2024 embedded VISION SUMMIT®

Transforming Enterprise Intelligence: The Power of Computer Vision and Gen AI at the Edge with OpenVINO™

Leila Sabeti

Americas AI Technical Sales Lead

Intel

OpenVINO intel.

Agenda

- Edge AI applications
- The OpenVINO™ toolkit: An open-standard for building AI at the edge, in the cloud, or locally
- Deep-dive: Enterprise intelligence and Intel®'s portfolio
- Flexible edge and cloud computing paradigms
- Q&A

© 2024 Intel

AI is Everywhere – from Edge to Cloud

Edge

- Real time data processing
- Wider reach
- Data sovereignty
- Cost efficiency

Frictionless Retail

Traffic Monitoring

Defect Detection

Depth Camera

IP Camera

NVR

Sensor

■■■

OpenVINO

oneAPI

© 2024 Intel

Enterprise Intelligence at the Edge with Intel®

Node
Fine-tuning, Inference

Cluster
Light Training, Tuning, Peak Inf.

Choose Your Compute

Light AI

Efficiency for sub-100 W designs

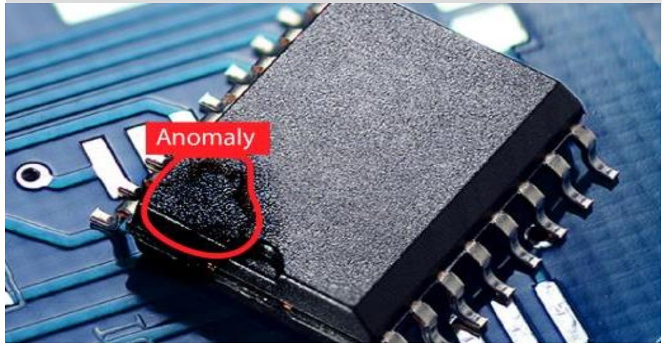
CPU AI, built-in GPU, built-in NPU

intel CORE

intel ATOM

+

intel iRIS Xe GRAPHICS



Medium AI

Scale up perf/W for diverse system designs

Discrete GPU & built-in CPU AI

intel XEON

intel CORE

+

intel ARC GRAPHICS

intel iRIS Xe GRAPHICS

Heavy AI

Optimize for peak perf and density

Optimize with high-end Discrete GPU

intel XEON

intel CORE

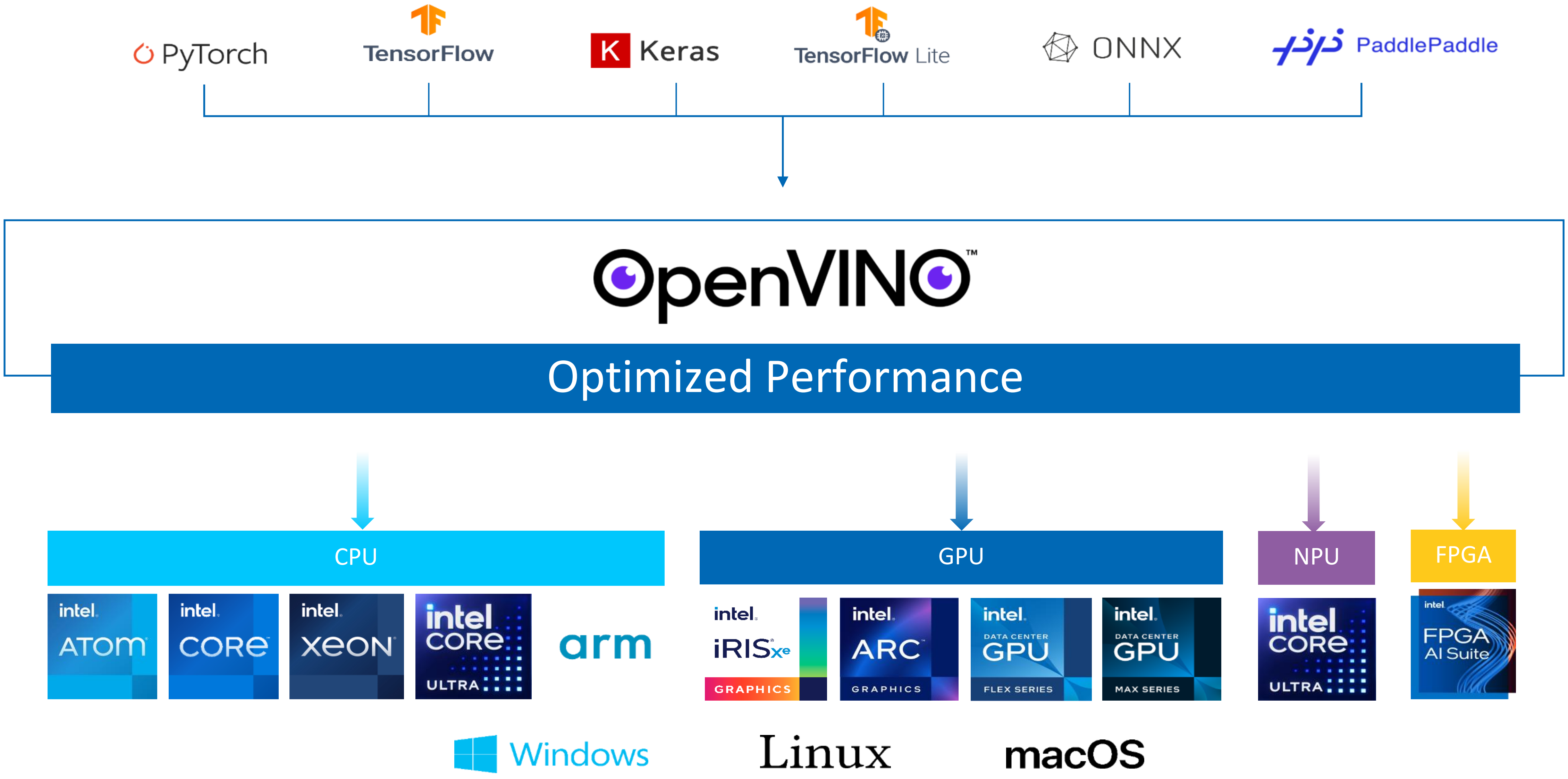
+

intel DATA CENTER GPU FLEX SERIES

intel ARC GRAPHICS

© 2024 Intel

OpenVINO: An Open Standard for Building AI at the Edge



PyTorch
TensorFlow
Keras
TensorFlow Lite
ONNX
PaddlePaddle

OpenVINO™

Optimized Performance

CPU
GPU
NPU
FPGA

Windows
Linux
macOS

© 2024 Intel

Unlocking Software Optimizations with Hardware Using the Intel® Edge AI Portfolio

Supports myriad AI use cases	Wide range of AI performance	Real-time and offline execution	Software Benefits with OpenVINO™
Open-source for AI, DL, Inference	Performance-optimized	Cross-platform Support	OpenVINO™ Toolkit Value Differentiators
Flexible video streams Ranges from few to many streams, and TOPS in low-end NVRs to TOPS for on-prem servers	**Temperature range** Ranges from -5 to +105 degree C	**TDP/Power** Ranges from sub 10 W to over 300 W	Hardware Benefits with Intel®'s Edge AI Portfolio

© 2024 Intel

Compute for AI: Intel®'s Platforms

Edge AI Platforms

intel CORE ULTRA

intel CORE

intel ARC™ A370M

Partner edge platforms using Intel® Arc™ GPU

ADVANTECH *Enabling an Intelligent Planet*

ADLINK TECHNOLOGY INC.

