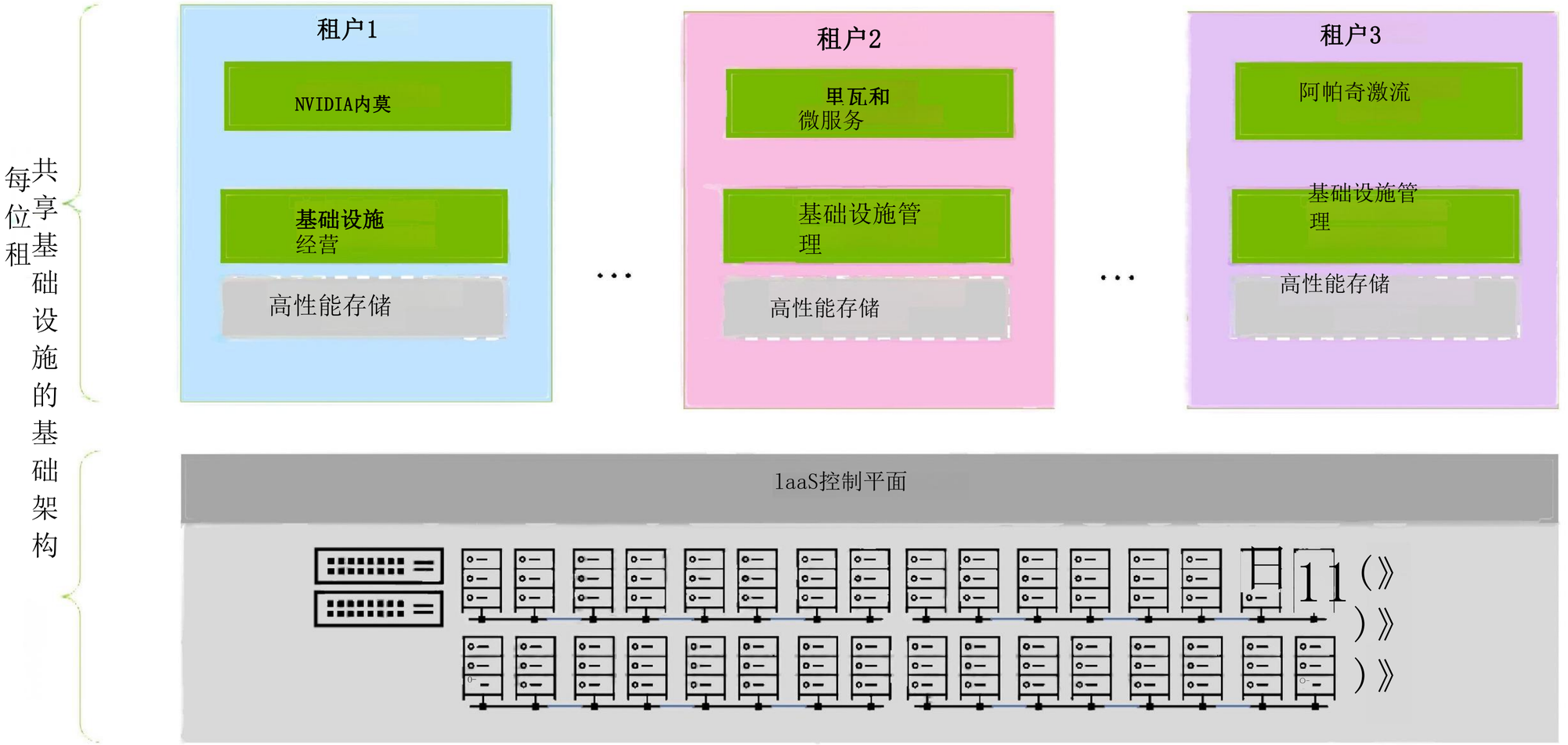
NVIDIA AI企业提供关键任务支持Bug修复

专业服务

NCP参考体系结构

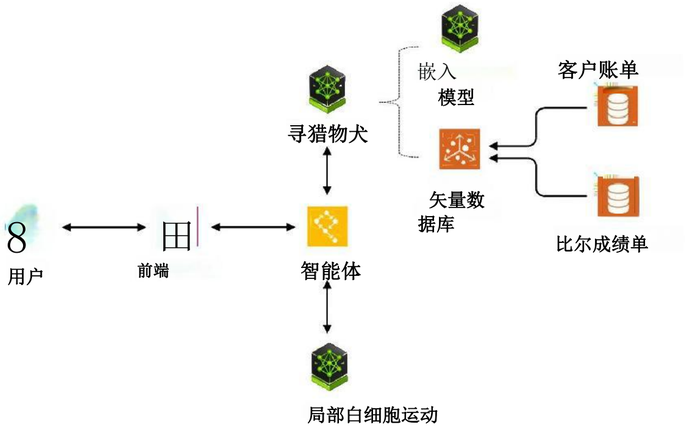
高级NCP示例服务

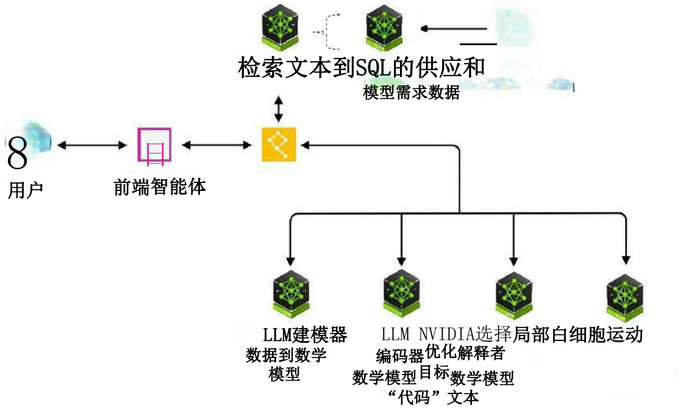
用于高性能的训练和推理

