Qualcomm

Nov 16, 2023

The future of AI is “on device”

Kyuwoong Hwang

Senior Director, Technology
Qualcomm Korea YH

@QCOMResearch

Today's agenda

Why on-device generative AI is key

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

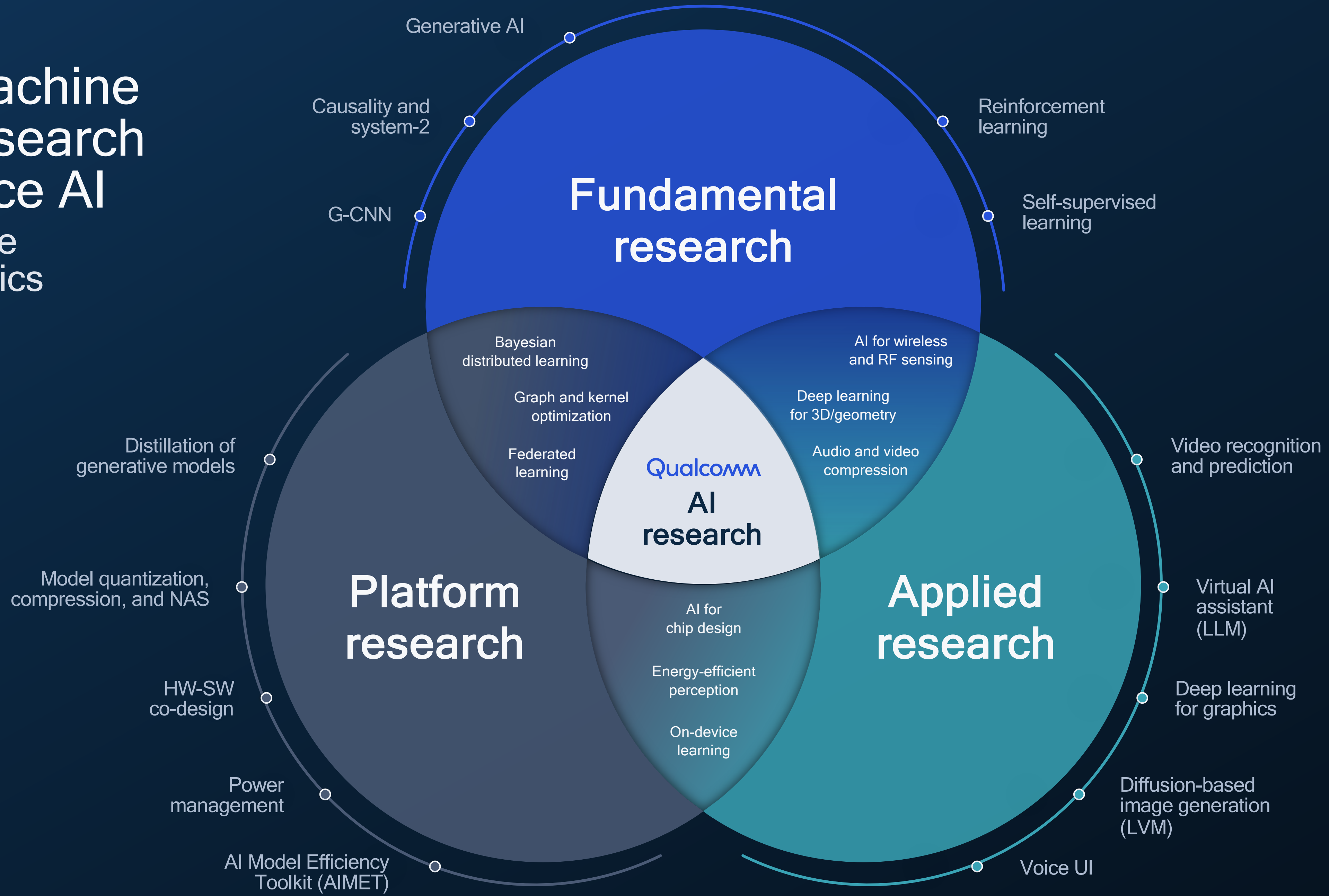
Full-stack AI optimizations for large language models – **Llama 2**

Hybrid AI technologies and architectures

Q&A

Leading machine learning research for on-device AI

across the entire spectrum of topics


Fundamental research
Platform research
Applied research
Qualcomm AI research
Generative AI
Causality and system-2
G-CNN
Reinforcement learning
Self-supervised learning
Bayesian distributed learning
Graph and kernel optimization
Federated learning
AI for wireless and RF sensing
Deep learning for 3D/geometry
Audio and video compression
AI for chip design
Energy-efficient perception
On-device learning
Distillation of generative models
Model quantization, compression, and NAS
HW-SW co-design
Power management
AI Model Efficiency Toolkit (AIMET)
Video recognition and prediction
Virtual AI assistant (LLM)
Deep learning for graphics
Diffusion-based image generation (LVM)
Voice UI

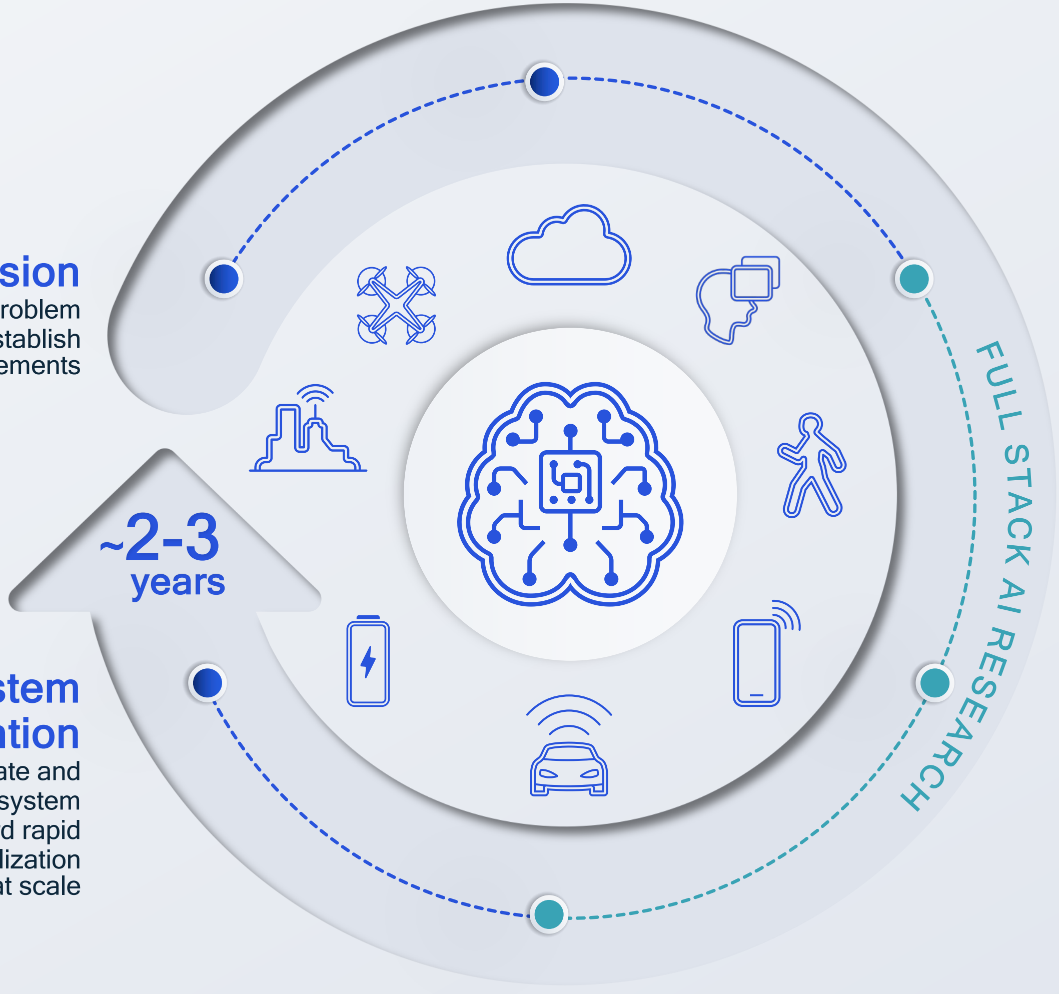
AIMET is a product of Qualcomm Innovation Center, Inc. Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
LLM: Large language mode; LVM: Language vision model

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

Vision
Identify a problem or need; establish requirements

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

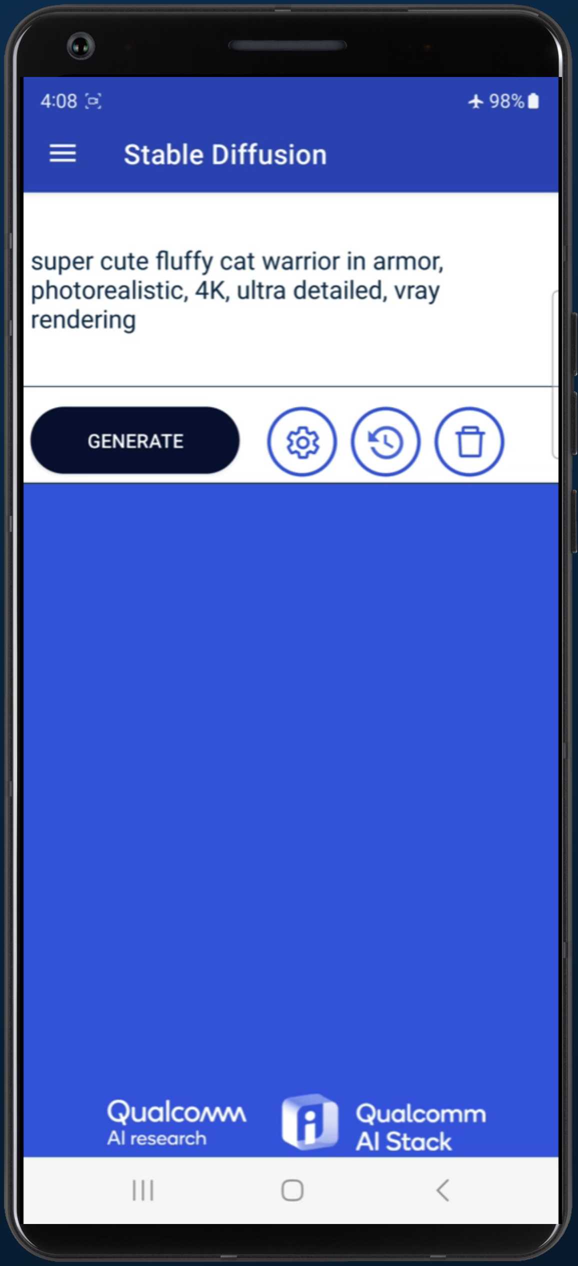
Software compilation
Develop tech & tools to improve graph-level and kernel-level software compilation performance

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

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

At MWC 2023

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



super cute fluffy cat warrior in armor, photorealistic, 4K, ultra detailed, vray rendering

1B+ parameter generative AI model runs efficiently and interactively

Full-stack AI optimization to achieve sub-15 second latency for 20 inference steps

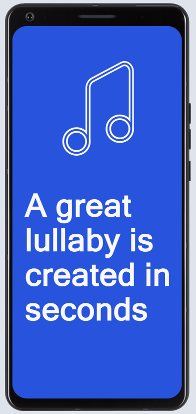
Enhanced privacy, security, reliability, and cost with on-device processing

Fast development enabled by Qualcomm AI Research and Qualcomm® AI Stack

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"



A great lullaby is created in seconds

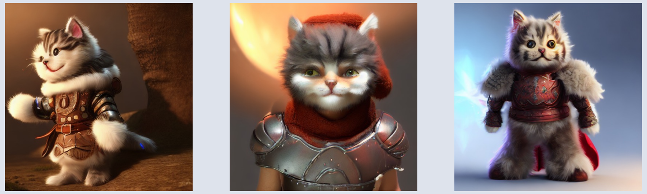
Real-life application of this platform

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

Image generation
(Stable Diffusion, MidJourney, etc.)

Input prompts

"Super cute fluffy cat warrior in armor"



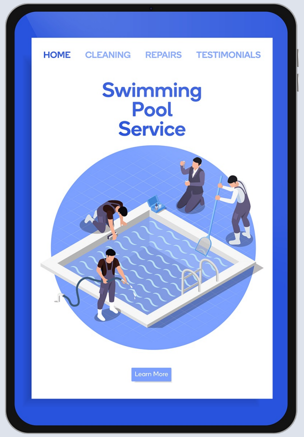
Real-life application of this platform

- 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

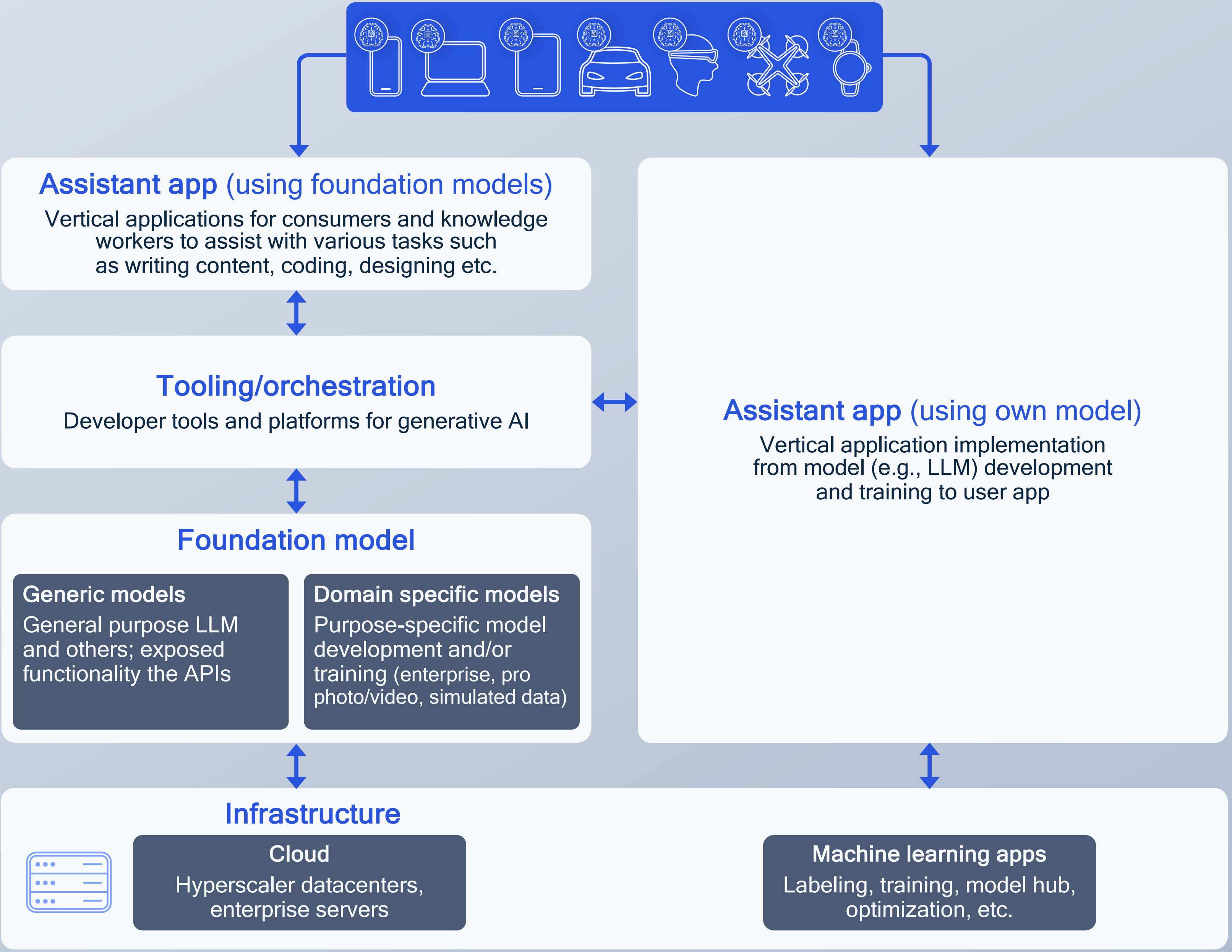
What is generative AI?

