企业 AI

生成式 AI 和面向企业的行业专用模型

利用专用的英特尔® AI 硬件和软件优化训练和部署
帮助您实现业务转型

intel®

目录

- 为什么在生成式 AI 领域与英特尔合作
- 生成式 AI 环境
  - 什么是生成式 AI 和大型语言模型
  - 生成式 AI 目前面临哪些挑战
- 行业专用模型
  - 为什么使用面向企业的行业专用模型
  - 面向企业的行业专用模型的优势以及与英特尔合作有何帮助
- 英特尔 AI 软件和硬件概述
- 英特尔大型语言模型产品
  - 英特尔® Gaudi® AI 加速器
  - 英特尔® 至强® 可扩展处理器
  - 英特尔® 酷睿™ Ultra
- 行为召唤
- 资源

为什么要与英特尔建立合作伙伴关系？

在英特尔，我们有机会改善地球上每个人的生活，提高每家企业的业绩

但我们并非孤军作战！

我们联手合作伙伴，**让 AI 遍布每个角落**，最大限度降低部署风险，为客户创造真正的价值

您与英特尔合作，就是选择了完整的 AI 生态系统

我们拥有广泛的 AI 技术组合，并与硬件、软件和系统集成商建立合作关系，紧密合作开发真实世界的解决方案，为各行各业、企业和社区提供差异化的业务成果。

帮助您发展业务。

加入我们的旅程，让 AI 遍布每个角落

利用英特尔® AI 解决方案为客户创造价值

英特尔的方法使广泛、开放的 AI 参与者生态系统能够提供满足企业特定生成式 AI 需求的解决方案

NAVER

开发功能强大的大型语言模型 (LLM)，在全球范围内从云端到设备端部署先进的 AI 服务。NAVER 已证实英特尔® Gaudi® 在执行大规模 Transformer 模型计算运算方面的基础能力以及出色的性能功耗比。

seekr

值得信赖的 AI 领导者，在英特尔® Tiber™ AI 云中，基于英特尔® Gaudi® 2、英特尔® Data Center GPU Max Series 和英特尔® 至强® 处理器运行生产工作负载，为 LLM 开发和生产部署提供支持。

BOSCH

探索智能制造的更多机会，包括生成制造异常情况合成数据集的基础模型，以提供稳健、均匀分布的训练集（例如自动光学检测）。

IFF

食品、饮料、香水和生物科学领域的全球领导者，将利用生成式 AI 和数字孪生技术建立集成的数字生物学工作流程，用于高级酶设计和发酵过程优化。

IBM

将第五代英特尔® 至强® 处理器用于其 watsonx.data™
数据存储，并与英特尔® 密切合作，验证用于英特尔® Gaudi® 加速器的 watsonx™ 平台。

airtel

借助英特尔尖端技术的强大功能，Airtel 计划利用丰富的电信数据增强其 AI 能力并提升客户体验。部署将符合 Airtel 始终站在技术创新前沿的承诺，并帮助在快速变化的数字化环境中开辟新的收入来源。

OLA

对印度首个具有 10 种语言生成功能的基础模型进行预训练和调优，性价比领先于市场上的解决方案。Krutrim 目前正在英特尔® Gaudi® 2 集群上预训练一个大型基础
模型。

Infosys

下一代数字服务和咨询行业的全球领导者宣布了一项战略合作，将包括第四代和第五代英特尔® 至强® 处理器、英特尔® Gaudi® 2 AI 加速器以及英特尔® 酷睿™ Ultra
处理器在内的英特尔® 技术引入 Infosys Topaz — 一套 AI 优先的服务、解决方案和平台，使用生成式 AI 技术加速实现业务价值。

企业 AI 价值主张

利用企业 AI 实现业务转型

在当今竞争异常激烈的环境中，**拥抱 AI 的企业正在迈步向前。**

各个行业的企业都在重新构想运营的各个方面，以了解 AI 如何增强或自动执行工作流程。

将 AI 嵌入企业结构是英特尔独有的专业知识。

从能够转变生产力的 AI 电脑，到多年积累的深刻理解哪些用例回报价值最大的专业知识，英特尔® 是您值得信赖的合作伙伴，可以安全、可靠地让 AI 遍布每个角落。

生成式 AI (GenAI) 创新预计将被各种规模的企业采用，采用速度将超越互联网时代、移动时代或云时代。

下一波 AI 平台将以经济实惠且灵活的方式迎接备受期待的现实。

采用全新视角来看待您的企业 AI。

本支持包将帮助您了解各个市场中的企业如何从生成式 AI（特别是行业专用模型）中获得巨大价值，从而取得长期成功

什么是生成式 AI 和大型语言模型？

生成式 AI (GenAI) 是 AI 的一部分，侧重于创建全新的原创内容。

它涉及训练和部署 AI 模型，以生成与训练数据集中的样本非常相似的图像、文本或音频等数据。

生成式 AI 算法使用深度学习和神经网络等先进技术来生成逼真、连贯的输出，从而实现图像合成、文本生成甚至创意艺术品等应用。

大型语言模型 (LLM) 是一种特定类型的自然语言处理模型，使用深度神经网络来处理和生成文本。LLM 以海量文本数据为基础进行训练，旨在生成连贯且有意义的输出。

了解详情

阅读更多内容

把握生成式 AI 的强大功能

企业将如何使用生成式 AI?

消费品和零售

- 虚拟试衣间
- 交付和安装
- 店内产品查找帮助
- 需求预测和库存规划
- 新颖的产品设计

医疗保健和医药

- 协助忙碌的一线员工
- 转录和总结医疗记录
- 回答医学问题的聊天机器人
- 为诊断和治疗提供信息的预测分析

制造

- 技术人员的专家助手
- 与机器的对话式交互
- 规范且主动的现场服务
- 自然语言故障排除
- 保修状态和文件
- 了解流程瓶颈，制定恢复策略

媒体与娱乐

- 智能搜索，量身定制的内容发现
- 标题和文案开发
- 内容质量实时反馈
- 个性化播放列表、新闻摘要、推荐
- 互动式观众选择叙事
- 定向优惠、订阅计划

金融服务

- 发现交易信号，提醒交易员注意弱势头寸
- 加速承保决策
- 优化和重建旧系统
- 对银行和保险模型进行逆向工程
- 监控潜在的金融犯罪和欺诈
- 自动收集数据以符合监管要求
- 从公司披露信息中提取洞察

来源：由 MIT Technology Review Insights 编制，所依据的数据来自“生成式 AI 时代的零售业”9、“大解锁：制造业中的大型语言模型”10、“生成式 AI 无处不在，无所不包”和“媒体和娱乐领域中的大型语言模型”12，Databricks，2023 年 4-6 月。