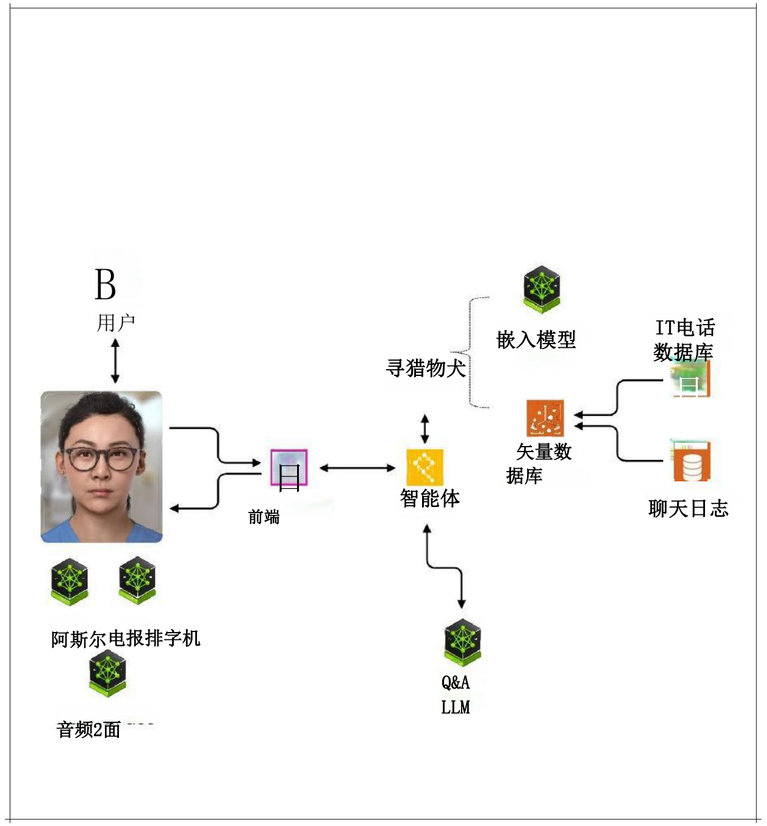


企业正在建立一个助理组织

智能应用程序可能需要许多A1模型和微服务





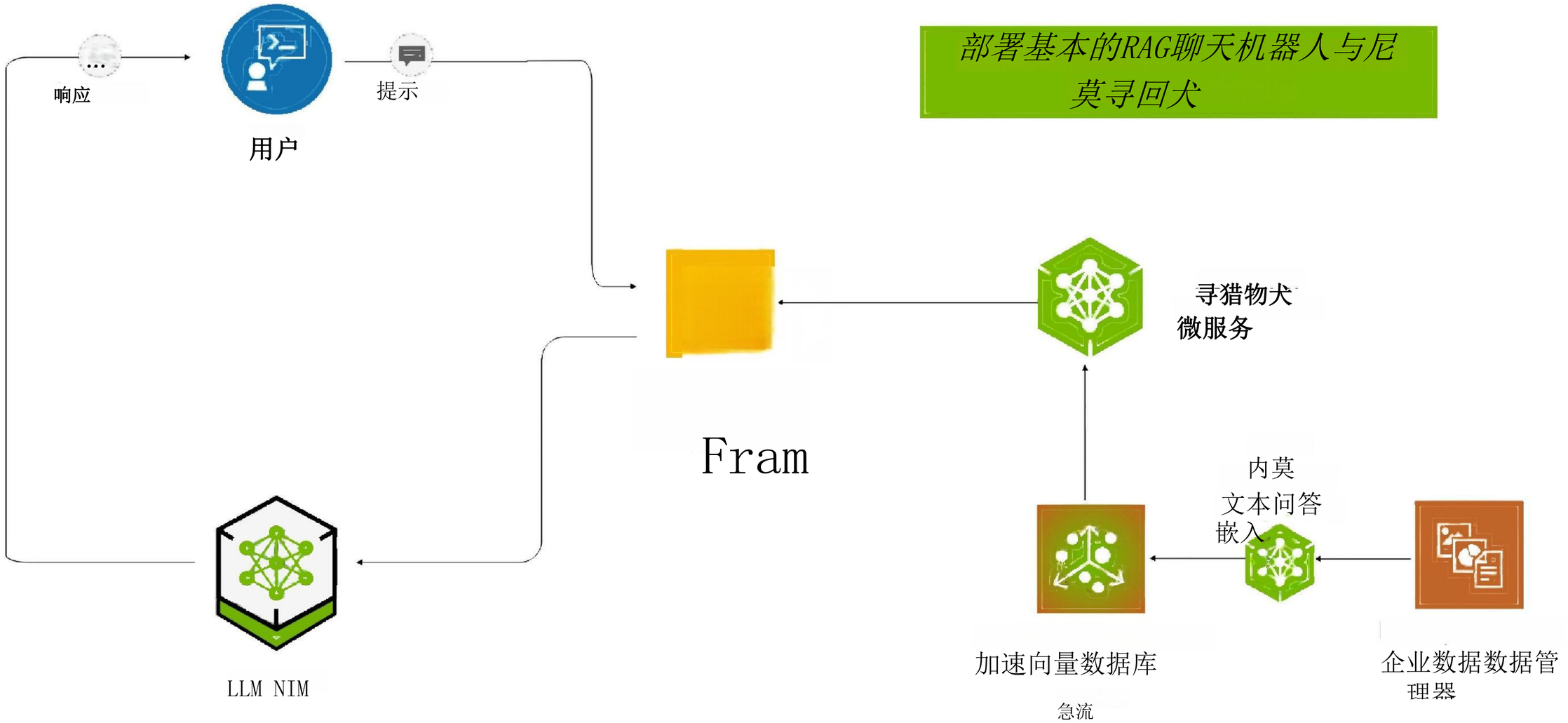


客服聊天机器人供应链副驾驶数字人类阿凡达

英伟达

人工智能虚拟助理

随着时间的推移，复杂性不断增加

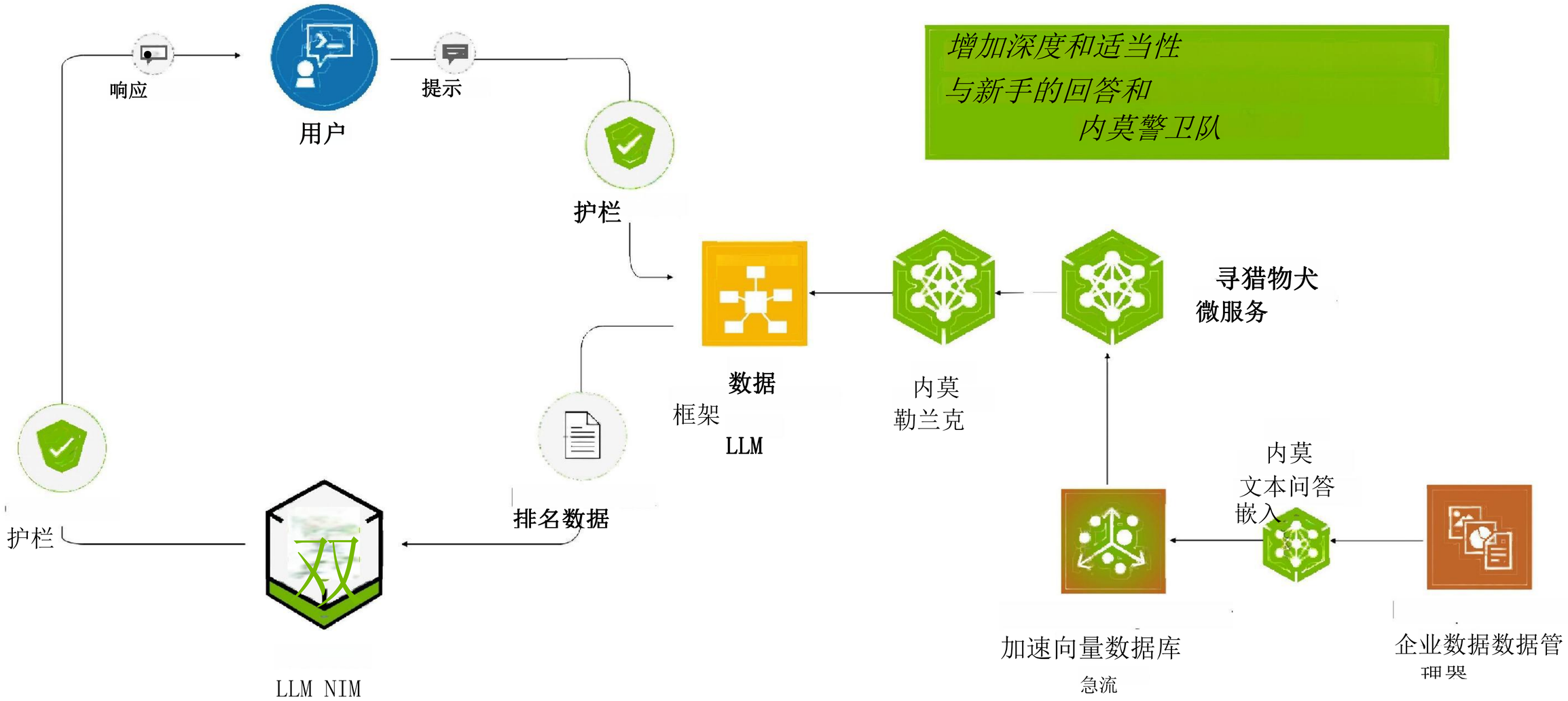


