

VMware vCloud Foundation Private AI

Peter Trawnicek
VMware Marketing Austria GmbH

September 2024

THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

Summer 1956

Proposed Topics:

1. Automatic Computers
2. How Can a Computer be Programmed to Use a Language
3. Neuron Nets
4. Theory of the Size of a Calculation
5. Self-improvement
6. Abstractions
7. Randomness and Creativity



@donalleniii | SORA



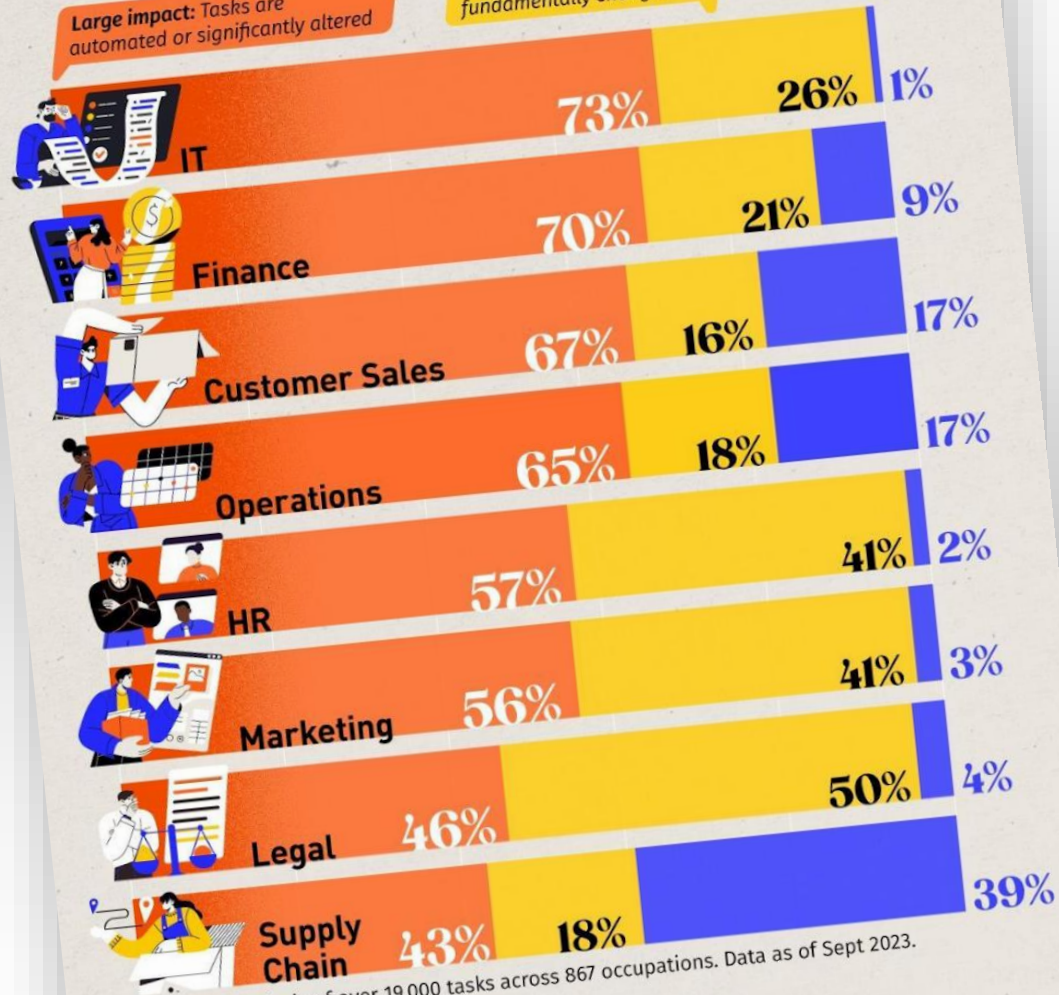
episode 5 presents the Bunny Armadillo

WHICH JOB DEPARTMENTS WILL AI Impact THE MOST

Large impact: Tasks are automated or significantly altered

Small impact: Tasks are not fundamentally changed

No impact



Based on analysis of over 19,000 tasks across 867 occupations. Data as of Sept 2023.

Source: Jobs of Tomorrow: Large Language Models and Jobs

**The Times Sues OpenAI
Over A.I. Use of Copyrighted**

Millions of articles
chatbots that

**JPMorgan pitches in-house chatbot as AI-based
research analyst**

About 50,000 staff are using a large language model designed by the bank to boost productivity

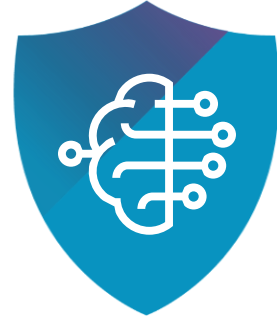
Powerful antibiotics discovered using AI

Machine learning spots molecules that work even against 'untreatable' strains of bacteria.

OpenAI over access to news titles

German publisher will allow content to be used to train artificial intelligence in 'first of its kind' agreement

JPMorgan executives told staff the large language model the bank used to summarize and summarising documents © Bloomberg



Private AI

Eine Architektur um die wirtschaftlichen und organisatorischen Möglichkeiten von KI mit den Datenschutz- und Compliance-Anforderungen in Einklang zu bringen

Warum *Private* KI?



Policy and
Control



Resource
Sharing



Lower
TCO



Centralized
Ops

Speed and Agility



Holen Sie sich die “Reference Architecture for VMware Private AI”



Falcon LLM



Llama 2



Hugging Face



MPT



NVIDIA NeMo



RAY



Kubeflow
VMware Distribution



PyTorch



VMware Cloud Foundation



Compute



Storage



Network/DPU



GPU

GENERALLY AVAILABLE

VMware Private AI Foundation with NVIDIA

Joint GenAI Platform

 NVIDIA Foundation Models

 NVIDIA Fine-Tuned Models

 Third Party and Community Models

NVIDIA NIM

NVIDIA NeMo Retriever

 NVIDIA AI Enterprise

NVIDIA NIM Operator

NVIDIA GPU Operator

Deep Learning VMs

Vector Database Postgres + pgvector

 VMware by Broadcom

Catalog Setup Wizard

GPU Monitoring



VMware Cloud Foundation



NVIDIA AI Enterprise

 Dell Technologies

 Hewlett Packard Enterprise

 Lenovo

 SUPERMICRO

 Hitachi Vantara

 Fsas Technologies

Choice of LLMs

Bare-Metal Performance

Faster Time-to-Value

VCF als Plattform für GenAI Workloads



Skalierbarkeit



Management
und Operations



Kosten & Kapazität



Individuelle
Dashboards



VMware Cloud Foundation™

Typische Anwendungsfälle unserer Kunden



S/W Entwicklung

Contact Centers
Problemlösung



IT Operations
Automatisierung

Advanced Information
Retrieval



PRIVATE CLOUD PLATFORM

VMware Cloud
Foundation 9



APP DELIVERY

VMware Tanzu
Platform 10.0



Innovation

PRIVATE AI

VMware Private AI
Foundation with
NVIDIA



SOFTWARE-DEFINED EDGE

Software-Defined
Edge Innovations



Danke!

peter.trawnicek@broadcom.com

+43 664 2101912

VMware Marketing Austria GmbH

Am Europlatz 5

1120 Wien

vmware[®]

by **Broadcom**

Copyright © 2024 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries.