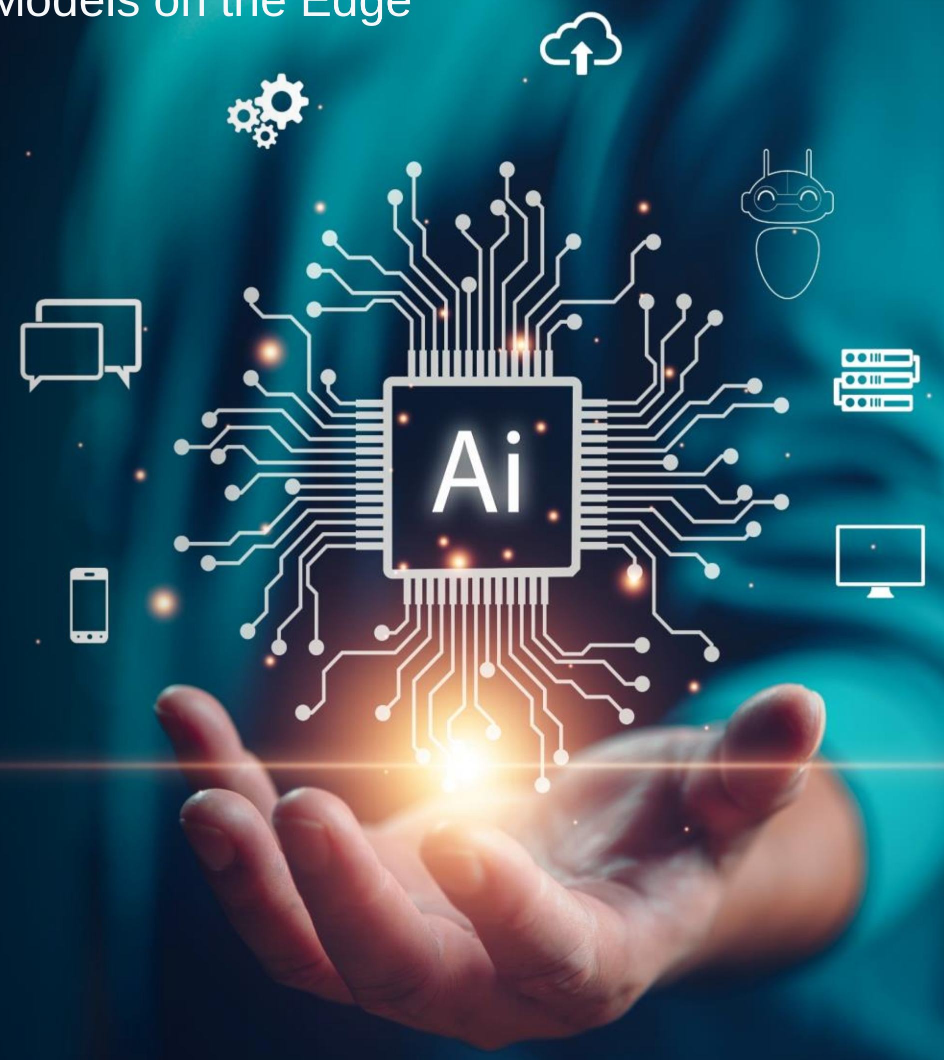
tinyAI Forum on Generative AI
and Foundation Models on the Edge

March 28, 2024

On-Device Generative AI



Fatih Porikli

Senior Director, Technology
Qualcomm Technologies, Inc.

@QCOMResearch

Today's agenda

Why on-device generative AI

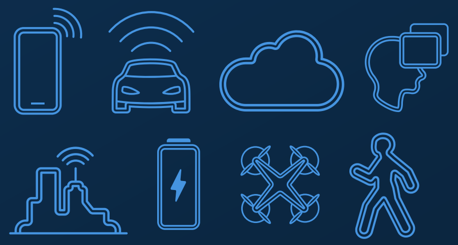
Full-stack AI optimizations for large vision models – **Stable Diffusion**

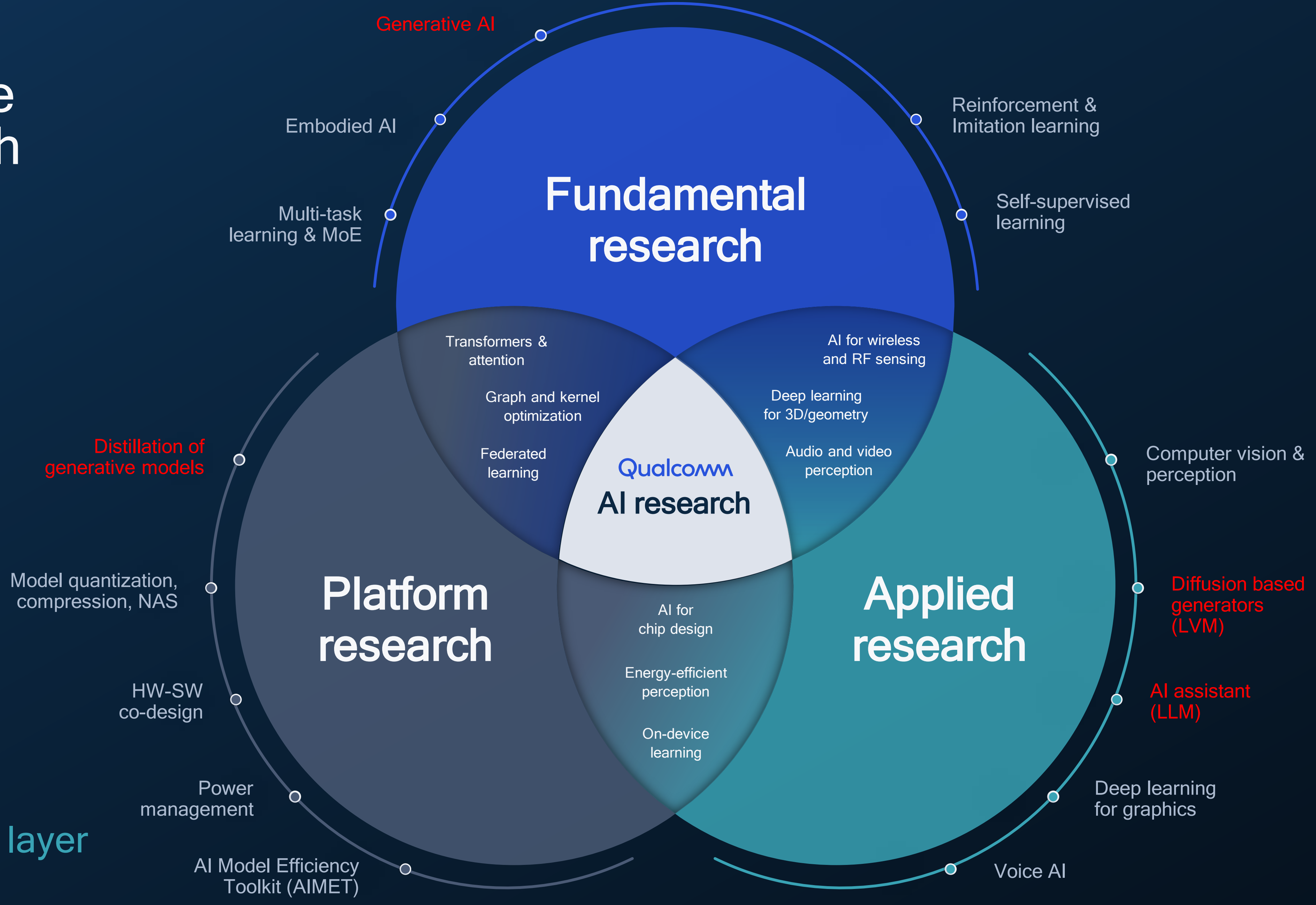
Full-stack AI optimizations for large language models – **Llama 2**