生成式 AI 和大型语言模型用例

聊天机器人和
虚拟助手

客户支持

代码生成和
调试 LLM

基于公司文件进行训练

情绪分析

评估客户满意度

文本分类和聚类

对大量数据进行分类
以识别趋势

语言译文

将公司网页转换为
其他语言

总结和转述

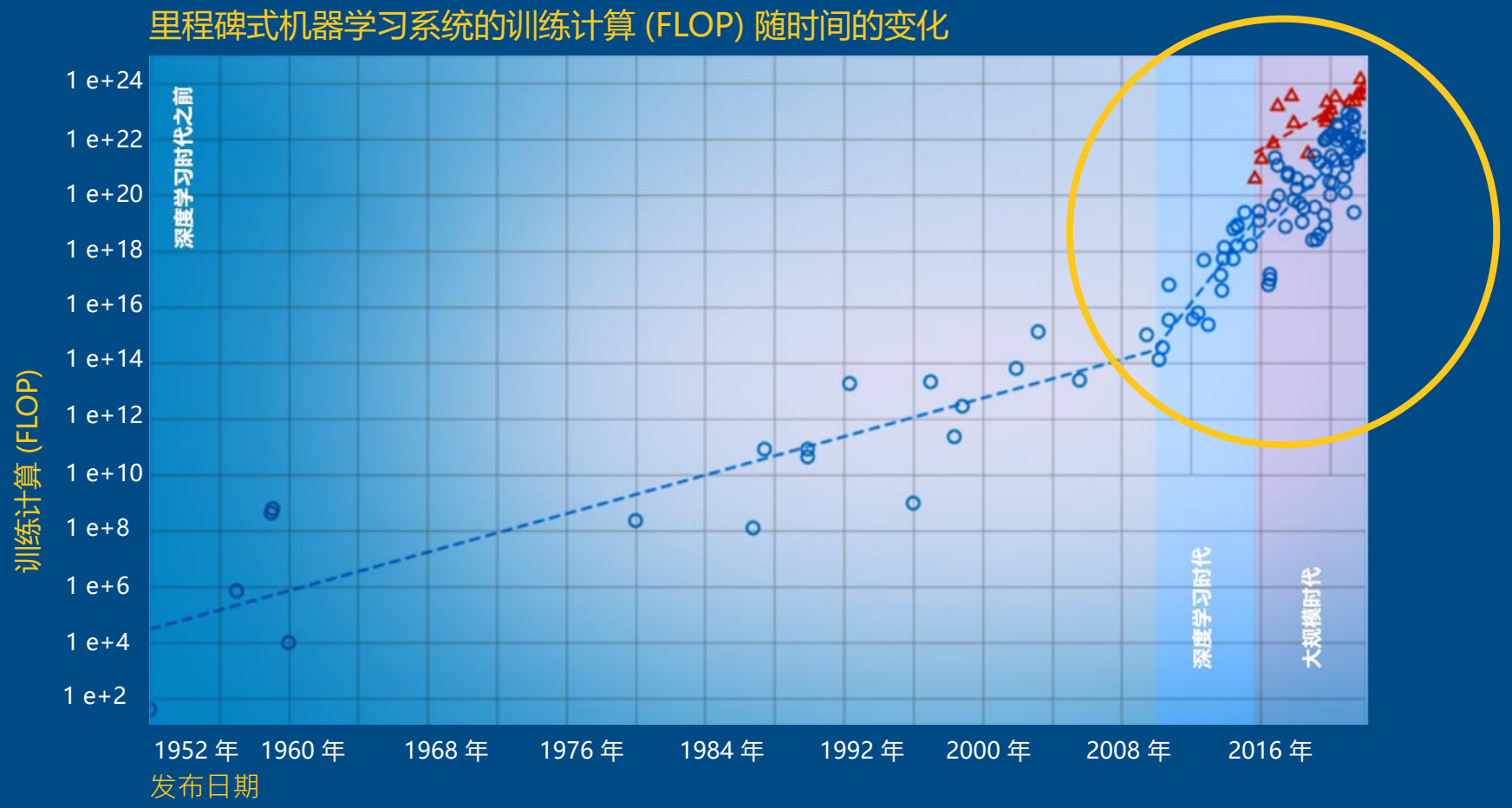
会议记录摘要

内容、图像、视频生成

电子邮件初稿、
创意生成、营销视觉
效果、短视频

来源：大型语言模型简明指南，Databricks，2023 年。

随着模型规模的增长，计算也在增长



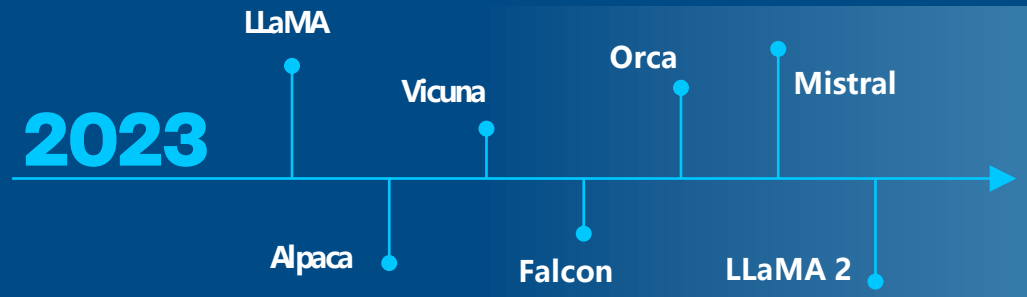
由 Epoch、阿伯丁大学、AI 治理中心、圣安德鲁斯大学、麻省理工学院、图宾根大学、马德里康普顿斯大学开展的研究

不仅仅是巨型模型

	巨型（第三方）	**对比**	小巧灵活（10-100 倍）
可解释性	专有模型	对比	**基于开源的模型**
准确度	**全能通用**	对比	定向、行业专用、定制
位置	基于云（即服务）	对比	本地运行推理；边缘、客户端和本地
成本	永久存在扩展成本	对比	成本管理
面市速度	设置快速（数秒）	对比	构建时间短（数小时/数天）