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



Assistant app (using foundation models)

Vertical applications for consumers and knowledge workers to assist with various tasks such as writing content, coding, designing etc.

Tooling/orchestration

Developer tools and platforms for generative AI

Foundation model

Generic models

General purpose LLM and others; exposed functionality the APIs

Domain specific models

Purpose-specific model development and/or training (enterprise, pro photo/video, simulated data)

Assistant app (using own model)

Vertical application implementation from model (e.g., LLM) development and training to user app

Infrastructure

Cloud

Hyperscaler datacenters, enterprise servers

Machine learning apps

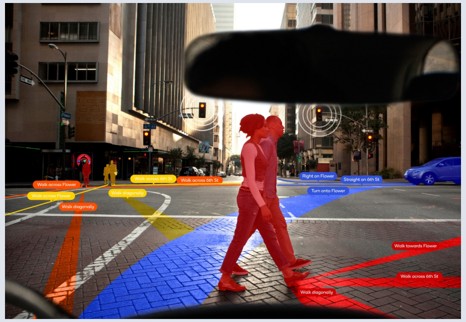
Labeling, training, model hub, optimization, etc.

XR

surreal world, floating jellyfish all around, beautiful waterfalls, mystical lake, towering mountains

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

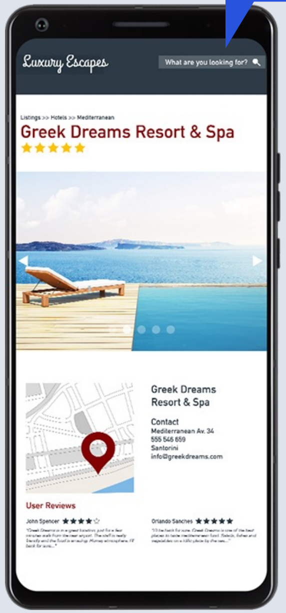
Automotive



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

Phone

“Make me reservations for a weekend getaway at the place Bob recommended”



Luxury Escapes

What are you looking for?

Listings >> Hotels >> Mediterranean

Greek Dreams Resort & Spa
★★★★★

Greek Dreams Resort & Spa

Contact
Mediterranean Av. 34
555 546 659
Santorini
info@greekdreams.com

User Reviews

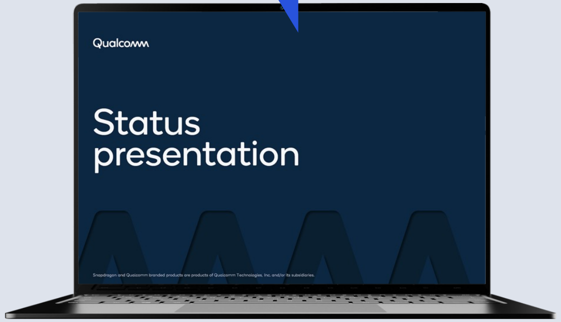
John Spencer ★★★★☆

Orlando Sanches ★★★★★

Gen AI can become a true digital assistant

PC

“Make me a status presentation for my boss based on inputs from my team”



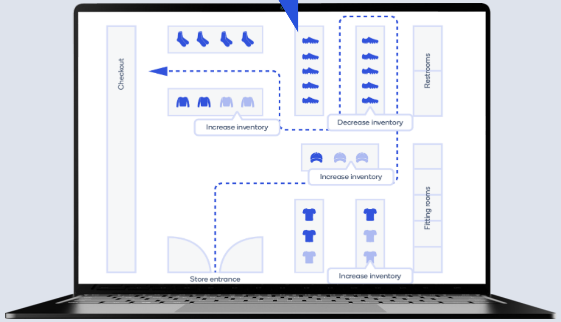
Qualcomm

Status presentation

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

IoT

“Suggest inventory and store layout changes to increase user satisfaction in the sports section”

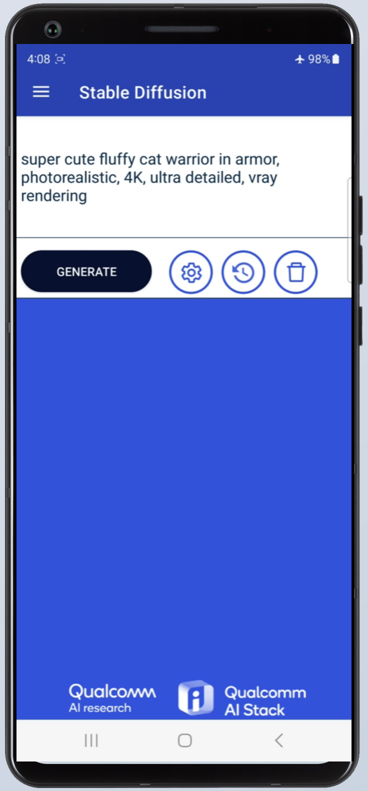


Gen AI can help improve customer and employee experience in retail, such as providing recommendations for inventory and store layout

Generative AI will impact use cases across device categories

Stable Diffusion

Denoising an image
with a diffusion model

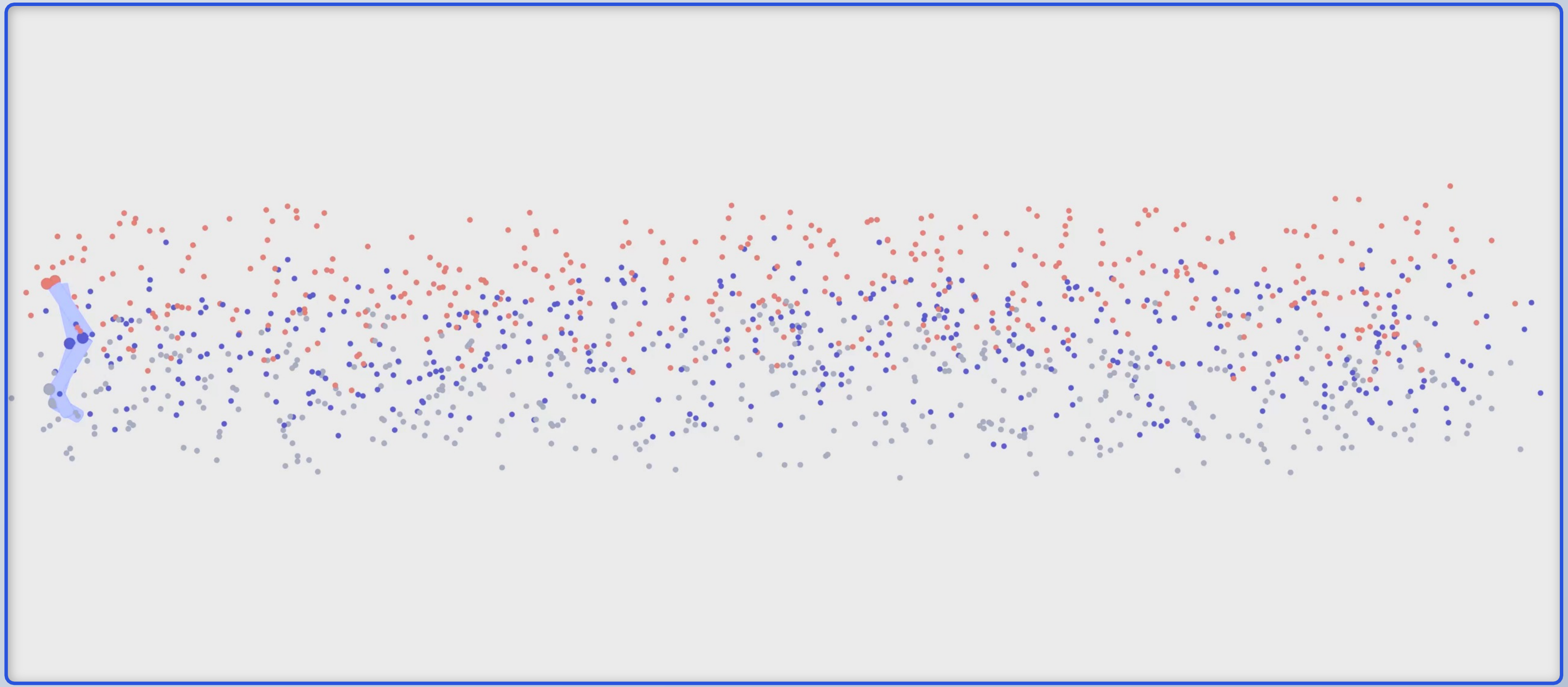


Stable Diffusion

super cute fluffy cat warrior in armor, photorealistic, 4K, ultra detailed, vray rendering

Generating robot trajectories

Instead of diffusing an image
we diffuse a robot trajectory

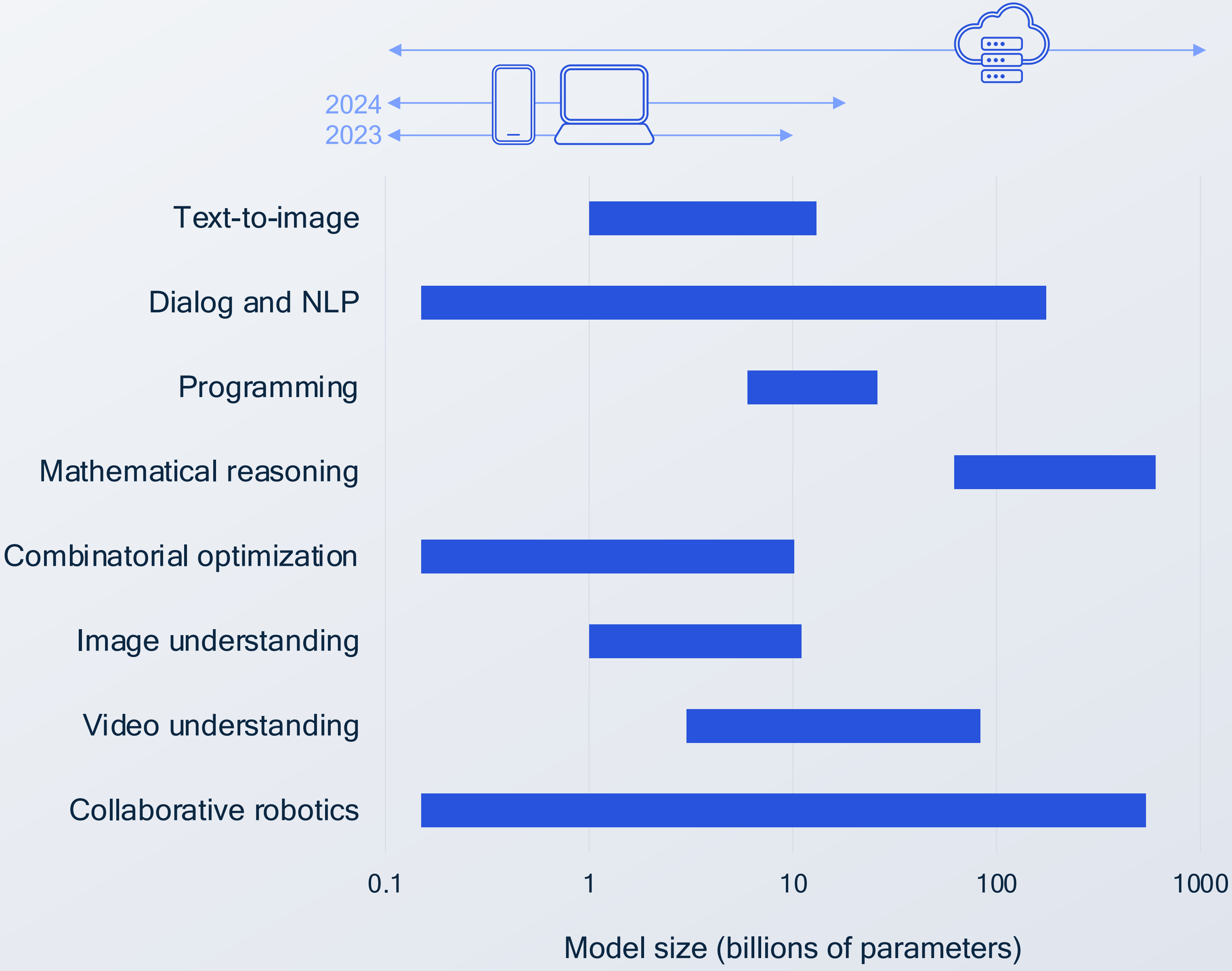


Generative AI with diffusion models for robotics path planning

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

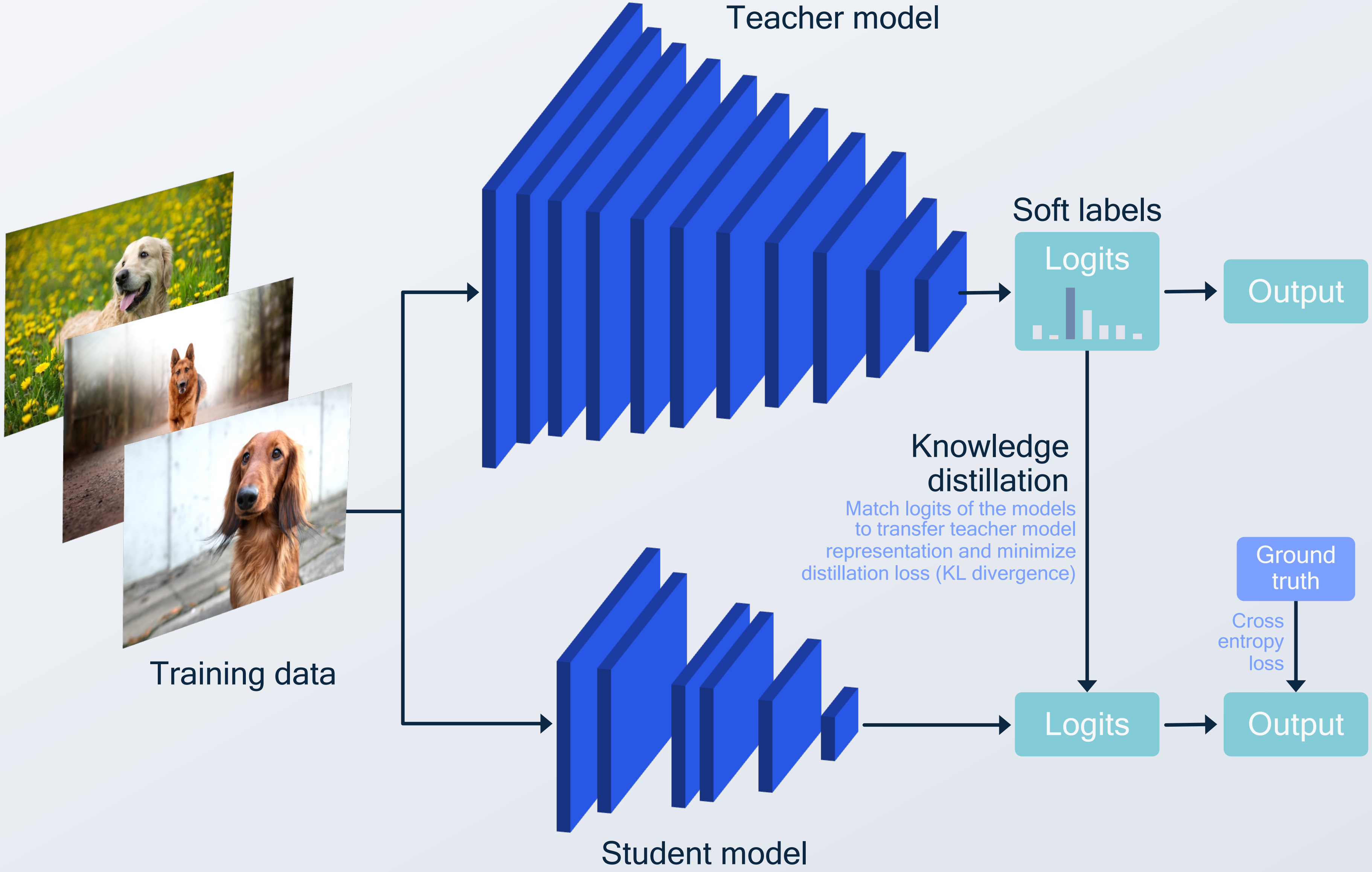
Training a smaller "student" model to mimic a larger "teacher" model