Q &A

Leading machine learning research for on-device AI

across the entire spectrum of topics





Full-stack AI research & optimization

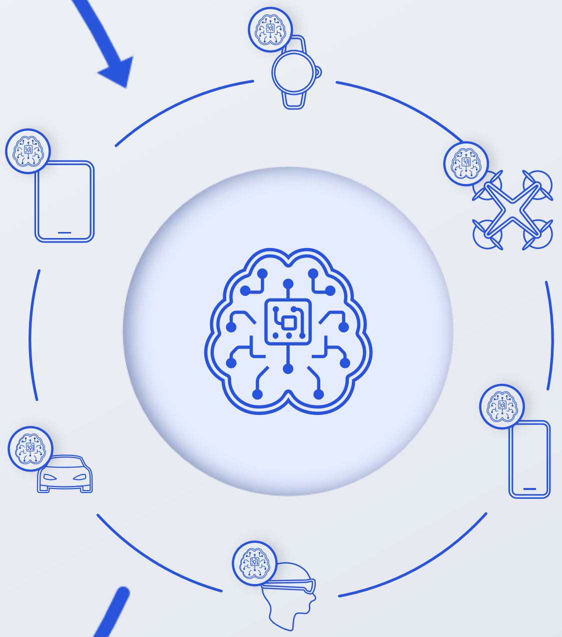
model, HW, SW innovation across each layer

LLM: Large language mode; LVM: Language vision model

AIMET is a product of Qualcomm Innovation Center, Inc. Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

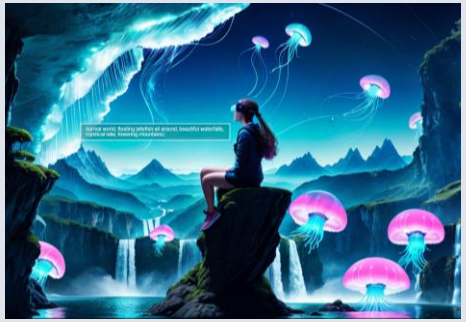
On-device intelligence is paramount

Process data closest to the source, complement the cloud



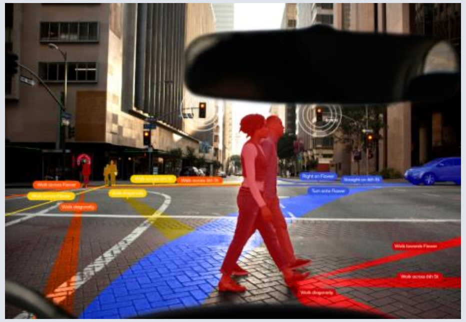
- Privacy
- Reliability
- Low latency
- Cost
- Energy
- Personalization

XR



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

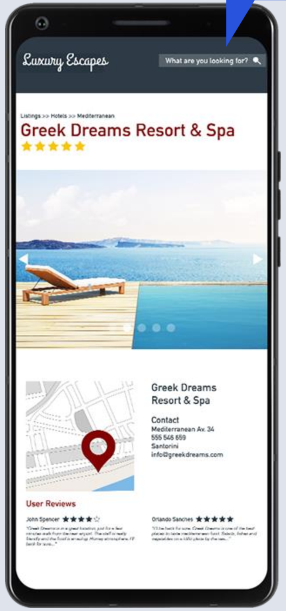
Automotive



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

Phone

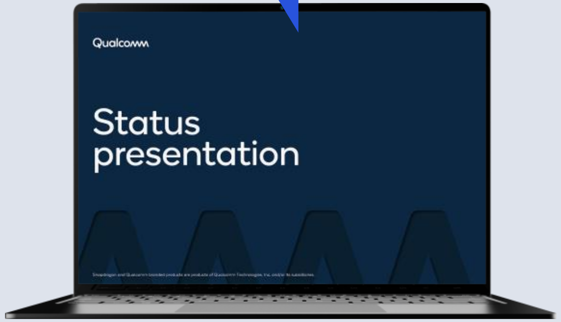
"Make me reservations for a weekend getaway at the place Bob recommended"



Gen AI can become a true digital assistant

PC

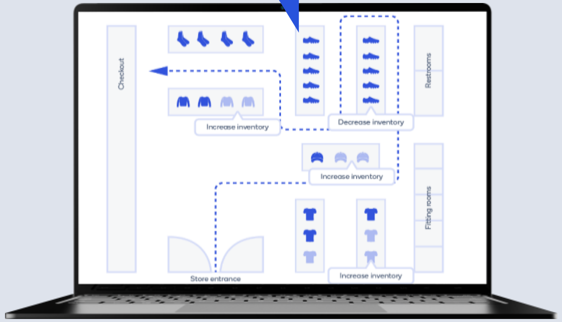
"Make me a status presentation for my boss based on inputs from my team"



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

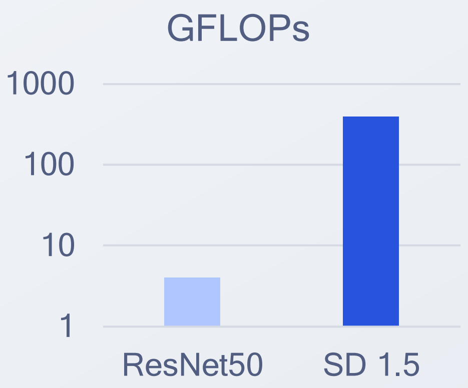
Generative AI will impact use cases across device categories

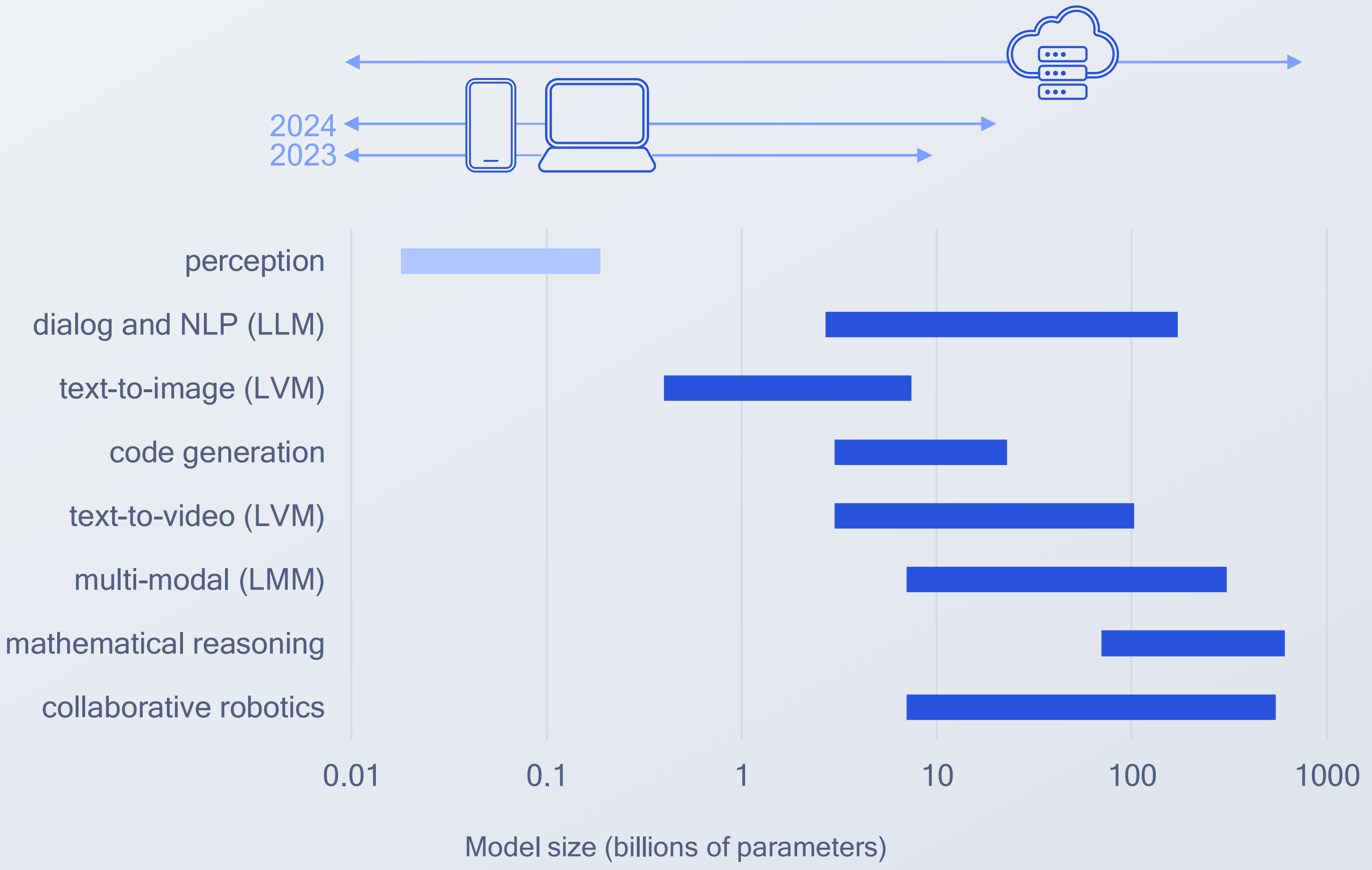
On-device AI to support a variety of Gen AI models

Compute: from GFLOPs to TFLOPs

Model size: from millions to billions of parameters

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





*Assuming INT4 parameters

What is diffusion?