多种小型模型的发展

6 个月内从 1000 亿参数到少于 200 亿参数



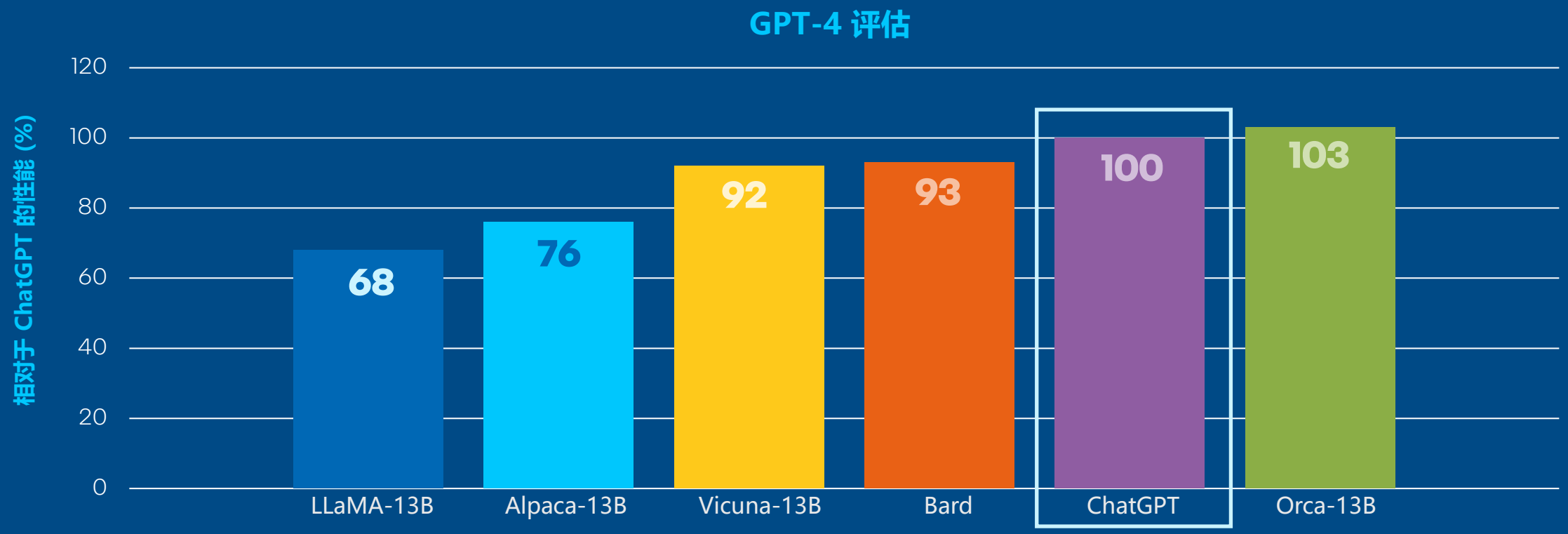
databricks

- 每周出现数十种小型模型
- 商业和开源许可证
- 事实表明，如果使用精心收集的数据进行训练，小型模型也可以复现大型模型的准确性

- 数千个行业专用商业模型和 AI 平台正在展示
- 可以在少量处理器上使用行业专用的数据调优模型

与 ChatGPT 相比，小型模型表现良好

证明小型模型是一个可行的选项，即使与 ChatGPT 等大型模型相比，仍然表现良好



根据 Vicuna 评估集的 GPT-4 评估，Orca 的表现优于包括 OpenAI ChatGPT 在内的多种基础模型

来源：Microsoft Research (2023)。Orca：从 GPT-4 的复杂解释轨迹中渐进学习

构建行业专用模型

1

第三方平台

API

大型基础模型

企业
数据库

企业矢量
数据库

调优

检索增强生成

2

企业平台

通用开放模型

企业
数据库

企业矢量
数据库

调优

检索增强生成

行业专用模型

金融聊天机器人

个性化营销内容

代码生成器

临床笔记转录

文本转语音生成器

行业专用模型可以为企业带来许多优势

小型定向模型可以提供同等或更优异的性能，通过减少时间和成本投资来提高投资回报率

更准确的输出

使用企业数据获得更高的行业专用准确性

降低成本

调优预训练模型，并/或使用 RAG，推理小型模型

在所选平台上的任何地方部署

本地运行推理；边缘、客户端和本地

安全且私密

满足数据安全和监管要求

可靠的 AI

使模型能够通过调优和 RAG 引用数据源

未来

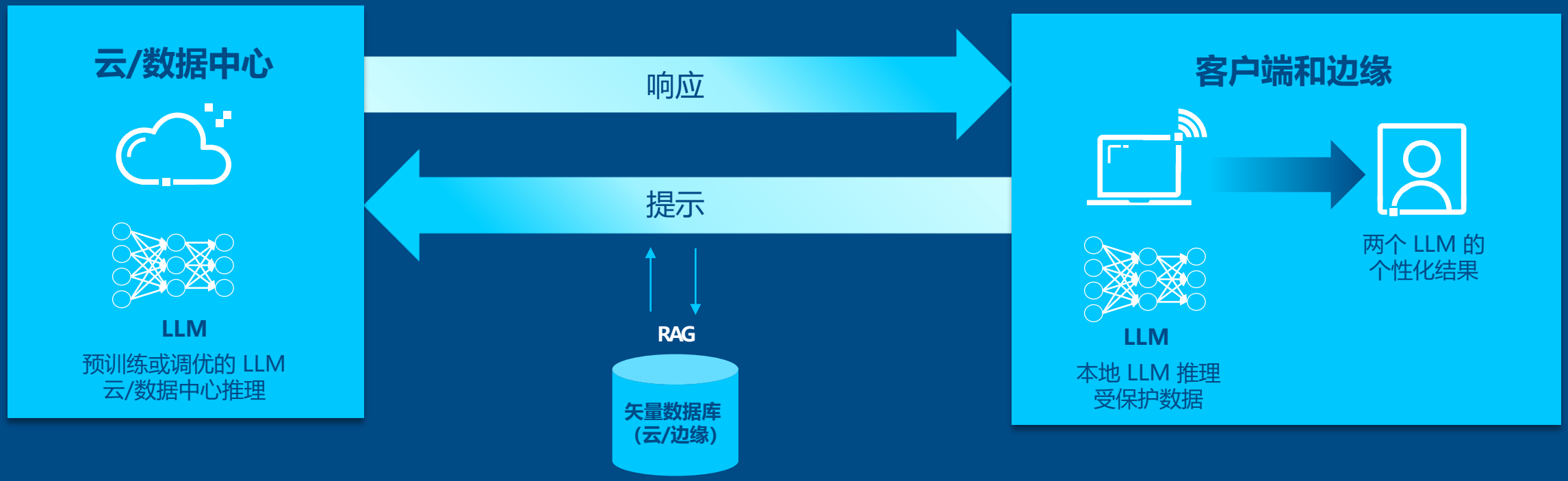
将有少量巨型模型和大量更灵活的小型 AI 模型嵌入到无数应用中[1]

[1]来源：适者生存：紧凑型生成式 AI 模型是大规模经济高效的 AI 的未来

无缝的云到边缘 AI 平台

英特尔 AI 产品组合

在云端训练和推理。使用 RAG 提高行业准确性。



生成式 AI — 生产年份

行业专用且高度智能的模型的使用正在增长

2022	2023	2024
实验	试点	生产

巨型模型铺平了道路

- 对于一般用途非常有效
- 训练和部署成本高昂
- 基于大型公共数据集构建
- 易于使用

小型、行业专用的模型

- 使用私有数据，获得业务特定的结果
- 部署在现有硬件上
- 提高效率、准确性、安全性和可追溯性
- 构建时间短

阅读博客

适者生存：紧凑型生成式 AI 模型是大规模经济高效的 AI 的未来

英特尔的行业专用模型方法

行业专用模型

优势

+ 在保持/提高准确性的同时，模型缩小 10 到 100 倍
+ 经济实惠地进行通用计算
+ 正确性；来源归因；可解释性
+ 利用私有/企业数据
+ 信息持续更新

挑战

- 任务范围缩小
- 只需少量调优和索引

英特尔目标

使用行业框架、预训练模型以及英特尔 AI 软件和工具，以最具成本效益且普遍存在的方法在英特尔硬件上调优和部署数万种模型

阅读更多内容

触手可及的生成式 AI

电子书 • 信息图表

企业 AI：帮助克服进入壁垒

要求	与英特尔® 合作有何帮助
面市速度	利用英特尔和 Hugging Face 的开发人员资源、Gaudi Developer Hub 以及 5 个参考套件，在生成式 AI 领域抢占先机
用户体验（准确性/延迟）	在英特尔® Gaudi® 加速器上对参数超过 100 亿的模型进行推理，在内置英特尔® AMX 的英特尔® 至强® 处理器上对参数少于 200 亿的小型模型进行推理，为用户提供实时体验[1]
计算可获得性	在全球 GPU 短缺的情况下，英特尔® 至强® CPU + 加速器是一种经济高效的替代方案。英特尔® Gaudi® 2 现已通过 SuperMicro 提供，英特尔® Gaudi® 3 将具有更广泛的可获得性。
熟悉的技术	小型模型的推理几乎可以在任何硬件上完成，包括可能已经成为您的计算环境一部分的普适解决方案[2]
大规模运营	英特尔® Gaudi® 2 具有近乎线性的可扩展性，每个加速器上集成了 24 个 100 GbE 端口。英特尔® 至强® 处理器已经存在于您的数据中心、现场以及云到边缘。65% 的数据中心推理在英特尔® 至强® 处理器上运行[3]
经济高效	在实际工作应用中，英特尔® 通过提供更好的性能、更低的价格和更均衡的 AI 推理平台，颠覆了这一行业并使 AI 得以普及。请参阅 NVIDIA 数据显示，英特尔® Gaudi 2 的每美元性能比其 H100 高 4 倍

[1]来源：实施生成式 AI 的四个障碍

[2]来源：适者生存：紧凑型生成式 AI 模型是大规模经济高效的 AI 的未来

[3]基于截至 2022 年 12 月运行 AI 推理工作负载的全球数据中心服务器用户的英特尔市场建模。

用于简化生成式 AI 训练和部署的软件资源

开源模型

- BLOOM（a BigScience initiative；176B params 59 languages Open-access）— 176B
- BioGPT — Domain 1.5B
- Stable Diffusion — 图片
- Llama2 GPT-J MPT Falcon — 7-65B LLM
- Stanford Alpaca — 调优的 7B LLM
- 知识 库

开放软件

- OpenVINO
- 英特尔® Extension for PyTorch (IPEX)（PyTorch）
- 英特尔® Extension for Transformers (ITREX)
- 英特尔® Extension for DeepSpeed (IDEX)（DeepSpeed）
- haystack
- fastRAG

生成式 AI 平台

- intel. GAUDI
- intel. DATA CENTER GPU MAX SERIES
- intel. XEON
- intel. CORE ULTRA

阅读更多内容

通过普遍存在的硬件和开放软件释放生成式 AI 的潜力

实现价值最大化

为什么英特尔的开放 AI 方法适合您的 AI 业务需求

避免受限于厂商
基于标准的开源软件

MODIN
SciPy
dmlc XGBoost
PyTorch
pandas
NumPy
scikit learn
TensorFlow
Spark
Numba
SigOpt AutoML
oneAPI

利用英特尔的硬件产品组合
面向 AI 用例进行优化

intel XEON
intel CORE ULTRA
intel ARC GRAPHICS
intel DATA CENTER GPU
intel GAUDI
intel ETHERNET

