Generative AI at the edge

Jinyang Guo
Institute of Artificial Intelligence
Beihang University

https://shalei120.github.io/course/LLM/LLM.html

Today's agenda

Why on-device generative AI is key

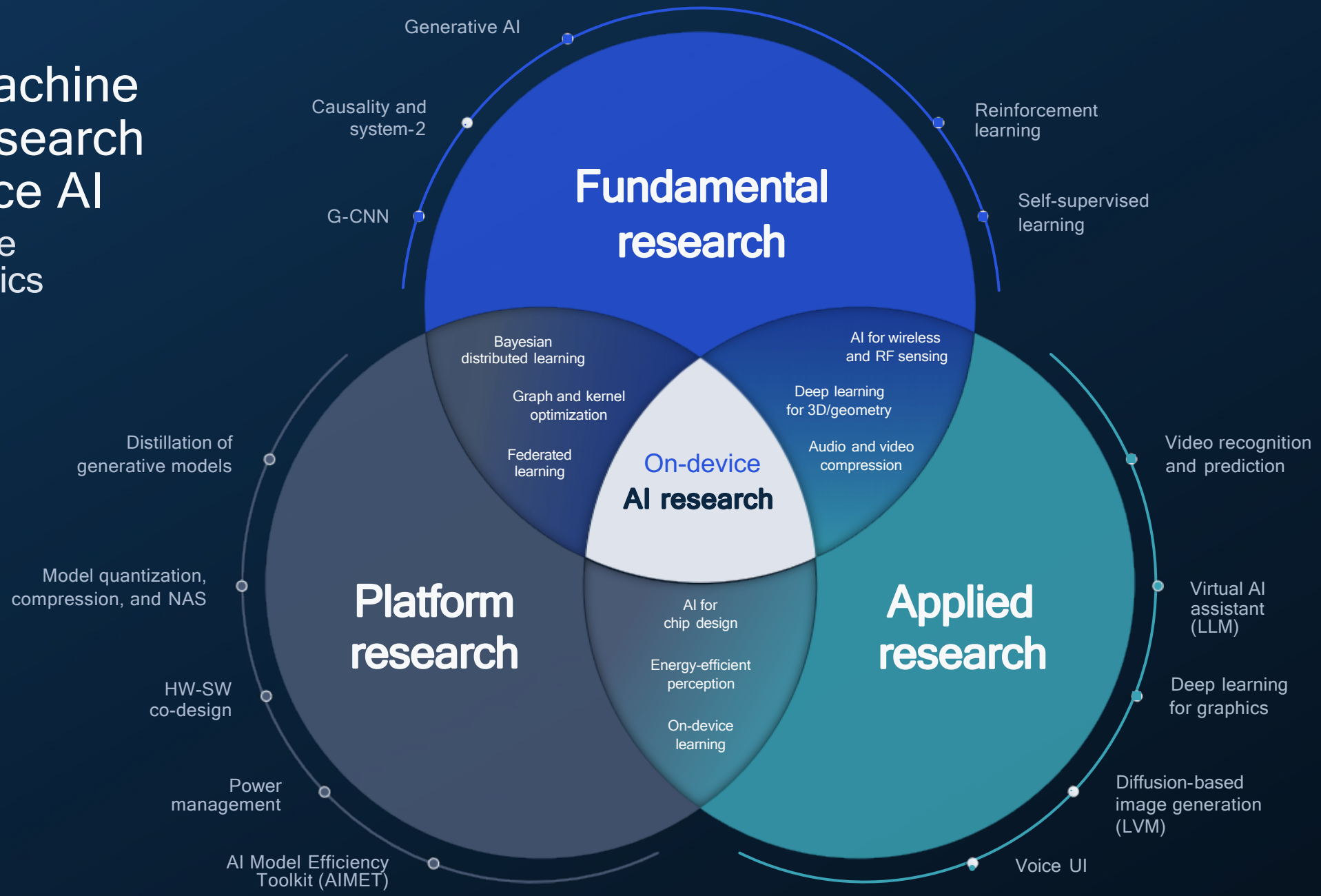
Full-stack AI optimizations for diffusion models – **Stable Diffusion**

Full-stack AI optimizations for large language models – **Llama 2**

Hybrid AI technologies and architectures

Leading machine learning research for on-device AI

across the entire spectrum of topics



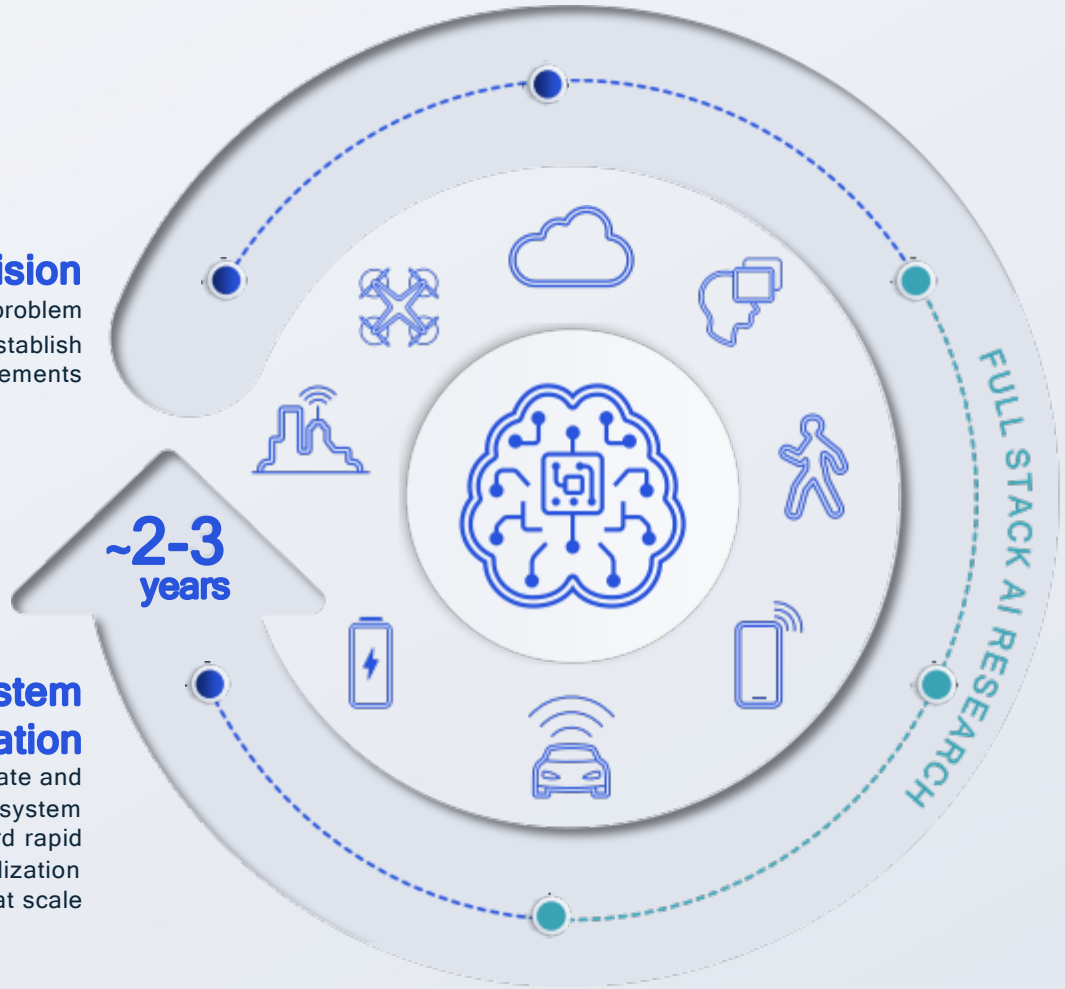
AIMET is a product of Qualcomm Innovation Center, Inc. Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
LLM: Large language mode; LVM: Language vision model

Full-stack AI research & optimization

Model, hardware, and software innovation across each layer to accelerate AI applications

Early R&D and technology inventions essential to leading the ecosystem forward

Transfer tech to commercial teams and influence future research with learnings from deployment



Invention
Invent new methods that set state-of-the-art

Model quantization & optimization
Develop tech & tools to quantize weights and modify architecture to run efficiently on hardware

Software compilation
Develop tech & tools to improve graph-level and kernel-level software compilation performance

Proof of concept
Target teams integrate models into final application for stable and intuitive demonstration

Ecosystem collaboration
Collaborate and drive the ecosystem toward rapid commercialization at scale

Vision
Identify a problem or need; establish requirements

Text generation
(ChatGPT, Bard, Llama, etc.)

Input prompts

"Write a lullaby about cats and dogs to help a child fall asleep, include a golden shepherd"

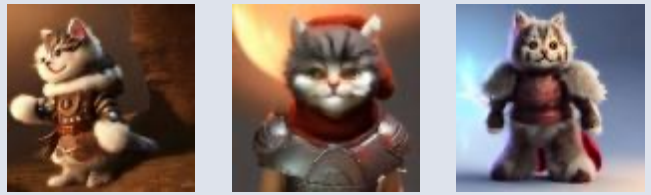
Real-life application of this platform

- Communications,
- Journalism,
- Publishing,
- Creative writing
- Writing assistance

Image generation
(Stable Diffusion, MidJourney, etc.)

Input prompts

"Super cute fluffy cat warrior in armor"



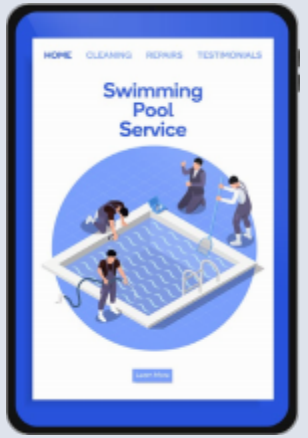
Real-life application of this platform

- Advertisements
- Published illustrations
- Corporate visuals
- Novel image generation

Code generation
(Codex, etc.)

Input prompts

"Create code for a pool cleaning website with tab for cleaning, repairs, and testimonials"



A beautiful website is created in seconds

Real-life application of this platform

- Web design
- Software development
- Coding
- Technology

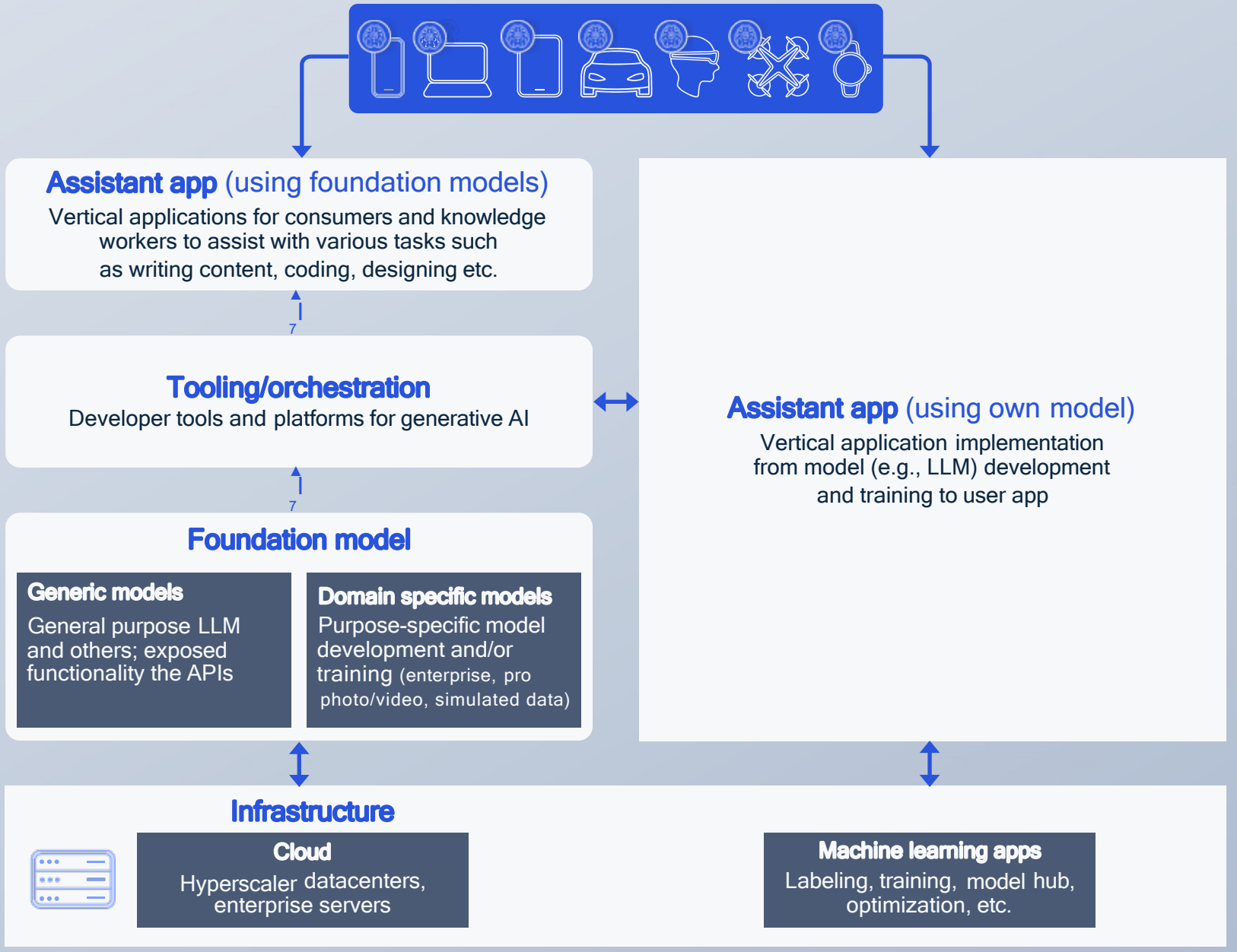
What is generative AI?