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

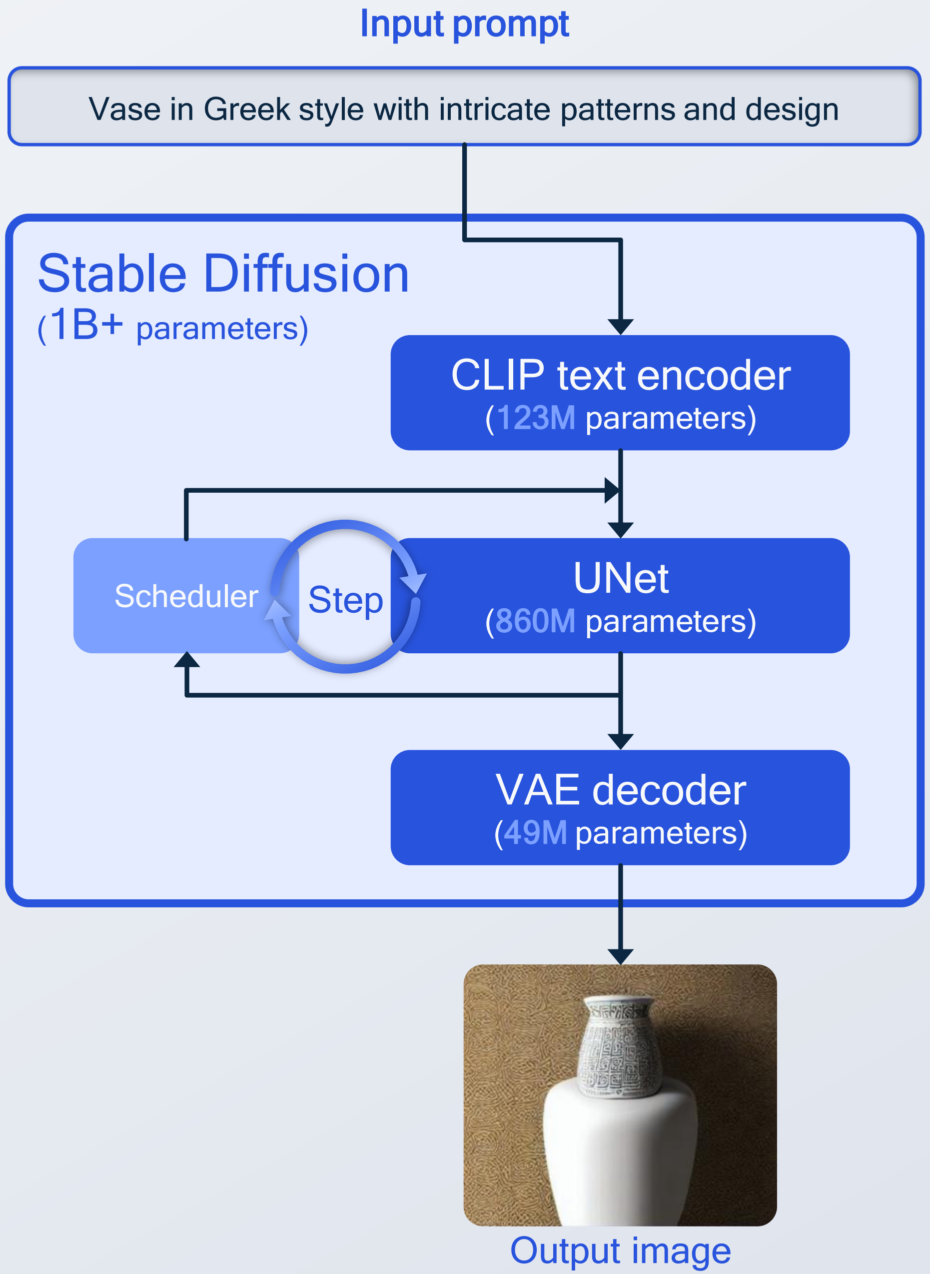
VAE: Variational Auto Encoder;
CLIP: Contrastive Language-Image Pre-Training

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



Knowledge distillation

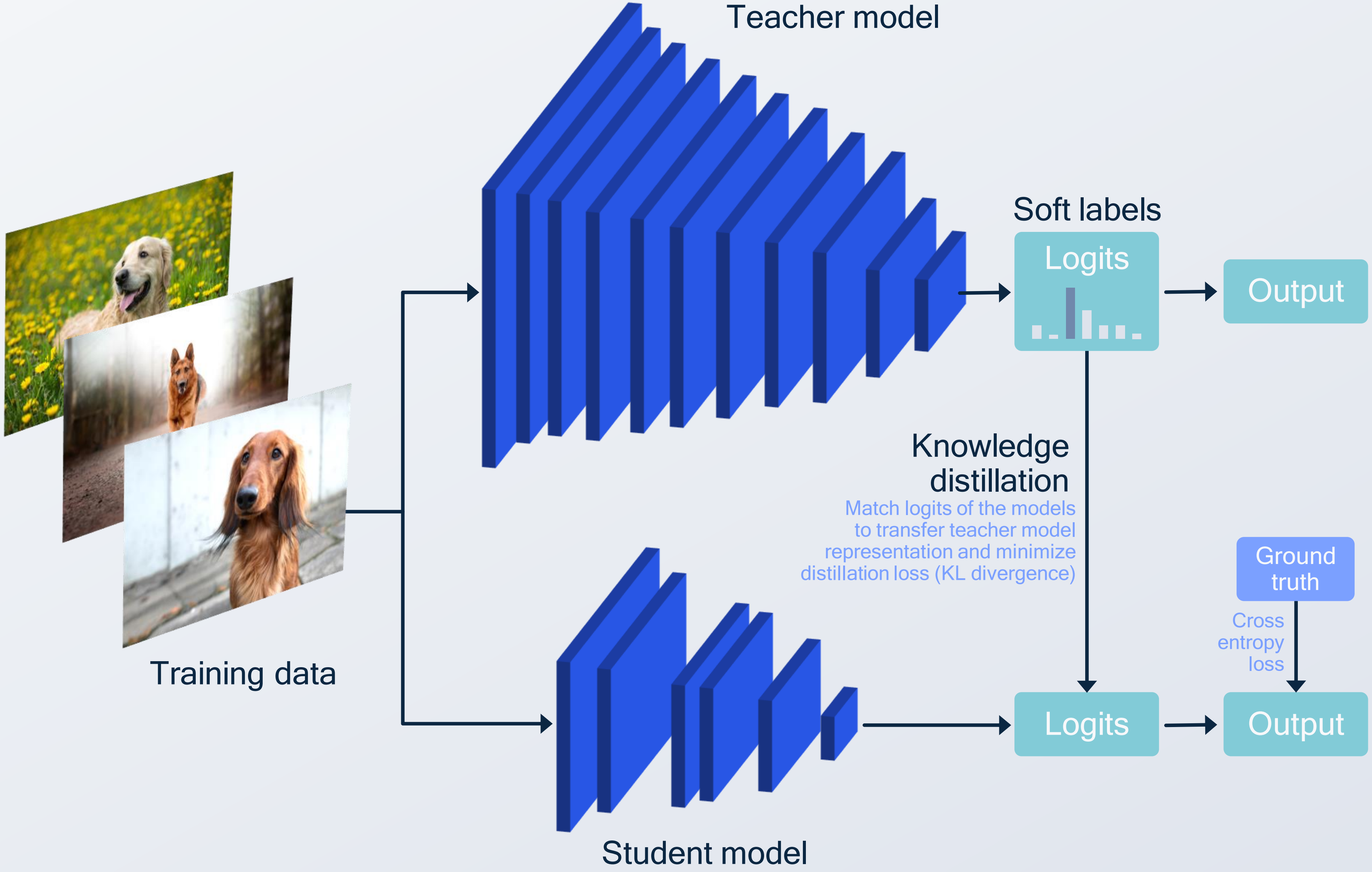
Training a smaller "student" model to mimic a larger "teacher" model