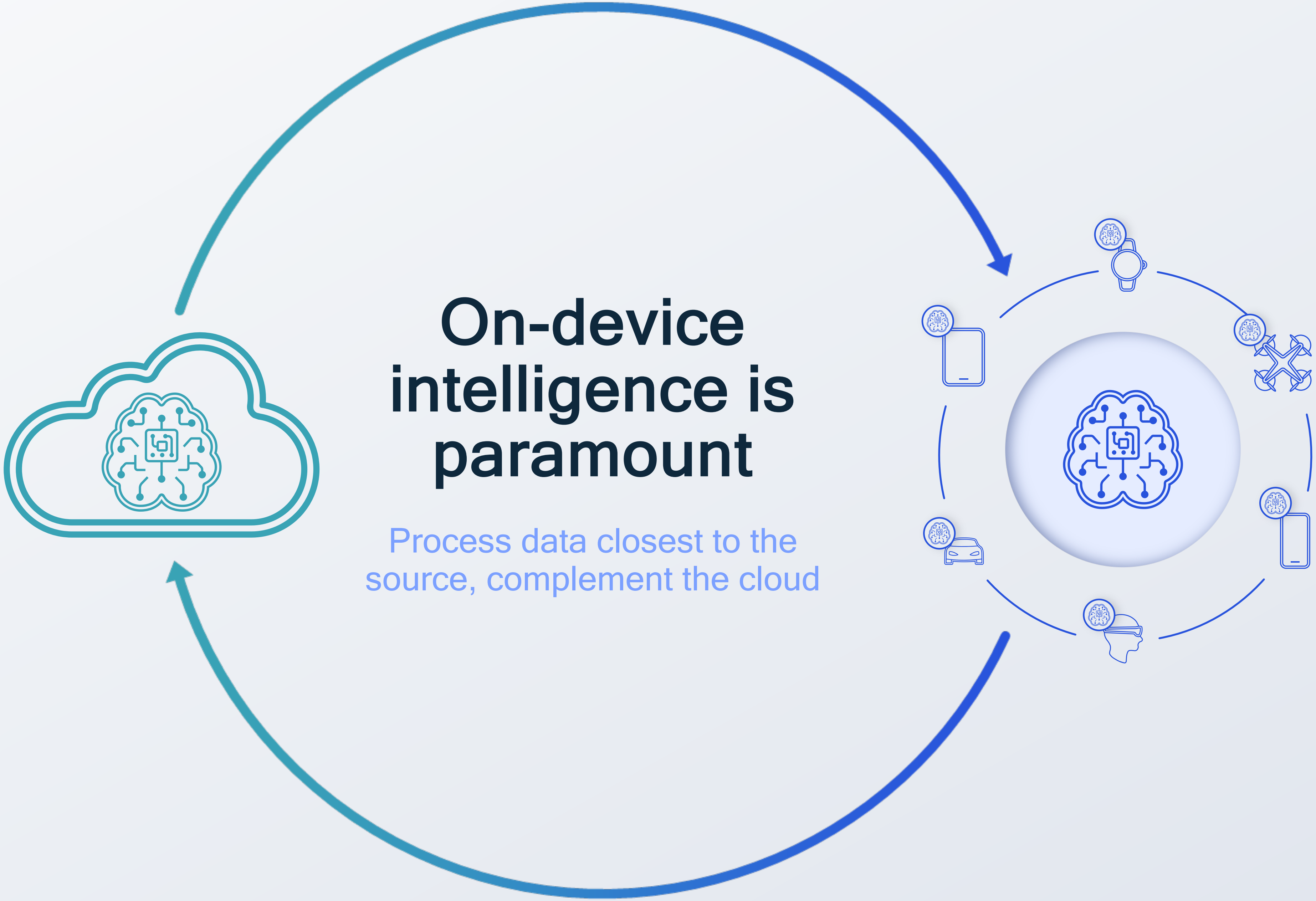
Create a smaller model with fewer parameters

Run faster inference on target deployment

Maintain prediction quality close to the teacher

Less training time





On-device intelligence is paramount

Process data closest to the source, complement the cloud

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

What is diffusion?

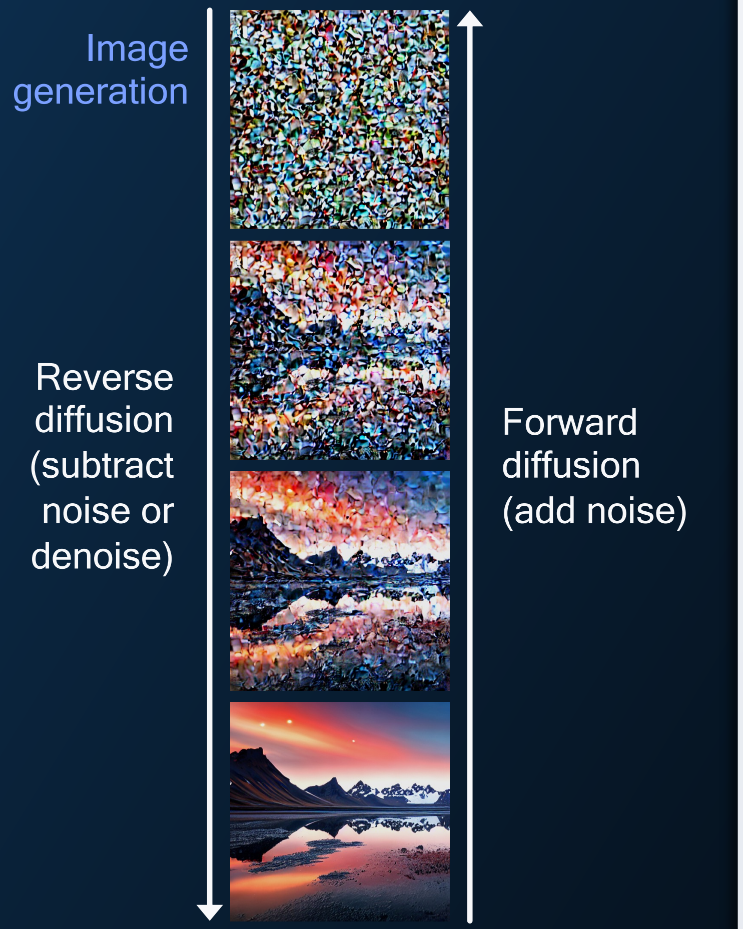


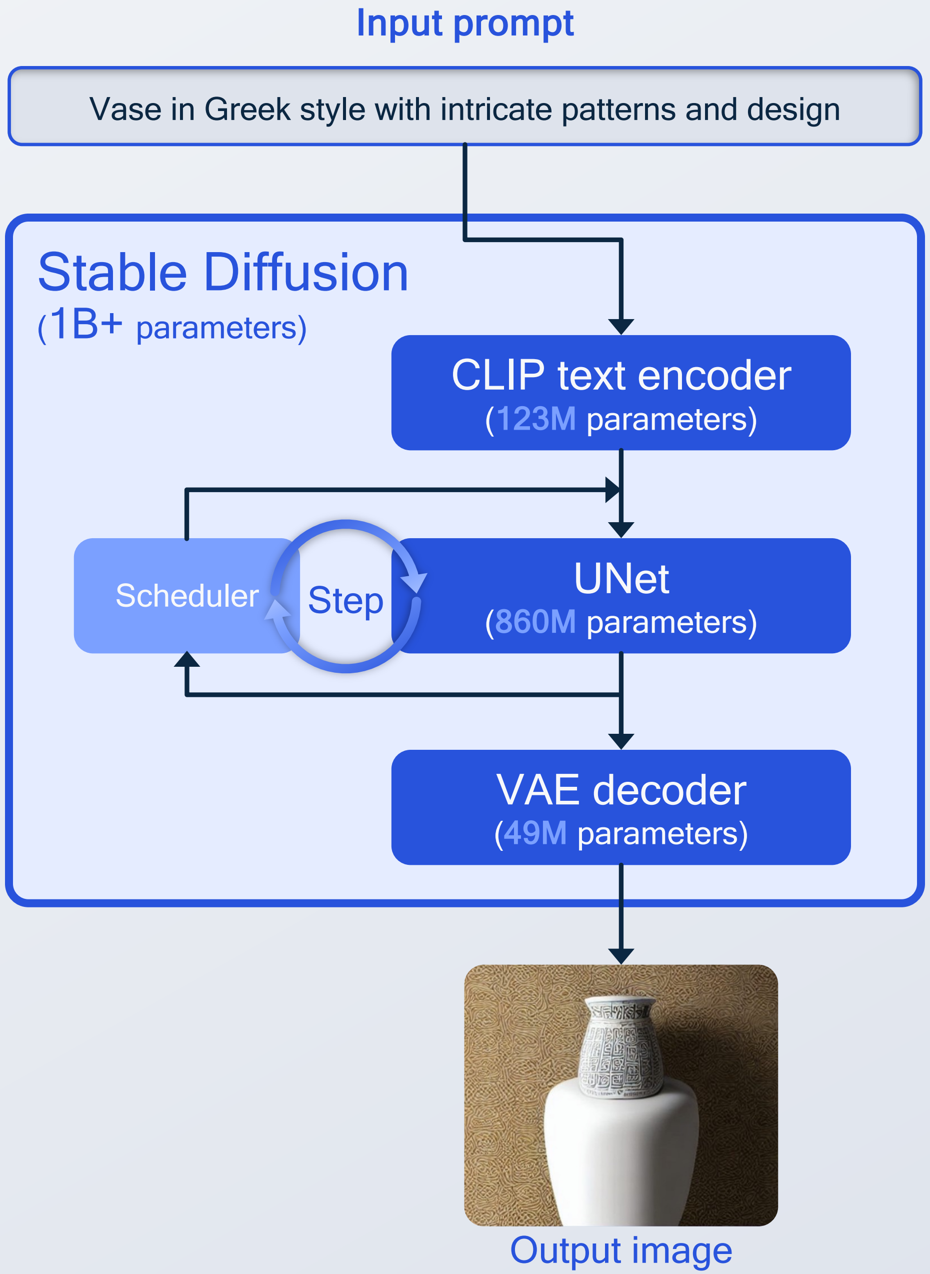
Image generation
Reverse diffusion (subtract noise or denoise)
Forward diffusion (add noise)

Stable Diffusion architecture

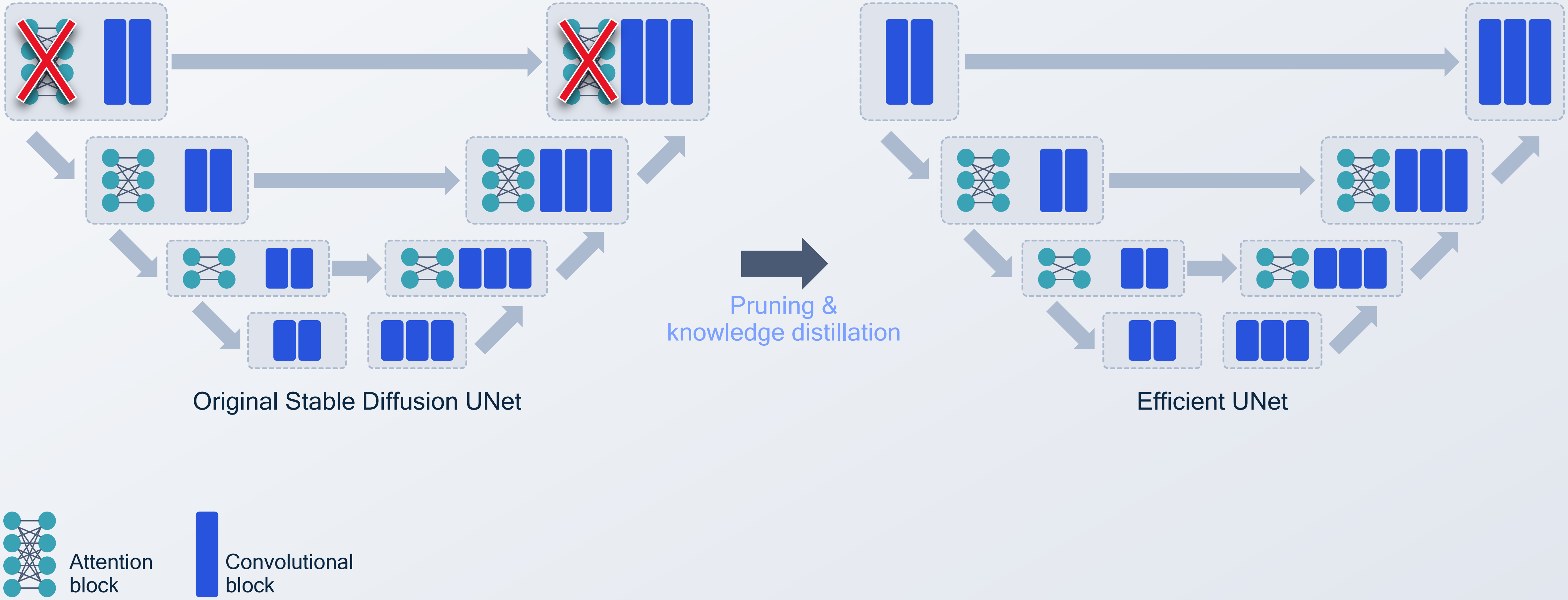
UNet is the biggest component model of Stable Diffusion

