

TECH TALK

Edge AI on Devices

Running AI models on phones and sensors

PRESENTED BY

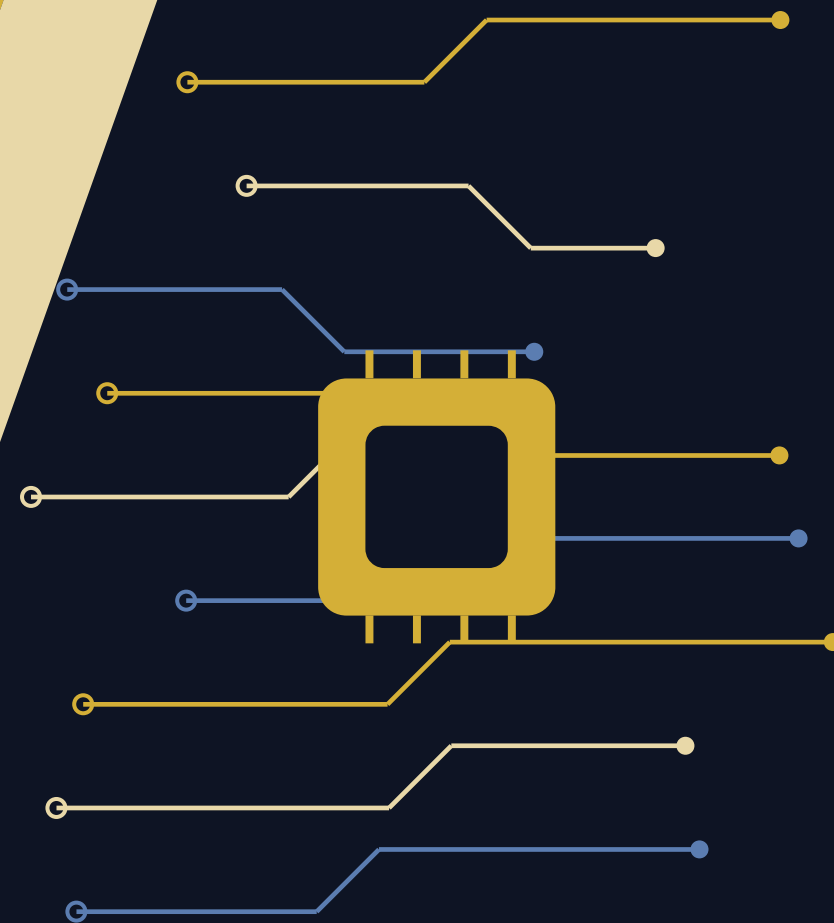
Presenter Name

DATE

2026

CONTACT

hello@yourdomain.in



Agenda

01

Introduction

02

Architecture

03

Trends

04

Adoption



**The fastest AI is the
one that never leaves
your device.**

Why it matters

Overview

- ▶ Edge AI runs models directly on devices
- ▶ It cuts latency and keeps data private
- ▶ Used in phones, cameras, cars and factories
- ▶ Small, efficient models make it possible

KEY IDEA

**Quantised models
can be four times
smaller.**

How it works

1

Train

In the cloud

2

Compress

Prune and quantise

3

Deploy

To the device

4

Update

Over the air

Core components

01

Models

Small, efficient networks.

02

Hardware

NPU and microcontrollers.