ASUS IoT

Intel® Edge AI Box

- Single NN Pipeline
- Multi NN Pipeline
- Data Fusion Pipeline

Seamlessly Integrated into Existing Camera and Video Deployments

© 2024 Intel

Challenges and Opportunities with AI at the Edge

© 2024 Intel

Intelligent Queue Management at the Edge with OpenVINO™ and YOLOv8

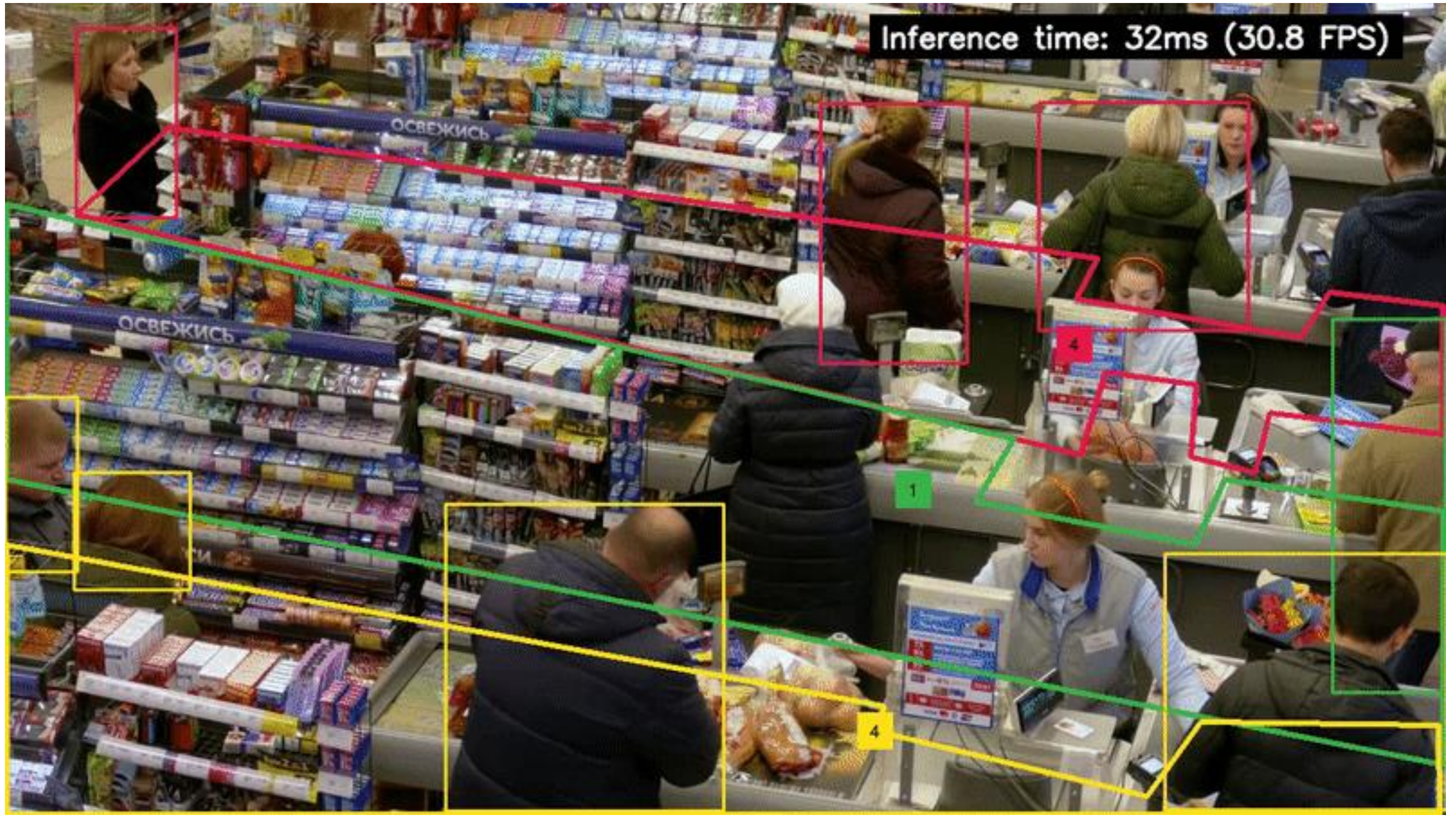
Objective: Optimize the queuing process and reduce wait times via object detection.

Challenges:

- Real-time scalability
- Device setup and calibration
- Model performance
- Low-power

Intel's solution:

- Fast and efficient inference with optimized YOLOv8 models using OpenVINO



© 2024 Intel

Mobile Multi-modal Assistant with MobileVLM and OpenVINO™

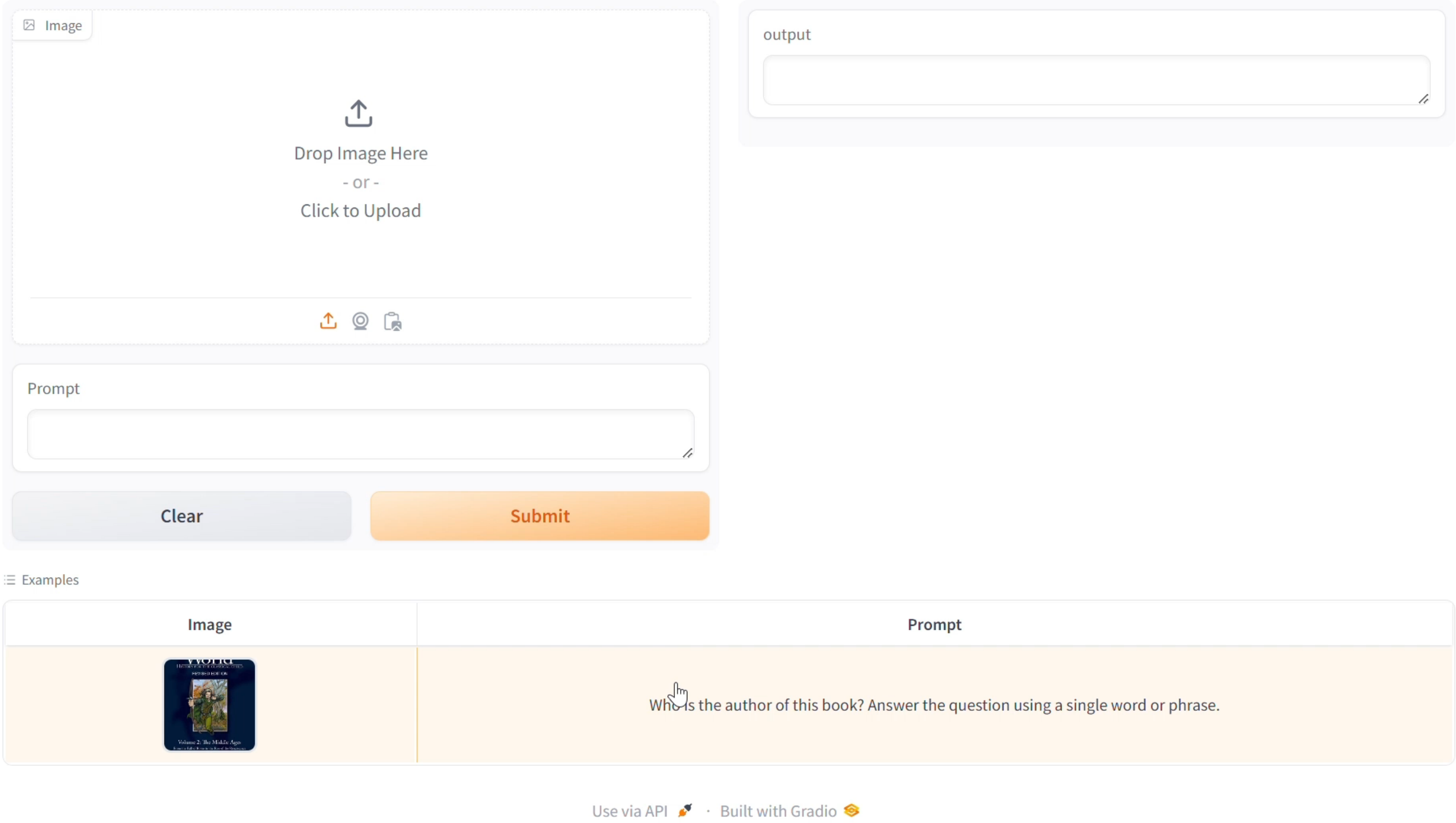
Objective: Use a mobile chatbot to answer questions about images

Challenges:

- Fast token generation
- Memory-efficiency
- Model size

Intel's solution:

- Compress and quantize LLM models for faster, efficient local inference

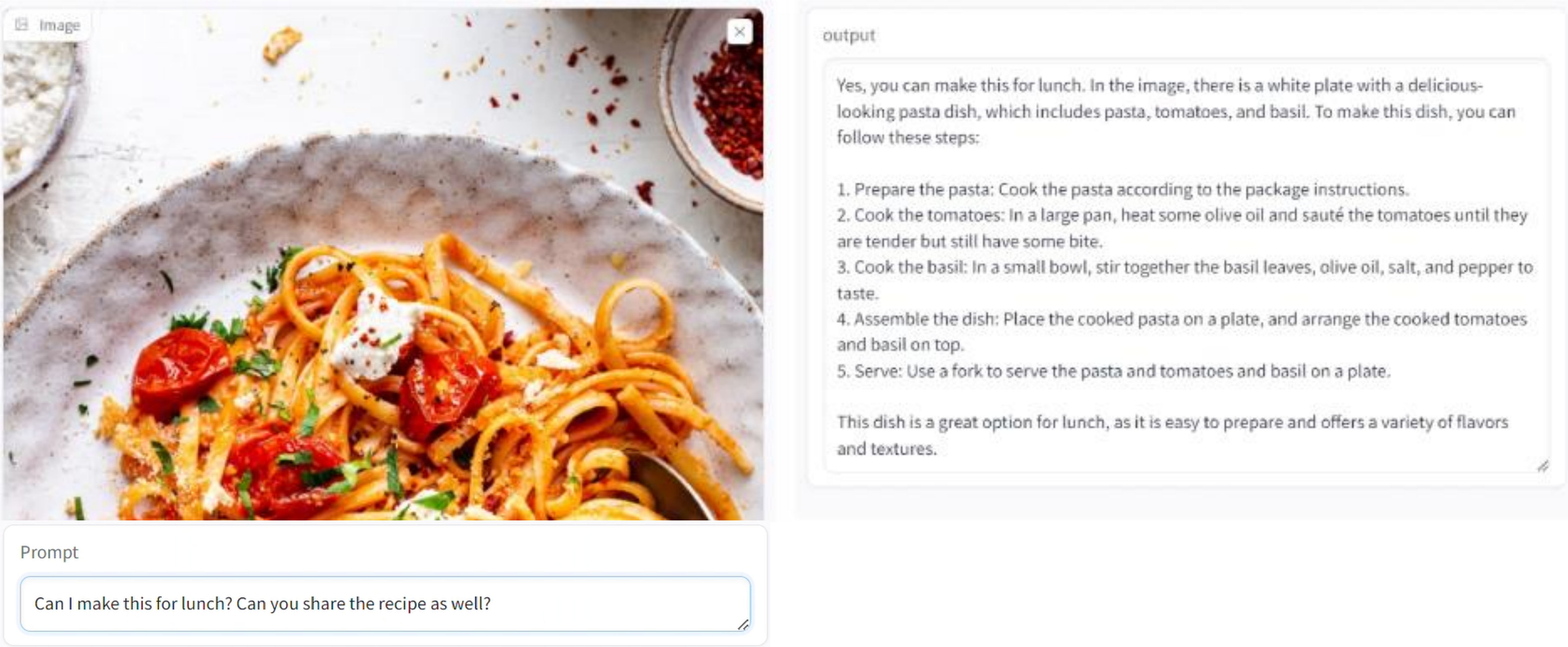


Who is the author of this book? Answer the question using a single word or phrase.

GPU

© 2024 Intel

Snapshot of MobileVLM Output



Prompt

Can I make this for lunch? Can you share the recipe as well?

output

Yes, you can make this for lunch. In the image, there is a white plate with a delicious-looking pasta dish, which includes pasta, tomatoes, and basil. To make this dish, you can follow these steps:

1. Prepare the pasta: Cook the pasta according to the package instructions.
2. Cook the tomatoes: In a large pan, heat some olive oil and sauté the tomatoes until they are tender but still have some bite.
3. Cook the basil: In a small bowl, stir together the basil leaves, olive oil, salt, and pepper to taste.
4. Assemble the dish: Place the cooked pasta on a plate, and arrange the cooked tomatoes and basil on top.
5. Serve: Use a fork to serve the pasta and tomatoes and basil on a plate.

This dish is a great option for lunch, as it is easy to prepare and offers a variety of flavors and textures.