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

Significant compute is required

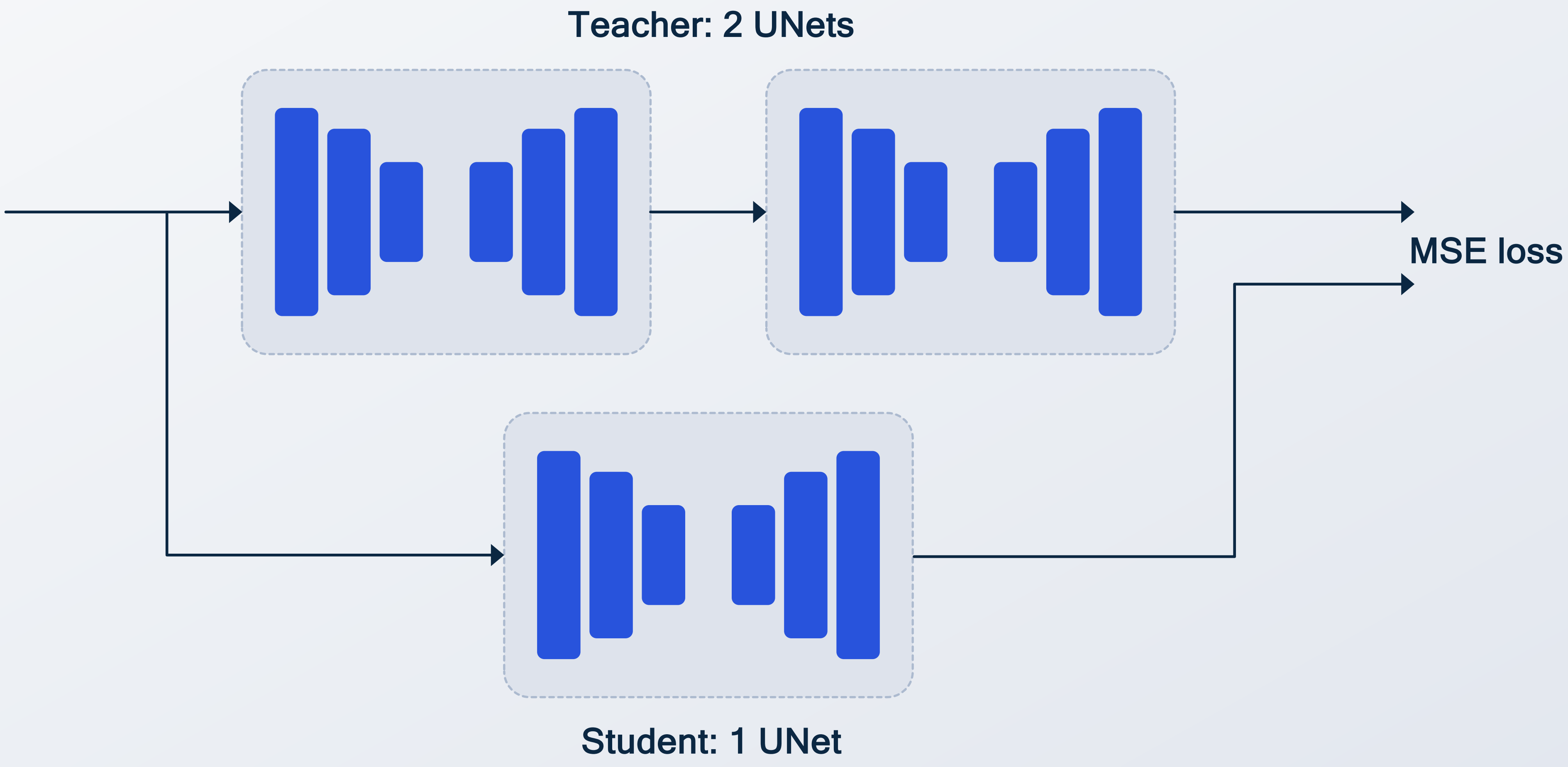


VAE: Variational Auto Encoder;
CLIP: Contrastive Language-Image Pre-Training



More efficient architecture design through pruning and knowledge distillation

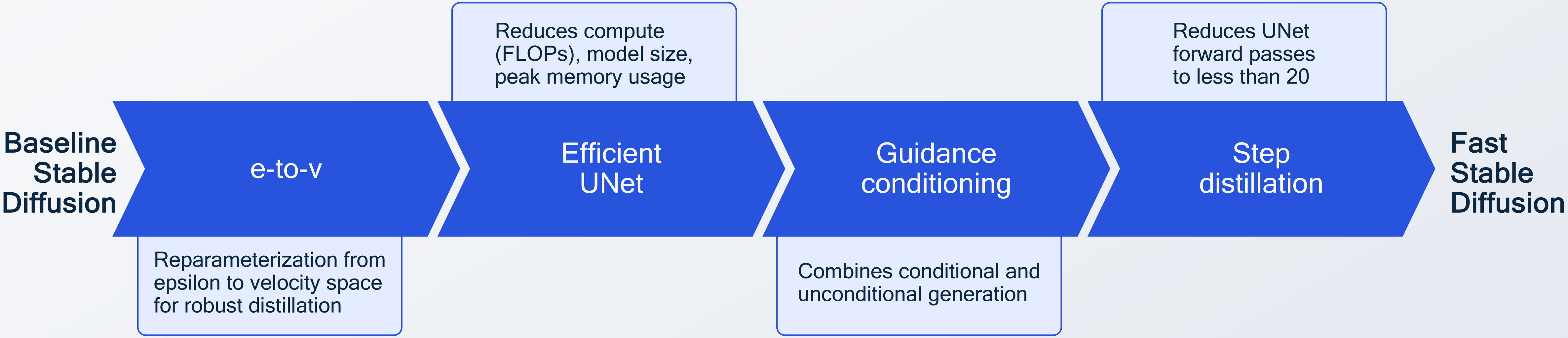
Reducing UNet compute (FLOPs), model size, and peak memory usage



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



Baseline Stable Diffusion

e-to-v

Reparameterization from epsilon to velocity space for robust distillation

Efficient UNet

Reduces compute (FLOPs), model size, peak memory usage

Guidance conditioning

Combines conditional and unconditional generation

Step distillation

Reduces UNet forward passes to less than 20

Fast Stable Diffusion

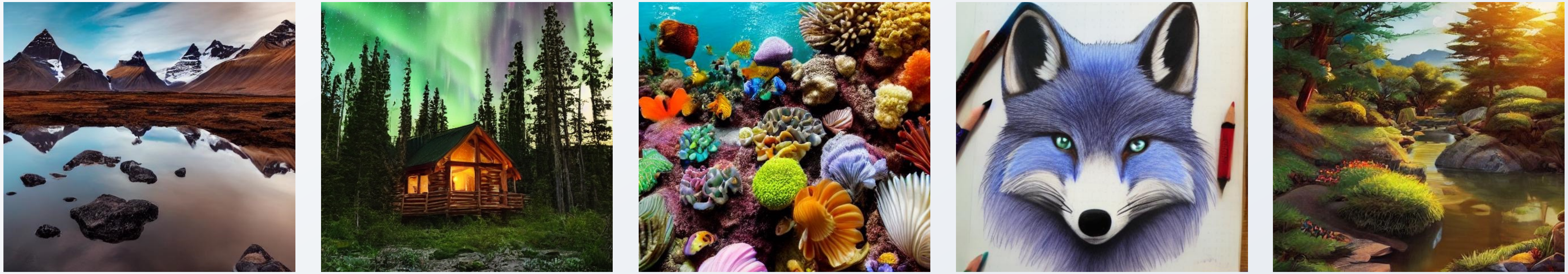
	FID↓	CLIP ↑	Inference latency
Baseline (SD-1.5)	17.14*	0.3037	5.05 seconds
Fast SD	20.08	0.3004	0.56 seconds

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

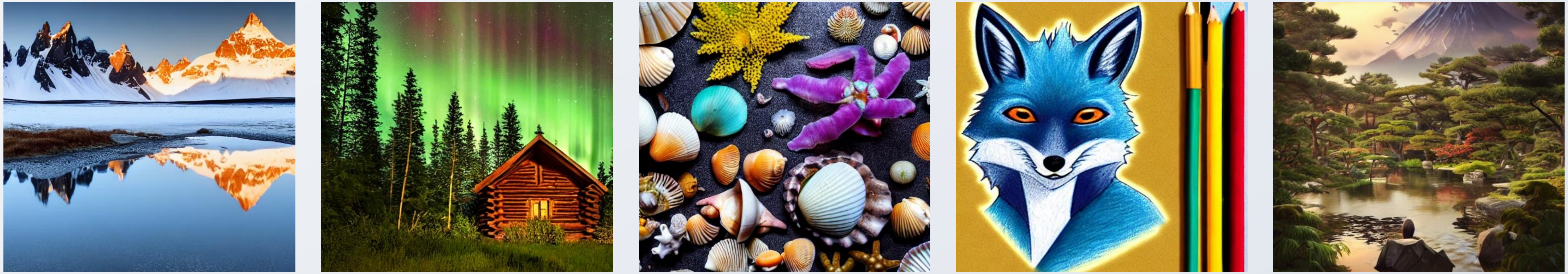
A hyper realistic photo of a beautiful cabin inside of a forest and full of trees and plants, with large aurora borealis in the sky

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

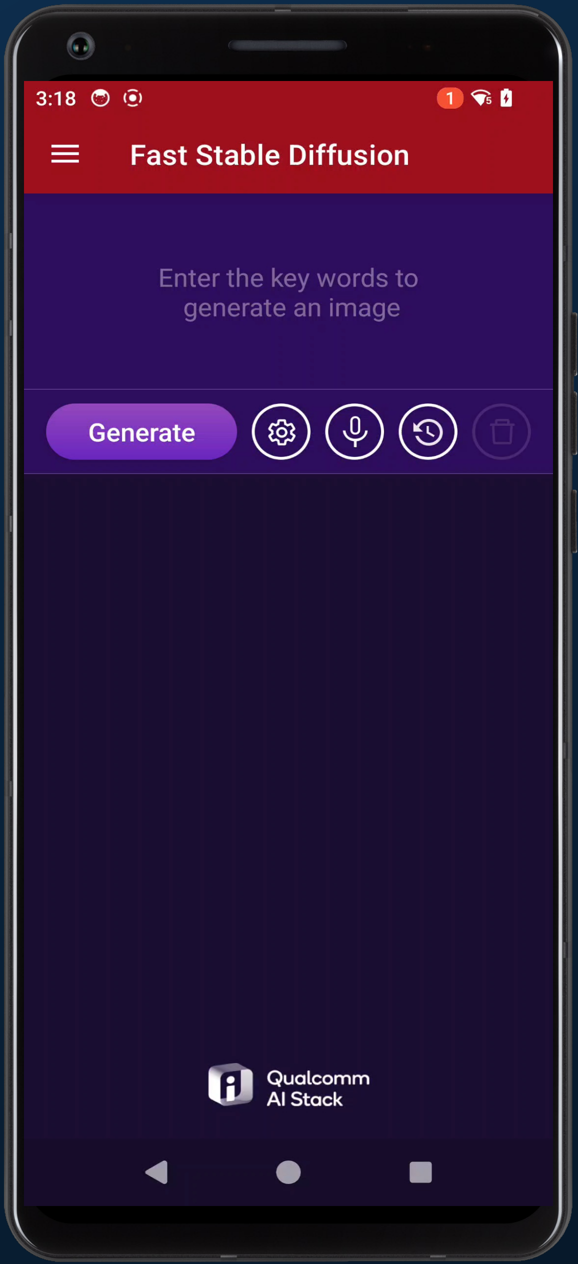
Japanese garden at wild life river and mountain range, highly detailed, digital illustration

Stable Diffusion V1.5



Similar image quality between our fast implementation and baseline model

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

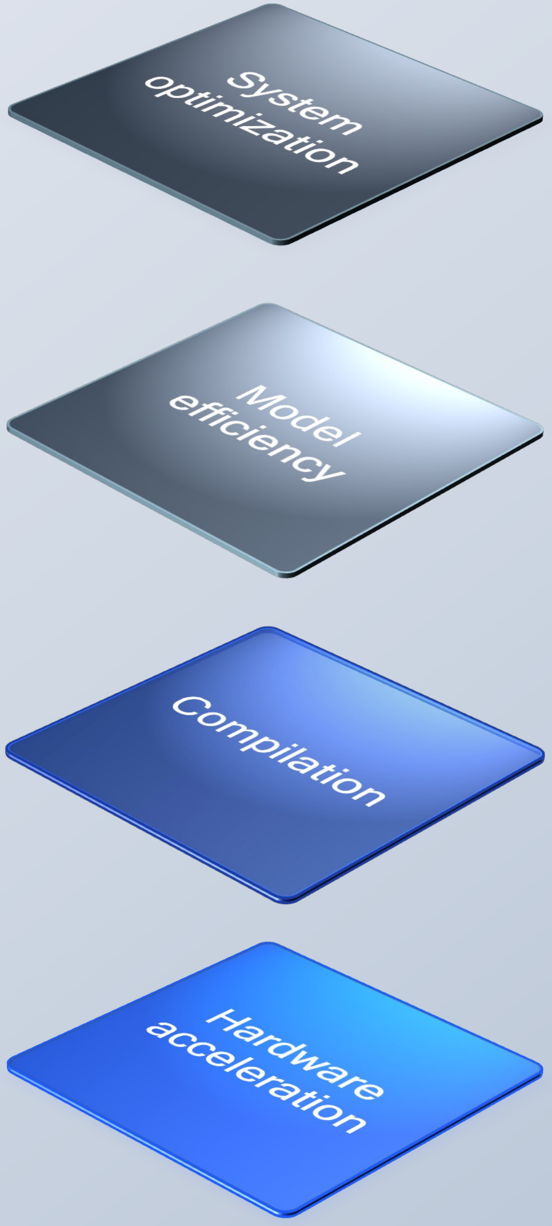
Full-stack AI optimization

for LVM

Runs completely on the device

Significantly reduces runtime latency and power consumption

Continuously improves the Qualcomm® AI Stack



System optimization

Designing an efficient diffusion model through knowledge distillation for high accuracy

Model efficiency

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

Compilation

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

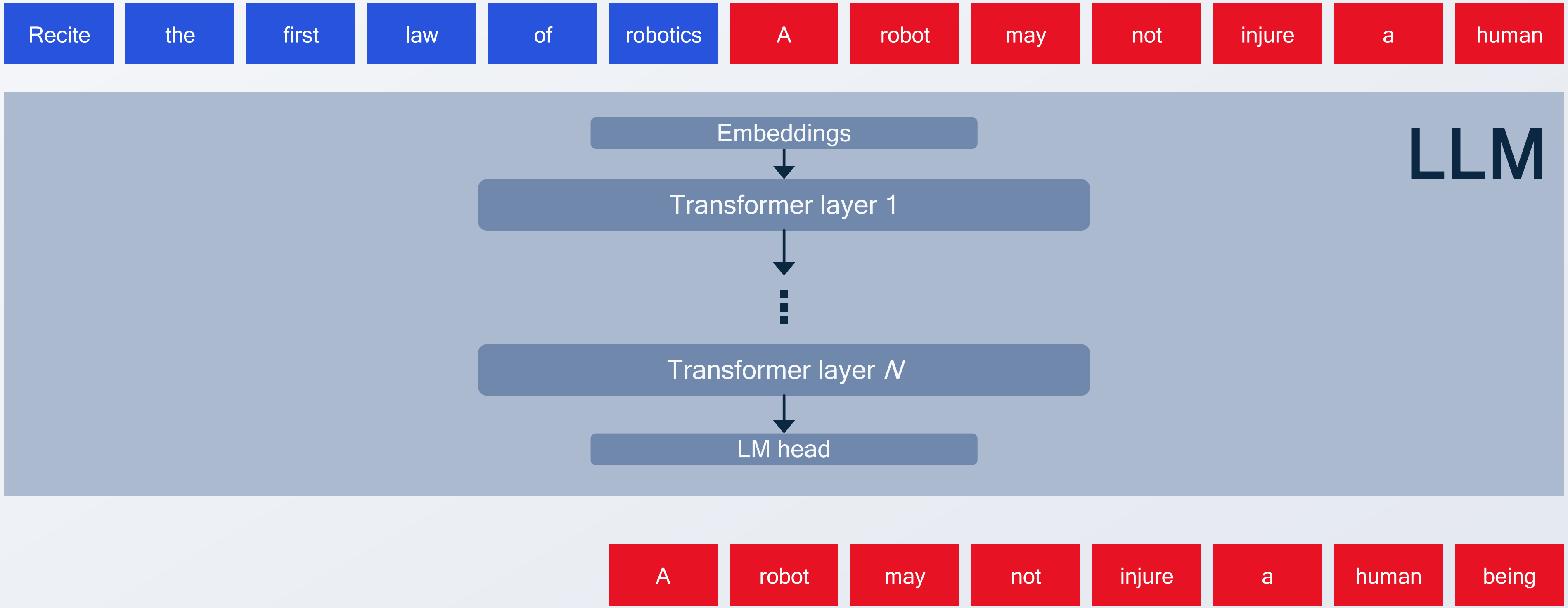
Hardware acceleration

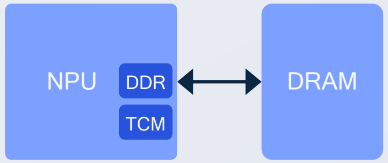
AI acceleration on the Qualcomm® Hexagon™ NPU of the Snapdragon® 8 Gen 3 Mobile Processor