AI models that create new and original content like text, images, video, audio, or other data

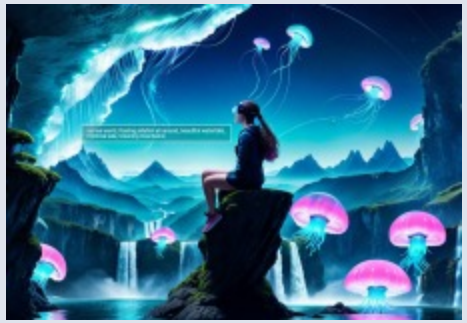
Generative AI, foundational models, and large language models are sometimes used interchangeably

The generative AI ecosystem stack

is allowing many apps to proliferate



XR

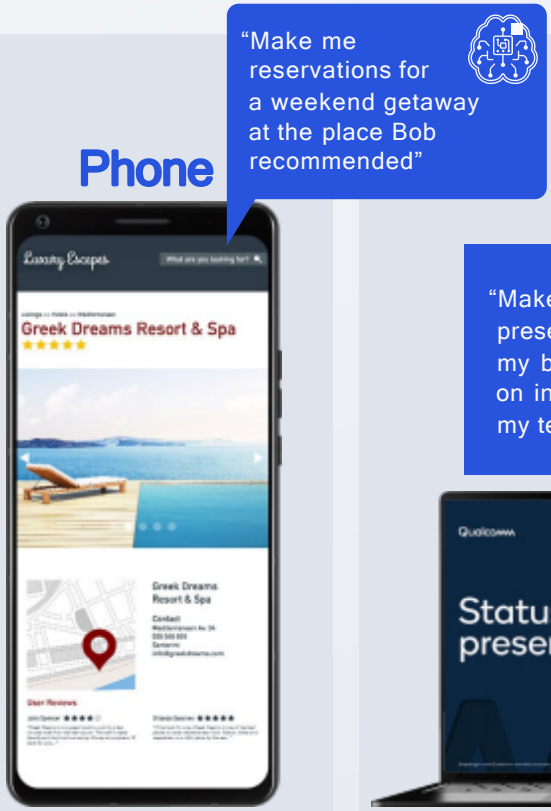


Gen AI can help create immersive 3D virtual worlds based on simple prompts

Automotive

Gen AI can be used for ADAS/AD to help improve drive policy by predicting the trajectory and behavior of various agents

Phone



Gen AI can become a true digital assistant

PC

Gen AI is transforming productivity by composing emails, creating presentations, and writing code

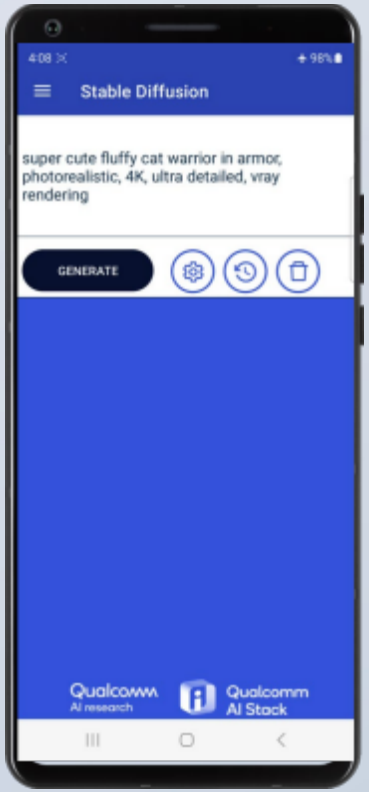
IoT

Gen AI can help improve customer and employee experience in retail, such as providing recommendations for inventory and store layout

Generative AI will impact use cases across device categories

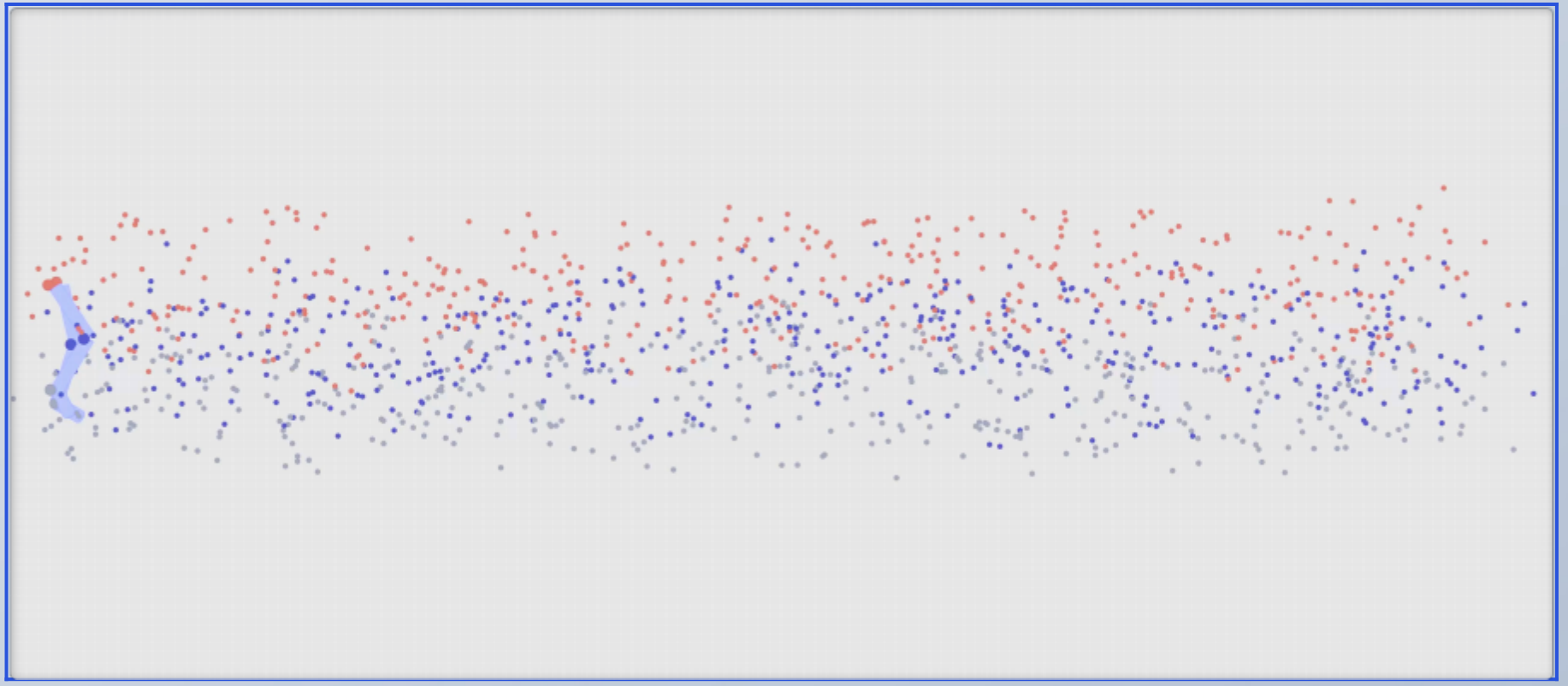
Stable Diffusion

Denoising an image with a diffusion model



Generating robot trajectories

Instead of diffusing an image we diffuse a robot trajectory



Generative AI with diffusion models for robotics path planning

What is diffusion?

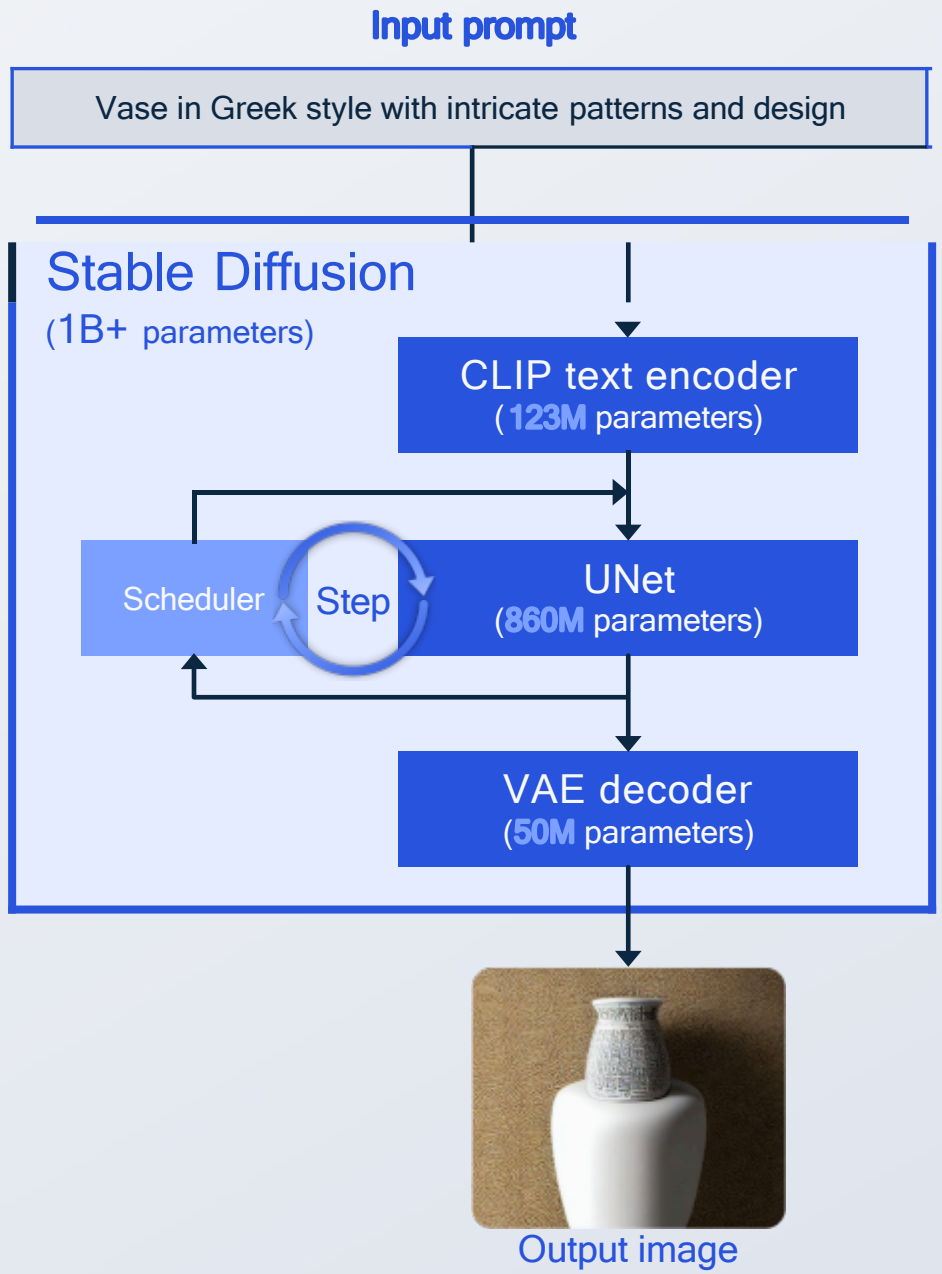
VAE: Variational Auto Encoder;
CLIP: Contrastive Language-Image Pre-Training

