

TECH TALK

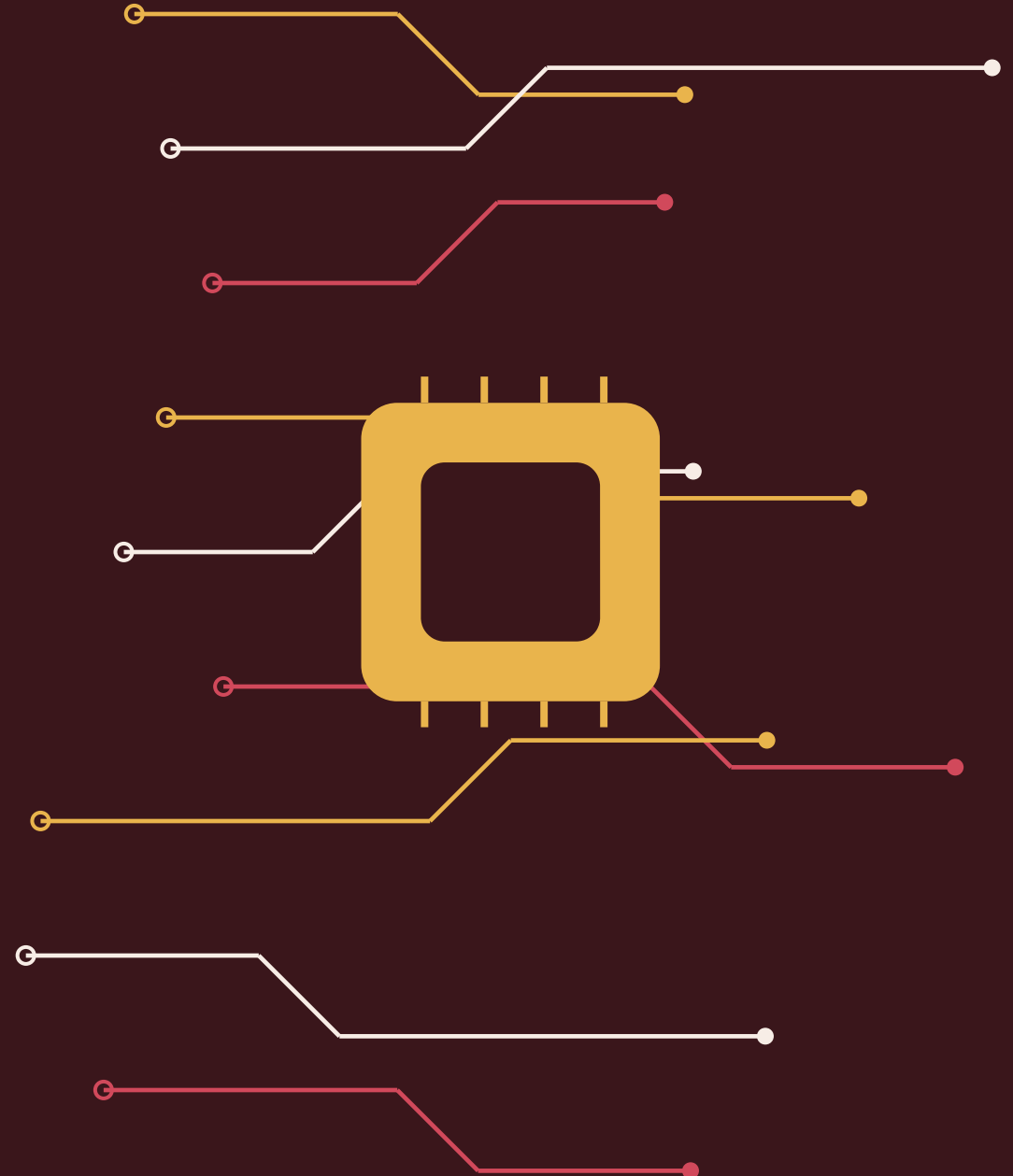
Edge AI on Devices

Running AI models on phones and sensors

PRESENTED BY
Presenter Name

DATE
2026

CONTACT
hello@yourdomain.in





02 | Agenda

01 Introduction

02 Architecture

03 Trends

04 Adoption



Image

SECTION 01

INTRODUCTION



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

Why it matters



05 | 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.**

SECTION 02

ARCHITECTURE



07 | How it works



Train

In the cloud



Compress

Prune and quantise



Deploy

To the device



Update

Over the air



08 | Core components

01

Models

Small, efficient networks.

02

Hardware

NPU and
microcontrollers.

03

Tooling

Converters and runtimes.