LVM: Language vision model

Illustration of autoregressive language modeling

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





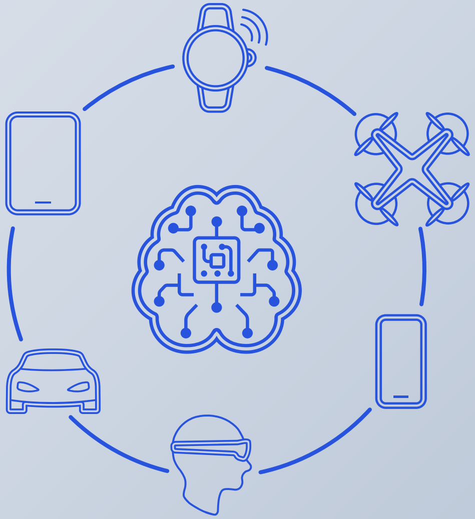
Huge bandwidth

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

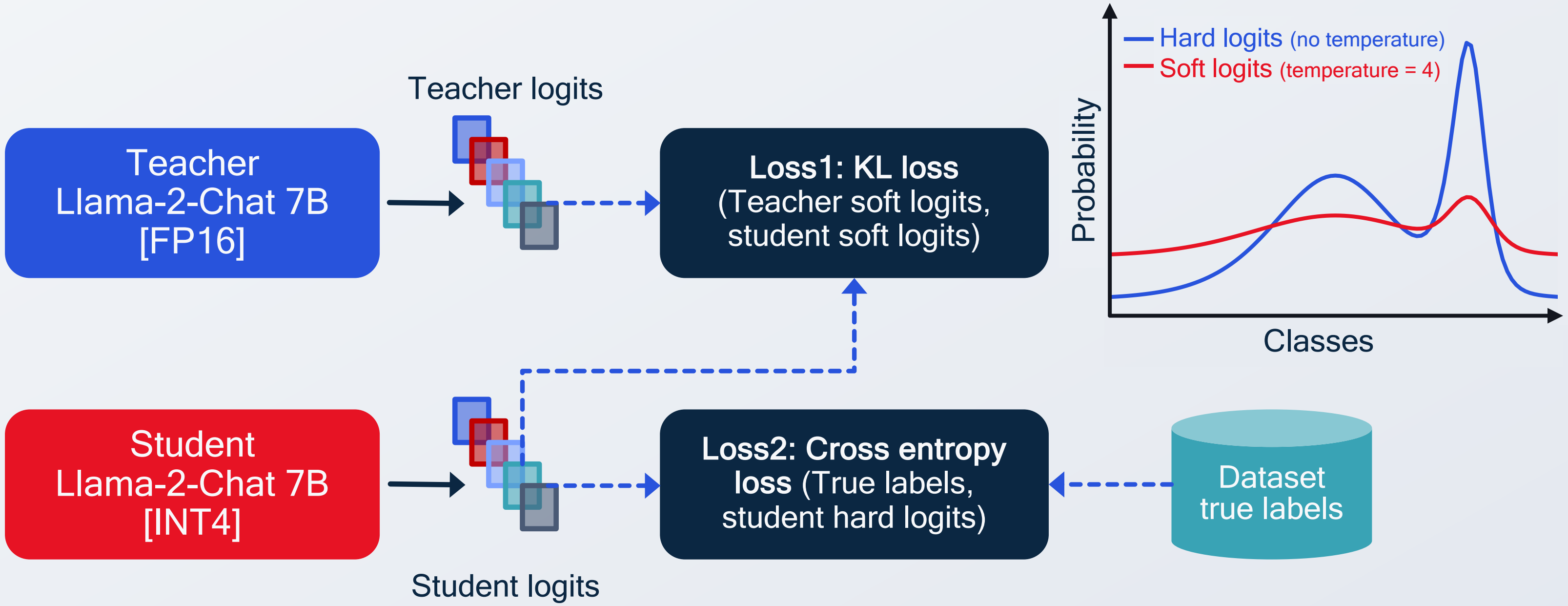
LLM quantization challenges

- 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



Teacher logits

Teacher
Llama-2-Chat 7B
[FP16]

Loss1: KL loss
(Teacher soft logits, student soft logits)

Student
Llama-2-Chat 7B
[INT4]

Loss2: Cross entropy loss (True labels, student hard logits)

Dataset true labels

Student logits

<1

Point increase in perplexity[1]

<1%

Decrease in accuracy

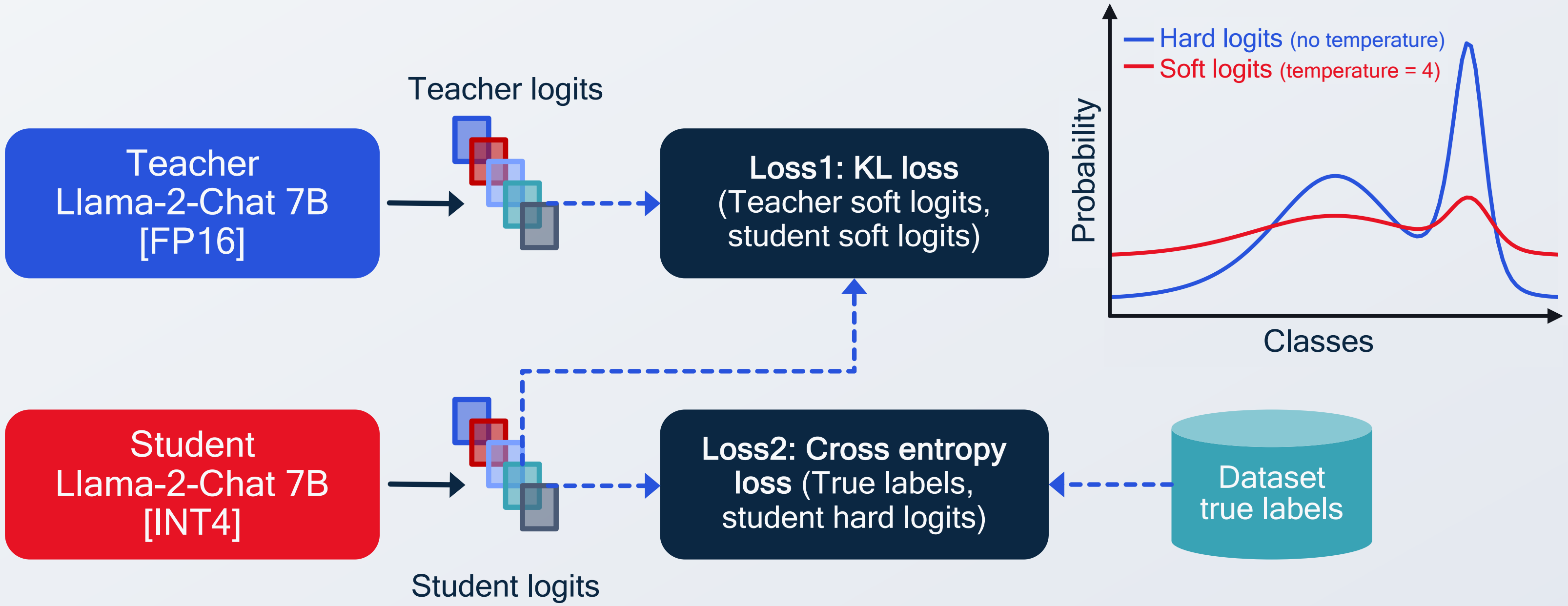
KD loss function combines the KL divergence loss and hard-label based CE loss

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

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


Teacher logits
Teacher
Llama-2-Chat 7B
[FP16]
Loss1: KL loss
(Teacher soft logits, student soft logits)
Student
Llama-2-Chat 7B
[INT4]
Loss2: Cross entropy loss (True labels, student hard logits)
Dataset true labels
Student logits
Hard logits (no temperature)
Soft logits (temperature = 4)
Probability
Classes

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

<1

Point increase in perplexity[1]

<1%

Decrease in accuracy

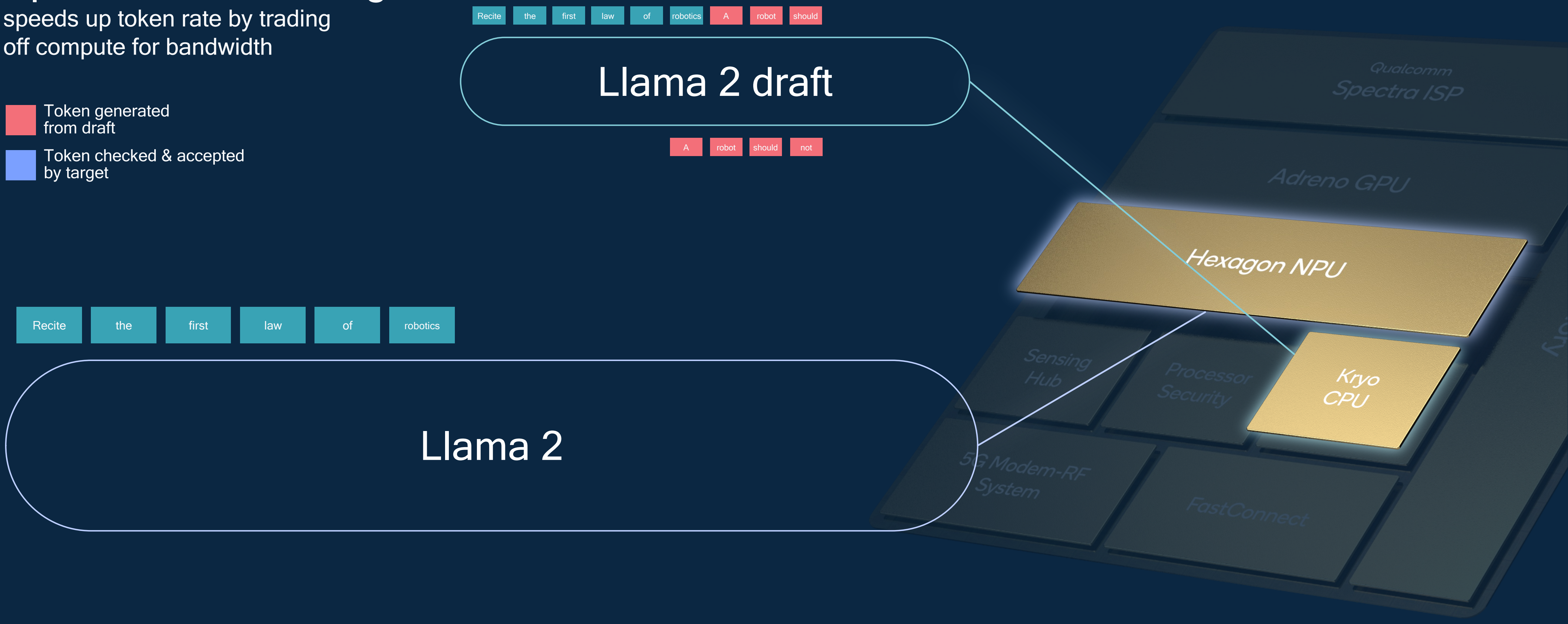
[1] Perplexity is average over several test sets, including wikitext and c4 (subset)

Speculative decoding

speeds up token rate by trading off compute for bandwidth

🟥 Token generated from draft

🟦 Token checked & accepted by target



Draft model generates a few speculative tokens at a time

Target model decides which to accept in one pass

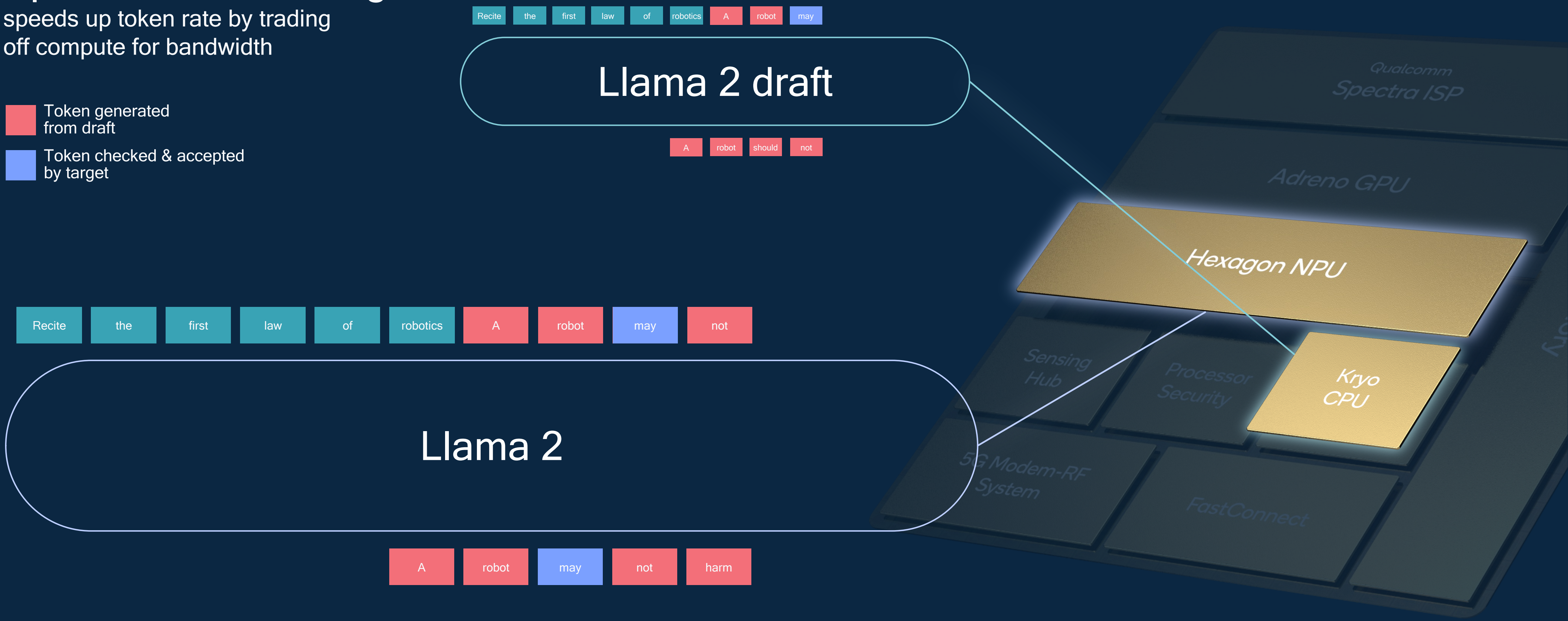
A good draft model predicts with a high acceptance rate