Stable Diffusion architecture

UNet is the biggest component model of Stable Diffusion

Many steps, often 20 or more, are used for generating high-quality images

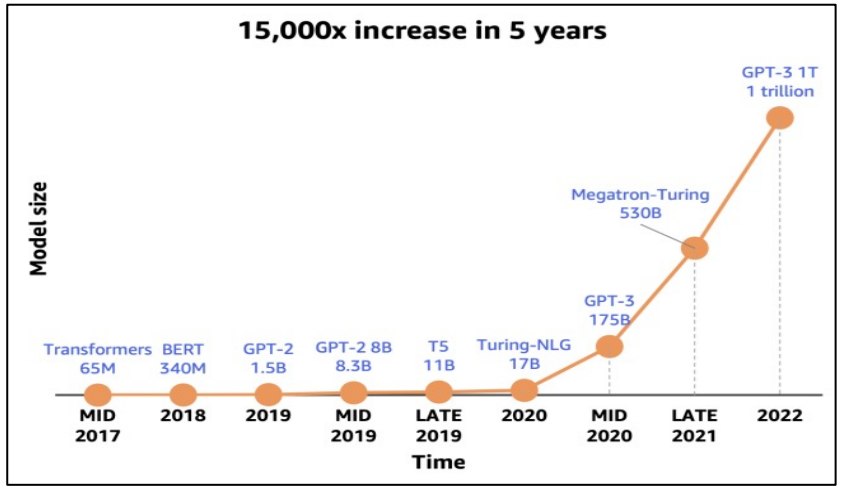
Significant compute is required

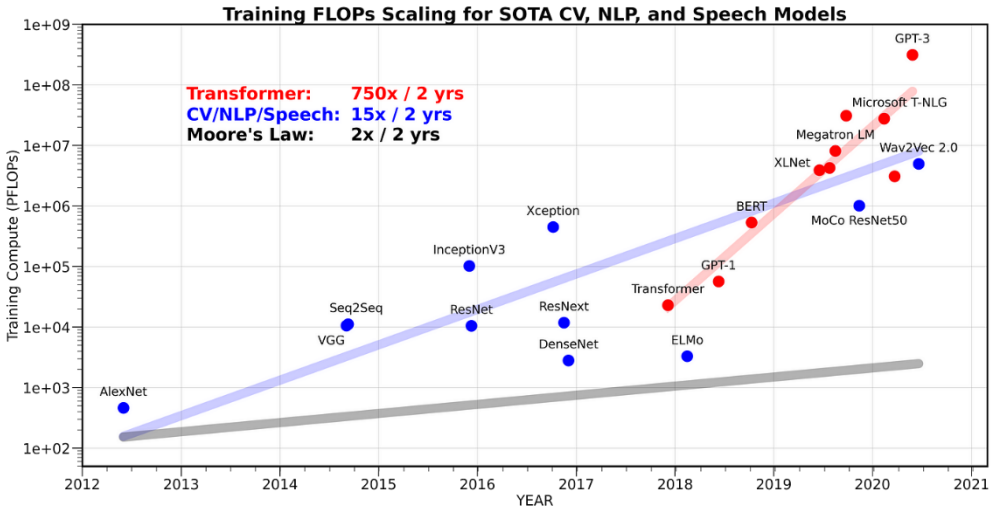


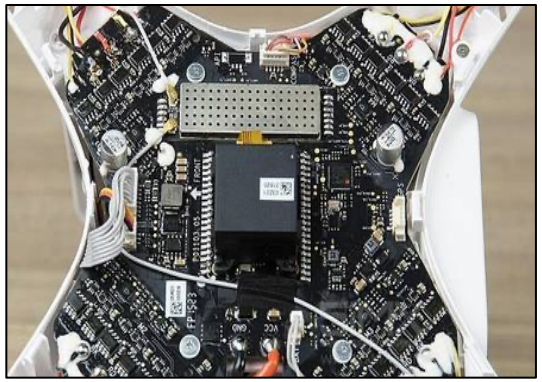
训练数据 **多**
模型参数 **多**

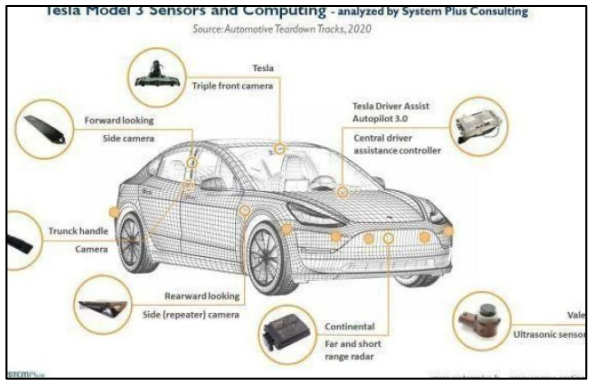
原型与部署使用存在巨大鸿沟

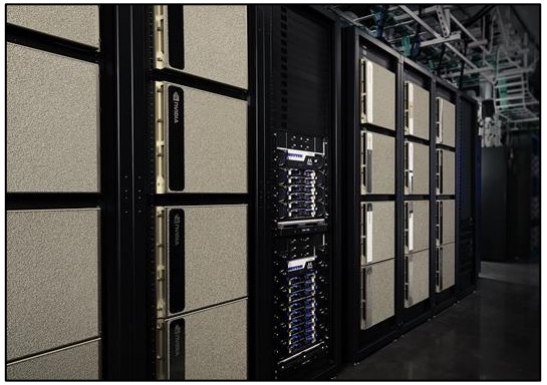
部署**场景多样**
片上**资源受限**

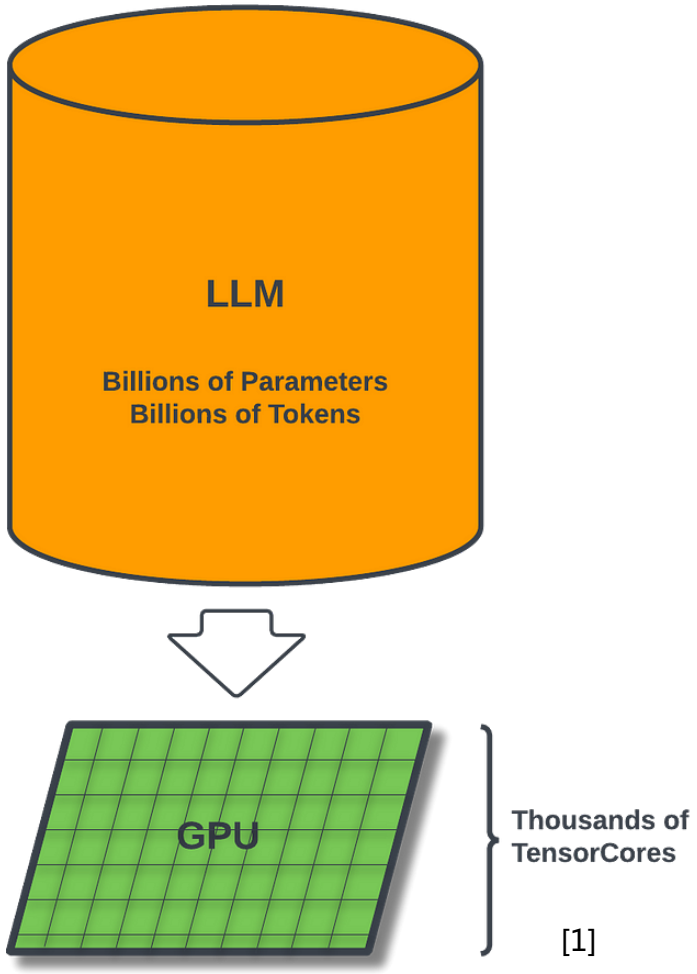








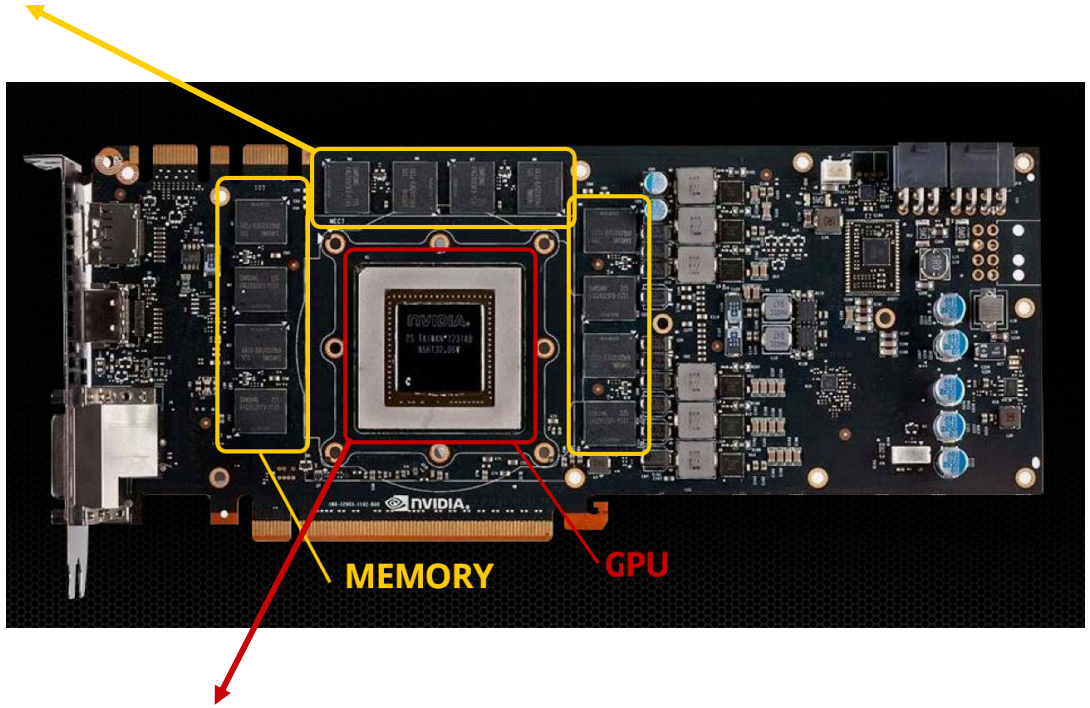




[1]

显存占用: **196.52 GB** (FP16/BF16, w/o datasets) *vs.* 141GB (Nvidia H200)

LLaMA-13B
hidden_size: 5120
model_max_length: 4096
intermediate_size: 13696
num_attention_heads: 40
num_hidden_layers: 40
batch_size: 1