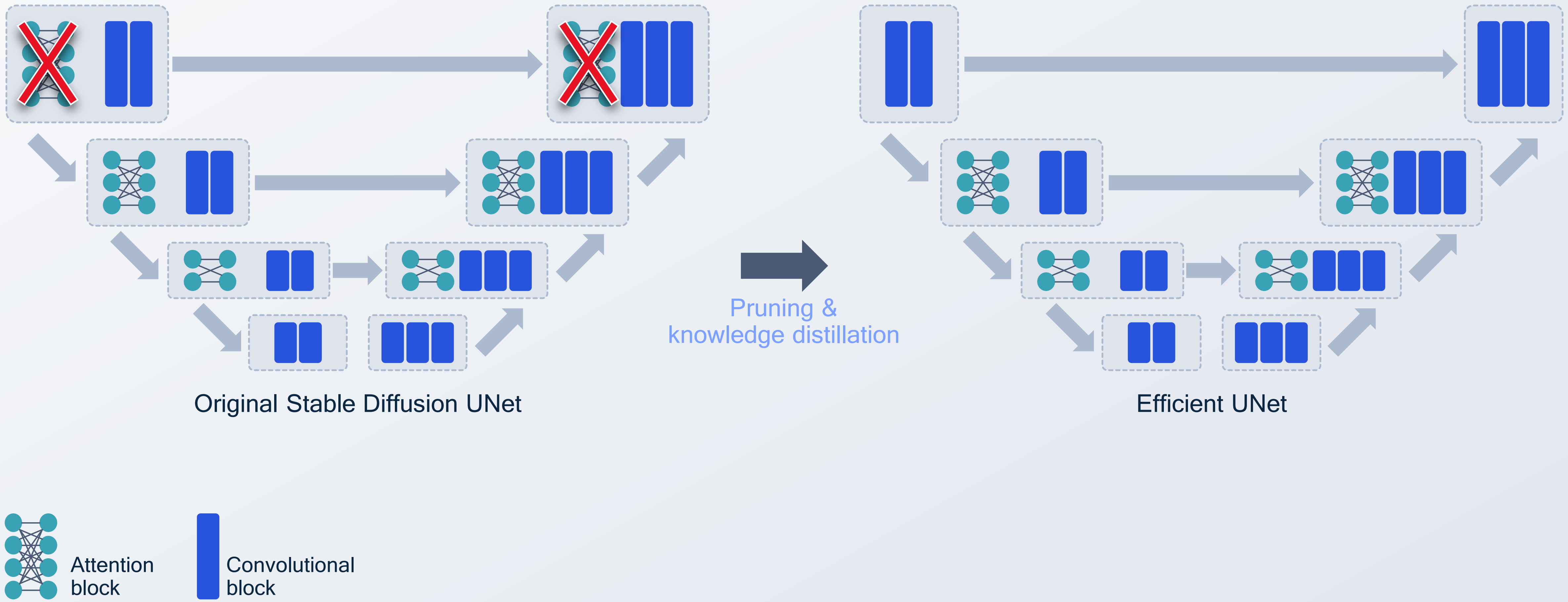
Create a smaller model with fewer parameters

Run faster inference on target deployment

Maintain prediction quality close to the teacher

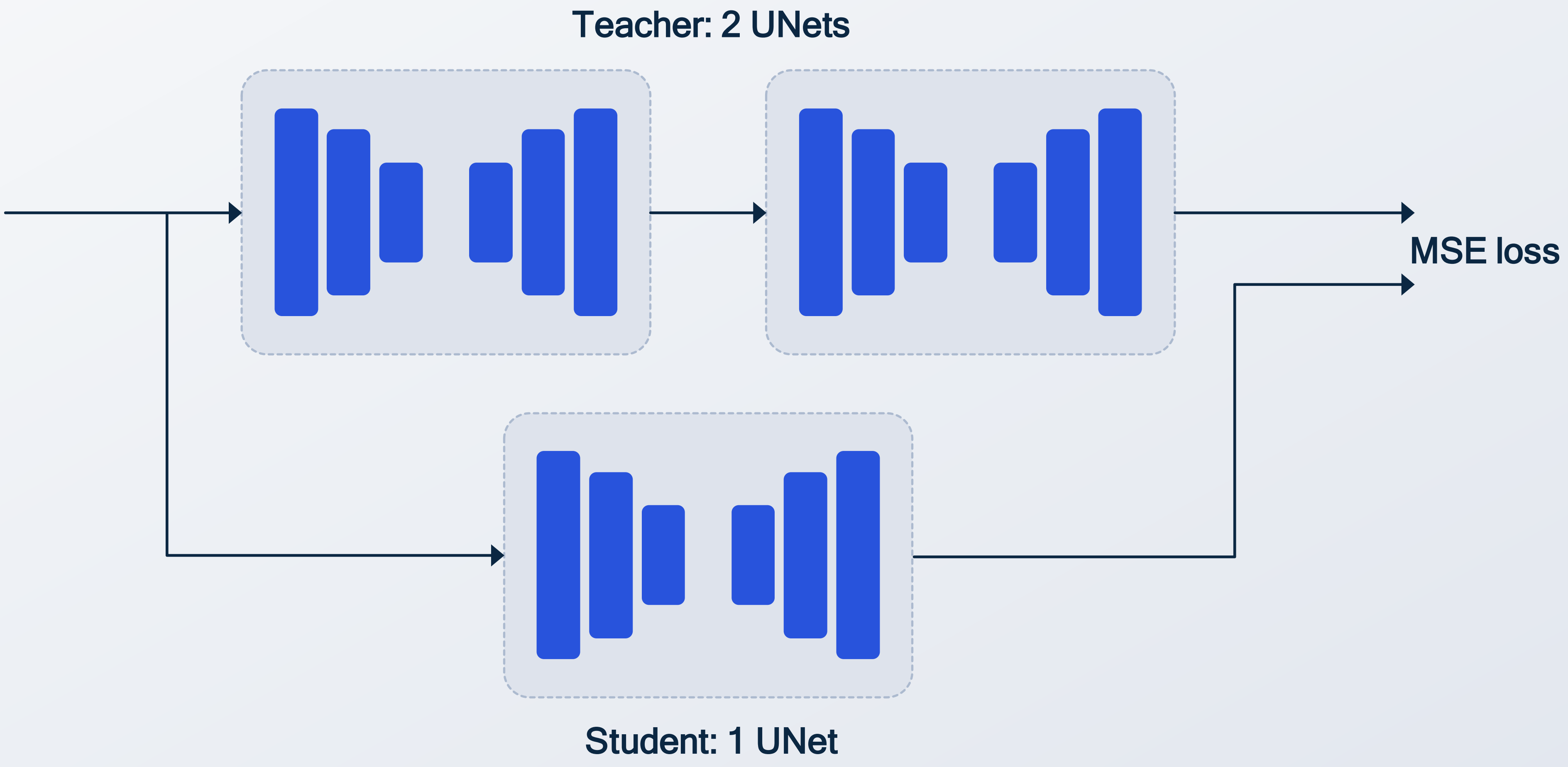
Less training time





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

Baseline Stable Diffusion → e-to-v → Efficient UNet → Guidance conditioning → Step distillation → Fast 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