利用面向未来 AI 的软件和开放标准优化的硬件，
从客户端和边缘到数据中心和云端，创造新的机会

英特尔® AI 软件目录

工程师数据 | 创建模型 | 优化和部署

MODIN · SciPy · dmlc XGBoost · PyTorch · ONNX RUNTIME · OpenVINO

pandas · NumPy · scikit learn · TensorFlow · DirectML

Write Once Deploy Anywhere

Spark · Numba · SigOpt AutoML · Intel Compressor

大规模数据分析†

机器学习和深度学习框架、优化和部署工具†

oneAPI

Intel® oneAPI Deep Neural Network Library

Intel® oneAPI Collective Communications Library

Intel® oneAPI Math Kernel Library

Intel®oneAPI Data Analytics Library

适用于 CPU、GPU 和其他加速器的开放式跨架构编程模型

云与企业

客户端和工作站

边缘

intel AI ANALYTICS TOOLKIT

加速端到端数据科学和 AI

英特尔® Tiber™ AI 云 和英特尔® Developer Catalog

试用最新的英特尔工具和硬件，访问优化的 AI 模型

cnvrg.io

全栈机器学习操作系统

英特尔® Geti

注释/训练/优化平台

Hugging Face

英特尔优化和微调配方、优化推理模型和模型服务

注：根据预期的 AI 使用模型，针对其他层的目标组件，优化堆栈每一层的组件，最右边一列的解决方案并非使用了每个组件

† 此列表包括针对英特尔硬件优化的流行开源框架

简化企业生成式 AI 的采用，并缩短强化、可信解决方案的生产时间

OPEA：

简化企业生成式 AI 的采用，并缩短强化、可信解决方案的生产时间

Open Platform for Enterprise AI

OPEA 合作伙伴

anyscale

CLOUDERA

DATASTAX

kx

DOMINO

Hugging Face

Red Hat

drant

MariaDB

MINIO

sas

Yellowbrick

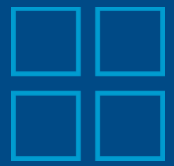
vmware

OPEA 价值

- 帮助企业利用生成式 AI（LLM、RAG）更快、更轻松地释放数据价值
- 降低分散生态系统的复杂性，并帮助解决方案在生产环境下扩展
- 与 Linux 基金会合作，激发行业领导者之间的协作和贡献

高效
利用现有基础设施、AI 加速器或您选择的其他硬件。

无缝
与企业软件集成，具有跨系统和网络的异构支持和稳定性。



开放
汇集同类最佳创新成果，并且不局限于专有供应商。

随时随地
通过为云、数据中心、边缘和 PC 构建的灵活架构在任何位置运行。

可信
具有安全的企业就绪管道和工具，可实现可靠性、透明度和可追溯性。

可扩展
接触充满活力的合作伙伴生态系统，以帮助构建和扩展您的解决方案。

与 Hugging Face 建立生成式 AI 合作关系

Hugging Face

为了促进生成式 AI 和语言 AI 的训练和创新，英特尔与 Hugging Face 这一热门的 AI 模型和数据集共享平台展开合作。最值得一提的是，Hugging Face 因其为 NLP 构建的 Transformer 库而闻名。

intel XEON

英特尔® 与 Hugging Face 合作构建了最先进的硬件和软件加速，以使用 Transformer 模型进行训练、调优和预测。

硬件加速由英特尔® 至强® 可扩展处理器驱动，而软件加速由我们优化的 AI 软件工具、框架和库组合提供支持。

intel GAUDI

英特尔® Gaudi® 深度学习加速器还通过 Optimum Habana 库与 Hugging Face 开源软件配合使用，使开发人员能够轻松
使用由 Hugging Face 社区优化的数千个模型。

Hugging Face 还发布了英特尔® Gaudi® 2 在以下生成式 AI
模型上的性能评估：Stable Diffusion、T5-3B、BLOOMZ 176B 和 7B 以及新的 BridgeTower 模型。

英特尔®、Articul8 和 BCG 合作提供安全的企业级生成式 AI

由英特尔 AI 超级计算机提供支持的开创性解决方案通过定制数据集释放商业价值，同时保持高水平的安全性和数据隐私

Articul8* 提供一站式生成式 AI 软件平台，该平台速度快、安全性高且成本效益高，可帮助大型企业客户实现 AI 的运营化和规模化。该平台在英特尔®硬件架构（包括英特尔®至强®可扩展处理器和英特尔®Gaudi®加速器）上推出并经过优化，但也将支持一系列混合基础设施替代方案。

intel GAUDI　intel XEON

继 Boston Consulting Group (BCG) 早期部署该技术后，团队已将该平台扩展到需要高水平安全性和专业领域知识的行业领域（包括金融服务、航空航天、半导体和电信）的企业客户。

阅读更多内容

Articul8 公告

Articul8 网站

面向企业的可靠 AI

挑战：

生成式 AI 模型是通过互联网上的大量数据学习而得来，其中可能包含社会中存在的偏见，并可能无意中应用这些偏见。LLM 可能被操纵来生成或传播错误信息、网络钓鱼电子邮件或社会工程攻击。

LLM 可能经常出现“幻觉”，并生成不准确的信息，这在医疗保健等行业尤其成问题，因为模型可能影响诊断和治疗决策，并可能对患者造成伤害。

了解详情

最大限度地降低生成式 AI 的风险

解决方案：

从事 AI 技术的公司和个人需要确保他们的软件开发和部署符合 AI 道德伦理

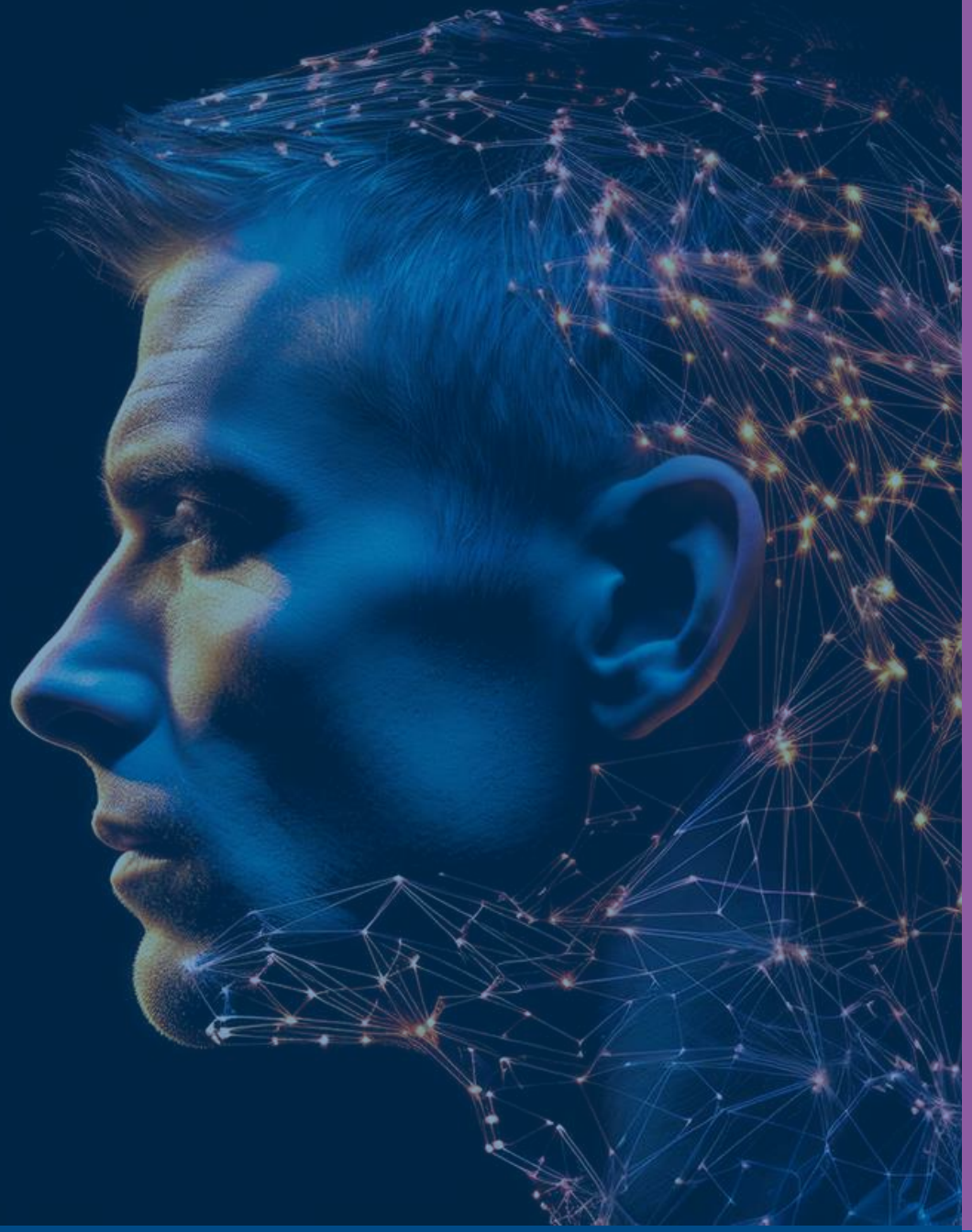
开源的英特尔® Explainable AI Tools 允许用户运行事后模型提炼和可视化，以检查 TensorFlow* 和模型的预测行为