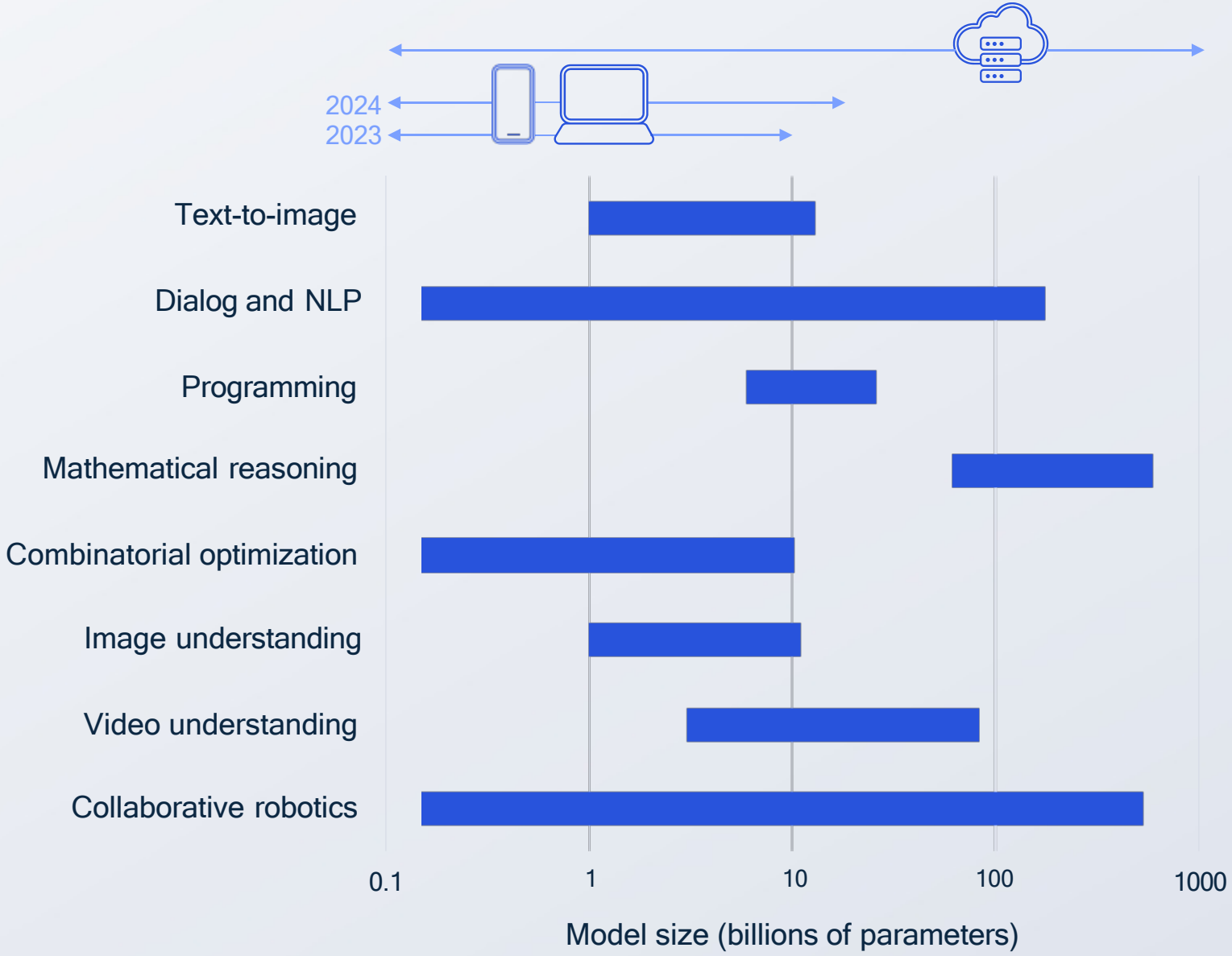
训练时间: **135,168 GPU 小时** (NVIDIA A100-80G) [2]

[1] Performance bottlenecks in deploying LLMs—a primer for ML researchers. 2023
[2] LLaMA: Open and Efficient Foundation Language Models. 2023

On-device AI can support a variety of Gen AI models

A broad number of Gen AI capabilities can run on device using models that range from **1 to 10 billion** parameters

We can run models with over **1 billion parameters on device today** and anticipate this growing to **over 10 billion parameters in the coming months**



Assuming INT4 parameters

Knowledge distillation

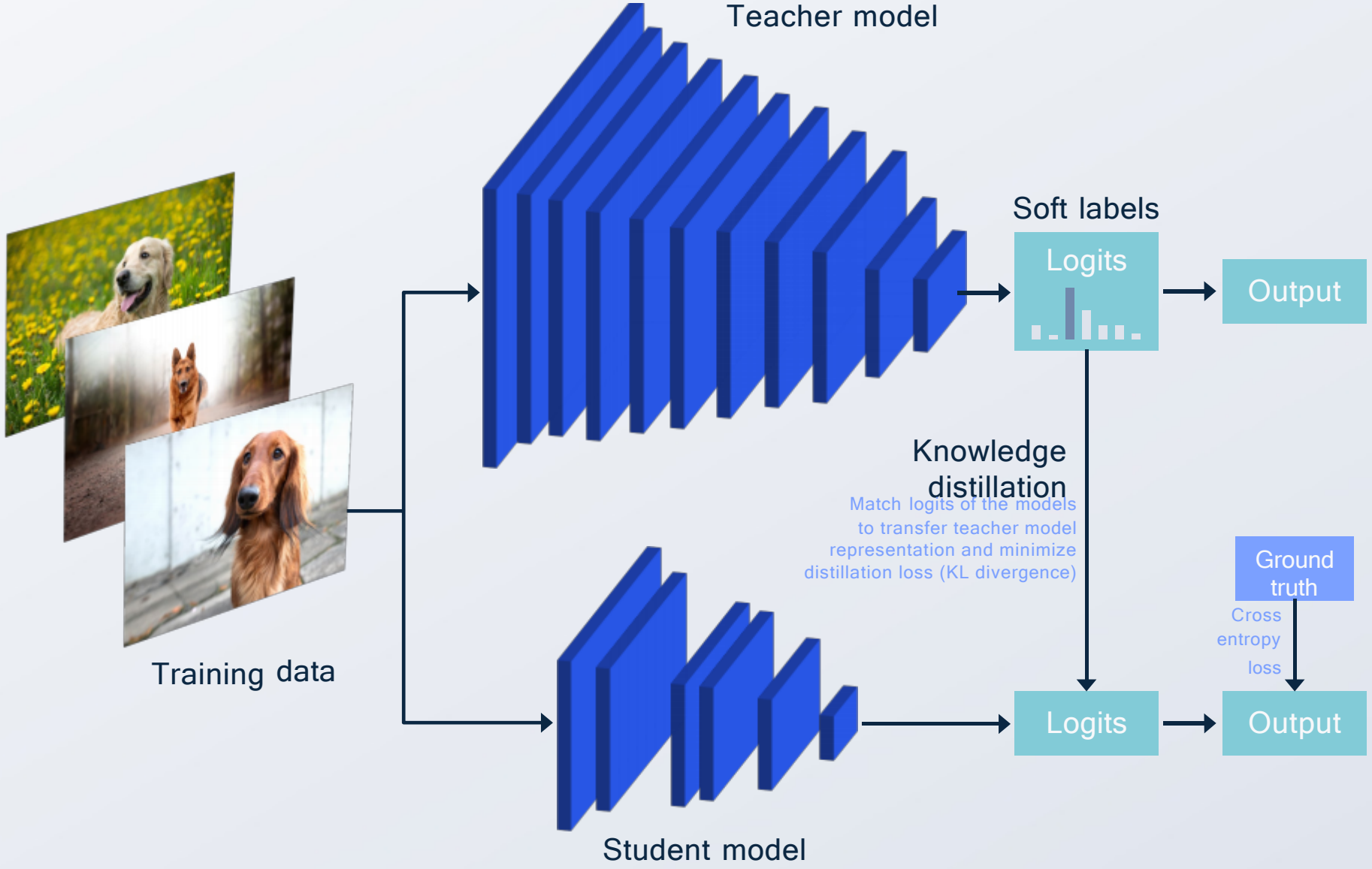
Training a smaller “student” model to mimic a larger “teacher” model

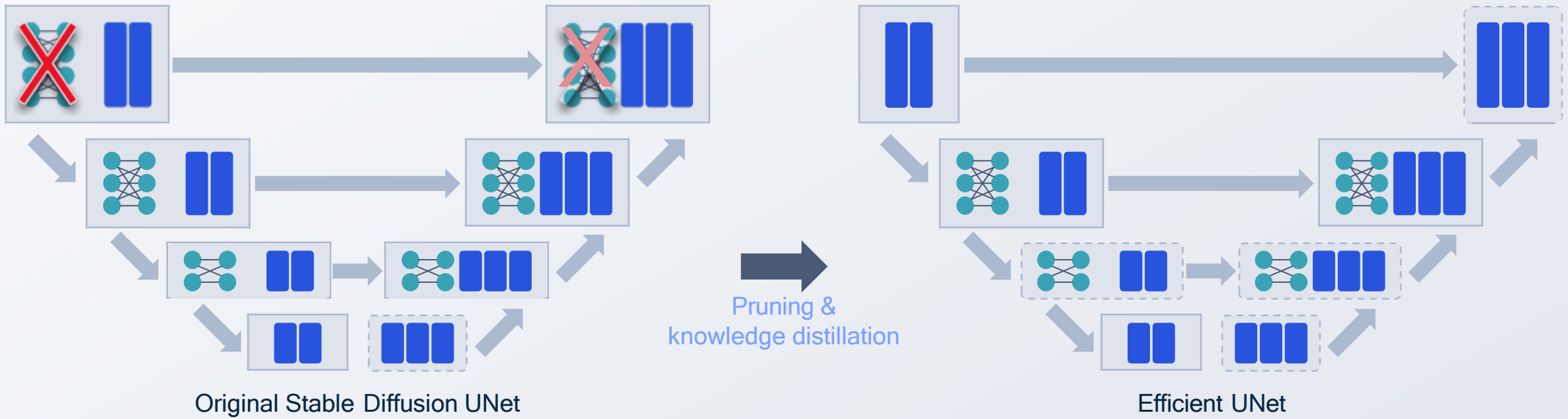
Create a smaller model with fewer parameters