	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

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

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

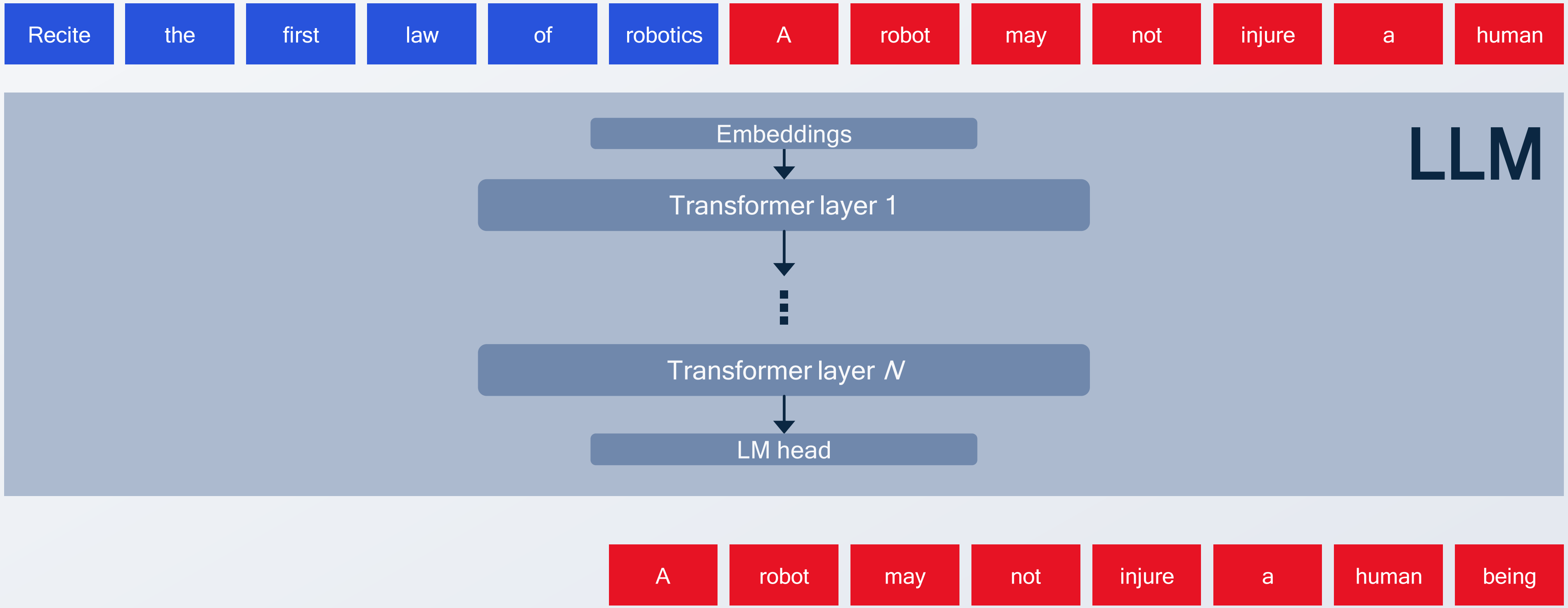
Qualcomm® AI Engine direct for improved performance and minimized memory spillage

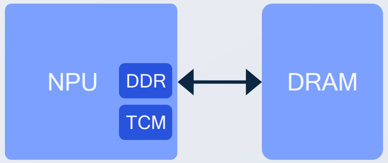
AI acceleration on the Qualcomm® Hexagon™ NPU of the Snapdragon® 8 Gen 3 Mobile Processor

LVM: Language vision model

Illustration of autoregressive language modeling

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



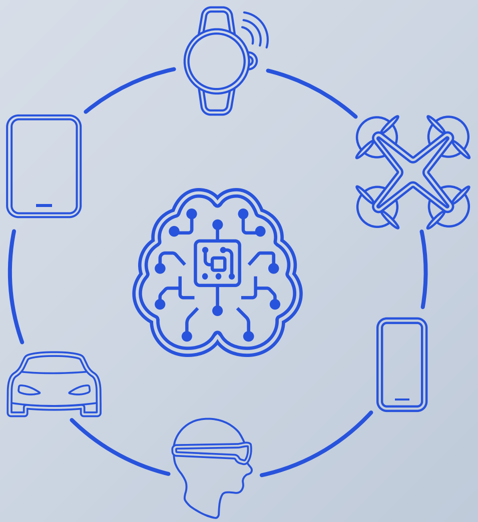


Huge bandwidth
Each parameter of the model must be read to generate each token (e.g., read 7B parameters for Llama 7B to generate a single token)