LLM 通常在大型公共数据集上训练，然后针对潜在敏感数据（例如财务和医疗数据）进行调优

英特尔 Open Federated Learning (OpenFL) 之类的技术引入了机密计算，使 LLM 可以安全地针对敏感数据进行调优，从而提高模型的普适性，同时减少“幻觉”和偏见

英特尔® 生成式 AI 产品

让 AI 遍布
每个角落



面向 AI 的可扩展系统

训练和调优

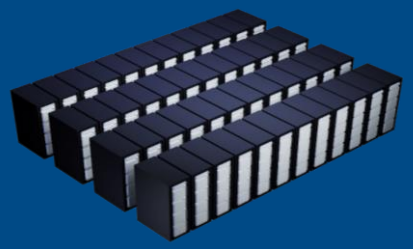
训练 | 峰值推理 | 主流推理/调优 | 基线推理 | 端点推理

推理和部署

AI
数据中心

边缘

客户端

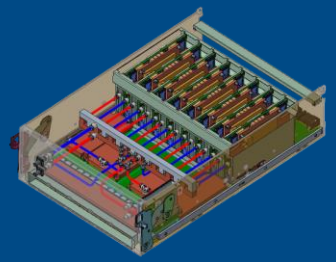


集群和数据中心扩展

intel. GAUDI

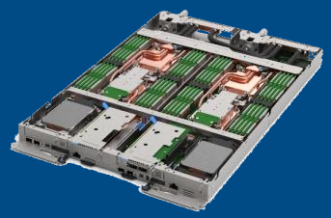
单机架多节点部署

intel. GAUDI intel. XEON MAX SERIES intel. XEON



多 GPU
或多插槽 CPU

intel. XEON intel. XEON MAX SERIES intel. DATA CENTER GPU FLEX SERIES



单插槽 CPU 或
单个 GPU

intel. XEON intel. DATA CENTER GPU FLEX SERIES intel. CORE ULTRA

客户端 CPU

intel. CORE ULTRA intel. ARC GRAPHICS

intel. ETHERNET

英特尔® NLP/LLM 产品

训练推理

GAUDI®2

英特尔® Gaudi® 2 AI 加速器专为加速 LLM 和 NLP 等大规模模型的训练和推理而设计。

使用英特尔® Gaudi® 2 加速生成式 AI 和大型语言模型

英特尔® Gaudi® 2 为 AI 训练提供领先的性能和最优的成本节省

新闻稿

英特尔® Gaudi® 2 深度学习加速器
在
深度学习训练和推理方面表现优异，
性能比 NVIDIA A100 提升高达 2.4 倍

新闻发布室 • 技术文章

立即观看

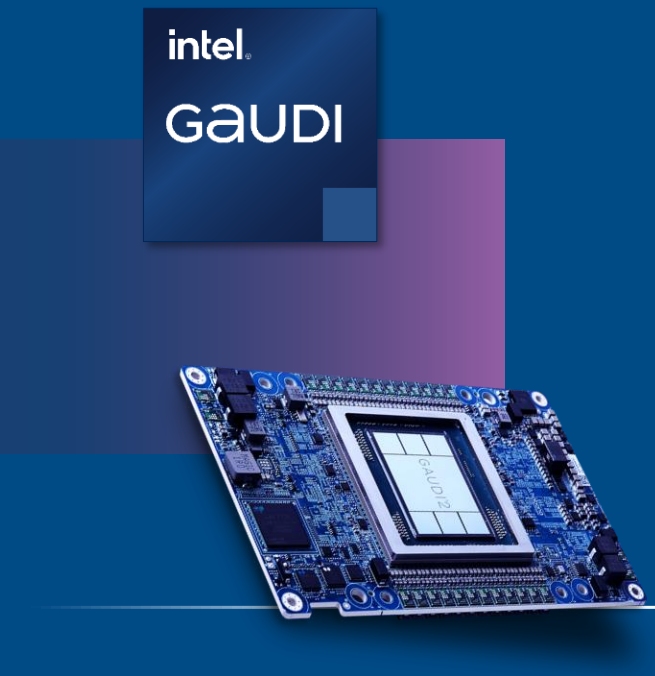
英特尔® 网络研讨会录像，讨论了英特尔® Gaudi® 2 AI 处理器在挖掘生成式 AI 和大型语言模型 (LLM) 潜力方面的尖端能力

英特尔® Gaudi® 2 仍然是唯一能与 NV H100 在生成式 AI 性能上相媲美的替代产品

[1]性能因用途、配置和其他因素而异；有关工作负载和配置的详细信息，请访问：intel.com/performanceindex 结果可能会有所不同。

在众多 LLM 中进行调优

Hugging Face 评估证实英特尔® Gaudi® 2 加速器的
LLM 性能与 NVIDIA A100 和 H100 对比如下



请访问 https://habana.ai/habana-claims-validation 了解工作负载和配置。结果可能会有所不同。

https://huggingface.co/blog/habana-gaudi-2-benchmark
https://huggingface.co/blog/bridgetower

GPT-J：英特尔® Gaudi® 2 结果

英特尔® Gaudi® 2 的 GPT-J 推理性能结果有力地验证了其具有竞争力的性能

- 英特尔® Gaudi® 2 针对服务器查询和离线样本的 GPT-J-99 和 GPT-J-99.9 **推理性能**分别为**每秒 78.58 和每秒 84.08**[1]
- **与 NVIDIA 的 H100 相比，英特尔® Gaudi® 2 提供了引人注目的性能**，H100 相比 Gaudi 2 仅有略微的性能优势，分别为 1.09 倍（服务器）和 1.28 倍（离线）[1]
- **英特尔® Gaudi® 2 的性能比 NVIDIA 的 A100 高 2.4 倍（服务器）和 2 倍（离线）**[1]
- 英特尔® Gaudi® 2 提交的数据采用了 FP8，在此新数据类型上达到了 **99.9% 的准确率**[1]

阅读更多内容

英特尔® Gaudi® 2 软件更新每六至八周发布一次，英特尔® 预计将继续在 MLPerf 基准测试中实现性能提升并扩大模型覆盖范围

新闻发布室文章

MLCommons 公告

[1]性能因用途、配置和其他因素而异；有关工作负载和配置的详细信息，请访问：https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/ 结果可能会有所不同。

英特尔® Gaudi® 2：基准测试结果

Supermicro 提供的
基准测试结果；
业界首个英特尔® Gaudi® 2
原始设备制造商

Gaudi 验证声明

使用英特尔® Gaudi® 2
AI 加速器进行
LLM 训练和推理

基准测试

更快的训练和推理：
英特尔® Gaudi® 2 对比
NVIDIA A100 80GB

基准测试

结果可能会有所不同。

英特尔® Gaudi® 2：基础模型训练和推理

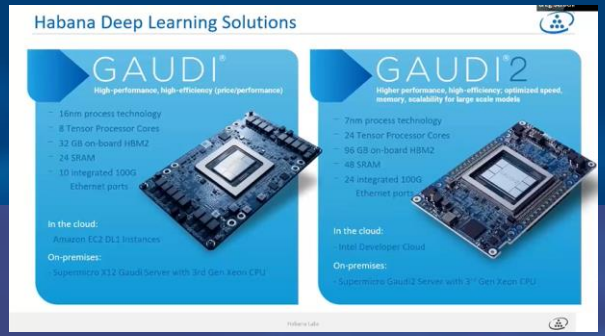
可以访问的支持 Gaudi 的模型请参见

开发人员目录

GAUDI®2

英特尔® Gaudi® 开发人员培训

入门：Gaudi 上的
深度学习和推理



最大限度地发挥
英特尔® Gaudi® 2 的能力：
加速生成式 AI 和大型语言模型

使用英特尔® Gaudi® 处理器
最大限度地提高模型性能：
获得最佳结果的先进工具和策略

英特尔® Gaudi® 软件（SynapseAI® 软件套件）

简化开发：您想要的开发方式

目标：轻松地将现有软件迁移到英特尔® Gaudi® AI 加速器，保留软件投资，并轻松构建新模型，**用于**训练和部署不断增长的众多模型，这些模型**定**义了深度学习、生成式 AI **和大型**语言模型。**通**过以下方式为数据科学家、开发人员、IT **和系**统管理员提供广泛支持：