© 2024 Intel

Document Visual Question Answering with Pix2Struct and OpenVINO™ on CPU

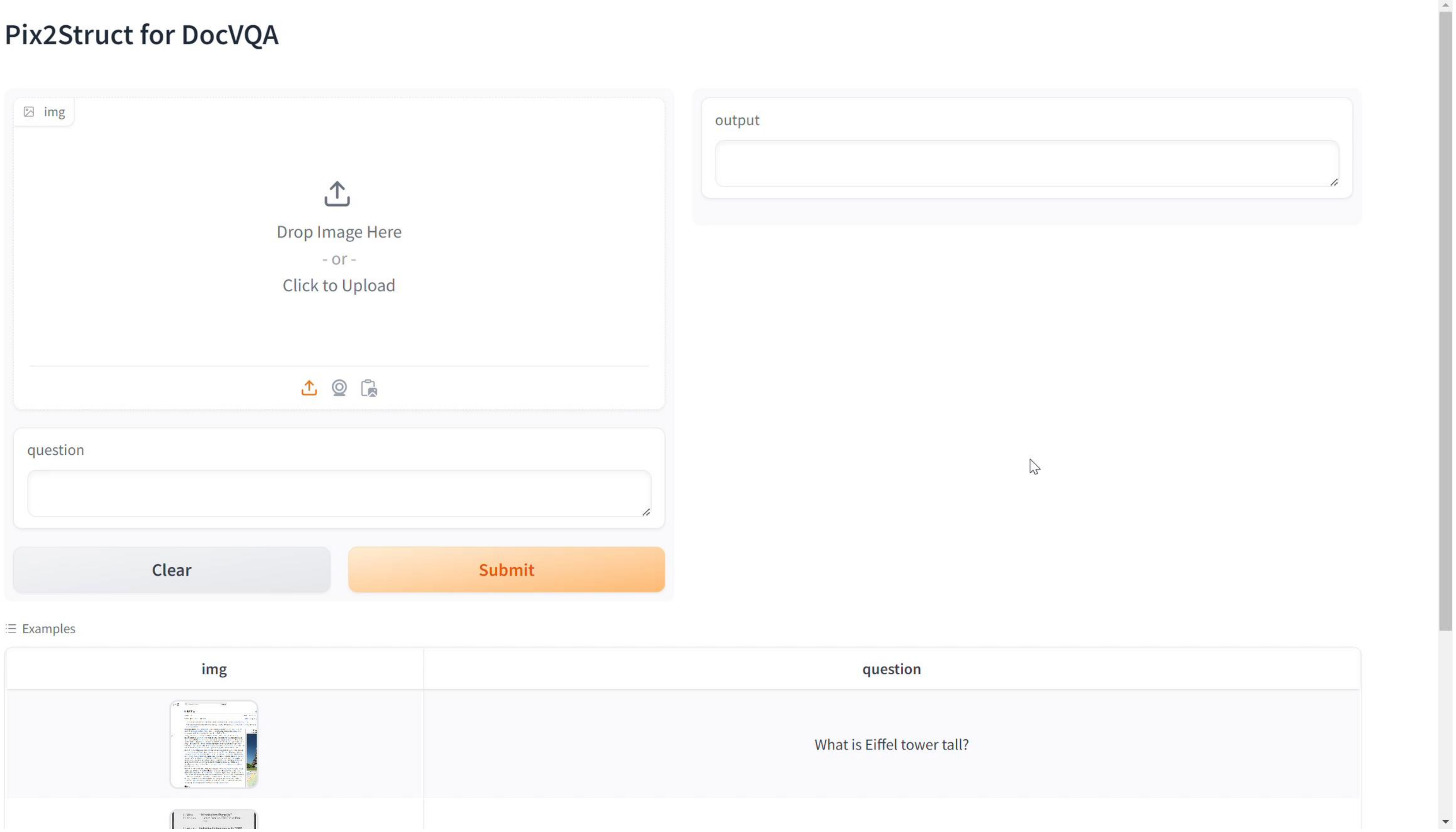
Objective: Use a low-power chatbot to answer questions about documents on the fly.

Challenges:

- Visual understanding
- Memory-efficiency
- Model size

Intel's solution:

- Compress and quantize mult modal models for faster, efficient local inference



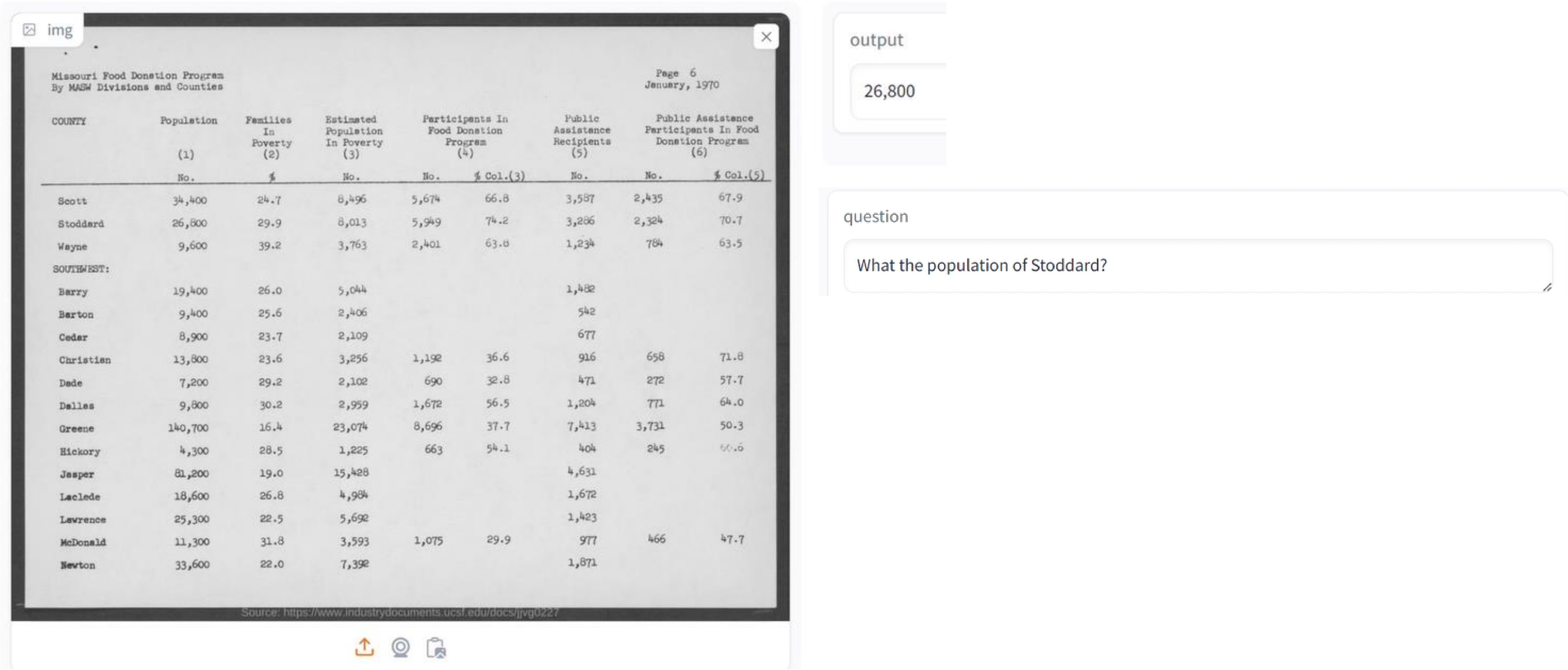
Pix2Struct for DocVQA

img	question
	What is Eiffel tower tall?