Run faster inference on target deployment

Maintain prediction quality close to the teacher

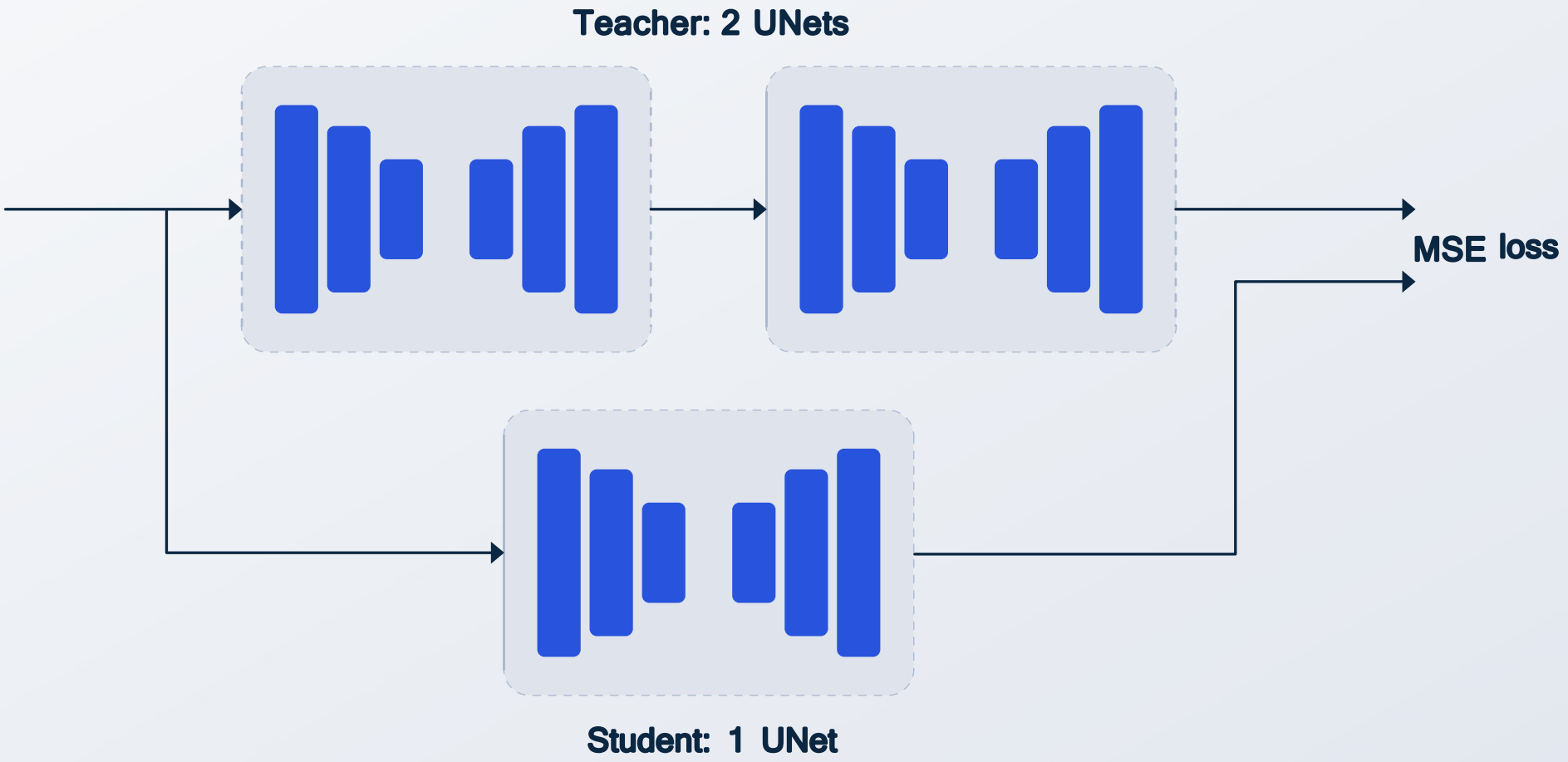
Less training time





More efficient architecture design through pruning and knowledge distillation

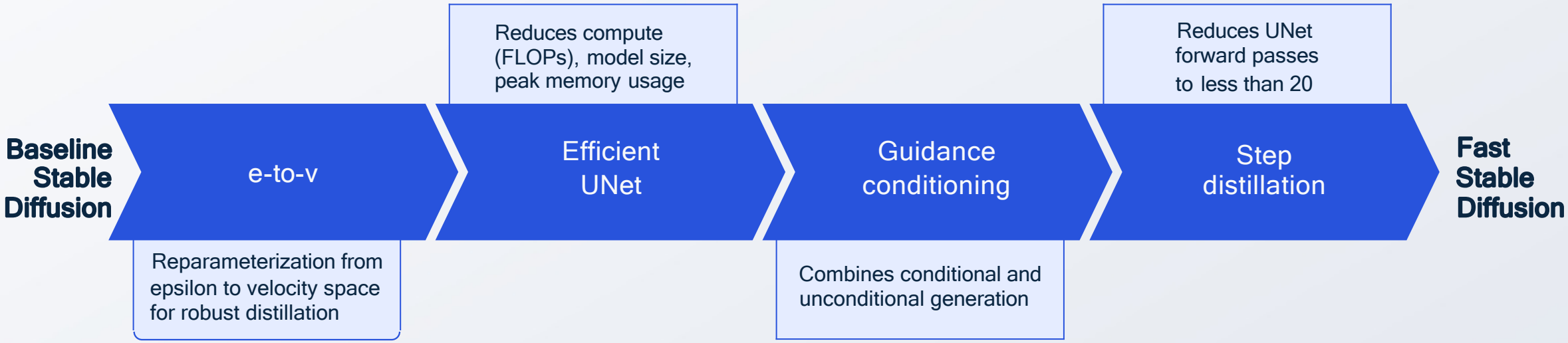
Reducing UNet compute (FLOPs), model size, and peak memory usage



Step distillation for the DDIM scheduler

Teach the student model to achieve in one step what the teacher achieves in multiple steps

DDIM: Denoising Diffusion Implicit Models; MSE: Mean-squared error



	FID↓	CLIP ↑	Inference latency
Baseline (SD- 1.5)	17.14*	0.3037	5.05 seconds
Fast SD	20.08	0.3004	0.56 seconds

9x
speedup vs baseline Stable Diffusion

Our full-stack AI optimization of Stable Diffusion significantly improves latency while maintaining accuracy

*: These results are not directly comparable since baseline Stable Diffusion was trained with over 20x larger dataset than fast Stable Diffusion. SD: Stable Diffusion

Fast Stable Diffusion

Panoramic view of mountains of Vestrahorn and perfect reflection in shallow water, soon after sunrise, Stokksnes, South Iceland, Polar Regions, natural lighting

A hyper realistic photo of a beautiful cabin inside of a forest and full of trees and plants, with large aurora borealis in the sky

Underwater world, plants, flowers, shells, creatures, high detail, sharp focus, 4k

High quality colored pencil sketch portrait of an anthro furry fursona blue fox, handsome eyes, sketch doodles surrounding it, photo of notebook sketch

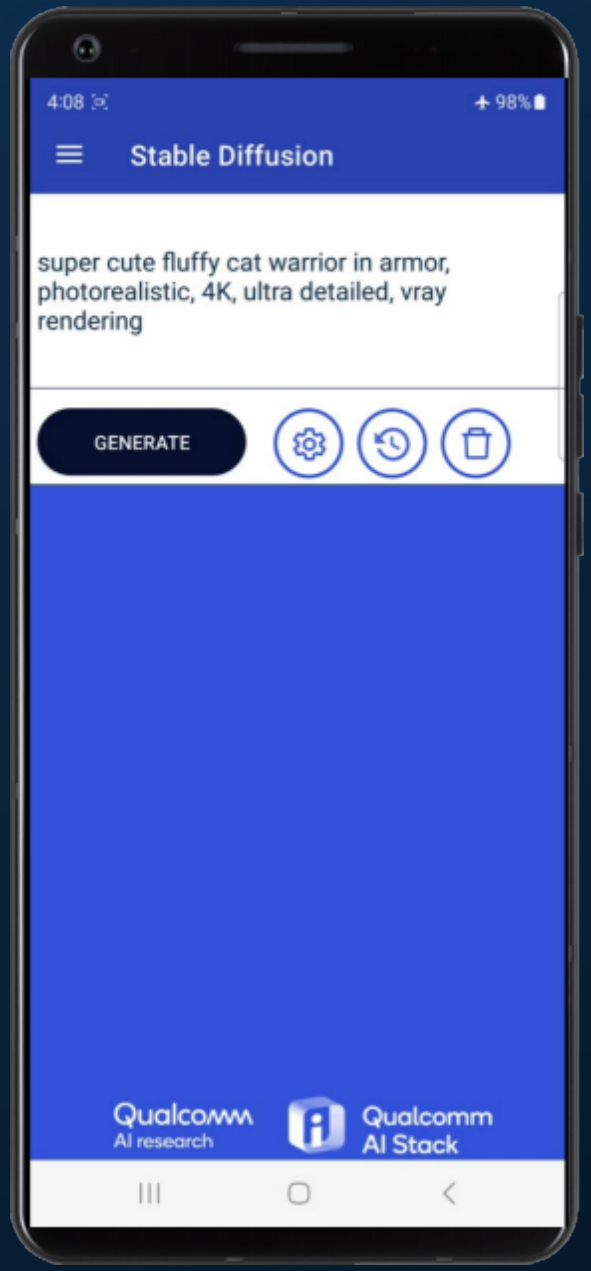
Japanese garden at wildlife river and mountain range, highly detailed, digital illustration

Stable Diffusion V1.5

Similar image quality between our fast implementation and baseline model

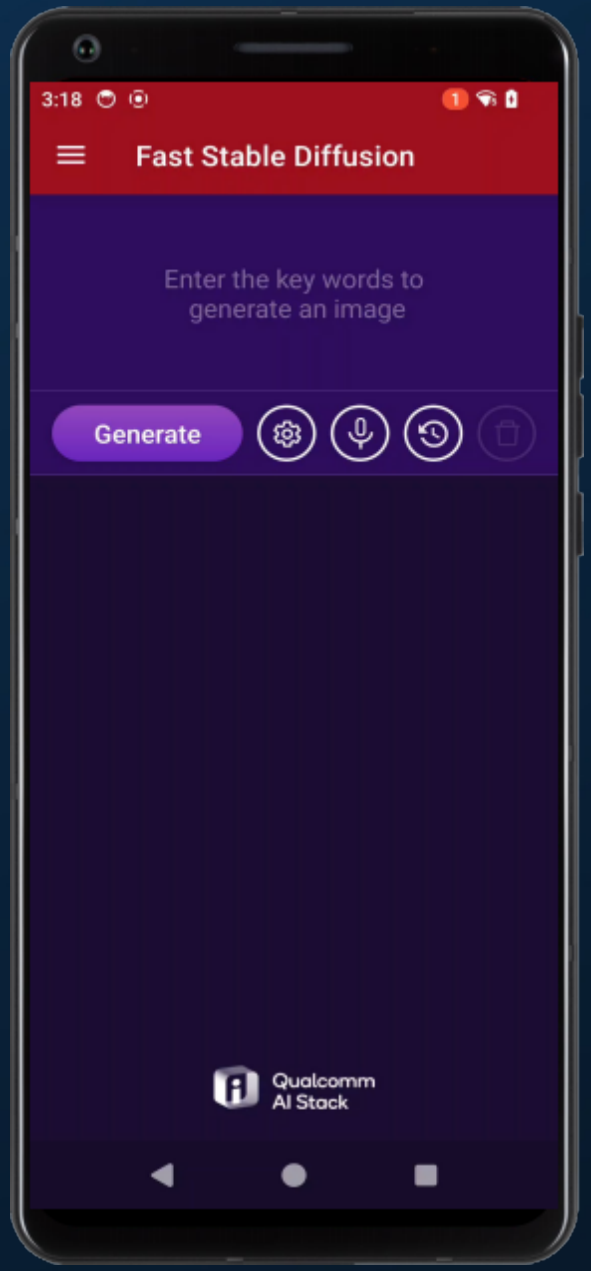
At MWC 2023

World's first on-device demo of Stable Diffusion running on an Android phone



- 1B+ parameter generative AI model runs efficiently and interactively
- Full-stack AI optimization to achieve sub-15 second latency for 20 inference steps
- Enhanced privacy, security, reliability, and cost with on-device processing
- Fast development enabled by Qualcomm AI Research and Qualcomm® AI Stack

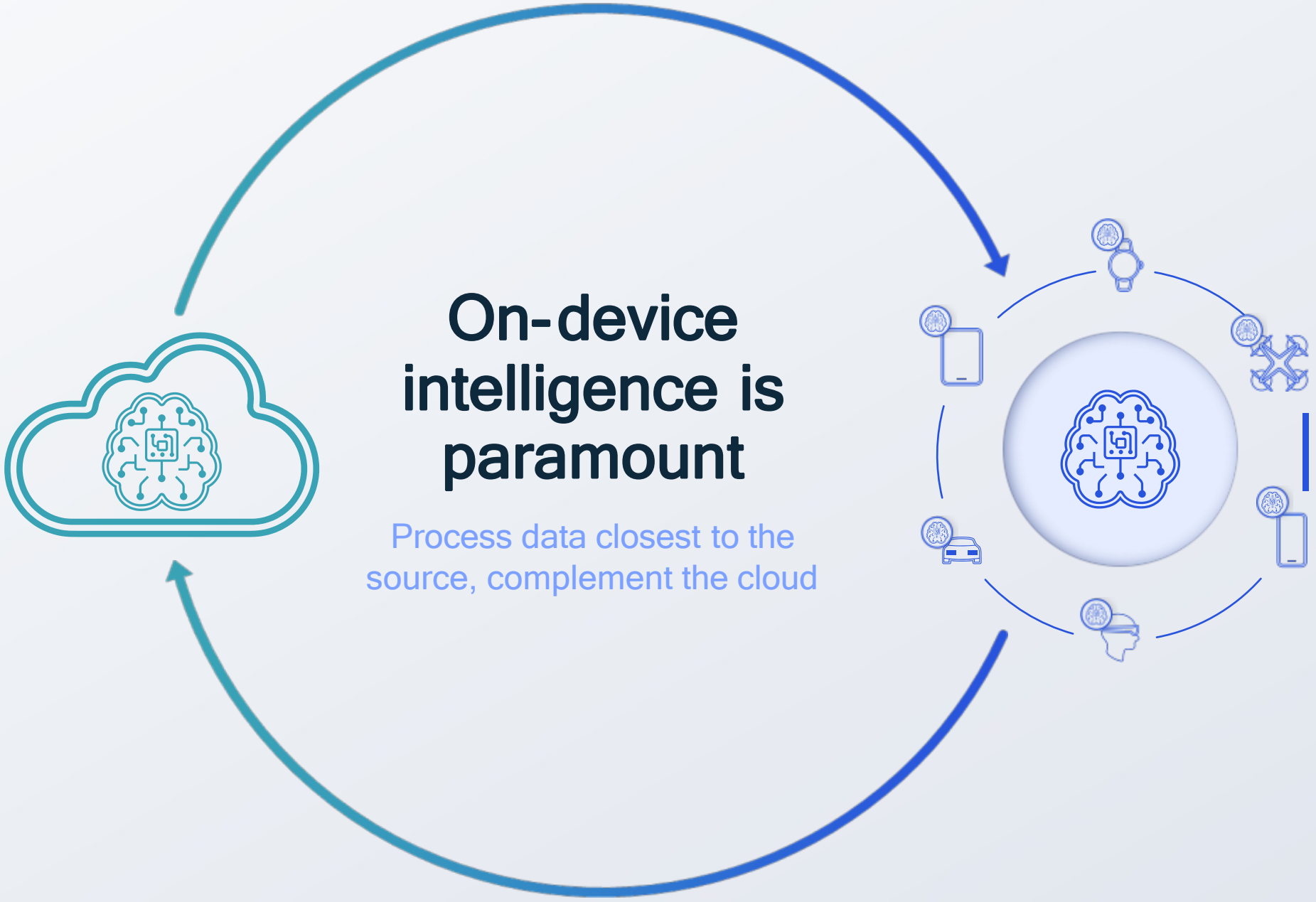
World's fastest AI text-to-image generative AI on a phone