09 | 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

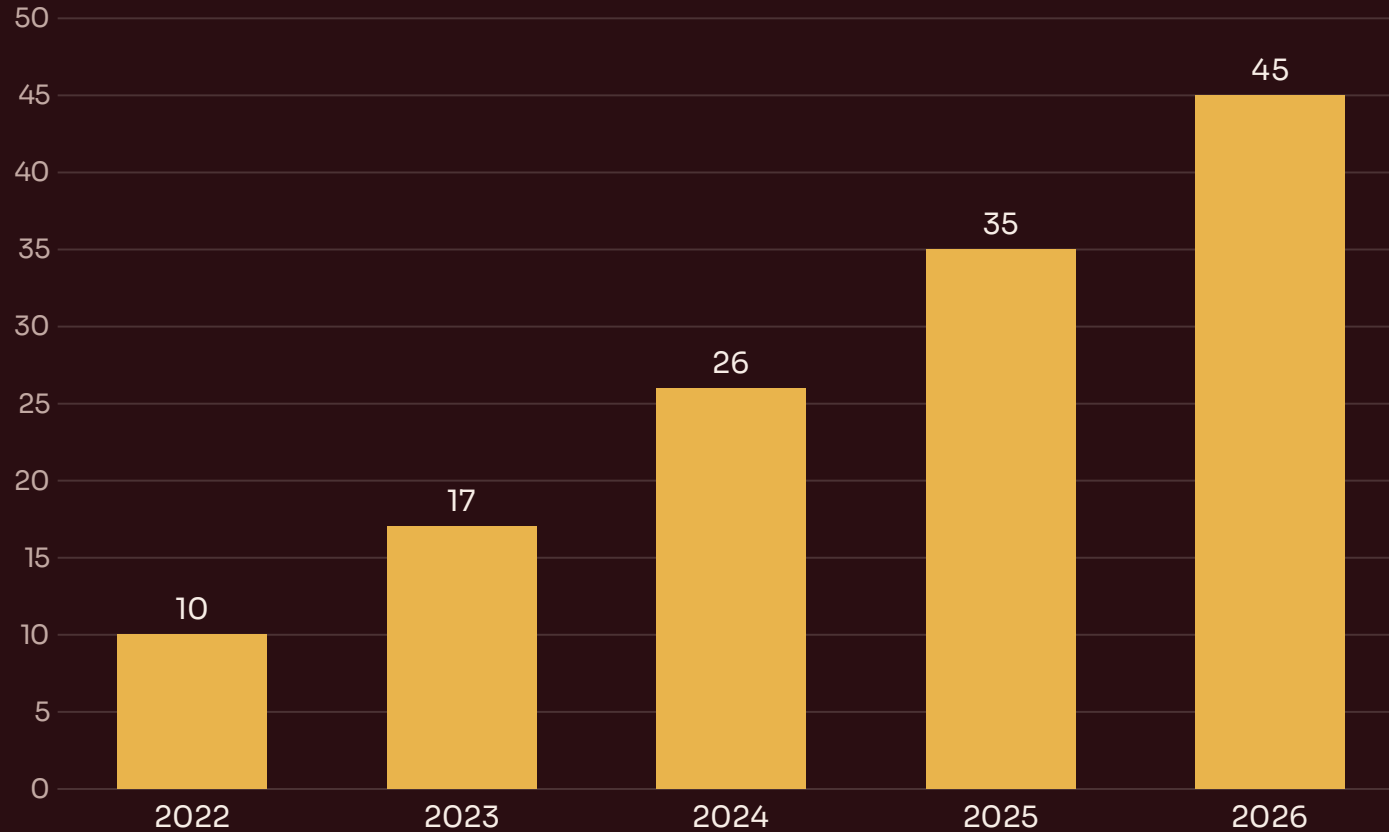
Generic examples; pick what fits your project.

SECTION 03

TRENDS



11 Adoption over time



Adoption keeps rising as tools mature (sample data).

Sample data for illustration — replace with your own figures.



12 | Use cases



Image 1

Smartphones

Photo and voice features



Image 2

Security

On-camera detection



Image 3

Industry

Defect inspection

SECTION 04

ADOPTION



14 | Benefits and risks

Benefits

- ✓ Low latency
- ✓ Privacy by design
- ✓ Works offline

Risks

- Limited compute
- Model updates are harder
- Battery impact



15 | Best practices

01

Measure first

Profile on real devices.

02

Compress smartly

Check accuracy after.

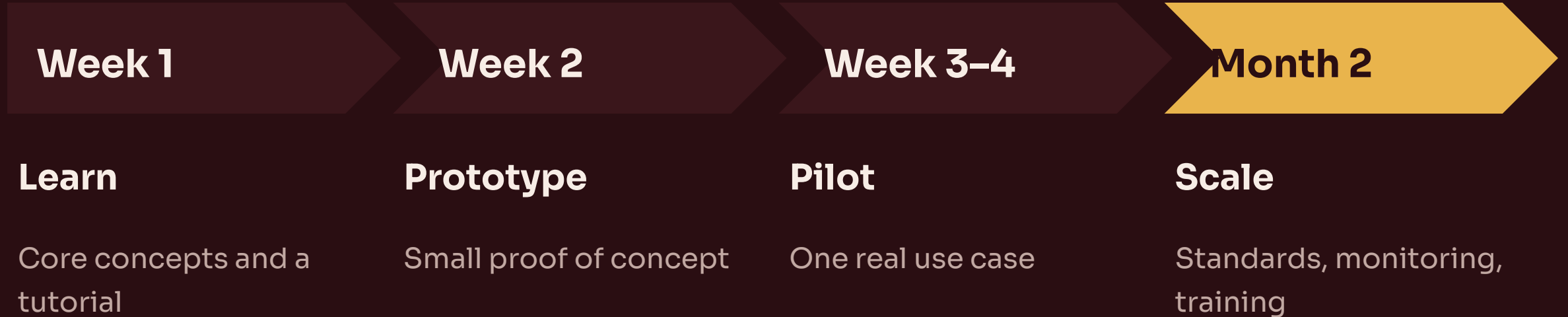
03

Plan updates

Version your models.



16 | Getting started roadmap



Questions?

Thanks for your time.

Presenter Name

hello@yourdomain.in