Intel® Core™ Ultra 9 processor CPU

© 2024 Intel

Snapshot of Pix2Struct Output



Missouri Food Donation Program
By MASW Divisions and Counties

Page 6
January, 1970

COUNTY	Population (1)	Families In Poverty (2)	Estimated Population In Poverty (3)	Participants In Food Donation Program (4)		Public Assistance Recipients (5)	Public Assistance Participants In Food Donation Program (6)	
	No.	%	No.	No.	% Col.(3)	No.	No.	% Col.(5)
Scott	34,400	24.7	8,496	5,674	66.8	3,587	2,435	67.9
Stoddard	26,800	29.9	8,013	5,949	74.2	3,286	2,324	70.7
Wayne	9,600	39.2	3,763	2,401	63.8	1,234	784	63.5
SOUTHWEST:								
Barry	19,400	26.0	5,044			1,482		
Barton	9,400	25.6	2,406			542		
Cedar	8,900	23.7	2,109			677		
Christian	13,800	23.6	3,256	1,192	36.6	916	658	71.8
Dade	7,200	29.2	2,102	690	32.8	471	272	57.7
Dallas	9,800	30.2	2,959	1,672	56.5	1,204	771	64.0
Greene	140,700	16.4	23,074	8,696	37.7	7,413	3,731	50.3
Hickory	4,300	28.5	1,225	663	54.1	404	245	60.6
Jasper	81,200	19.0	15,428			4,631		
Laclede	18,600	26.8	4,984			1,672		
Lawrence	25,300	22.5	5,692			1,423		
McDonald	11,300	31.8	3,593	1,075	29.9	977	466	47.7
Newton	33,600	22.0	7,392			1,871		

Source: https://www.industrydocuments.ucsf.edu/docs/jjvg0227

output

26,800

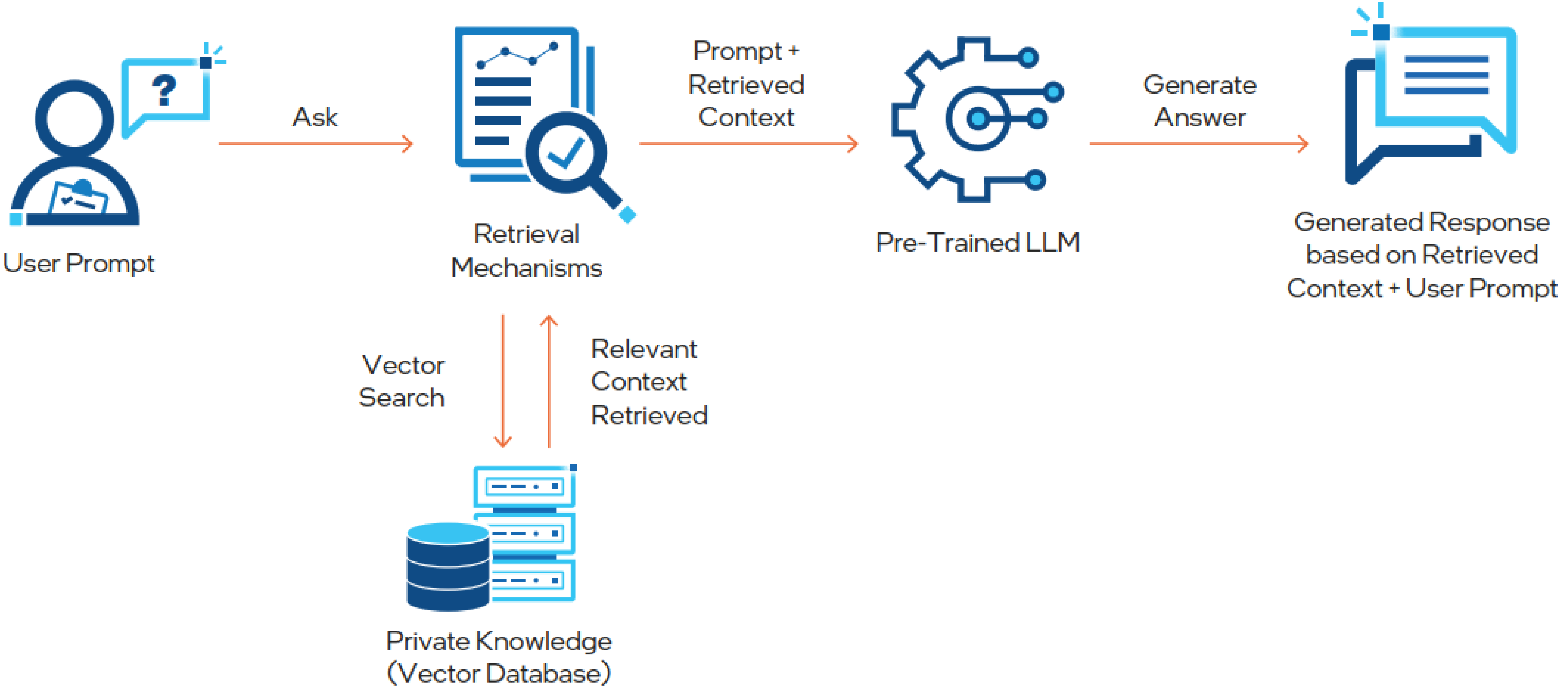
question

What the population of Stoddard?

© 2024 Intel

Enterprise Intelligence with LLMs using RAG

Connect knowledge bases to LLMs with Retrieval Augmented Generation (RAG)