- 开发人员网站
- GitHub

英特尔® Gaudi® AI 加速器

深度学习软件生态系统汇集了领先的软件提供商、工具和代码，以加速开发基于以下框架的最先进的深度学习模型：PyTorch、TensorFlow、PyTorch Lightning 和 DeepSpeed frameworks

cnvrg.io

PyTorch Lightning

准备好使用英特尔® Gaudi® 软件了吗？

英特尔® Gaudi® 2 AI 加速器现已推出！由 Denvr Cloud 独家提供

英特尔® Gaudi® 2
软件生态系统

英特尔® Gaudi® 2
AI 加速器 (7nm)

英特尔®Gaudi®2—非常适合生成式 AI 的需求

- 现已推出！Denvr Cloud 上的 Gaudi 2 集群
- 多达 8 个 Gaudi 2 节点试用
- 英特尔客户的优先 VIP 定价
- Denvr Dataworks 高度定制化商业服务与支持
- 无缝迁移到 Denvr Cloud 上的 Gaudi 2 集群
- 独家优先布置 Denvr Cloud 上 Gaudi 3 集群 —即将推出！

立即开始

即将推出

intel GAUDI

训练推理

英特尔® Gaudi® 3

凭借为更多客户提供更多选择的性能、可扩展性和效率，英特尔® Gaudi® 3 加速器可帮助企业获得洞察、实现创新并增加收入

即将推出 — 英特尔® Gaudi® 3 AI 加速器

为生成式 AI 带来性能、可扩展性和效率的选择

英特尔® Gaudi® 3 将为寻求大规模部署生成式 AI 的全球企业带来 AI 训练和推理的重大飞跃

新闻稿

英特尔® Gaudi® 3 加速器与 NVIDIA H100 性能对比

预计英特尔® Gaudi® 3 会将以下两种模型的平均训练时间缩短

50%（平均训练时间缩短幅度[3]）：7B 和 13B 参数的 Llama2 模型以及 GPT-3 175B 参数模型

预计英特尔® Gaudi® 3 的性能将超越 H100：

50%：加速器推理吞吐量提升幅度[1]

40%：推理能效提升幅度[2]

Llama 7B 和 70B 参数以及 Falcon 180B 参数模型

阅读更多内容

白皮书

英特尔® Gaudi® 3 将从 2024 年第 2 季度开始向原始设备制造商提供，包括：

Lenovo DELL Hewlett Packard Enterprise SUPERMICRO

[1]NV H100 的比较基于 https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8，报告的数字是每个 GPU 的数字。对比英特尔® Gaudi® 3 针对 LLAMA2-7B、LLAMA2-70B 和 Falcon 180B 的预测。结果可能会有所不同。

[2]NV H100 的比较基于 https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8，报告的数字是每个 GPU 的数字。对比英特尔® Gaudi® 3 针对 LLAMA2-7B、LLAMA2-70B 和 Falcon 180B 的预测。NVIDIA 和 Gaudi 3 的能效均基于内部估算。结果可能会有所不同。

[3]NV H100 的比较基于：https://developer.nvidia.com/deep-learning-performance-training-inference/training “大型语言模型”标签对比英特尔® Gaudi® 针对 LLAMA2-7B、LLAMA2-13B 和 GPT3-175B 的预测，截至 2024 年 3 月 28 日。结果可能会有所不同

英特尔® NLP/LLM 产品

推理

第四代和第五代英特尔® 至强® 可扩展处理器通过英特尔® 深度学习加速、英特尔® AMX 和英特尔® AVX-512 加速 NLP。专为高性能计算而设计，
可用于加速 NLP 工作负载。可以处理大量线程、大内存容量和高内存带宽，适合语言翻译、文本摘要和文本转语音等 NLP 工作负载。

第五代英特尔® 至强®：专为 AI 设计的处理器

第五代英特尔® 至强® 处理器的内核皆有 AI 加速功能，无需添加独立加速器，
就能满足苛刻的端到端 AI 工作负载要求

更出色的 AI 推理性能

与上一代相比[1]

提升高达 42%

通用计算性能提升

与上一代相比[1]

平均 21%

更快的自然语言处理

与上一代相比[1]

提升高达 23%

英特尔执行副总裁兼数据中心
与人工智能事业部总经理
Sandra Rivera

"我们的第五代英特尔® 至强® 处理器专为 AI 而设计，可为在云端、网络和边缘用例中部署 AI 功能的客户提供更高的性能。我们在经过验证的基础上，推出第五代英特尔® 至强®，有助于以更低的总体拥有成本，实现快速采用和扩展，这正是我们与客户、
合作伙伴和开发人员生态系统长期合作的结果。

更多信息

网站

产品简介

有关工作负载和配置的信息，请访问 intel.com/processorclaims：第五代英特尔® 至强® 可扩展处理器。结果可能会有所不同。

英特尔® 至强®：在真实世界的 AI 应用中，CPU 性能遥遥领先

在实际工作应用中，英特尔为 AI 推理提供了性能更强、价格更低、
更均衡的平台，从而颠覆了行业，实现了 AI 的民主化：

- 更大的高速缓存有助于数据局部性，
  更大的内存容量可解决更大的问题
- 更高的内核频率、多个标量端口和无序执行，
  有助于加速单线程或多线程标量计算
- 英特尔®高级矢量扩展 512（英特尔®AVX-512），
  有助于非 DL 矢量计算
- 英特尔®高级矩阵扩展（英特尔®AMX），
  内置 AI 加速硬件支持

技术全文

信息图表

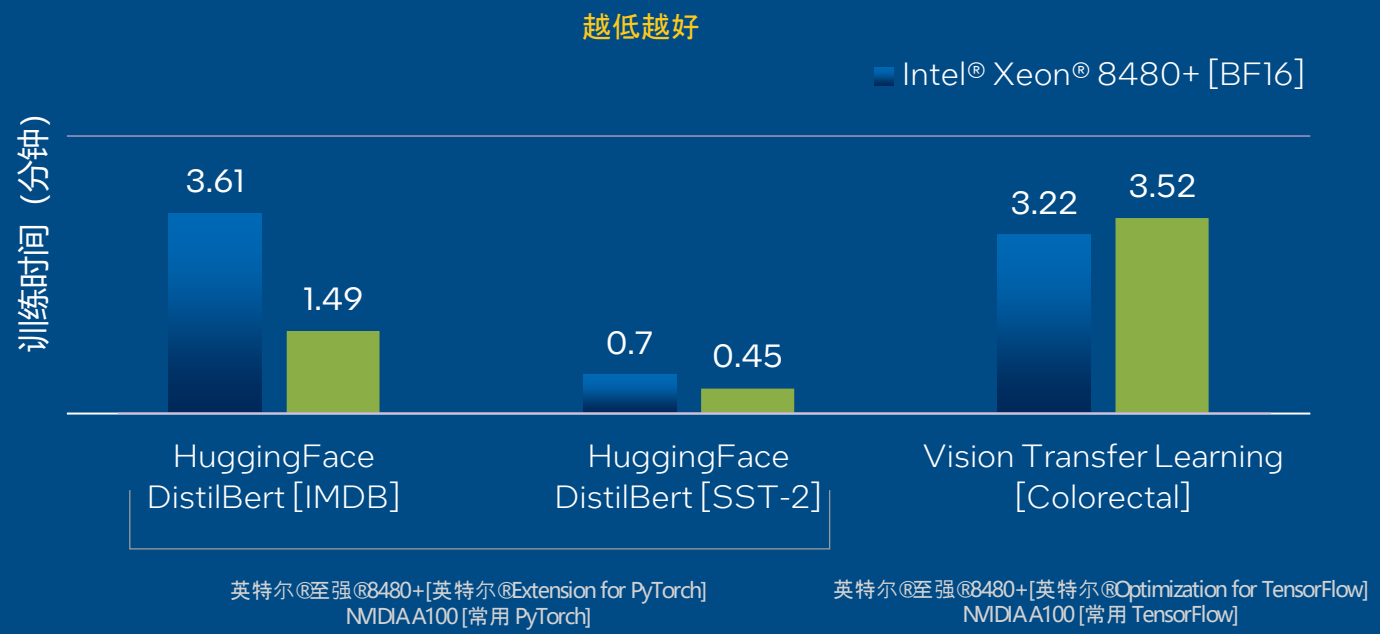
intel XEON

揭开 GPU 的神秘面纱：内置加速器的 CPU 如何彻底改变 AI

利用英特尔® 至强® 可扩展处理器，不到 4 分钟即可完成调优模型[1]