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

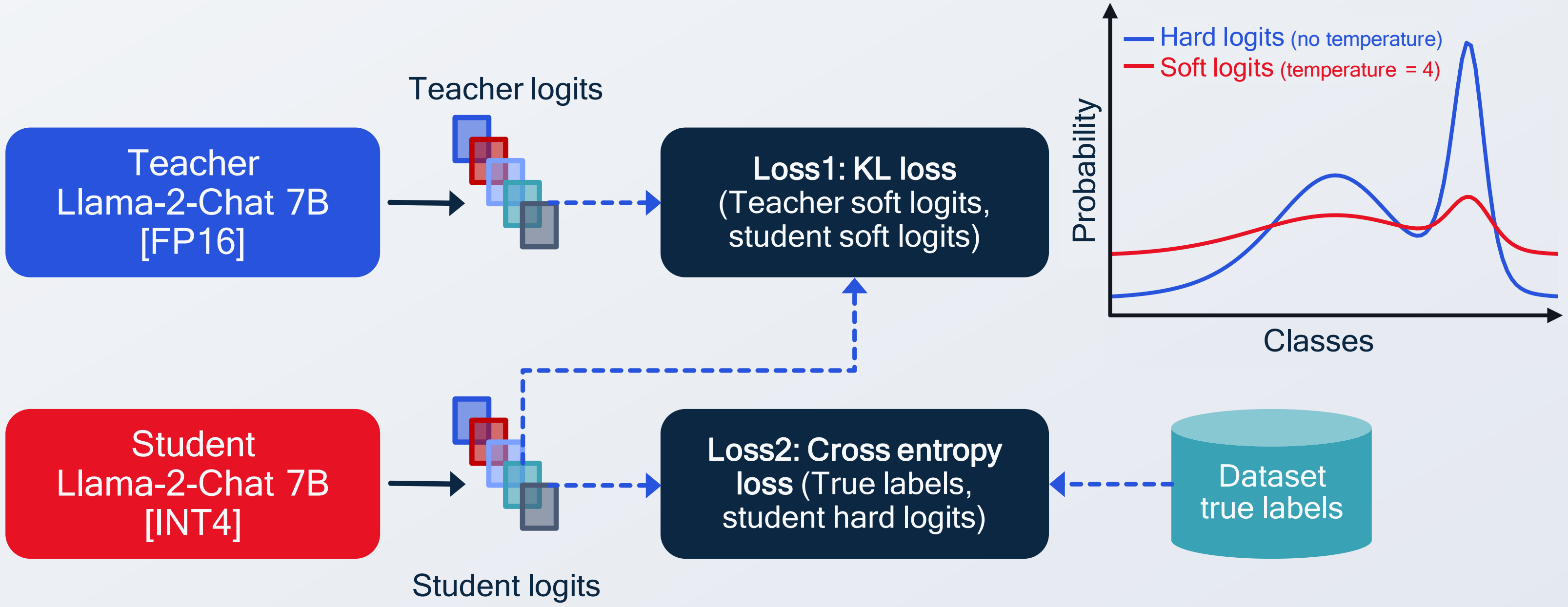
LLM quantization challenges

- 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

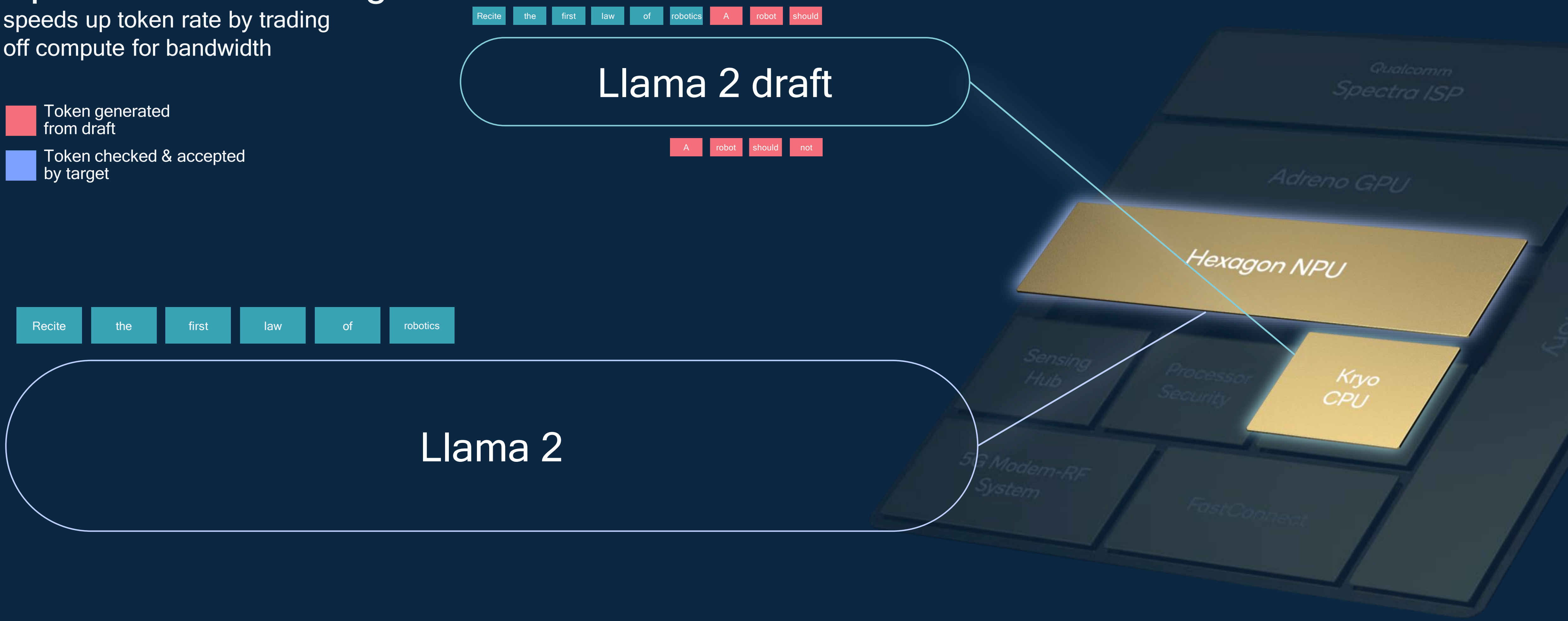
<1
Point increase in perplexity[1]

<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



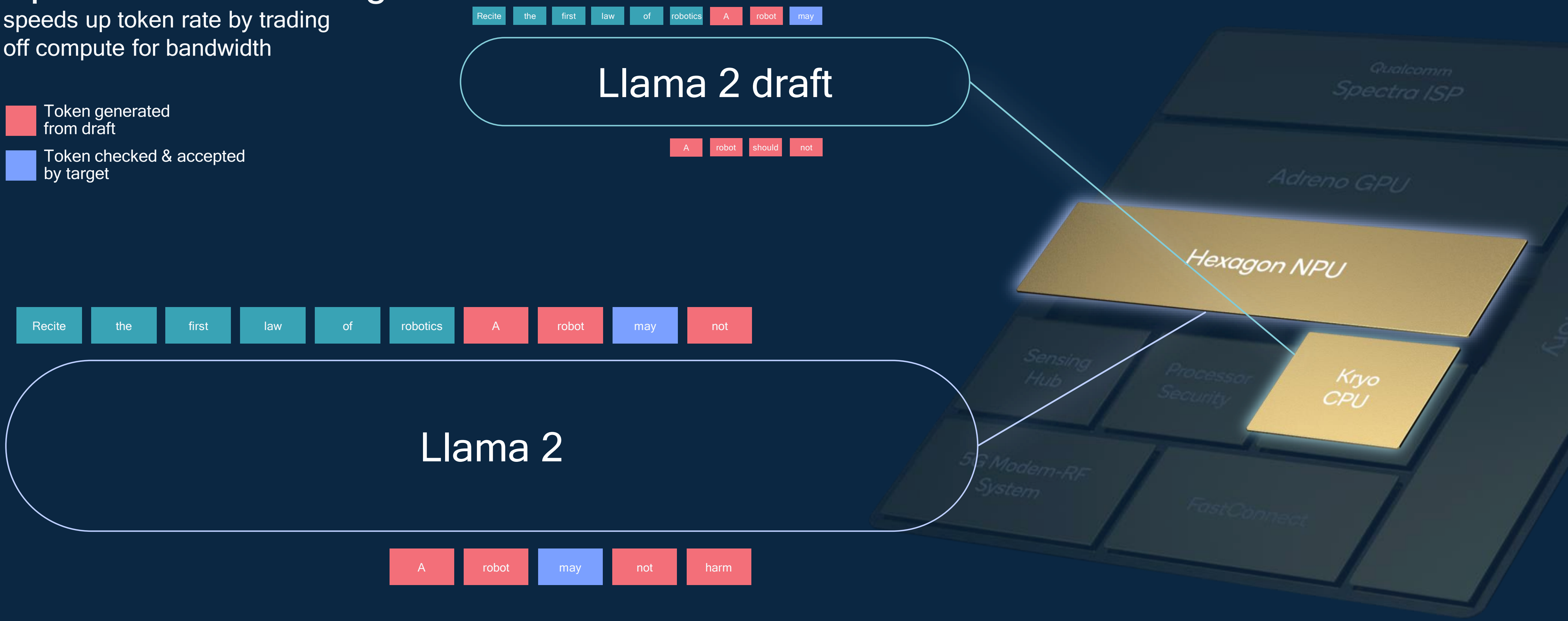
Draft model generates a few
speculative tokens at a time