User Prompt

Ask

Retrieval Mechanisms

Prompt + Retrieved Context

Pre-Trained LLM

Generate Answer

Generated Response based on Retrieved Context + User Prompt

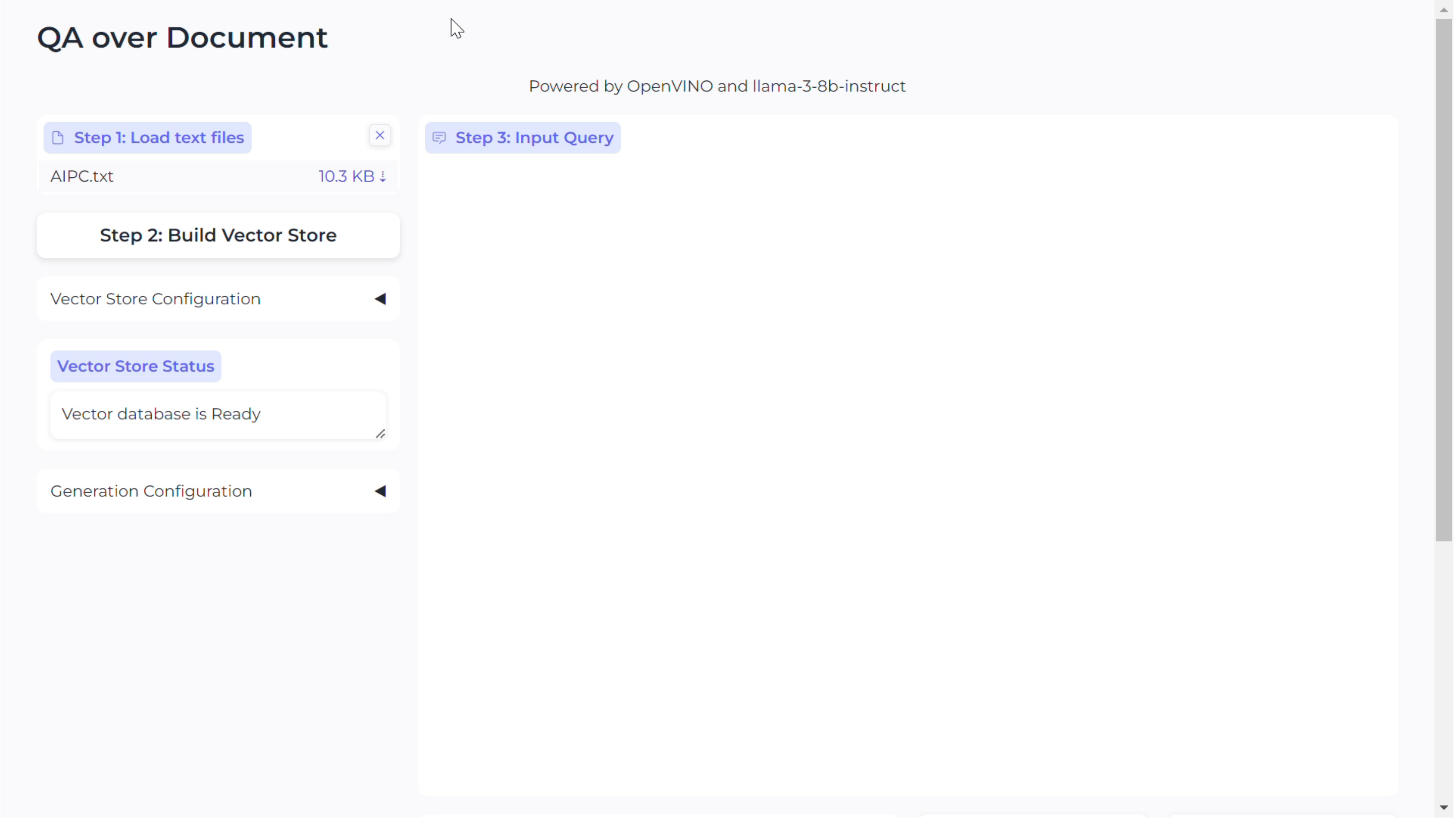
Vector Search

Relevant Context Retrieved

Private Knowledge (Vector Database)

© 2024 Intel

Running LLM + RAG with OpenVINO™ and LangChain on iGPU for the edge



QA over Document

Powered by OpenVINO and llama-3-8b-instruct

Step 1: Load text files

AIPC.txt 10.3 KB

Vector Store Status: Vector database is Ready

Enable enterprise intelligence through knowledge-based search

© 2024 Intel

Enterprise Data Protection at the Edge

Intel® Software Guard Extensions (Intel® SGX)

Secure Access Service Edge (SASE)

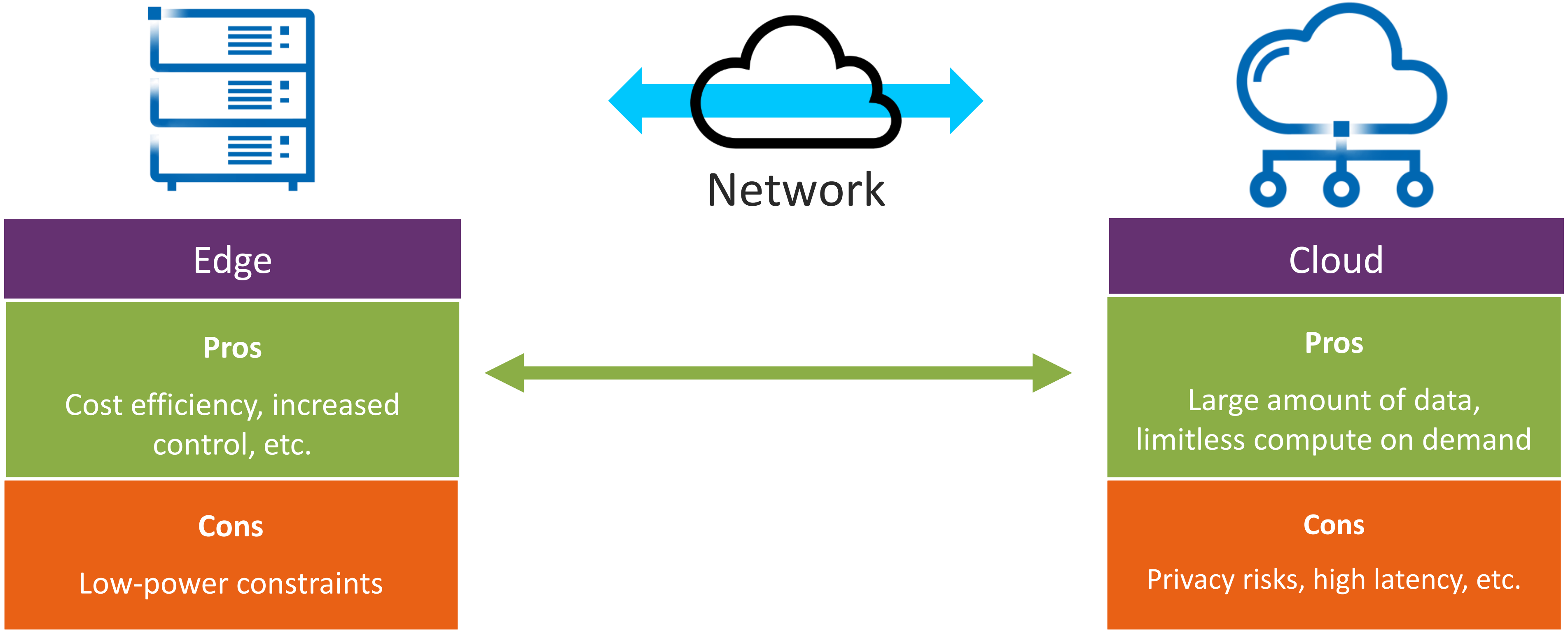
Intel® QuickAssist Technology (Intel® QAT)

© 2024 Intel

Edge to Cloud Paradigms

© 2024 Intel

Edge to Cloud: Flexibly Using Compute



Network

Edge

Pros

Cost efficiency, increased control, etc.