nvida。

人工智能虚拟助理

随着时间的推移，复杂性不断增加

增加深度和适当性
与新手的回答和
内莫警卫队

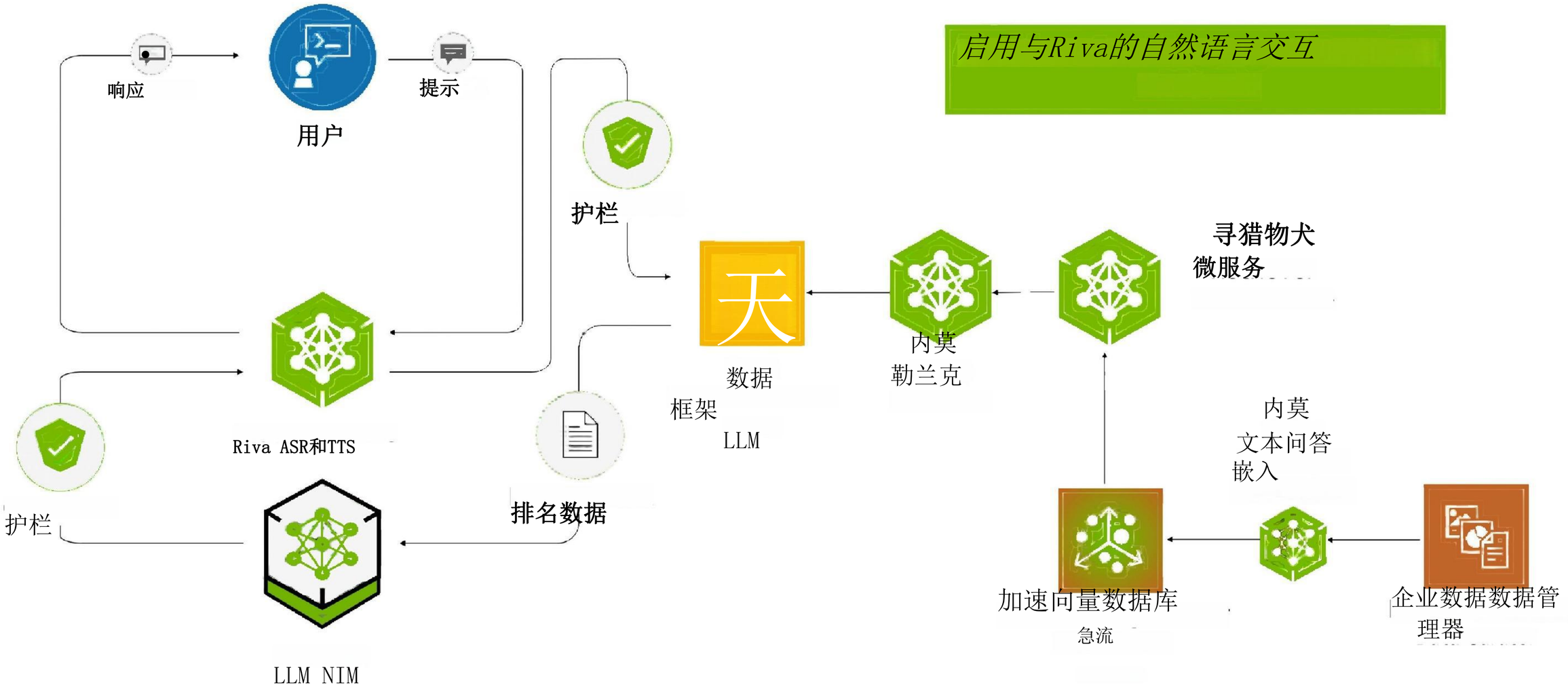


英伟达

人工智能虚拟助理

随着时间的推移，复杂性不断增加

启用与Riva的自然语言交互

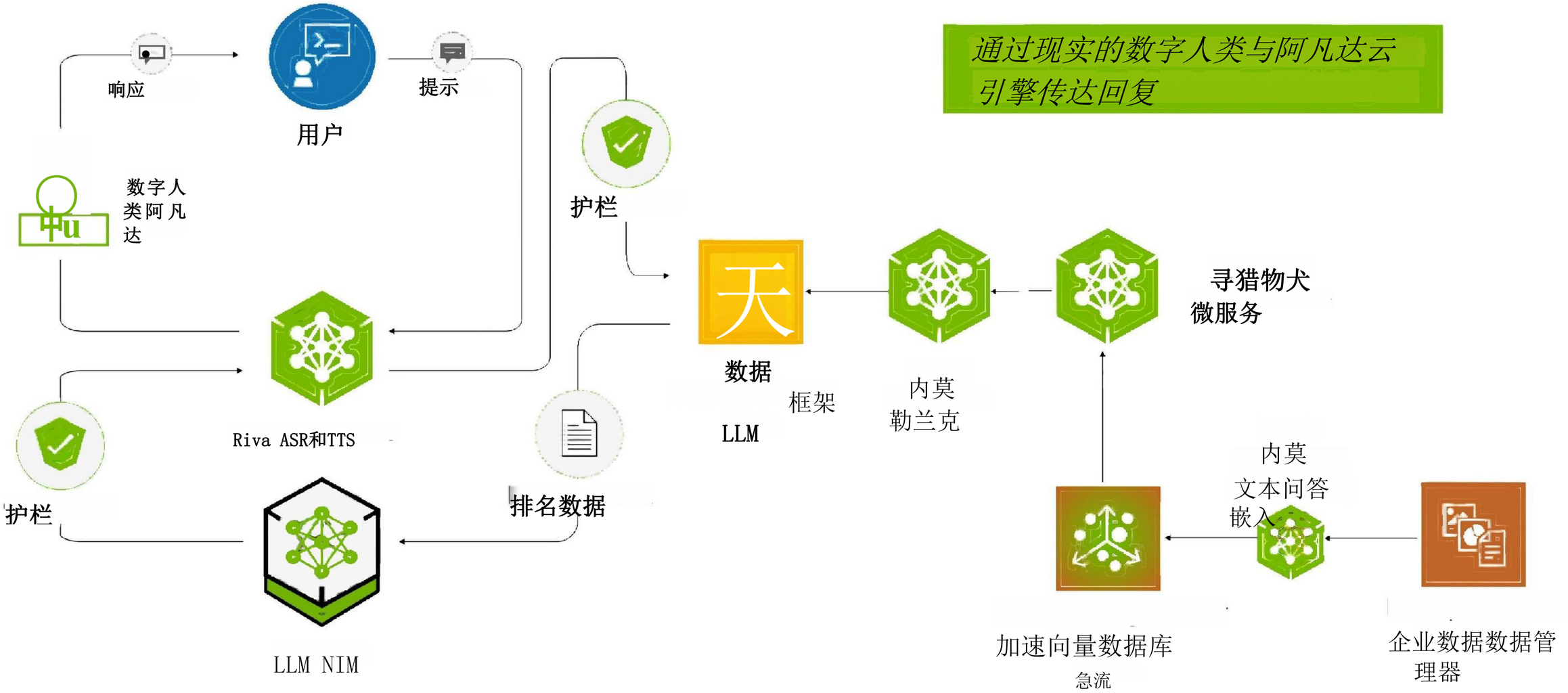


英伟达

人工智能虚拟助理