Hugging Face

训练时间性能调优，英特尔® 至强® Platinum 8480+ 处理器对比 NVIDIA A100 GPU



另请参阅：

更好的性能：Numenta 采用英特尔® CPU 对比 NVIDIA GPU

[1]请参阅第四代英特尔至强可扩展处理器的性能指标的 [A221]。结果可能会有所不同。

第四代英特尔® 至强® 处理器上的 LLM

人工智能 (AI) 聊天机器人技术作为一种与客户交互和改善客户服务的方式，越来越受到企业和组织的青睐，但是为特定用例构建、优化和维护聊天机器人的成本高昂，可能会让许多组织望而却步

更多信息

第四代英特尔® 至强® 可扩展处理器上的 AI 调优指南

指南链接 ›

第四代英特尔® 至强® 处理器通过英特尔® Advanced Matrix Extensions 供改进的数据管理和高效计算，当与通过英特尔® Extension for PyTorch 提供的自动混合精度 (AMP) 功能相结合时，该技术堆栈在迁移学习和从头开始训练中小型模型等工作负载方面表现出强大的竞争力

技术指南文章

面向生成式 AI、搭载第五代和第四代英特尔® 至强® 处理器的 Cisco UCS

越小越好：Q8-Chat LLM 是英特尔® 至强® 处理器上高效的生成式 AI 体验

LLM 需要大量计算能力（通常由高端 GPU 提供）来为搜索或对话应用等低延迟用例提供足够快的预测。遗憾的是，对于许多组织来说，相关成本可能过高，使其难以在应用中使用最先进的 LLM。

了解有助于减少 LLM 大小和推理延迟的优化技术，帮助他们在英特尔® CPU 上高效运行。

技术指南文章 ›

Hugging Face

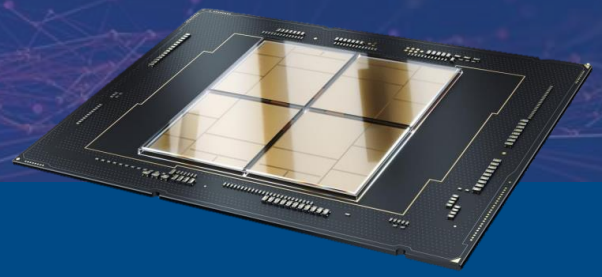
“更多公司最好专注于更小、训练和运行成本更低的特定模型。”

开始使用第四代英特尔® 至强® 与 Hugging Face

面向 LLM 的英特尔® 至强® 处理器

总结

- **非常适合推理**专用领域 LLM
- 满足迁移学习用例的需求
- **在英特尔® 至强®** 处理器上使用开源软件部署 LLM，轻松实现最佳性能



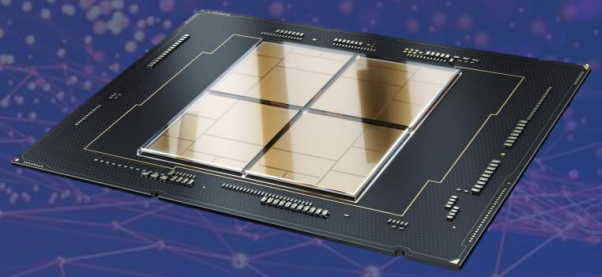
面向 LLM 的英特尔® 至强® 可扩展处理器

非常适合使用最流行的 AI 框架和库来构建和部署通用 AI 工作负载

- **利用**现有基础设施**来推理行**业专用 LLM
- 满足迁移学习用例的需求
- **在英特**尔®**至强**®处理器上使用开源软件部署 LLM，轻松实现最佳性能

英特尔® 至强®
真实世界 AI 应用中的 CPU 性能领先地位

技术文章 • 信息图表



GPT-J
第四代英特尔® 至强® 结果

2 离线模式下的每秒段落数[1]

1 实时服务器模式下的每秒段落数[1]

新闻发布室文章 • MLCommons 公告

揭开 GPU 的神秘面纱：内置加速器的 CPU 如何彻底改变 AI

内置英特尔® AMX 的第四代英特尔® 至强® 上的阿里巴巴 NLP 案例研究

阅读更多内容

[1]性能因用途、配置和其他因素而异；有关工作负载和配置的详细信息，请访问：https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/ 结果可能会有所不同。

英特尔® NLP/LLM 产品

客户端上的小规模推理

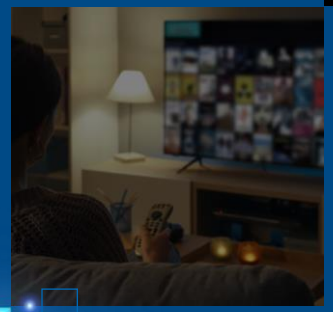
英特尔® 酷睿™ Ultra 引领 AI 电脑时代

英特尔® 酷睿™ Ultra 处理器针对轻薄且功能强大的高端笔记本电脑进行了优化，具有 3D 性能混合架构、先进的 AI 功能，并可提供内置英特尔锐炫™ GPU。英特尔® 酷睿™ Ultra 处理器采用全新英特尔® 4
工艺打造，实现了性能与能效的最佳平衡，可满足游戏、内容创作和移动办公的需求。

用例：电脑上的 AI

创作者：照片视频搜索和编辑

更快、更自然的过滤器、
更高质量的预览和更快的导出时间，
以及更快的自动搜索。



协作/直播

全新的 AI 功能，可用于下一代视频会议、直播和协作，延长电池续航时间。

主流游戏

用于游戏内 3D 动画的
全新 AI 功能，增添了现实感、
转录和聊天翻译。

电脑上的 AI

工作效率

用于写作、创作、编码和离线功能（例如文本和语法预测）的 AI 助理。

可用性

AI 辅助视听功能满足用户的不同需求，
让用户更轻松在电脑上创作，
提高工作效率。

创作者：文本转图像

全新的 AI 效果和功能，仅需几个
描述词（营销、广告、设计），
即可创建图像。

”让平凡之事造就不平凡“

*英特尔 2023 年 12 月 14 日“广布 AI” 活动新闻发布会

面向生成式 AI 的英特尔® 酷睿™ Ultra

节能表现出众的英特尔客户端处理器引领 AI PC 时代

效率和性能的重大改进

AI 效率

高达 **70%**

生成式
AI 性能提升幅度[2]

节能

高达 **25%**

功耗降低幅度[3]

阅读更多内容 公告 • 产品简介 • 网站

英特尔®酷睿™ Ultra 配备首款英特尔客户端片上 AI 加速器（神经处理单元 (NPU)），使 AI 加速达到全新的能效水平，与上一代相比，能效提升 2.5 倍[1]

英特尔®酷睿™ Ultra H 和 U 代芯片均包含两个全新的低功率岛 (LP-E) 核心，用于低强度工作负载，英特尔 AI NPU 内的两个神经计算引擎旨在处理生成式 AI 推理。

[1]基于在英特尔® 酷睿™ Ultra 7 165H NPU 与英特尔® 酷睿™ i7-1370P GPU 上运行 int8 模型时 UL Procyon AI 基准测试的性能功耗比测量结果。
[1,2,3]请访问 www.intel.com/PerformanceIndex 了解工作负载和配置。结果可能会有所不同。

加速 AI 创新

英特尔® 正与业界领先的 ISV 合作，优化您的 AI 体验。

AI 电脑加速计划旨在将独立硬件开发商 (IHV) 和独立软件开发商 (ISV) 与英特尔® 资源联系起来，包括人工智能 (AI) 工具链、培训、联合工程、软件优化、硬件、设计资源、技术专业知识、联合推广和销售机会。