Cons

Low-power constraints

Cloud

Pros

Large amount of data, limitless compute on demand

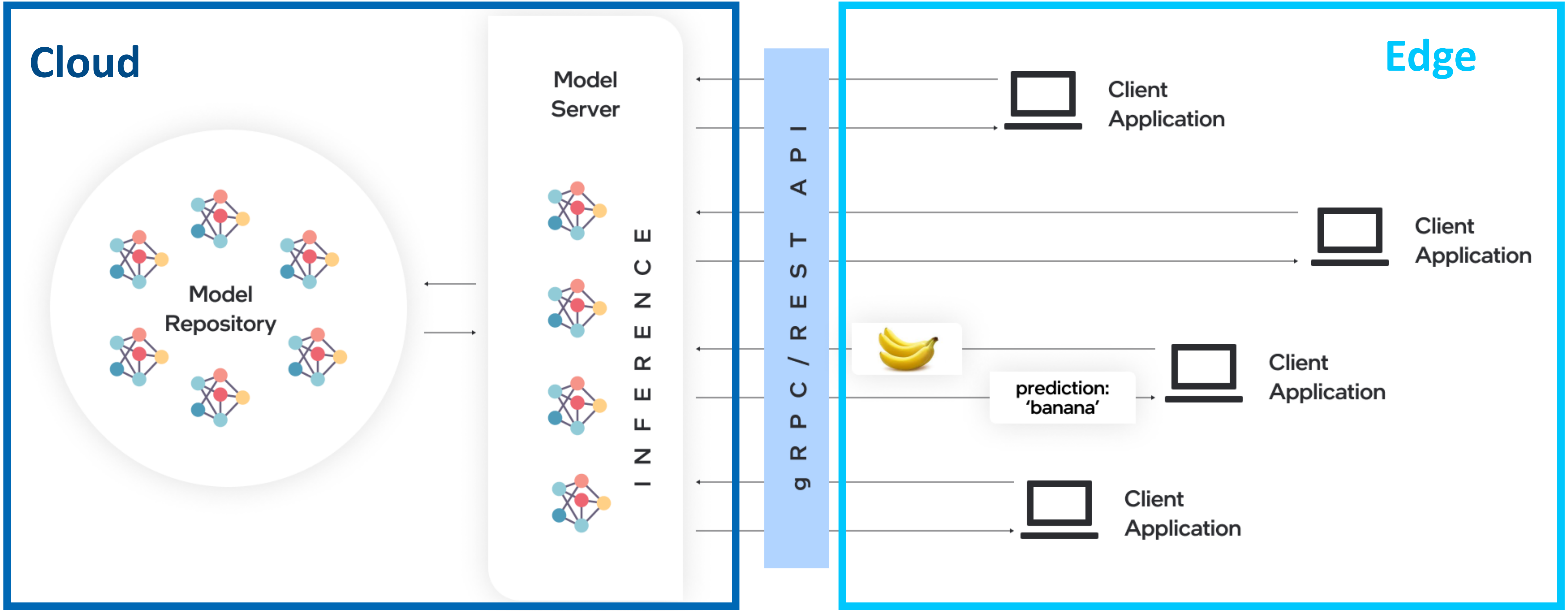
Cons

Privacy risks, high latency, etc.

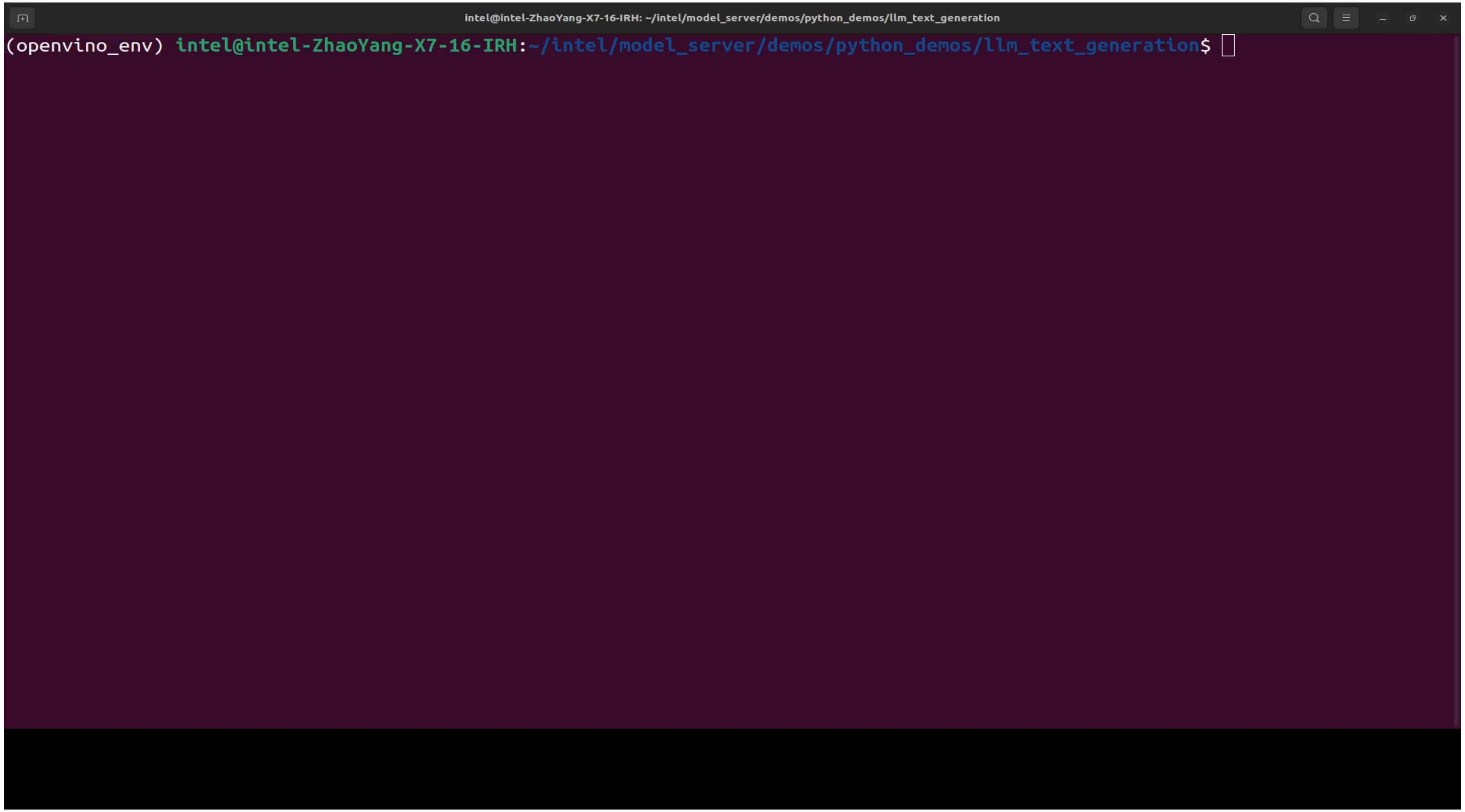
© 2024 Intel

Edge to Cloud with OpenVINO™ Model Server

Move workloads across the edge and cloud
Powered by OpenVINO™ Runtime



LLM Assistants:
OpenVINO™ Model Server with INT8 Compression


intel@intel-ZhaoYang-X7-16-IRH: ~/intel/model_server/demos/python_demos/llm_text_generation

```
(openvino_env) intel@intel-ZhaoYang-X7-16-IRH:~/intel/model_server/demos/python_demos/llm_text_generation$ 
```

Deploying a Quantized Tiny-llama model across client and server

© 2024 Intel

Intel®'s AI Hardware Portfolio

Edge AI Platforms

intel CORE ULTRA

intel CORE

intel ARC™ A370M

Partner edge platforms using Intel® Arc™ GPU

ADVANTECH
Enabling an Intelligent Planet

ADLINK TECHNOLOGY INC.