随着时间的推移，复杂性不断增加

通过现实的数字人类与阿凡达云引擎传达回复

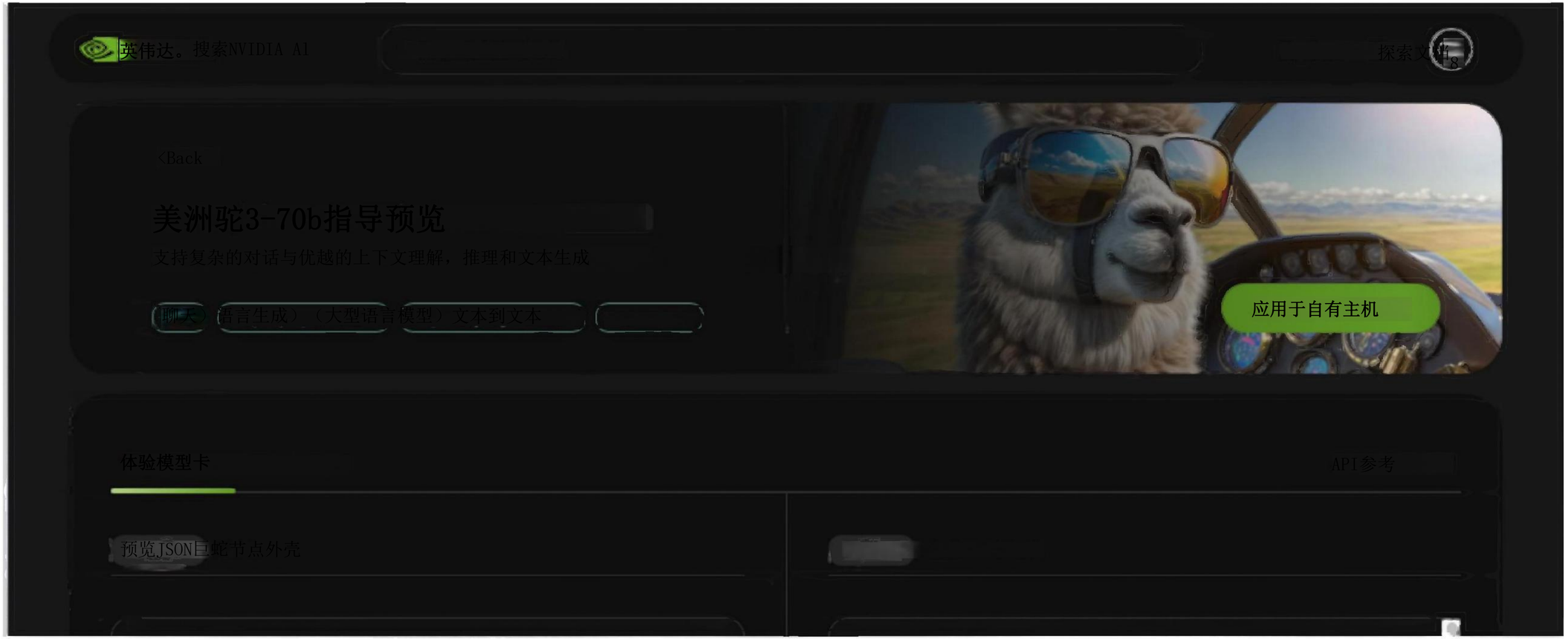


英伟达

下一步

NVIDIA API目录

免费试用NIM，然后下载并在本地运行

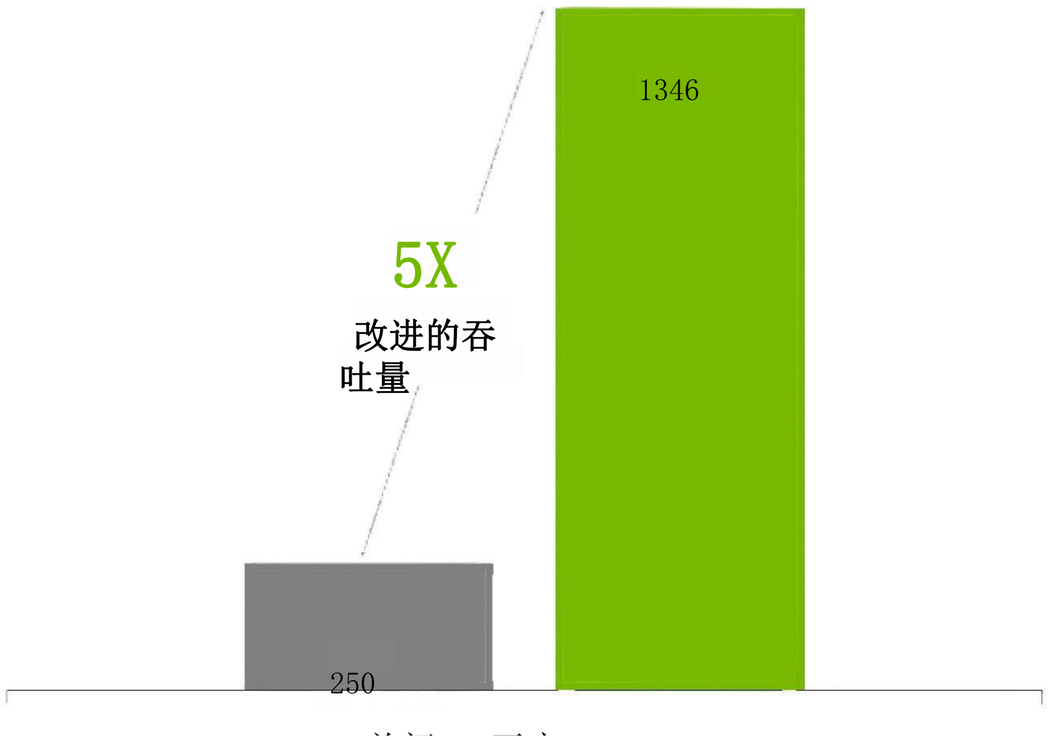


ai.nvIC1a.通信

nvida。

改进效率的效率