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



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

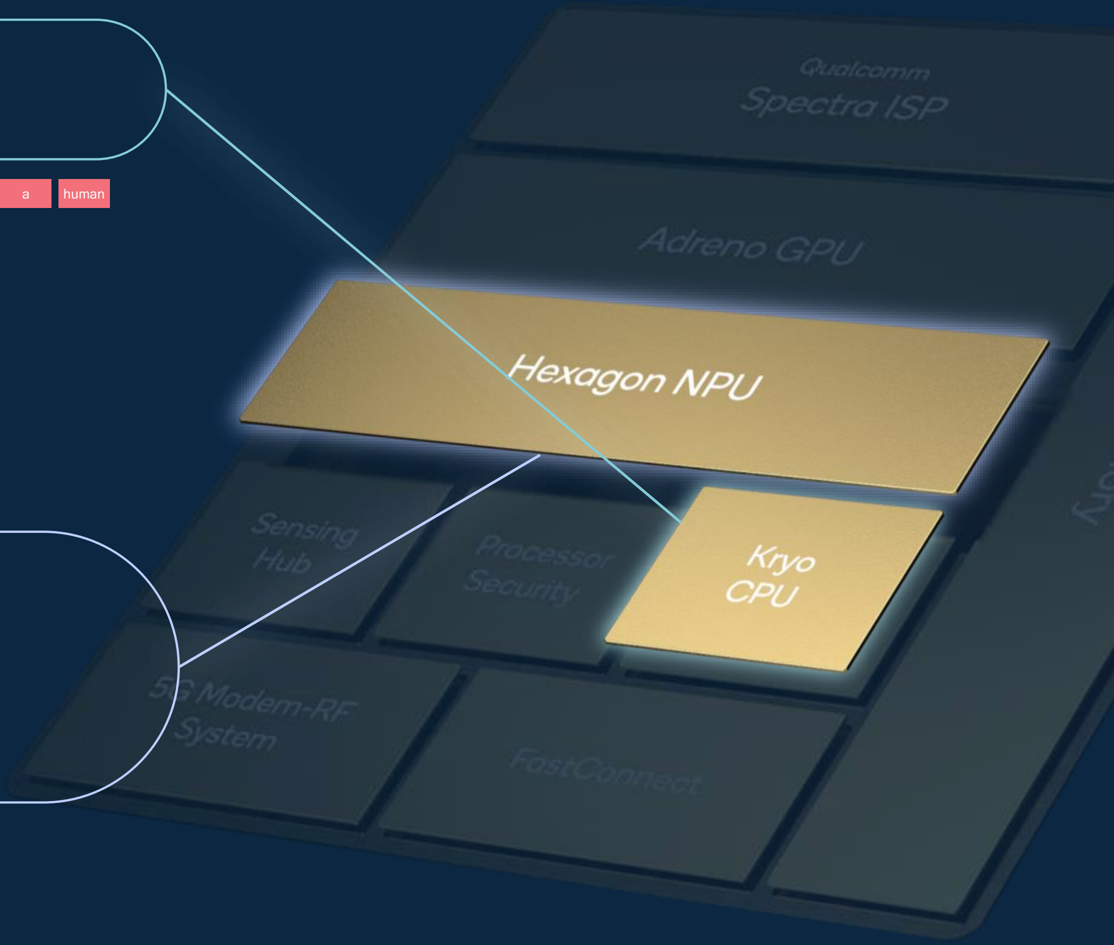
Recite | the | first | law | of | robotics | A | robot | may | not | injure | a

Llama 2 draft

not | injure | a | human

Recite | the | first | law | of | robotics | A | robot | may

Llama 2



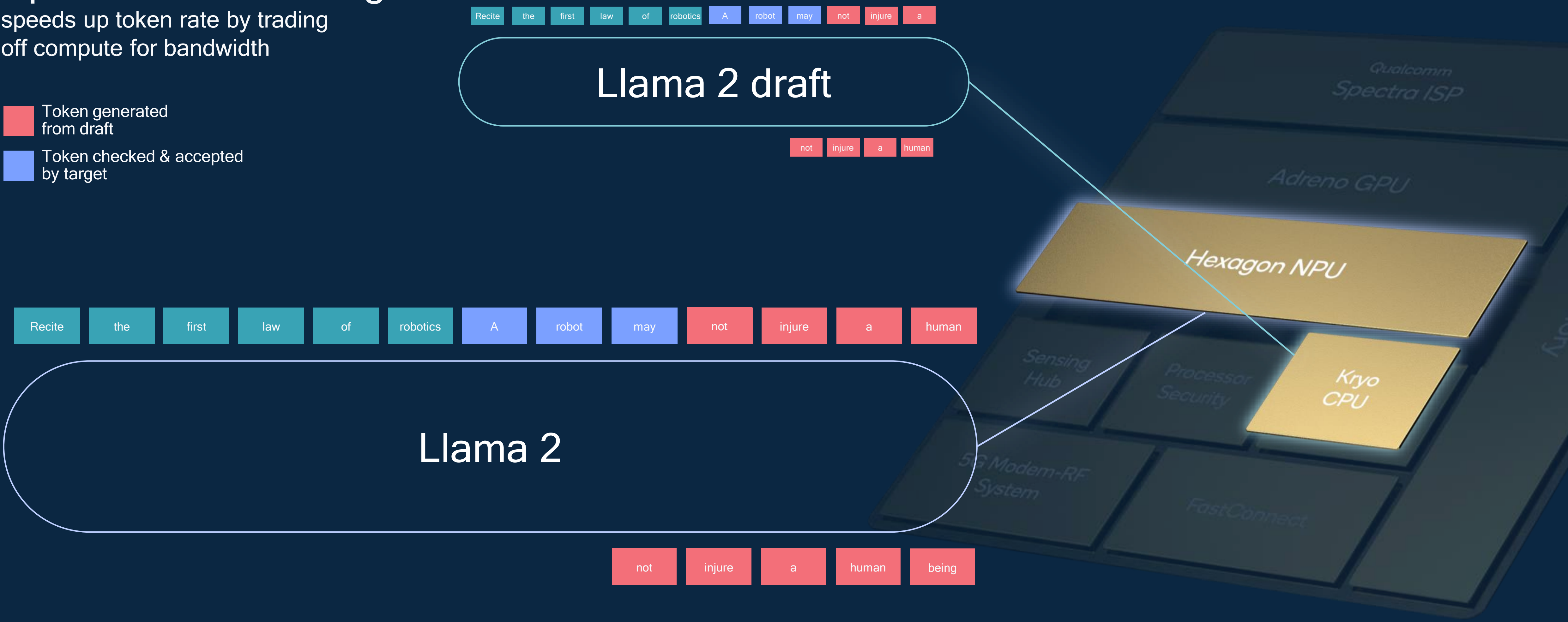
Draft model generates a few speculative tokens at a time

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



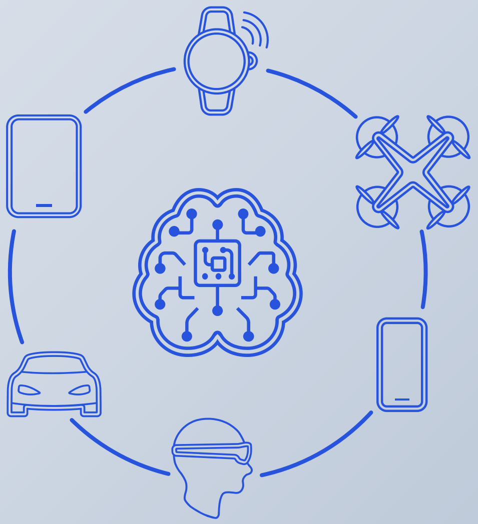
Draft model generates a few speculative tokens at a time