ASUS IoT

Cloud Platforms

intel XEON

intel DATA CENTER GPU FLEX SERIES

intel GAUDI

Client Platforms

intel ARC GRAPHICS

intel CORE ULTRA

Desktops

Laptops

© 2024 Intel

Conclusion

- AI at the edge is transforming enterprise intelligence
- But not without several challenges: scalability, setup, AI performance, etc.
- At Intel®, we see the full end-to-end stack as key for optimizing AI at the edge, and across the cloud to edge
- OpenVINO is an open standard, ready-to-use for building AI and Gen AI
- Try It Yourself: openvino.ai

© 2024 Intel

Resources

Resources

- openvino.ai
- intel.com/edgeai
- Demos: intel.com/openvinonotebooks
- Enterprise Security Solutions at the Edge with Intel

2024 Embedded Vision Summit

May 23rd (12:00 pm – 12:30 pm)

“Identifying and Mitigating Bias in AI”

May 23rd (1:30 pm – 2:00 pm)

“Intel’s Approach to Operationalizing AI in the Manufacturing Sector”

© 2024 Intel

intel. innovation

AI: The New Age

Solving the World's Toughest Challenges, **Together**.

Calling All Developers & Technologists!

From front-end, web, app devs to back-end, full-stack, database & DevOps to data scientists & researchers, and more:
Learn, collaborate, and solve at Intel Innovation –
an event for developers by developers.

- ✓ Hear from leading industry luminaries, technologists & start-up entrepreneurs in the field of AI.
- ✓ Get the latest AI development tools, hands-on experience & join on-site Hackathons to optimize your AI code & workflows.
- ✓ Learn the breadth of future technology advancements in AI through keynotes, sessions, birds of a feathers, and hands-on labs.
- ✓ Share unique ideas and perspectives and collaborate with your peers.

Save the Date:
September 24-25, 2024
San Jose Convention Center, CA

Opt-In for Early Access When Registration Opens!

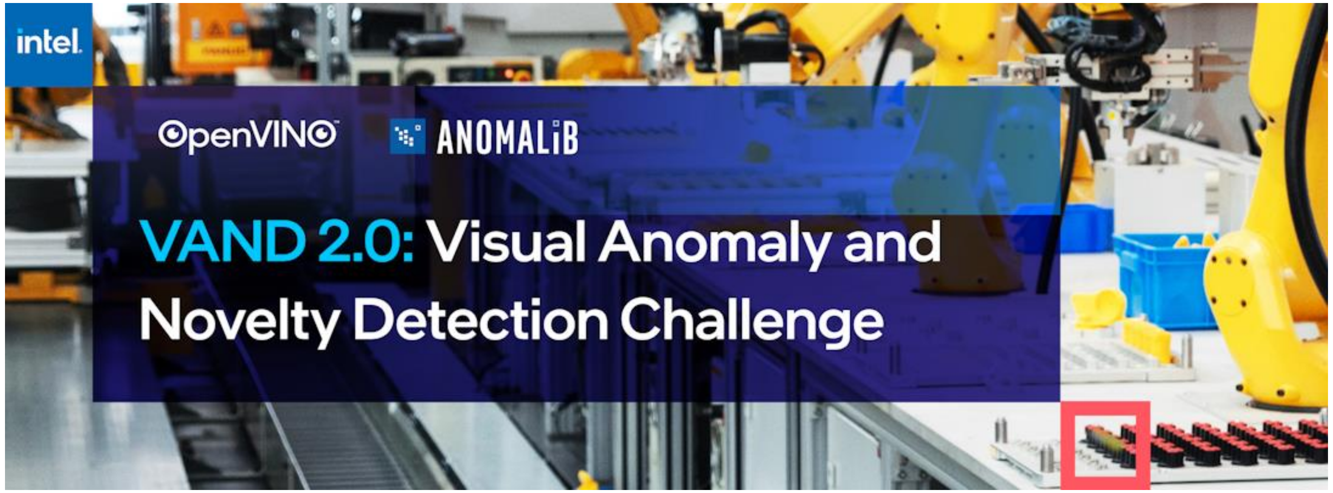
www.intel.com/innovation

Join us at CVPR!

Hackster OpenVINO Challenge

Ends June 1st

https://www.hackster.io/contests/openvino2024/



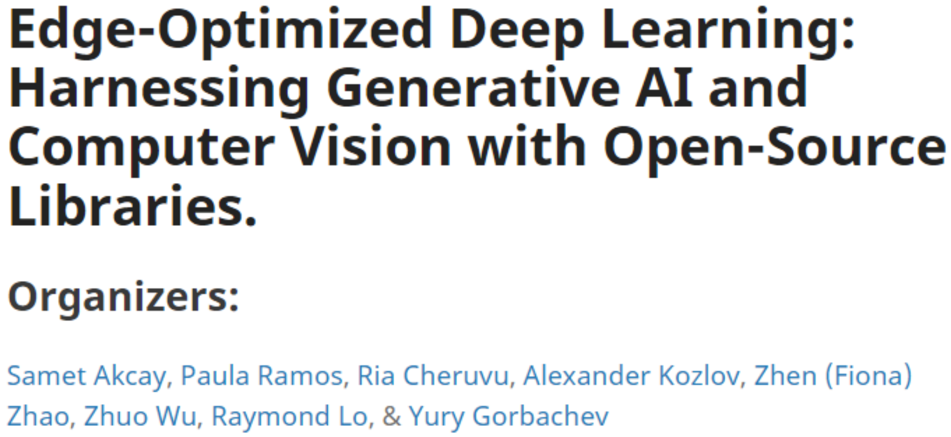
OpenVINO™ at CVPR

Tutorial Date: June 17th

https://paularamo.github.io/cvpr-2024/

CVPR JUNE 17-21, 2024 SEATTLE, WA

OpenVINO



Edge-Optimized Deep Learning: Harnessing Generative AI and Computer Vision with Open-Source Libraries.

Organizers:

Samet Akcay, Paula Ramos, Ria Cheruvu, Alexander Kozlov, Zhen (Fiona) Zhao, Zhuo Wu, Raymond Lo, & Yury Gorbachev