Takes less than 0.6 seconds for generating 512x512 images from text prompts

Efficient UNet architecture, guidance conditioning, and step distillation

Full-stack AI optimization to achieve this improvement



On-device intelligence is paramount

Process data closest to the source, complement the cloud

- Privacy
- Reliability
- Low latency
- Cost
- Energy
- Personalization

Full-stack AI optimization
for LVM

Runs completely
on the device

Significantly reduces
runtime latency and
power consumption

Continuously improves
the Qualcomm® AI Stack

Designing an efficient diffusion model through knowledge distillation for high accuracy

Knowledge distillation for pruning and removing of attention blocks, resulting in accurate model with improved performance and power efficiency

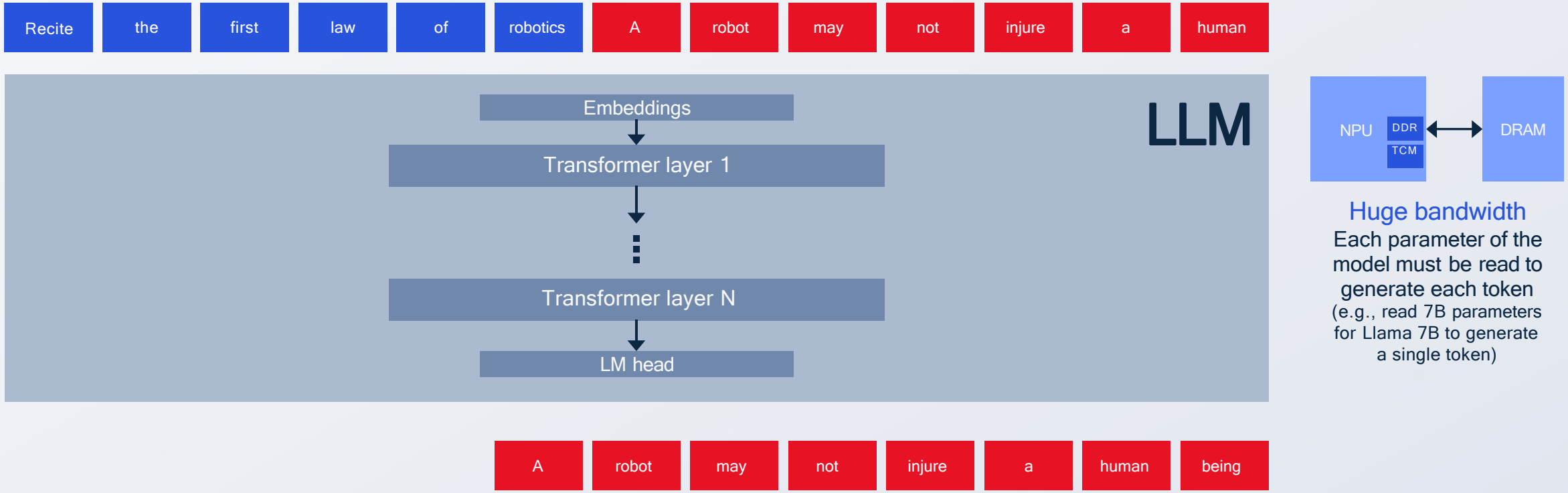
Qualcomm® AI Engine direct for improved performance and minimized memory spillage

AI acceleration on the Qualcomm® Hexagon™ NPU of the Snapdragon® 8 Gen 3 Mobile Processor

LVM: Language vision model

Illustration of autoregressive language modeling

Single-token generation architecture of large languages models results in high memory bandwidth



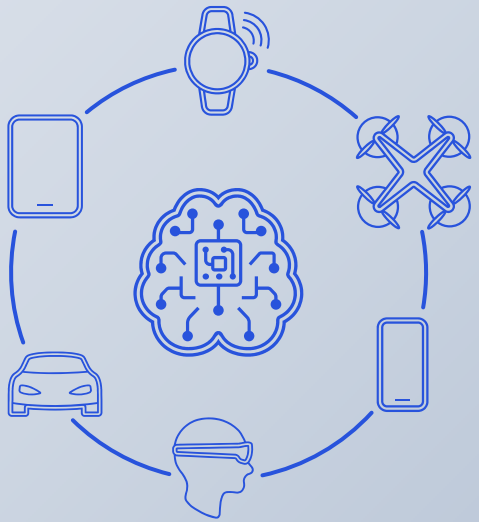
Huge bandwidth

Each parameter of the model must be read to generate each token (e.g., read 7B parameters for Llama 7B to generate a single token)

LLMs are highly bandwidth limited rather than compute limited

LLM quantization motivations

- A 4x smaller model (i.e., FP16 -> INT4)
- Reduce memory bandwidth and storage
- Reduce latency
- Reduce power consumption



Shrinking an LLM for increased performance while maintaining accuracy is challenging

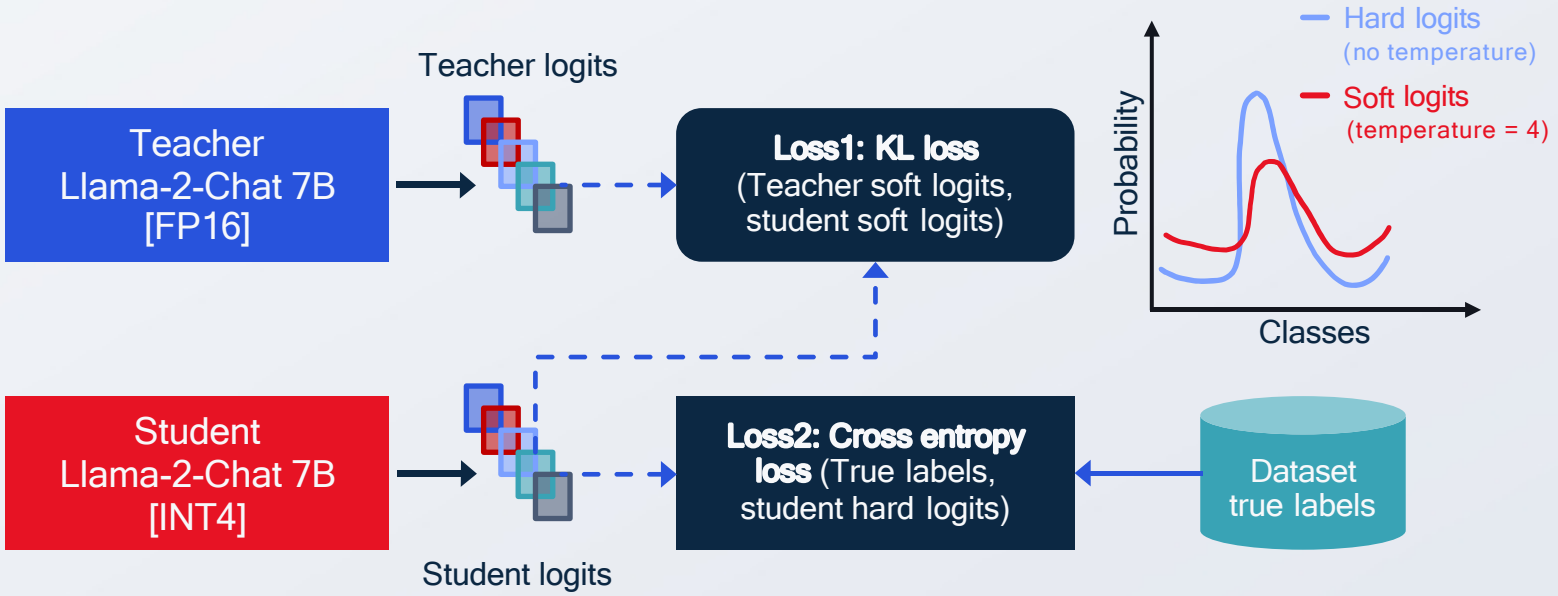
LLM quantization challenges

- Maintain accuracy of FP published models
- Post-training quantization (PTQ) may not be accurate enough for 4-bit
- The training pipeline (e.g., data or rewards) is not available for quantization aware training (QAT)

Quantization-aware training with knowledge distillation

Reduces memory footprint while solving quantization challenges of maintaining model accuracy and the lack of original training pipeline

Construct a training loop that can run two models on the same input data



KD loss function combines the KL divergence loss and hard-label based CE loss

<1
Point increase in perplexity[1]

<1%
Decrease in accuracy

1: Perplexity is average over several test sets, including wikitext and c4 (subset)

Speculative decoding

speeds up token rate by trading off compute for bandwidth

Recite the first law of robotics A robot may not injure a human

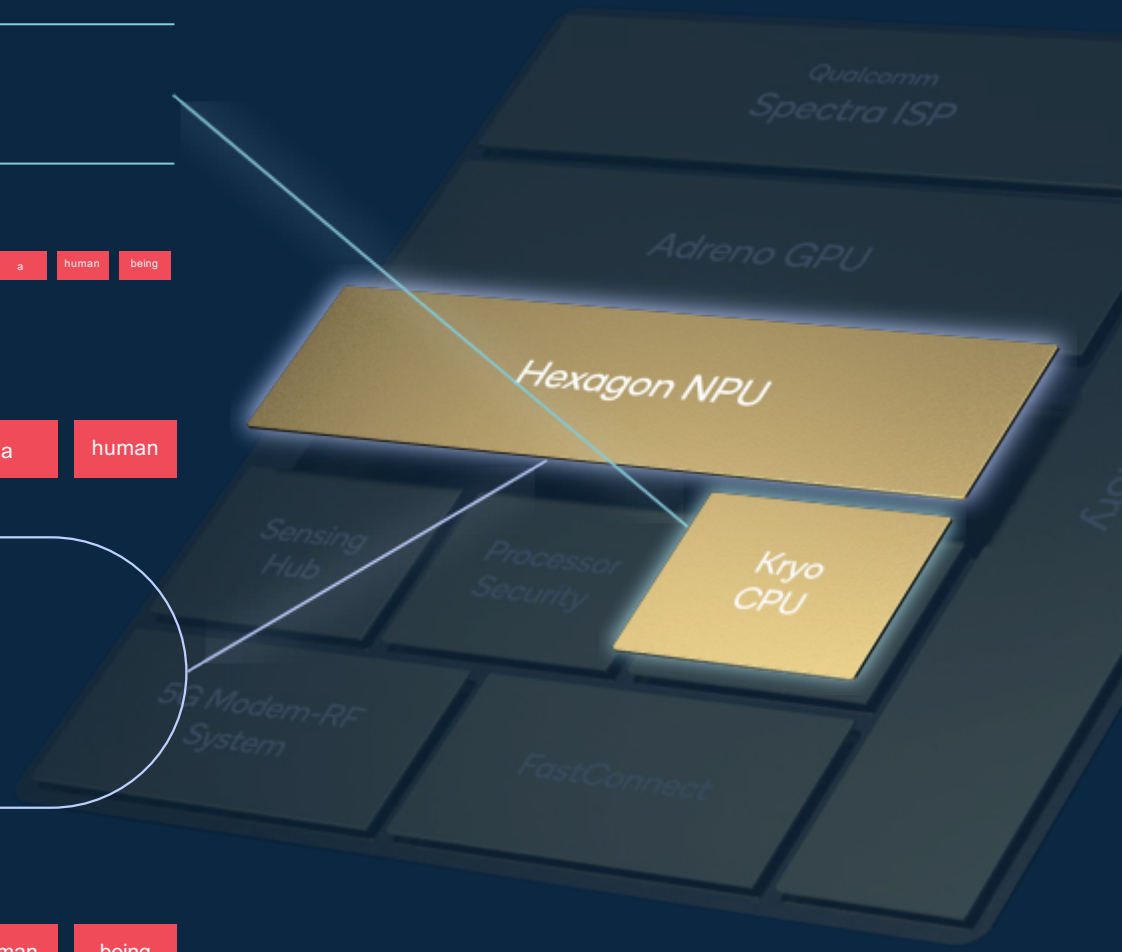
Llama 2 draft

A robot may not injure a human being

Recite the first law of robotics A robot may not injure a human

Llama 2

A robot may not injure a human being



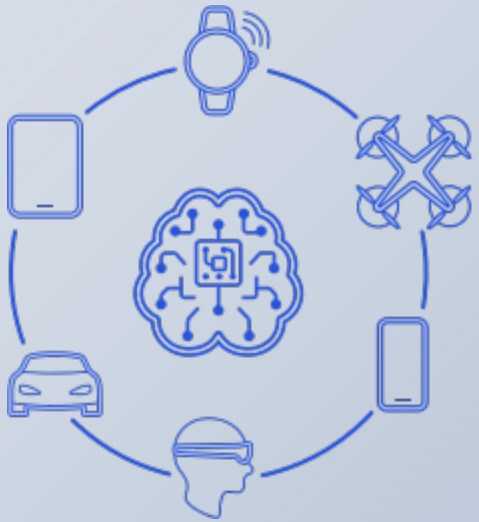
Draft model generates a few speculative tokens at a time