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

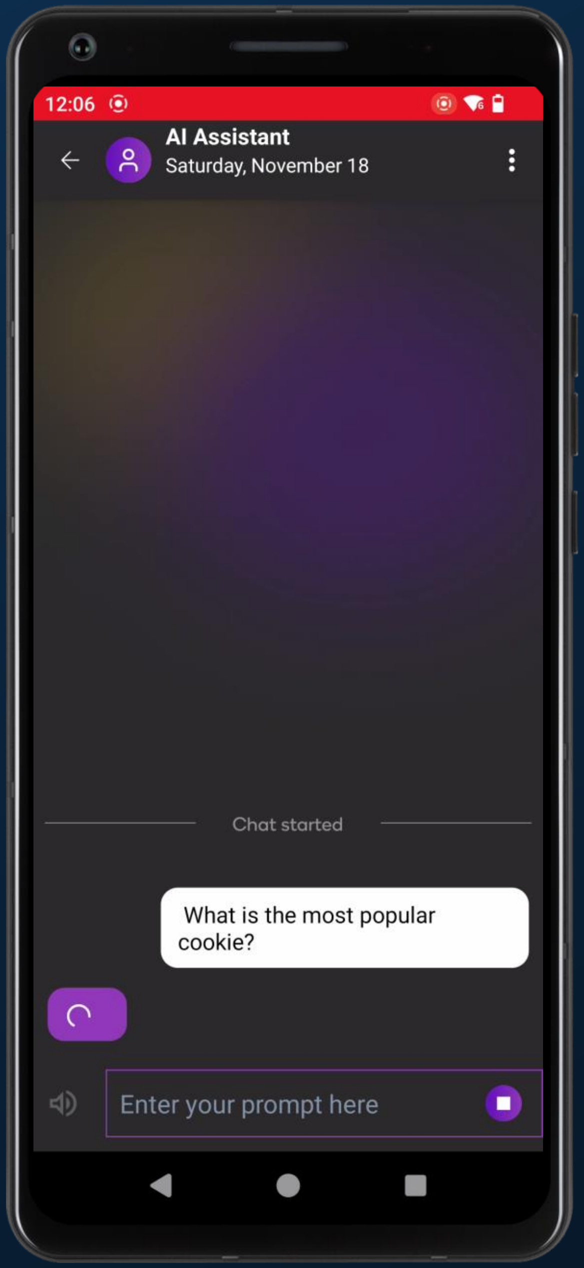
Up to

20

tokens per second

At Snapdragon Summit 2023

World’s fastest Llama 2-7B on a phone



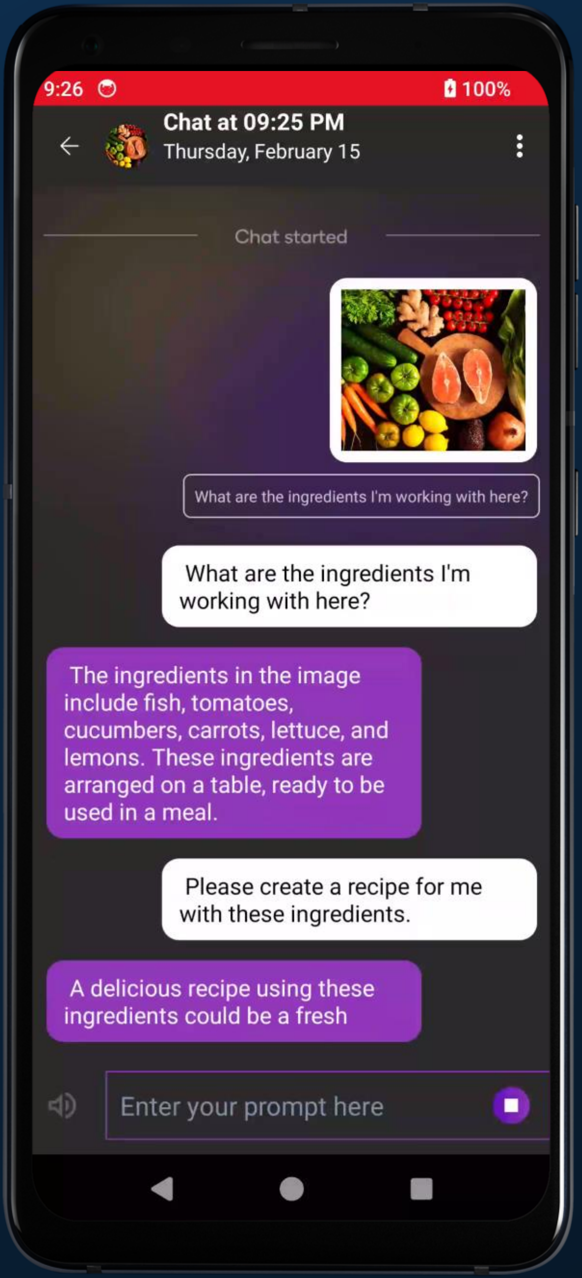
Up to 20 tokens per second

Demonstrating both chat and application interaction on device

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

At MWC 2024

World's first large multimodal model (LMM) on an Android phone



LLMs can now see

7+ billion parameter LMM, LLaVA, with text, speech, and image inputs

Multi-turn intuitive conversations about an image at a responsive token rate

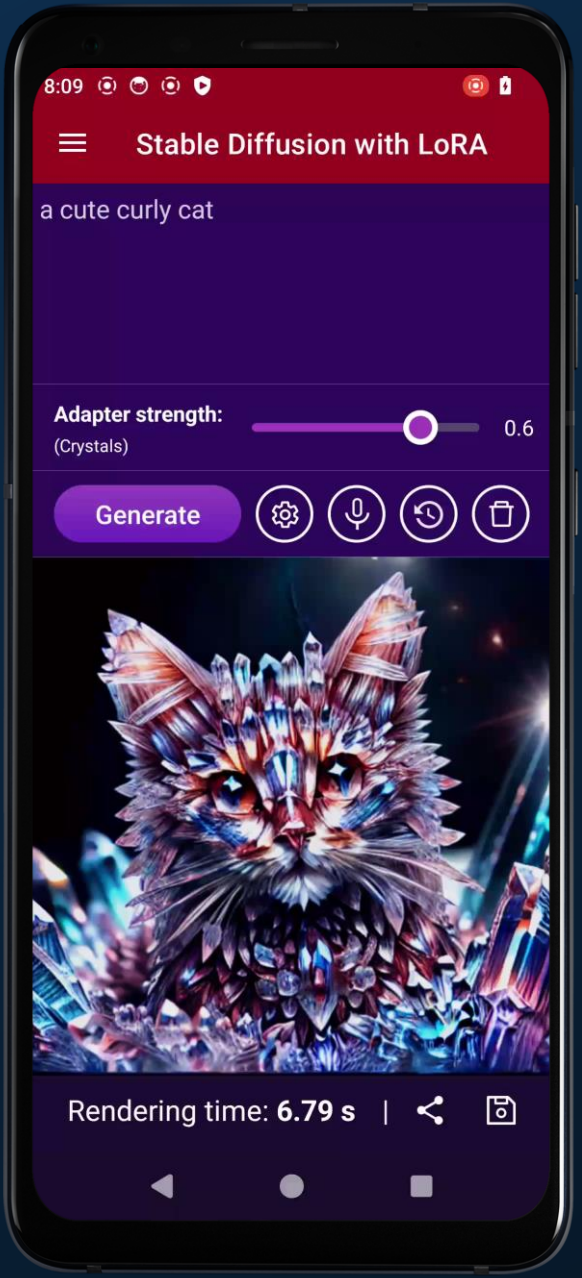
Full-stack AI optimization to achieve high performance at low power

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

LLM: Large Language Model; LLaVA: Large Language and Vision Assistant

At MWC 2024

Our first low rank adaptation (LoRA) on an Android phone



1+ billion parameter Stable Diffusion with LoRA adapter for customized experiences

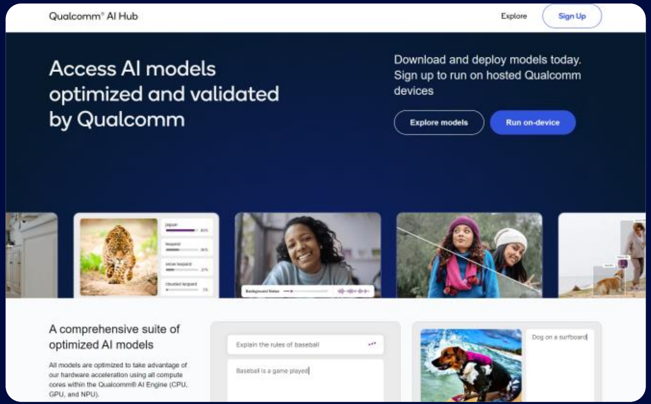
LoRA enables scalability and customization of on-device generative AI across use cases

Full-stack AI optimization to achieve high performance while fast switching between adapters and minimizing memory need

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

Qualcomm AI Hub

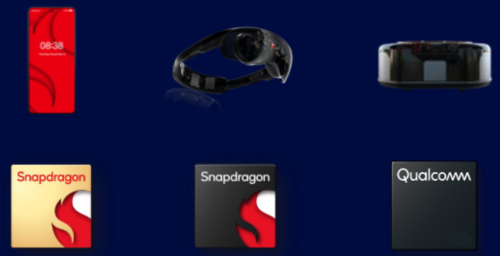
Library of fully optimized AI models for deployment across Snapdragon and Qualcomm platforms



AIHUB.QUALCOMM.COM

Model Selection — Target Platform — Test and validate — Deployment

Qualcomm AI Stack



```
import qai_hub as hub

# select device
device = hub.get_device("qualcomm-snapdragon-8gen2")

# produce model
job – hub.submit_compile_and_profile_job(torce_model
name="MyDetector",
device=device,
input_shapes=[(3, 720, 1024)]

# deploy to device
model = job.download_target_model ()
```

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

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

Qualcomm

Follow us on:

For more information, visit us at:

qualcomm.com & qualcomm.com/blog