03

Tooling

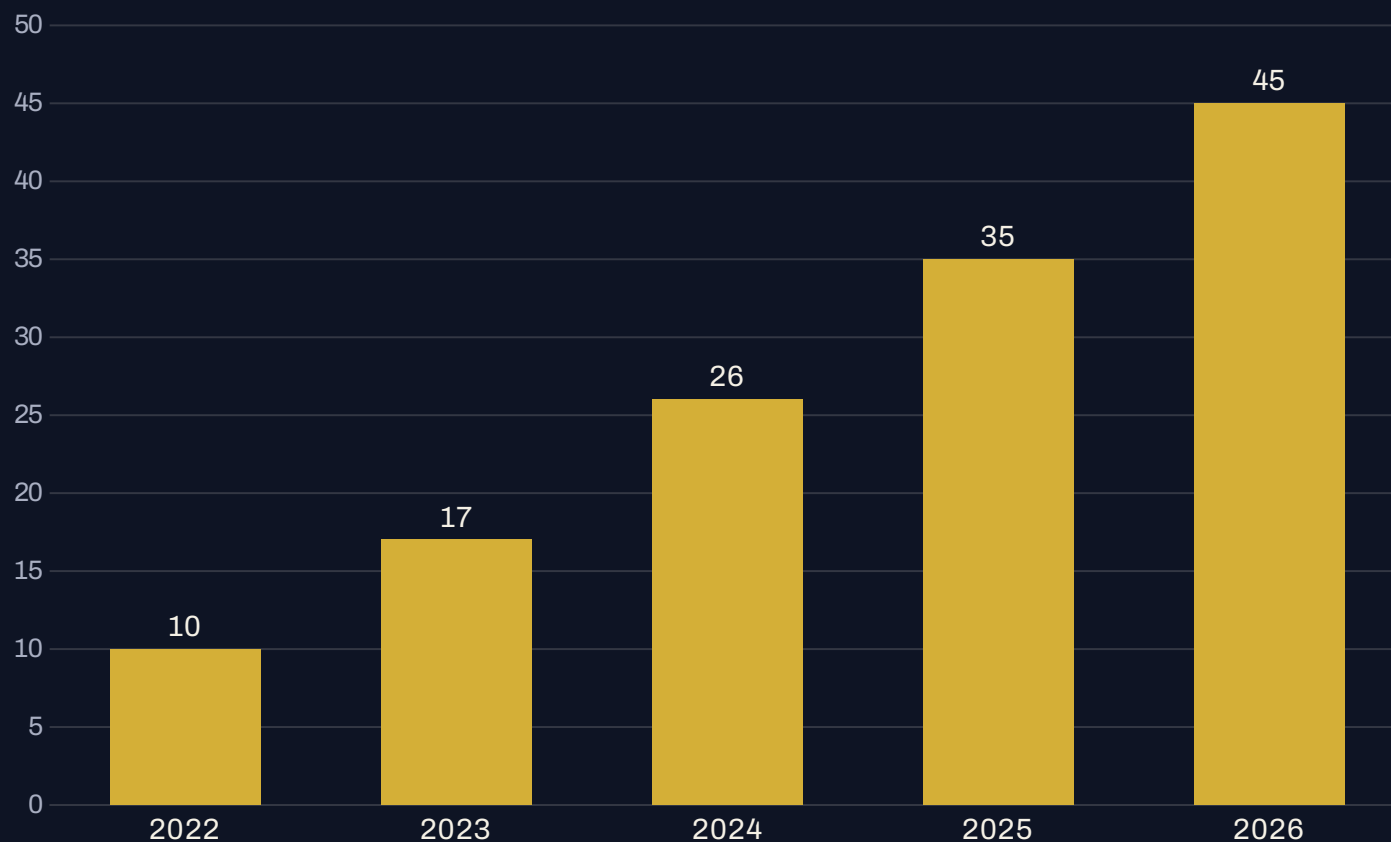
Converters and runtimes.

Tools and options

Layer	Popular choices	Notes
Runtimes	Mobile and embedded runtimes	Optimised inference
Hardware	NPU, GPU, MCU	Match to power budget
Techniques	Quantisation, pruning	Smaller, faster models

Generic examples; pick what fits your project.

Adoption over time



Adoption keeps rising as tools mature (sample data).

Sample data for illustration — replace with your own figures.

Use cases



Image 1

Smartphones

Photo and voice features



Image 2

Security

On-camera detection



Image 3

Industry

Defect inspection

Benefits and risks

Benefits

- Low latency
- Privacy by design
- Works offline

vs

Risks

- Limited compute
- Model updates are harder
- Battery impact

Best practices

01

Measure first

Profile on real devices.

02

Compress smartly

Check accuracy after.

03

Plan updates

Version your models.

Getting started roadmap

Week 1

Learn

Core concepts and a tutorial

Week 3–4

Pilot

One real use case

Week 2

Prototype

Small proof of concept

Month 2

Scale

Standards, monitoring,
training

Questions?

Thanks for your time.

Presenter Name

hello@yourdomain.in