Target model decides which to accept in one pass

A good draft model predicts with a high acceptance rate

Small draft model motivations

- 10x smaller draft model than target model
- Fast results
- Reduce memory bandwidth, storage, latency, and power consumption



Train a significantly smaller draft LLM for speculative decoding while maintaining enough accuracy is challenging

Small draft model challenges

- The training pipeline (e.g., data or rewards) is not available
- Cover multiple families, e.g., 7B and 13B models
- Match the distribution of the target model for higher acceptance rate

Speculative decoding provides speedup with no accuracy loss

Using our research techniques on Llama 2-7B Chat, we achieved

AI assistant enables basic chat and chat-assisted apps on device

Orchestration across different tasks based on user query

Powered by Llama 2 Chat (7B)

Voice UI with Snapdragon Voice Activation and Whisper-Small (244M)

User interface
Voice/text/browser

Snapdragon Voice Activation & Whisper-Small
(~244M param)

Task classification

miniLM
(~33M param)

Orchestrator

Llama 2 Chat
(7B param)

Travel planner API

LM: Language model

AI Assistant based on Llama 2

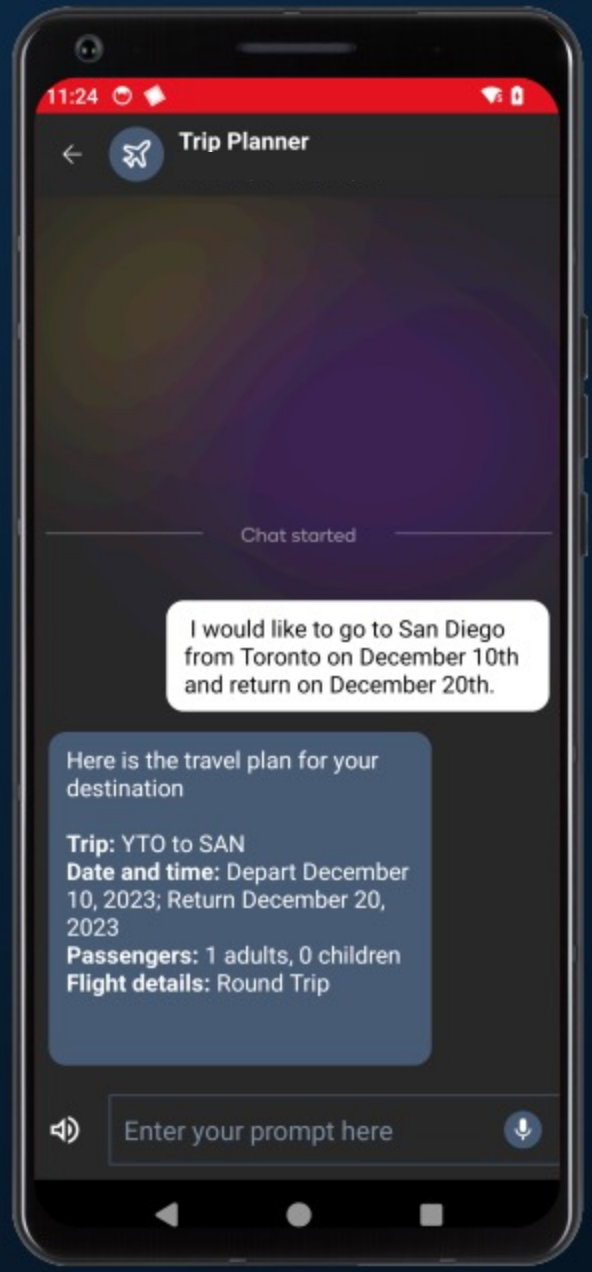
At Snapdragon Summit 2023

World's fastest Llama 2-7B on a phone

Up to 20 tokens per second

Demonstrating both chat and application interaction on device

World's first demonstration of speculative decoding running on a phone



Full-stack AI optimization
for LLM

Runs completely
on the device

Significantly reduces
runtime latency and
power consumption

Continuously improves
the Qualcomm® AI Stack

Designing a good draft model for given target model through knowledge distillation for high acceptance and no accuracy loss

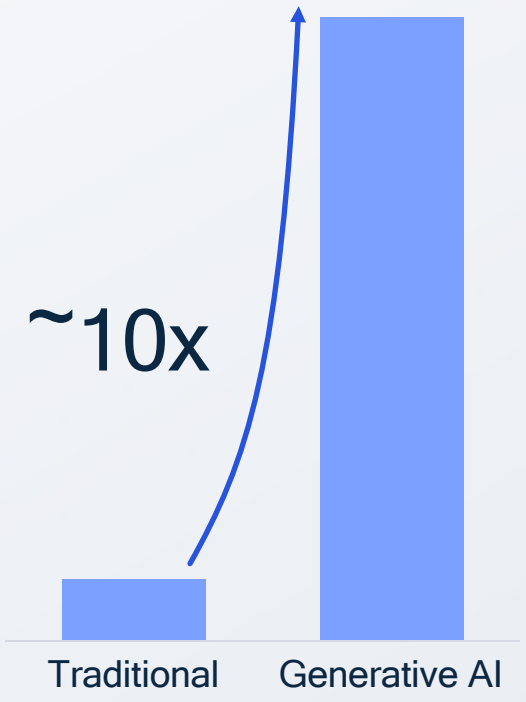
QAT with knowledge distillation for accurate INT4 target LLM for improved performance and power efficiency

Qualcomm AI Engine direct for improved performance and minimized memory spillage

AI acceleration on the Qualcomm® Hexagon™ NPU of the Snapdragon® 8 Gen 3 Mobile Processor

QAT: Quantization-aware training; LLM: Large language model

Cost per query[1] (e.g. web search) × Gen AI applications × Billions of users



Personal assistant

Web search

Image & video creation

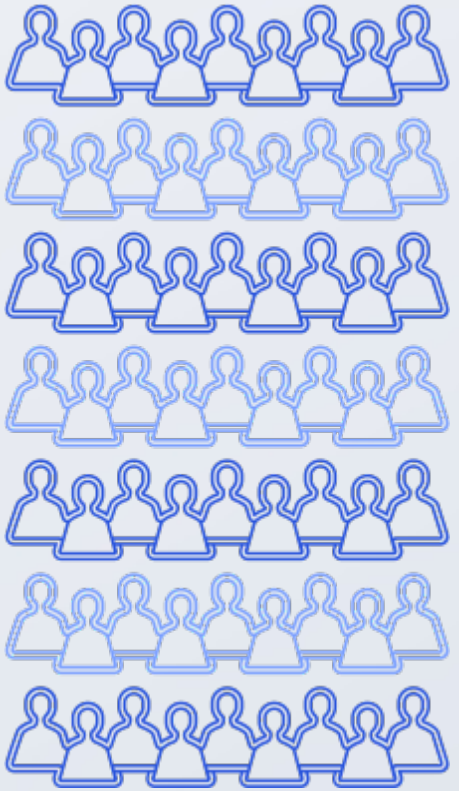
Coding assistant

Text summarization

Conversational chatbots

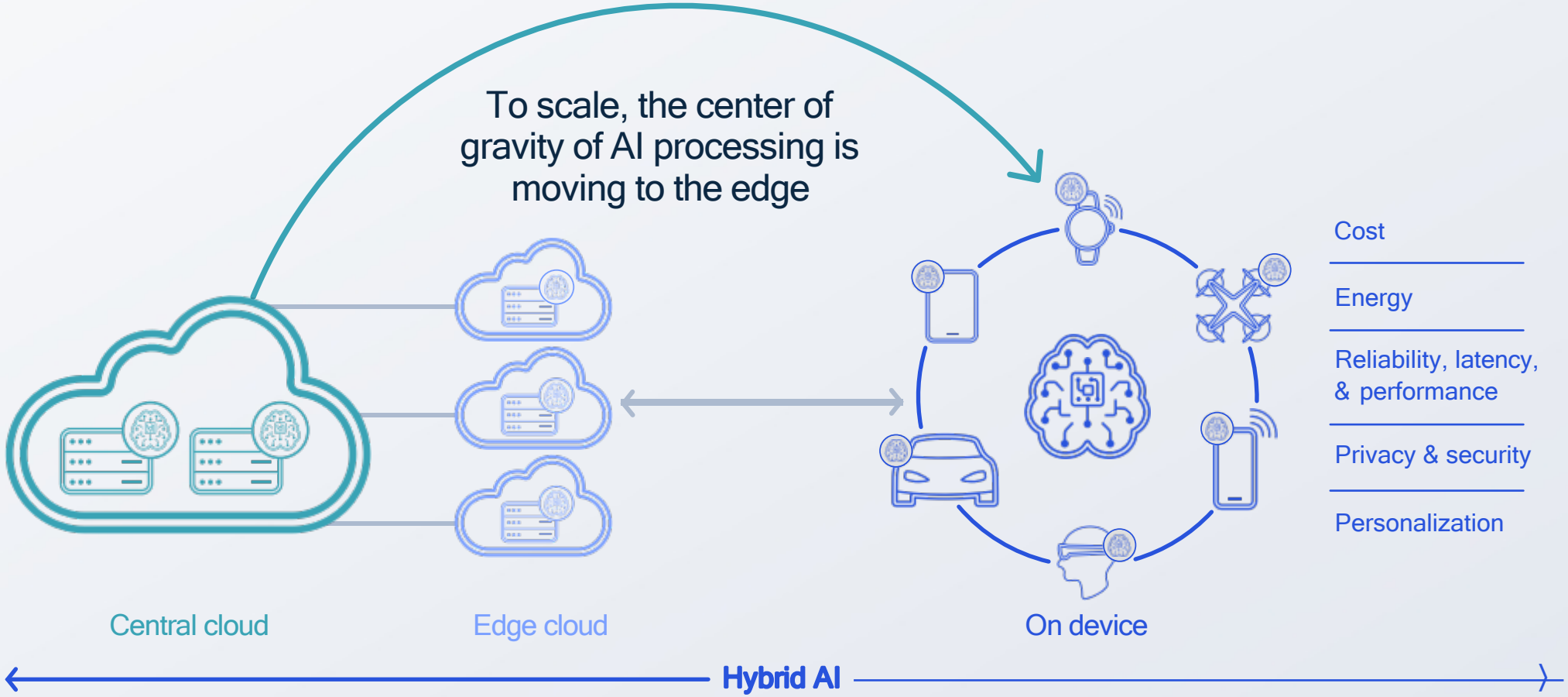
Copy creation

...