Speculative decoding

speeds up token rate by trading
off compute for bandwidth

🟥 Token generated from draft

🟦 Token checked & accepted by target



Recite the first law of robotics A robot may

Llama 2 draft

A robot should not

Recite the first law of robotics A robot may not

Llama 2

A robot may not harm

Draft model generates a few speculative tokens at a time

Target model decides which to accept in one pass

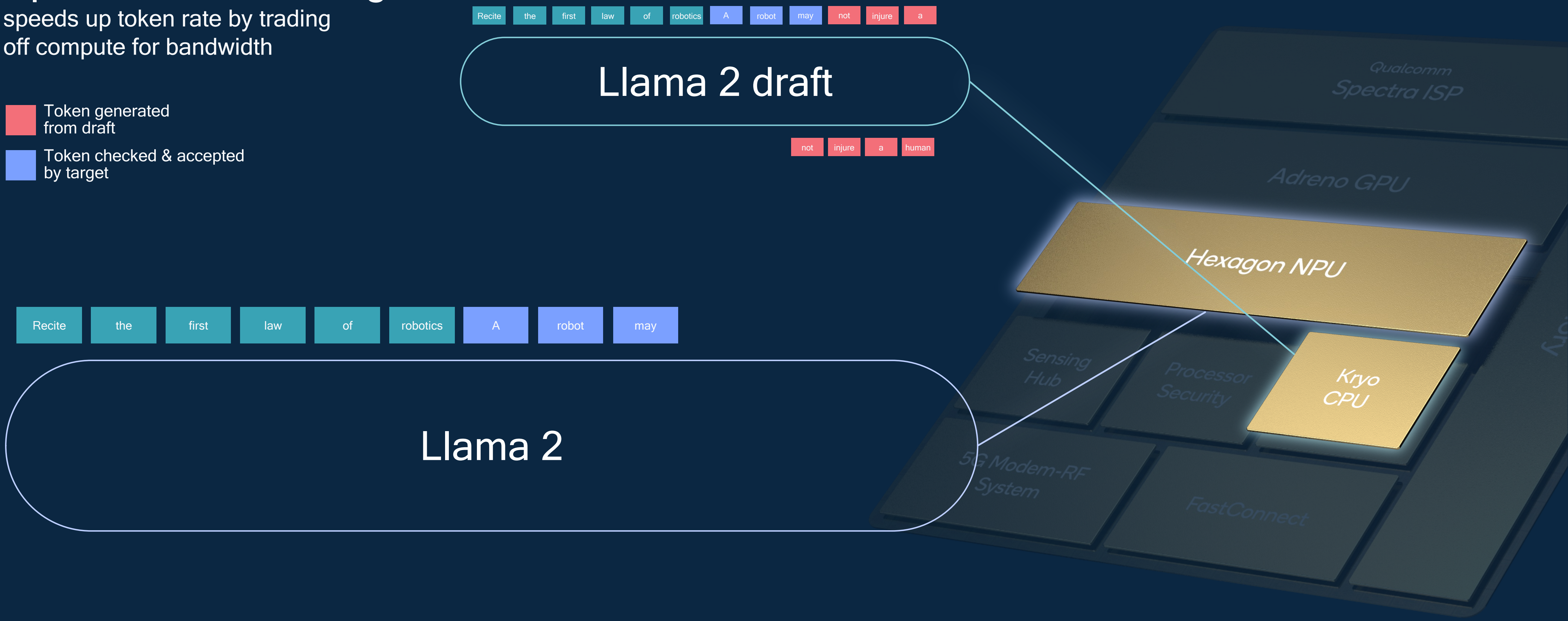
A good draft model predicts with a high acceptance rate

Speculative decoding

speeds up token rate by trading off compute for bandwidth

🟥 Token generated from draft

🟦 Token checked & accepted by target



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

Llama 2 draft

not injure a human

Recite the first law of robotics A robot may

Llama 2

Draft model generates a few speculative tokens at a time

Target model decides which to accept in one pass

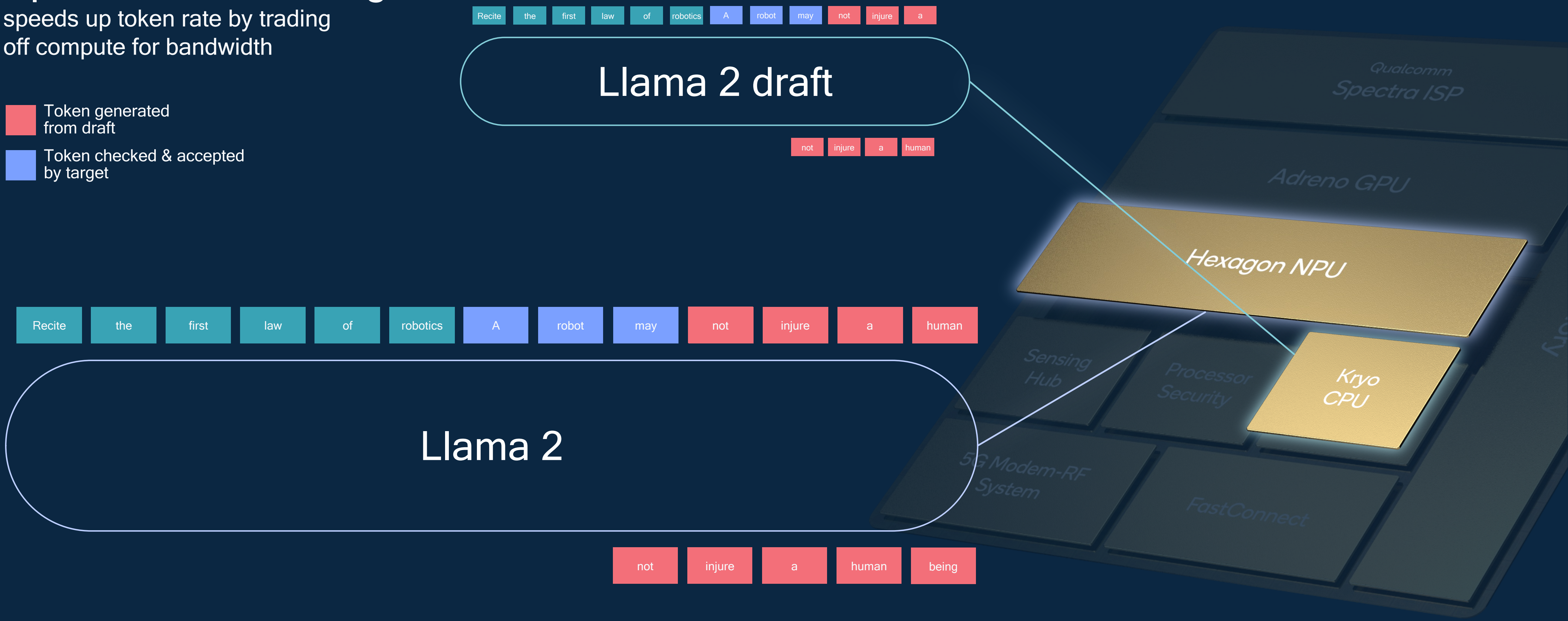
A good draft model predicts with a high acceptance rate

Speculative decoding

speeds up token rate by trading
off compute for bandwidth

🟥 Token generated from draft

🟦 Token checked & accepted by target



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

Llama 2 draft

not injure a human

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

Llama 2

not injure a human being

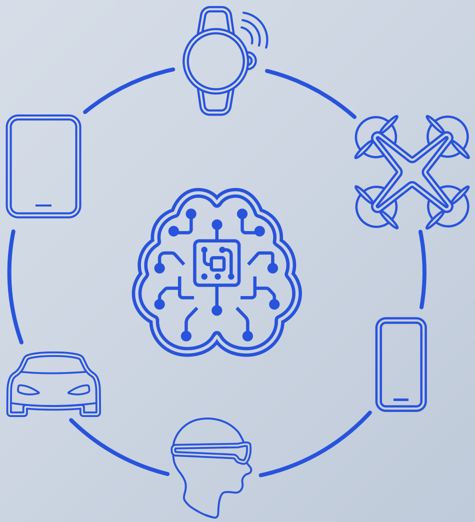
Draft model generates a few speculative tokens at a time

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

Up to

20

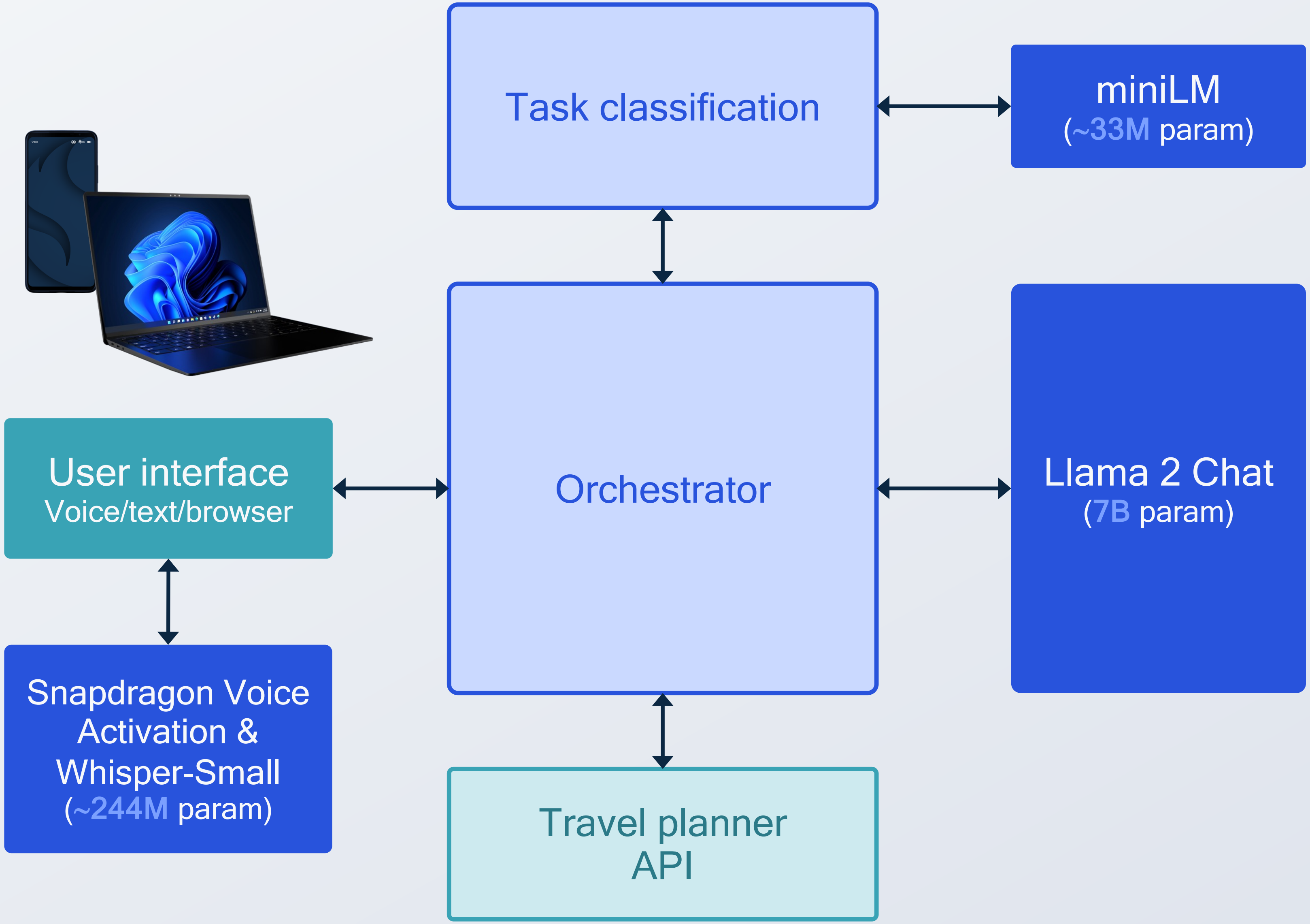
tokens per second

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)



LM: Language model

AI Assistant based on Llama 2

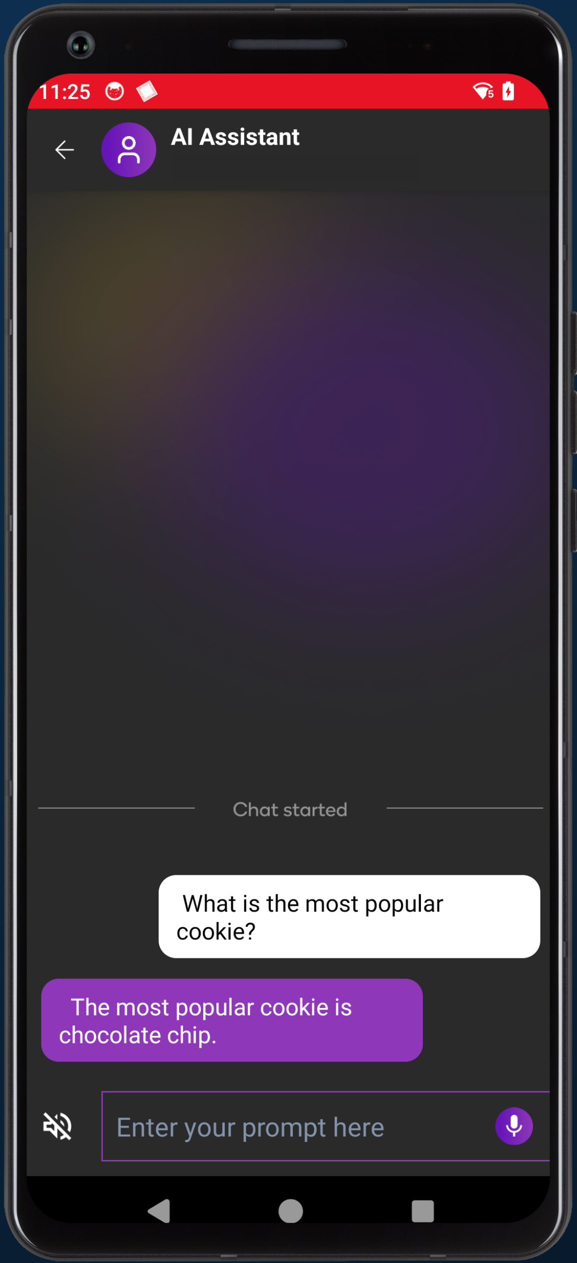
At Snapdragon Summit 2023

World’s fastest Llama 2-7B on a phone

Up to 20 tokens per second

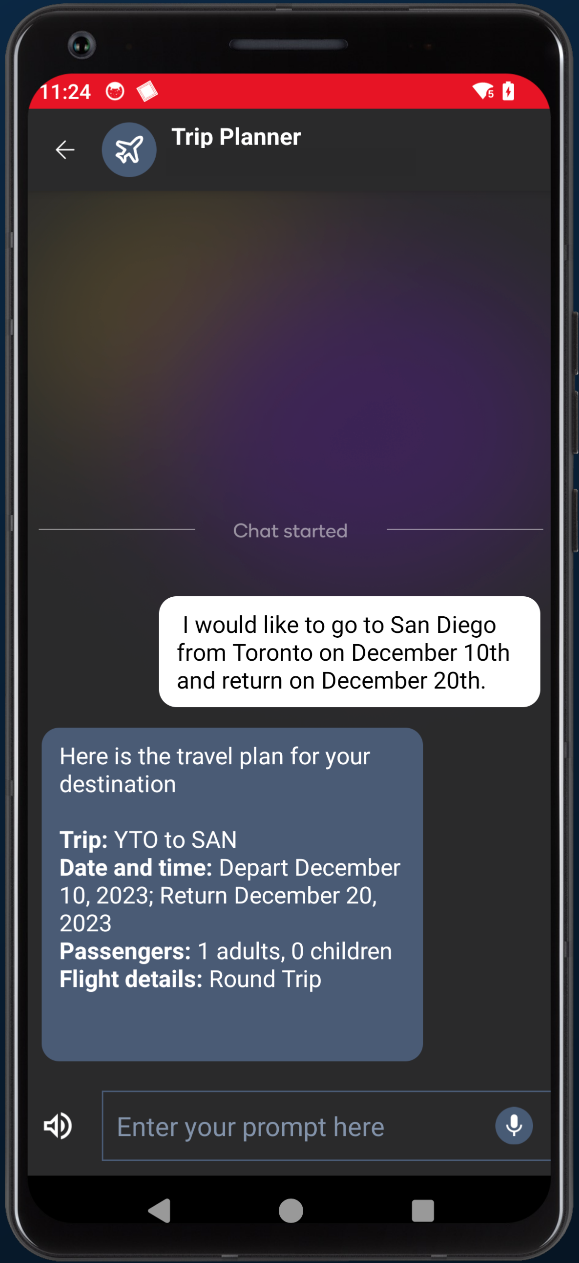
Demonstrating both chat and application interaction on device

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



What is the most popular cookie?