最先进的吞吐量降低了解决方案的总体成本



Llama 370B NIM提供了5倍更高的吞吐量

Llama370B-推理端点	裸金属H100例子	英伟达最佳化的H100例子
价格	$4/hr	$5/hr
吞吐量	1	5x
#1小时内的令牌	900,000	4,845,600
每10万代币的成本	$0.44	$0.10

每个代币的成本降低了近23%
在NVIDIA优化的H100实例上

英伟达

确保成功的资源

NVIDIA如何支持您

规定性指导、培训、市场营销

参考体系结构

来自NVIDIA合作伙伴的服务

销售培训

NVIDIA学院的技术培训

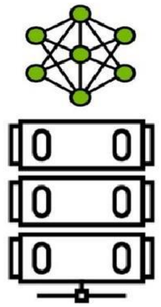
支持市场活动和需求生成

访问NVIDIA启动板

英伟达

开始工作

客户获取A1解决方案



构建

使用参考体系结构来构建集群

发展

使用参考设计

创建A1服务

装备

卖方培训和支持

POC

客户

POC

销售

客户获取

英伟达

有两种出售A1的方法

NVIDIA A1企业作为一个服务或套件

销售托管人工智能服务

针对特定用例的受管理的AI解决方案

- 只需添加数据和定制

在众多产品中进行选择
客户解决方案
需要

RAG驱动的聊天机器人，
路线优化，
数据分析、虚拟助手等...

提供帮助要部署的客户阿尔

？
口

无需开始使用部署所要求的A1解决方案的专业知识

更快的

为他们的目标用例寻求一个专家，需要超出规范的定制

销售NVIDIA AI企业

客户评估并部署他们自己的解决方案

发送客户到ai。英伟达。com

客户尝试NVIDIA人工智能企业免费

销售NVIDIA在NCP基础设施上部署的AI企业许可证

客户已准备好进行部署

客户解决方案
NIM和参考
设计完成

出售NVIDIA AI企业要在上部署的许可证
NCP基础设施

英伟达

英伟达