了解详情

利用英特尔® Tiber™ AI 云加速企业 AI 开发

在最新的英特尔® 硬件和软件集群上学习、原型设计、测试和运行应用及工作负载

在此开发环境下使用最新的硬件和软件创新**加速和扩展 AI。**
获得更多计算能力和选择来**调优您的软件和生成式 AI。**

与英特尔一起开始

亲身体验最新的英特尔® 产品。借助英特尔技术增强您的 AI 技能。

早期技术访问

评估预发布的英特尔® 平台和相关的英特尔优化软件堆栈。

大规模部署 AI

利用英特尔的最新机器学习工具套件以及英特尔® Tiber™ AI 云上托管的库，加快 AI 部署。

阅读技术文章 ›
开始 ›

行为召唤

教育

了解如何将英特尔®技术用于生成式 AI 和行业专用模型，以及英特尔®至强®和英特尔®Gaudi®产品线如何帮助您赢得更多业务

开始

参与

开始使用

英特尔®Tiber f AI 云

在此开发环境下使用最新的硬件和软件创新加速和扩展 AI

以及

使用 AI 参考套件

联系人

请联系您的英特尔®代表，了解更多信息

如何访问英特尔® 合作伙伴联盟客户支持

英特尔 Virtual Assistant

这个聊天机器人位于每个**合作伙伴**联盟网页的右下角，提供可解**答大多数**问题的**自助服**务或实时支持代理**的快速**链接。

“获取帮助”大型条幅广告

提交在线支持请求。**此**链接位于合作伙伴联盟网站**内****大多数**页面的页脚处。

合作伙伴联盟“获取帮助”页面

获取帮助页面提供了有关合作伙伴联盟成员可用的**大多数工具和**权益的详细自助指南。

AI 激活区

数字优先的 AI 工作空间，汇集关键资源、工具和优势 — 激励合作伙伴基于英特尔® 技术构建、营销和销售解决方案

技术支持

销售和营销支持

技术支持

销售和营销支持

技术支持

销售和营销支持

AI 参考套件

利用这些参考套件，组织可以大幅缩短解决问题的时间并获得显著的性能提升

金融和保险

欺诈检测

GitHub • 博客 • 蓝图

医疗和生命科学

疾病防护

GitHub • 博客

制造和公用事业

异常检测

GitHub • 博客

车队管理

预测性维护

GitHub

流程自动化

文档自动化

GitHub • 博客 • 蓝图

工作流程

- 深度学习迁移学习
- HF 调优和推理优化
- 深度学习分布式压缩
- 分布式经典机器学习工作流程
- 使用英特尔® 加速器进行深度学习预训练
- 使用 DGL 和 PyG 进行图形分析和 GNN
- Big-DL 上的分布式 Trang/推理
- Ray 上的 LLM 预训练和调优

工具

- 英特尔® Distribution for Python
- 英特尔® Optimized Modin
- 英特尔® Optimized XGBoost
- 英特尔® Extension for Scikit-Learn
- 英特尔® Optimized Tensorflow (ITEX)
- 英特尔® Optimized PyTorch（stock 和 IPEX）
- 英特尔® Neural Compressor
- SigOpt Python SDK 和 CLI
- CNVRG Python SDK 和 CLI
- 英特尔优化的 Horovod
- DeepSpeed

行业套件

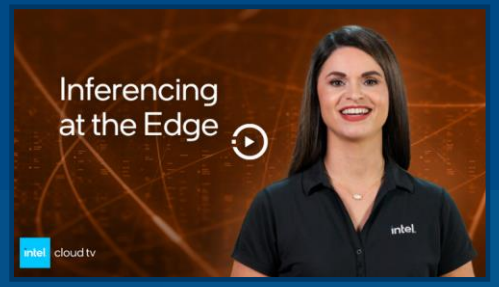
- 时序
- PPML
- 迁移学习
- Transformer/NLP

参考套件以**容器**的形式提供，可以在**主要的云上以及本地使用。**
参考套件基于**工作流程**和**行业工具套件**分层，可以独立用于支持**多个行业的更广泛用例。**

Cloud TV

英特尔®Cloud TV 将探讨云计算新闻、趋势和战略，助力您取得成功

英特尔® Gaudi® AI 加速器为您带来生成式 AI 机遇



使用边缘数据推理获取洞察

利用云端 AI 创造竞争优势

利用云技术进行 AI 推理

云端 AI

迈上快速路径，随时随地扩展 AI

培训

让 AI 遍布每个角落 — 生成式 AI 企业用例

生成式 AI 不仅仅适用于互联网聊天机器人。许多企业正在思考如何利用生成式 AI 和大型语言模型的力量来协助日常运营。本课程将探讨生成式 AI 在企业中的用例，并提供您的组织如何将其应用到日常运营中的相关注意事项。

注册 ›

简化用于数据生成和大型语言模型的 AI

注册 ›

将 AI 集成到组织的工作负载或扩展现有基础设施是一项技能要求高的计算密集型工作，需要开发在海量数据集上训练的稳健模型，并使用强大的 GPU 来充分运行它们。并非每个组织都拥有完成此任务所需的资源。

此课程重点介绍一个解决方案：Accenture* 和英特尔® 的开源 AI 参考套件集合，旨在让组织更容易使用 AI，并为缩短训练和推理时间进行了优化。

其他培训

技术

资料类型	标题和链接
能力课程	云端 AI 能力课程
网络研讨会	借助 Hugging Face 为英特尔® 硬件优化 AI
网络研讨会	如何设置基于云的分布式训练来调优 LLM
培训课程	通过提示优化和情境学习改进 LLM
培训课程	简化用于数据生成和大型语言模型的 AI
培训课程	自然语言处理
培训课程	采用 TensorFlow* 的应用深度学习
培训课程	小巧灵活 — 通往企业生成式 AI 的捷径
培训课程	生成式 AI 的下一波浪潮 — 行业专用 LLM
指南	生成式 AI 入门开发人员指南：用例特定的方法
培训课程	将英特尔® 至强® 处理器上的 AI 引入解决方案领域

其他培训

非技术

资料类型	标题和链接
视频系列	采用生成式 AI
培训课程	小巧灵活 — 通往企业生成式 AI 的捷径
培训课程	生成式 AI 的下一波浪潮 — 行业专用 LLM
培训课程	AI 无处不在的原理能力课程
培训课程	AI 软件和生态系统原理能力课程
培训课程	参与 AI 生态系统：利用软件赢得胜利、利用 SI 扩展并销售解决方案
培训课程	面向真实世界的生成式 AI 和大型语言模型

更多资源

资料类型	标题和链接
网络研讨会	生成式 AI 网络研讨会系列
网络研讨会	让生成式 AI 遍布每个角落
播客	Copilot、ChatGPT、Stable Diffusion 和生成式 AI 将如何改变我们的开发、工作和生活方式
业务简介	随处部署 AI
博客系列	使用第四代英特尔至强处理器对生成式 AI 进行调整和推理
解决方案简介	使用联想 ThinkSystem SR650 V3/第四代英特尔至强处理器部署和扩展生成式 AI 推理 全新的英特尔和 VMware 技术助力联想 ThinkAgile VX V3 系统实现性能飞跃
技术文章	利用英特尔® AI 硬件和软件优化加速 Llama 2
研究新闻稿	10% 的受访组织在 2023 年将生成式 AI 解决方案投入生产
炉边谈话视频	迎接生成式 AI 的计算和可持续性挑战
播客	Hugging Face 和英特尔：推动实用、更快、民主化和合乎道德的 AI 解决方案
Twitter/X 对话	民主化大型语言模型如何推动 AI 开发
Supermicro 基准测试	Habana 验证声明
Hugging Face 基准测试	基准测试
培训/网络研讨会	云解决方案架构师 (CSA) 技术讲座：AI 与 Habana
白皮书	“企业 AI 关乎开发人员”白皮书
信息图表	CPU 是企业 AI 的关键

更多资源

资料类型	标题和链接
解决方案简介	利用英特尔企业 AI 与 Red Hat® OpenShift® AI 简化 AI 的采用和部署
指南	AI 指南
参考套件	AI 非结构化文本数据生成
白皮书	Zoho 正在优化和加速视频 AI 工作负载
白皮书	Seekr 开发值得信赖的 AI 筛选系统
解决方案简介	教育安全：AI 和机密计算有助于实现安全的远程考试
案例研究和视频	Nature Fresh Farms 利用 AI 实现从种子到商店的全流程
案例研究	QMed Asia 提升早期癌症检出率
案例研究和视频	MetaApp 改造基于 AI 的推荐系统
解决方案简介	优化 AI 模型训练和改进以用于自动光学检查 (AOI)
博客	借助提示词提升 LLM 效率

法律通告与免责声明

通告和免责声明.


intel