The most popular cookie is chocolate chip.



I would like to go to San Diego from Toronto on December 10th and return on December 20th.

Here is the travel plan for your destination

Trip: YTO to SAN
Date and time: Depart December 10, 2023; Return December 20, 2023
Passengers: 1 adults, 0 children
Flight details: Round Trip

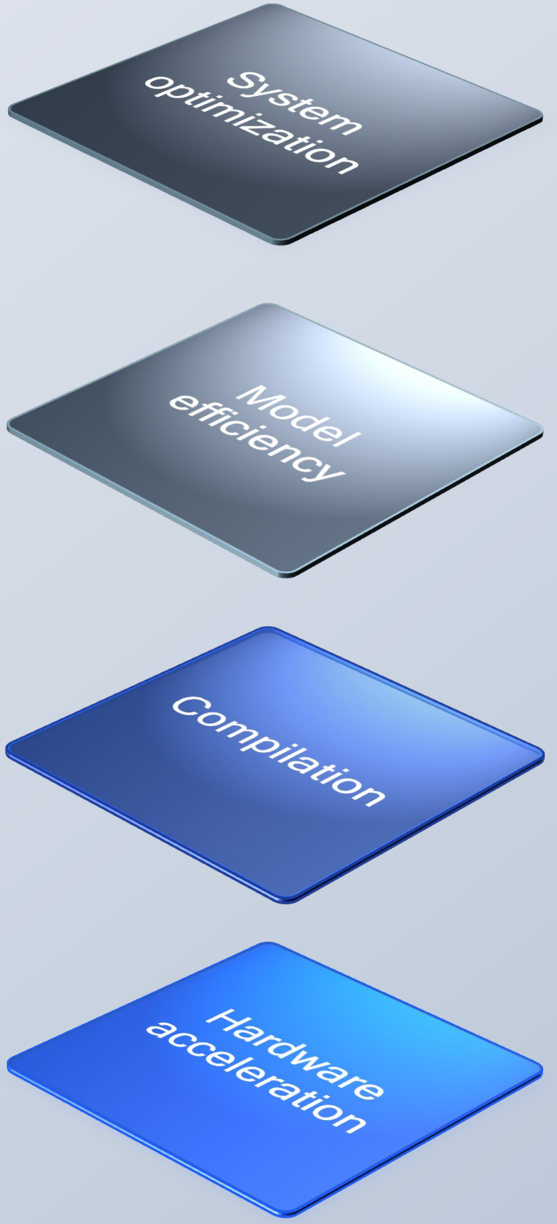
Full-stack AI optimization

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

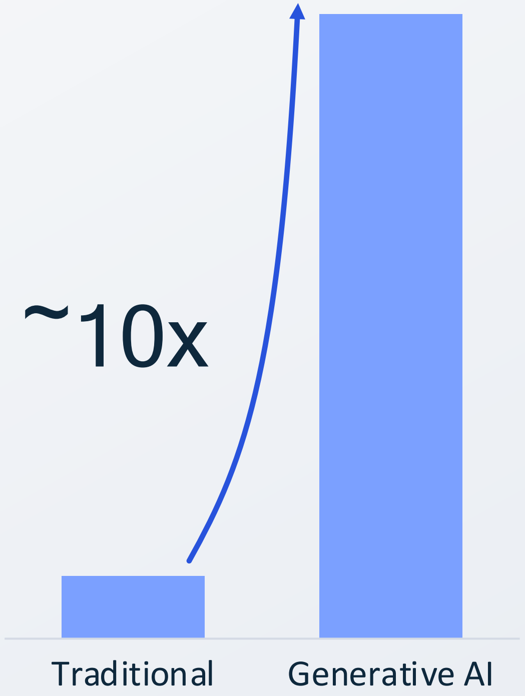
Qualcomm AI Engine direct for improved performance and minimized memory spillage

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] × Gen AI applications × Billions of users

(e.g. web search)



Personal assistant

Web search

Image & video creation

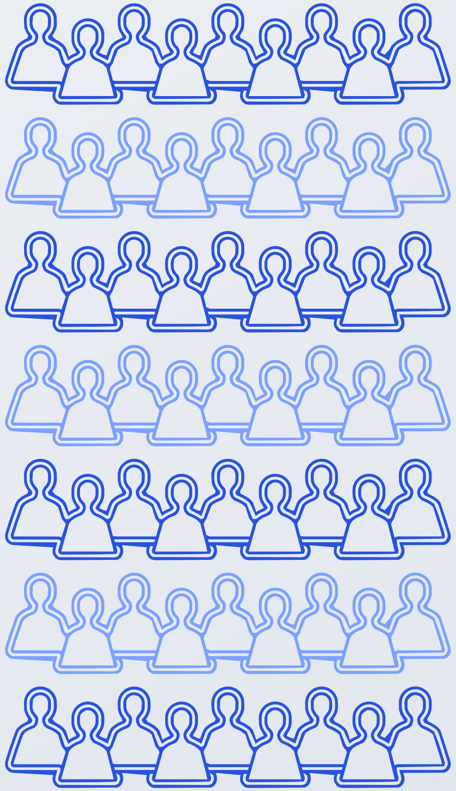
Coding assistant

Text summarization

Conversational chatbots

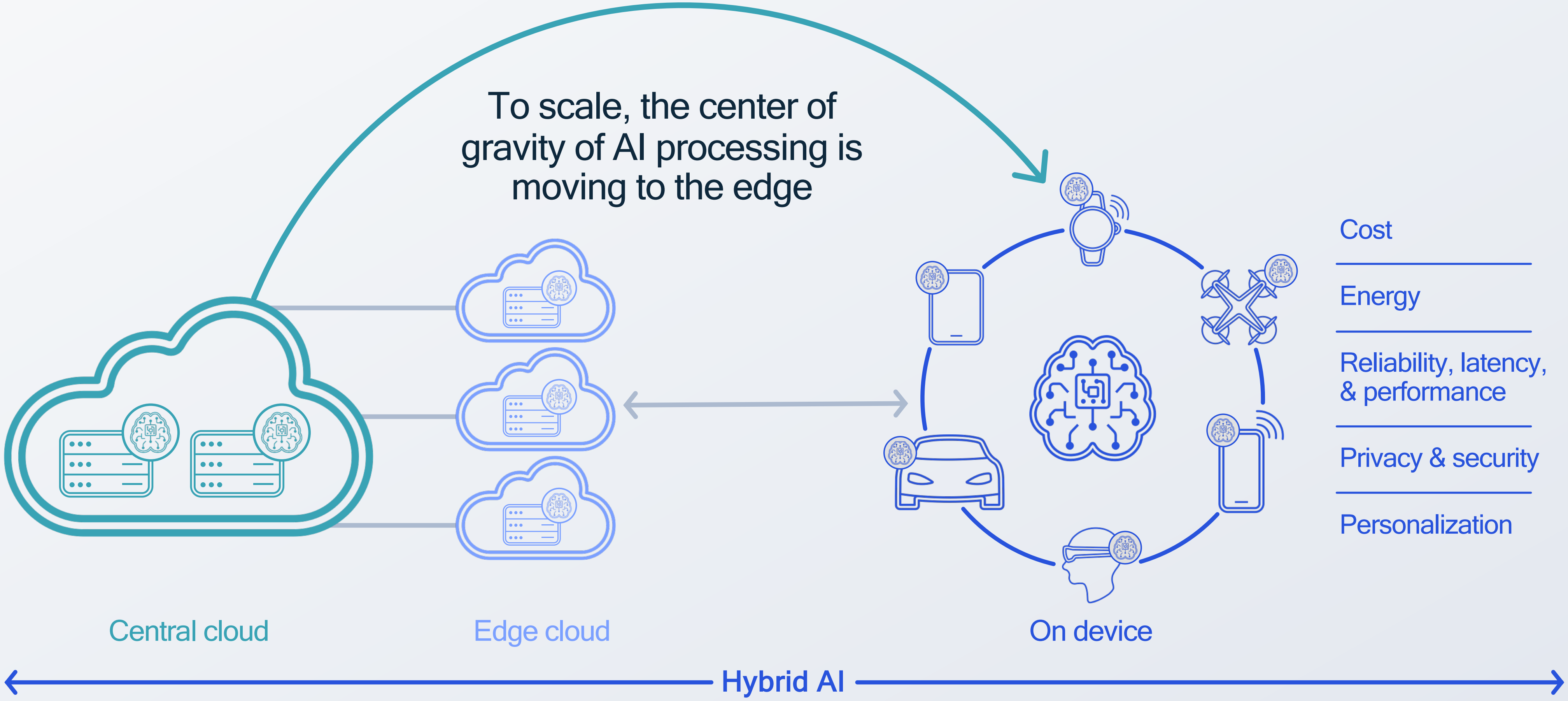
Copy creation

...