Cloud economics will not allow generative AI to scale

1: Reuters 2023



We are a leader in the realization of the hybrid AI

Convergence of:

Wireless connectivity
Efficient computing
Distributed AI

Unlocking the data that will fuel our digital future and generative AI

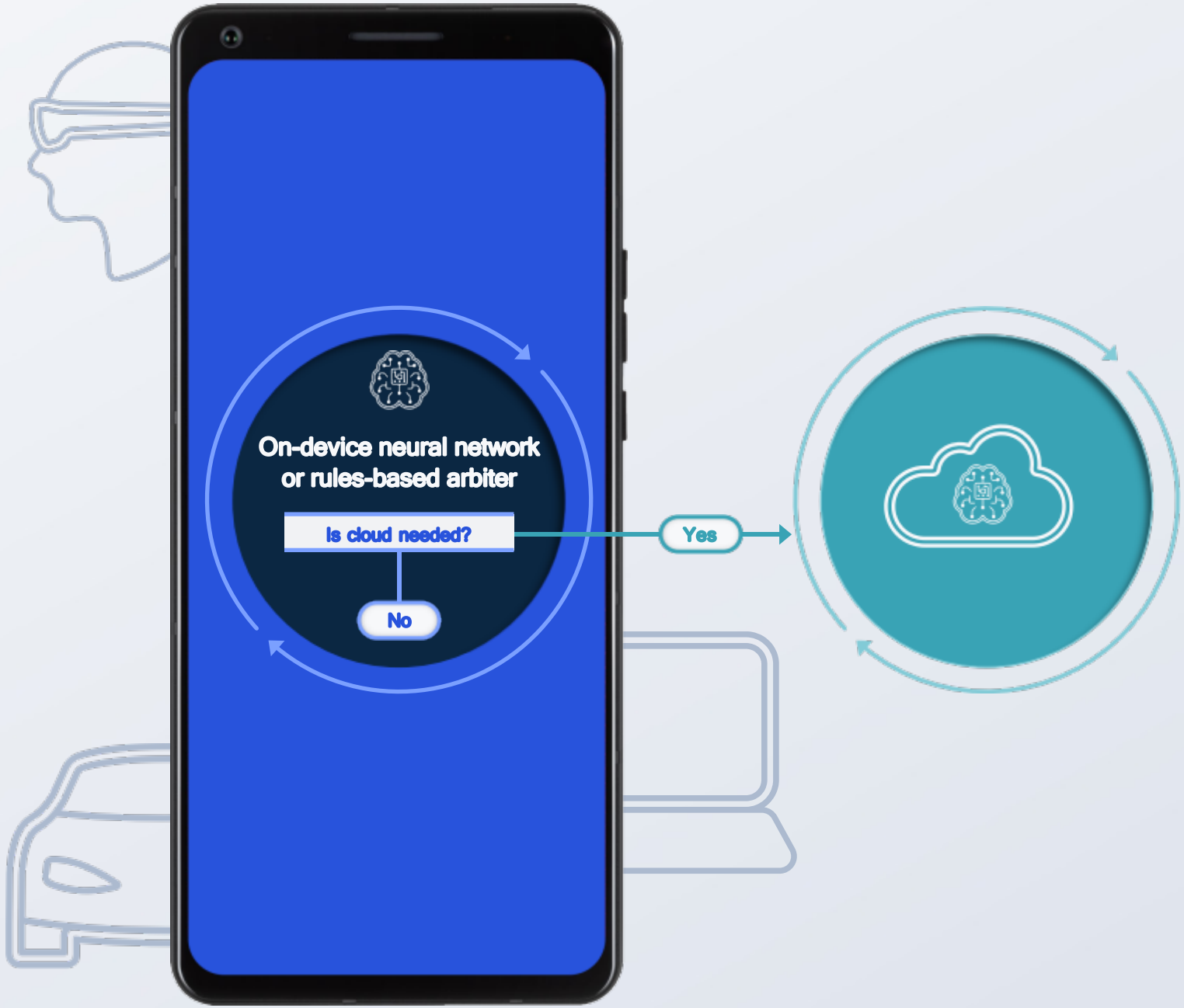
Device-centric hybrid AI

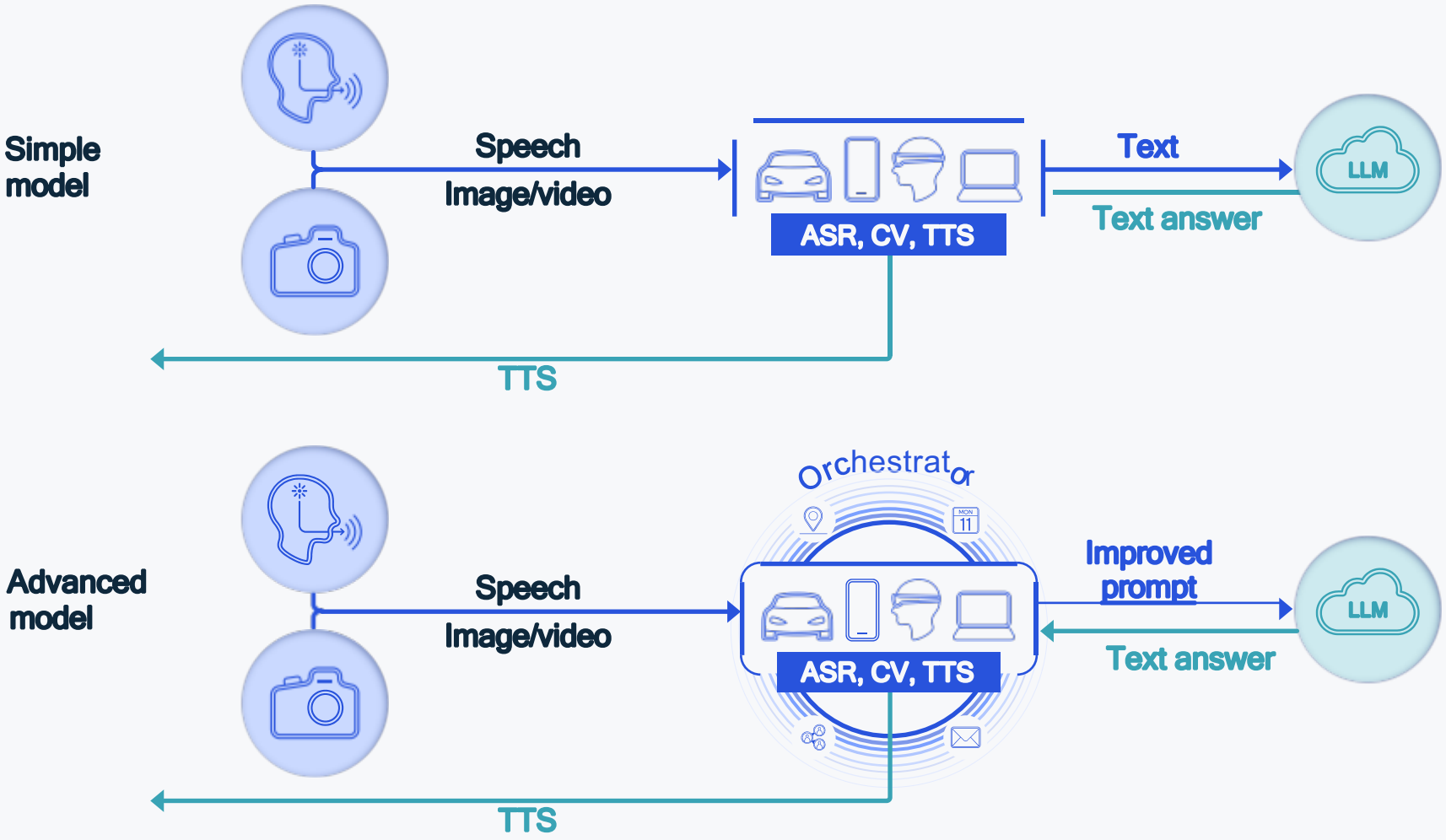
The device acts as the anchor point

On-device neural network or rules-based arbiter will decide where to run the model

More complex models will use the cloud as needed

It will be seamless to the user



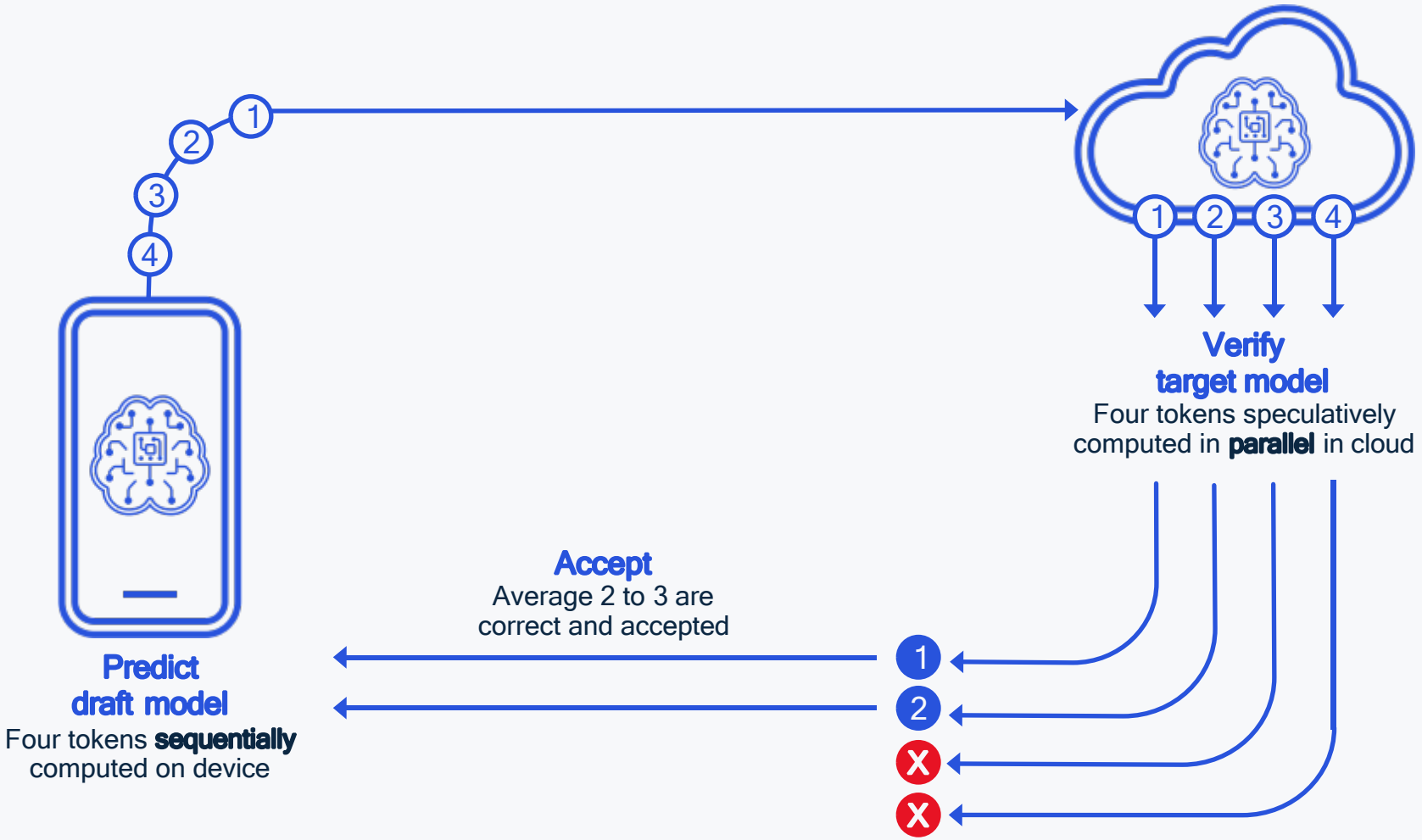


- Sensor and human-machine interface processing run on device
  - ASR, CV, TTS
- LLM runs in the cloud
- For advanced version, an on-device orchestrator uses on-device learning and personal data to provided improved prompts to the LLM

Device-sensing hybrid AI

The device acts as the eyes and ears

ASR: Automatic speech recognition; CV: Computer vision; TTS: Text to speech



- LLMs are memory-bound and produce a single token per inference, reading in all the weights
- The smaller draft model runs on device, sequentially
- The larger target model runs on the cloud, in parallel and speculatively
- The good tokens are accepted
- Results in net speedup in tokens per unit time and energy savings

Joint-processing hybrid AI

Multi-token speculative decoding as an example

Summary

- On-device generative AI offers many benefits
- Generative AI is happening now on the device
- On-device AI leadership is enabling generative AI to scale
- Hybrid AI is the future

Thank you