Cloud economics will not allow generative AI to scale

[1] Reuters 2023



To scale, the center of gravity of AI processing is moving to the edge

Central cloud

Edge cloud

On device

Cost

Energy

Reliability, latency, & performance

Privacy & security

Personalization

Hybrid AI

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

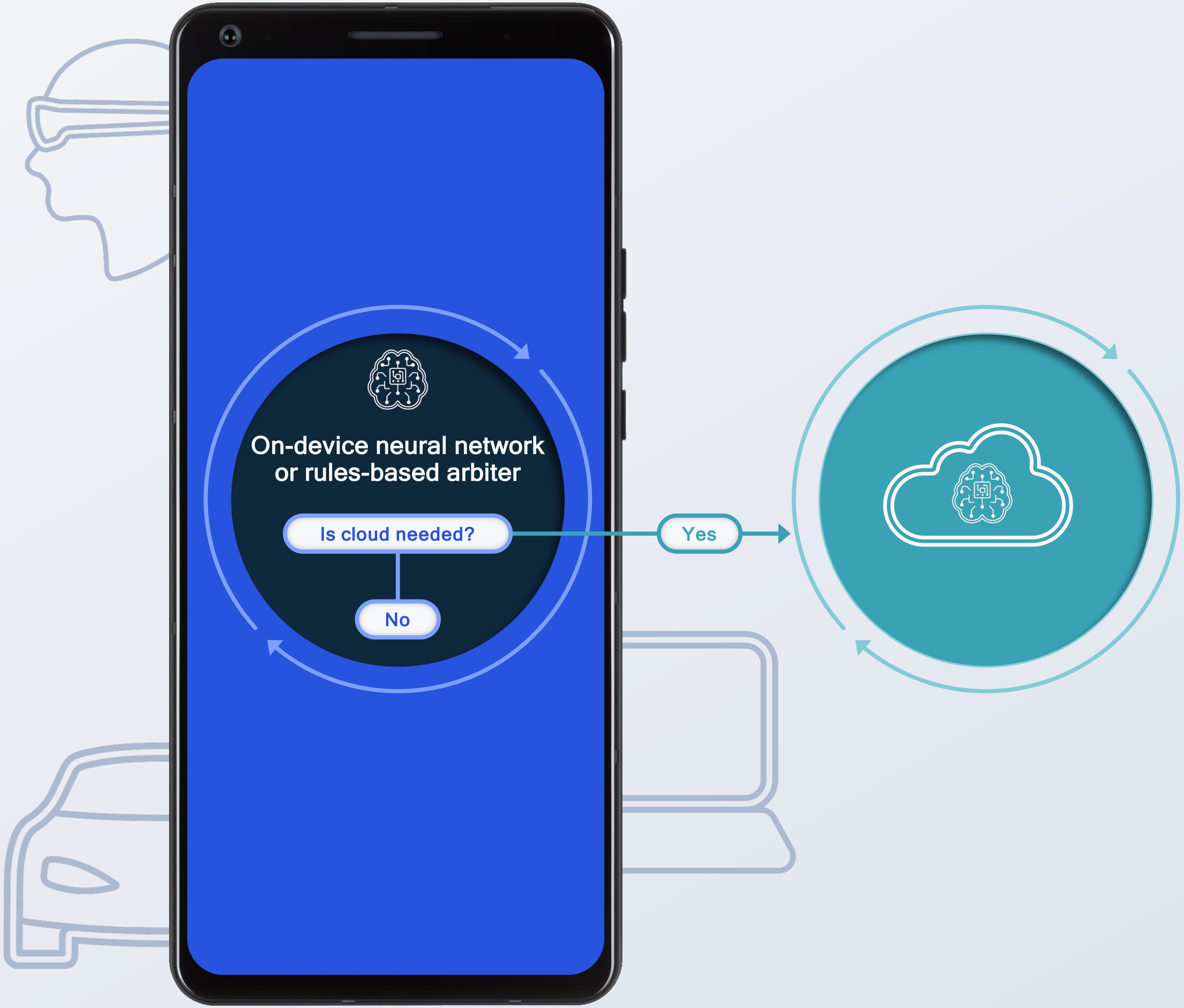
Device-centric hybrid AI

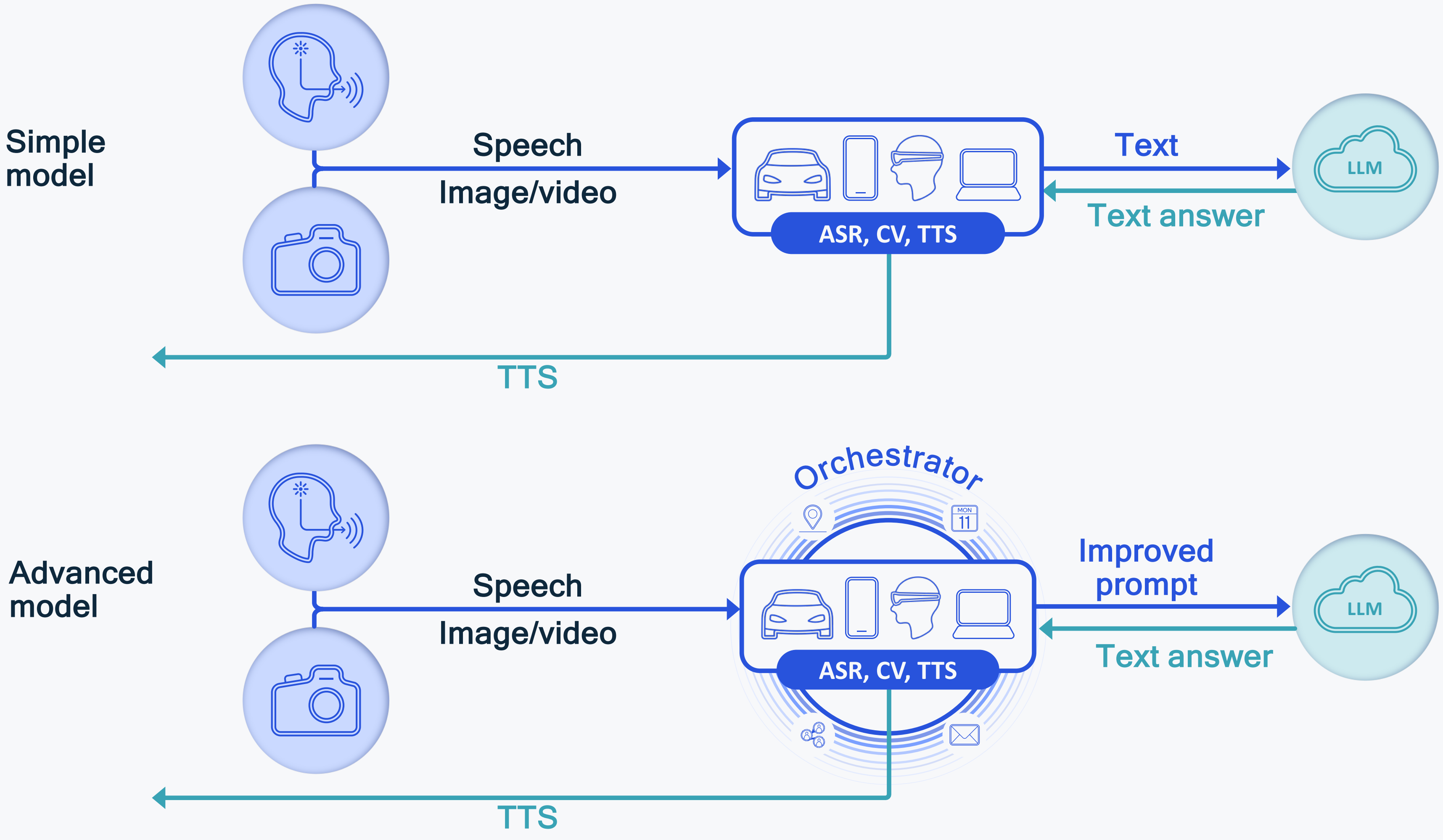
The device acts as the anchor point

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



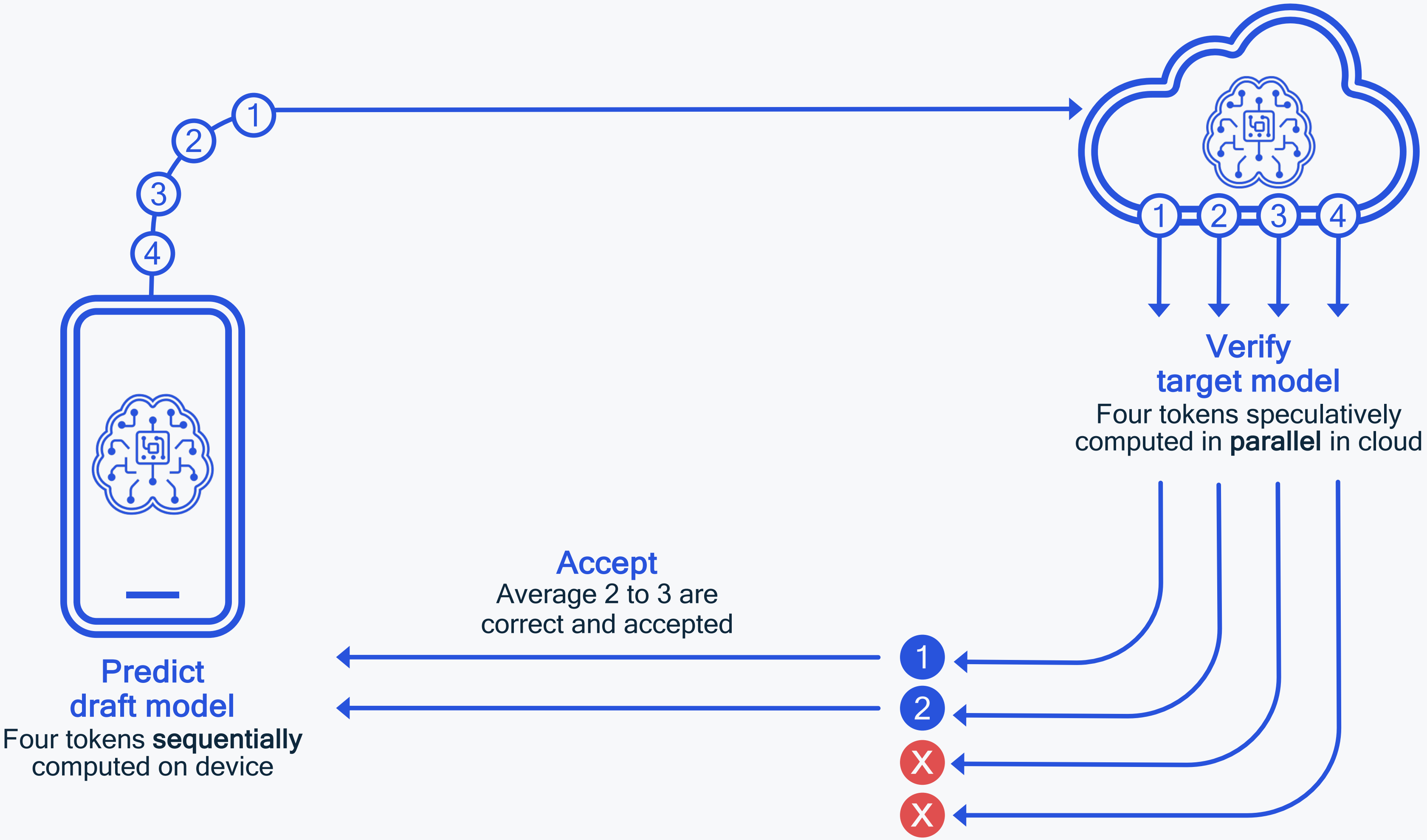


- Sensor and human-machine interface processing run on device
  - ASR, CV, TTS
- LLM runs in the cloud
- For advanced version, an on-device orchestrator uses on-device learning and personal data to provided improved prompts to the LLM

Device-sensing hybrid AI

The device acts as the eyes and ears

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



Verify
target model
Four tokens speculatively computed in **parallel** in cloud

Accept
Average 2 to 3 are correct and accepted

Predict
draft model
Four tokens **sequentially** computed on device

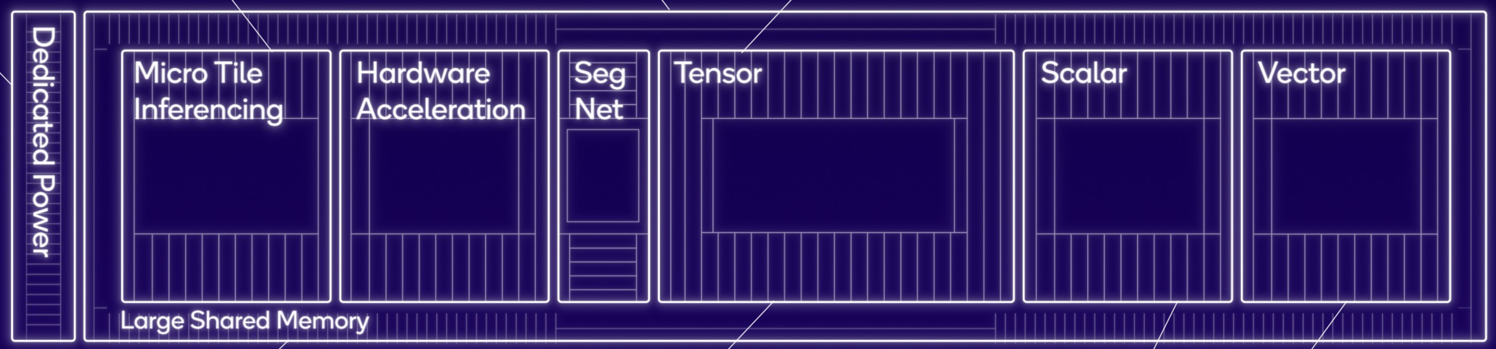
- LLMs are memory-bound and produce a single token per inference, reading in all the weights
- The smaller draft model runs on device, sequentially
- The larger target model runs on the cloud, in parallel and speculatively
- The good tokens are accepted
- Results in net speedup in tokens per unit time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

Joint-processing hybrid AI

Multi-token speculative decoding as an example

Qualcomm
Spectra ISP
Adreno GPU
Hexagon NPU
Sensing Hub
Processor Security
Kryo CPU

Qualcomm® Hexagon™ NPU



Upgraded micro tile inferencing

Micro-architecture upgrade

Dedicated power rails for accelerators

Peak performance cores

2X bandwidth

Higher bandwidth into tensor accelerator

Higher clock speed

Qualcomm® Sensing Hub

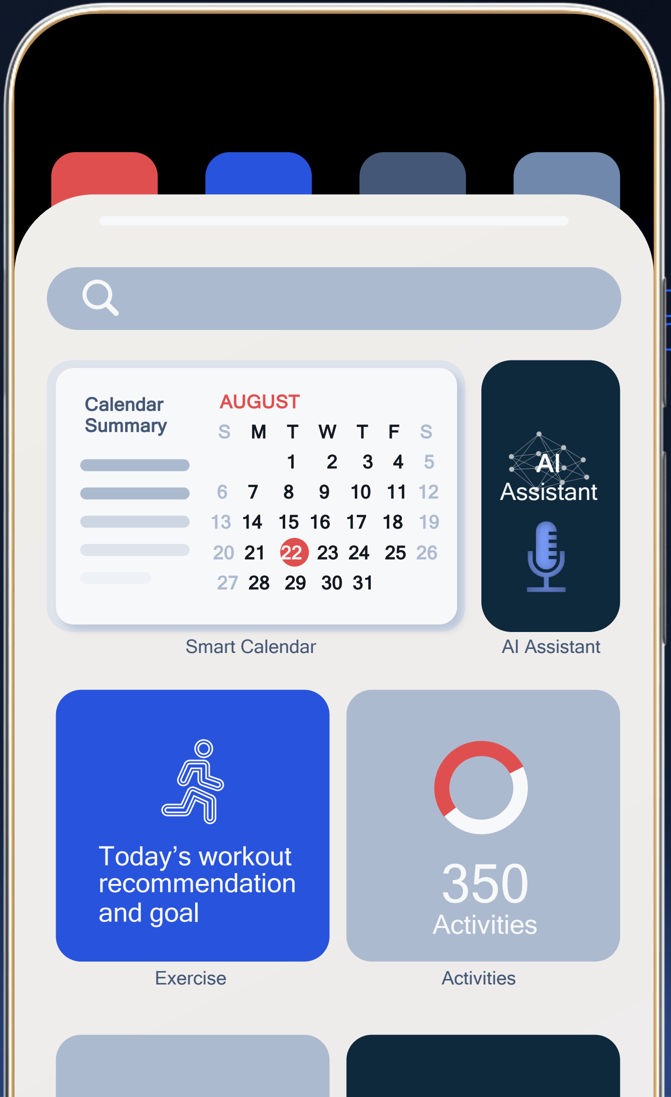
3.5X

more powerful in AI performance

30%

more memory

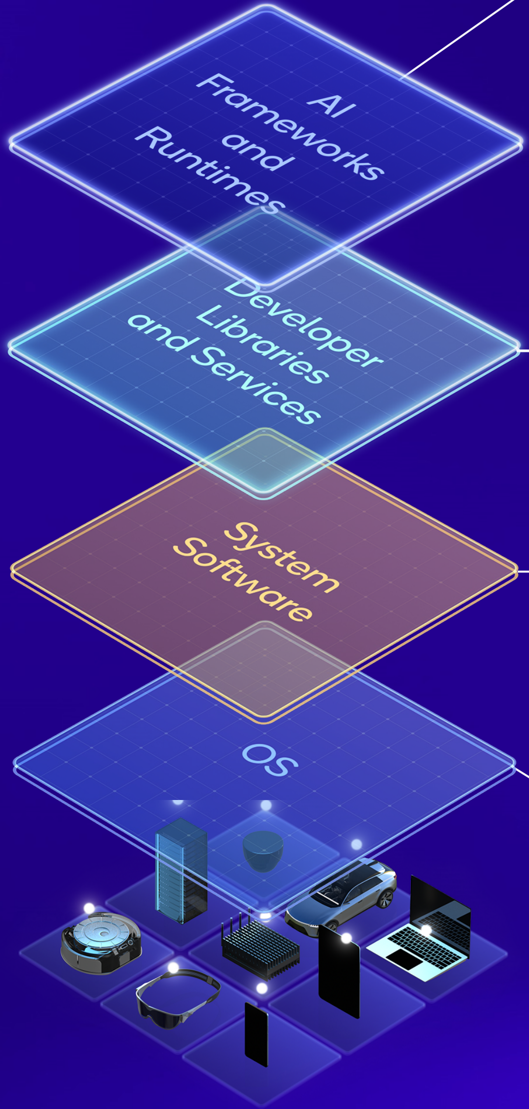
Contextual personalization



Personal profile information to support "better input prompt engineering" means a better end consumer experience..

Qualcomm AI Stack

Qualcomm AI Studio



AI Frameworks

TensorFlow PyTorch ONNX Keras

AI Runtimes

Qualcomm® Neural Processing SDK ONNX RUNTIME TF Lite Micro Direct ML TF Lite

Qualcomm® AI Engine direct

Math Libraries | Compilers | Virtual platforms

Profilers & Debuggers | Programming Languages | Core Libraries

System Interface | SoC, accelerator drivers | Emulation Support

android Windows Linux Zephyr ubuntu® CentOS QNX

Qualcomm AI Stack

Qualcomm

On-device generative AI offers many benefits

Generative AI is happening now on the device

Our on-device AI leadership is enabling generative AI to scale

Hybrid AI is the future

Questions

Connect with us

www.qualcomm.com/research/artificial-intelligence

www.qualcomm.com/news/onq

www.youtube.com/c/QualcommResearch

@QCOMResearch

https://assets.qualcomm.com/mobile- computing-newsletter-sign-up.html

www.slideshare.net/qualcommwirelessevolution

Thank you

Follow us on:

For more information, visit us at:

qualcomm.com & qualcomm.com/blog

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes our licensing business, QTL, and the vast majority of our patent portfolio. Qualcomm Technologies, Inc., a subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of our engineering, research and development functions, and substantially all of our products and services businesses, including our QCT semiconductor business.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

TINY ASIA
Seoul, South Korea
November 16, 2023

Copyright Notice