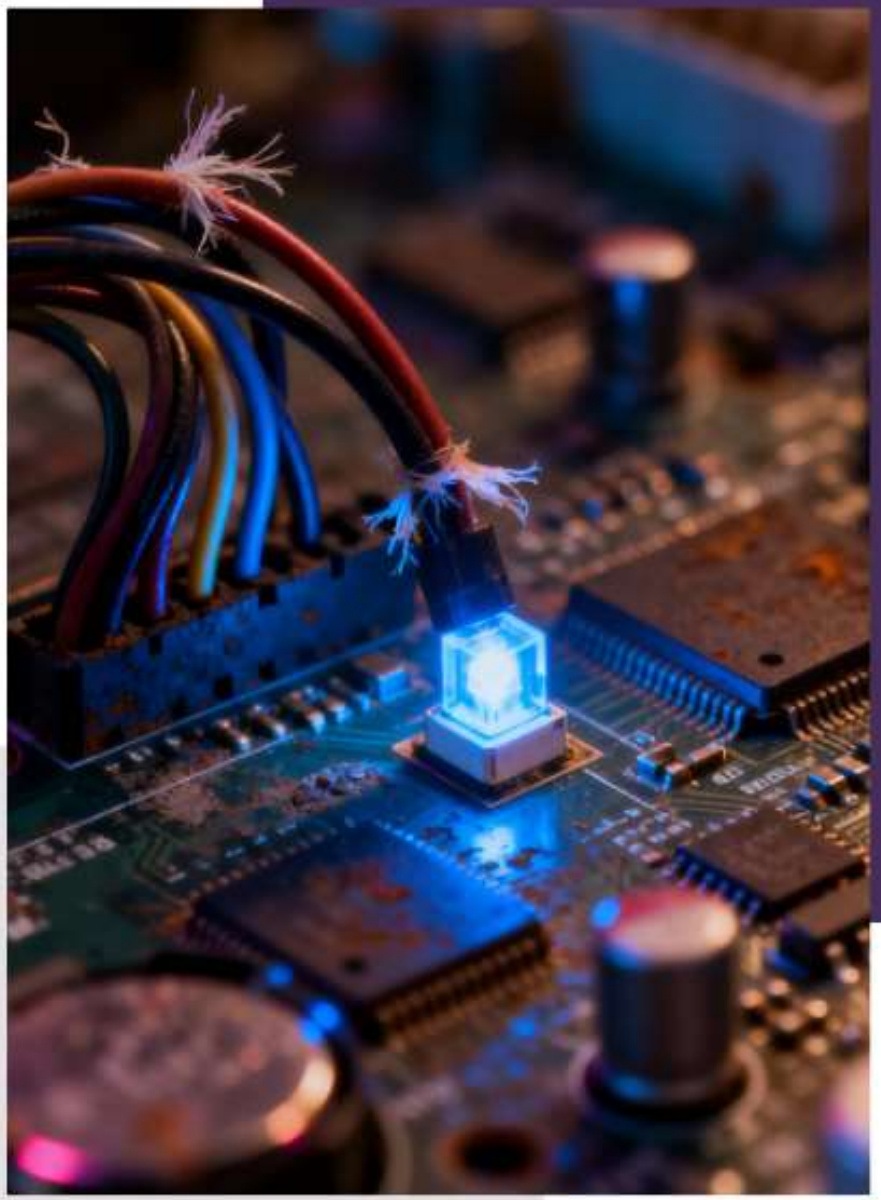
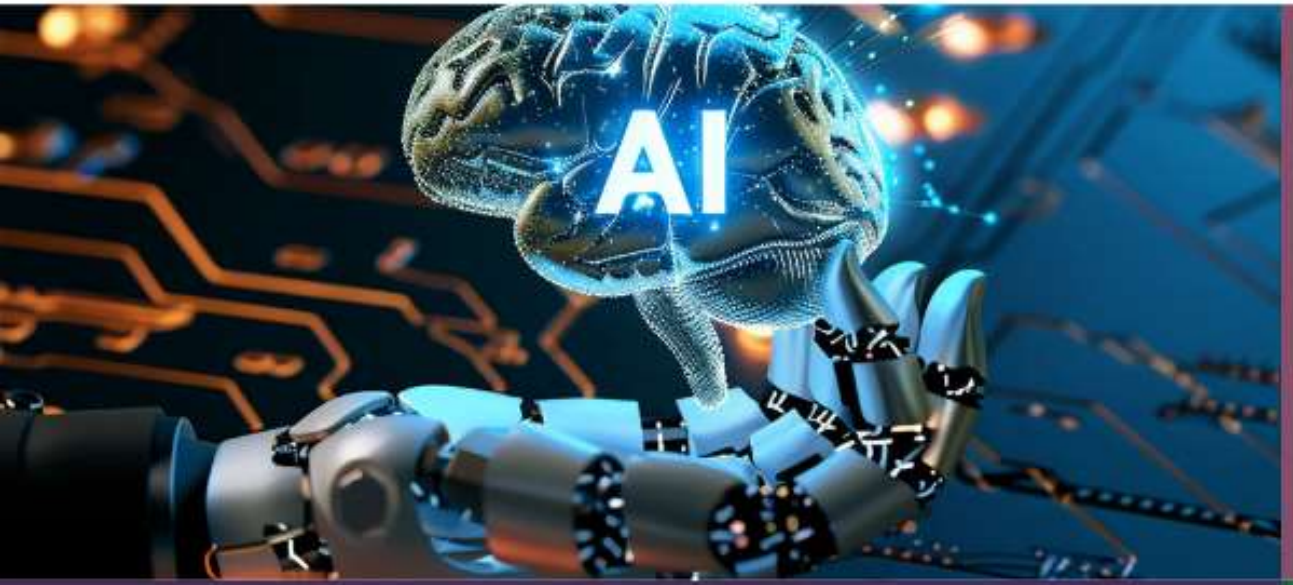


PathShala AI: Offline AI Architecture

Democratizing education in rural schools through high-performance edge computing and 100% offline AI assistance

Bridging the digital divide with quantized Llama-3.2 models and WebGPU acceleration.

PathShala AI - Hack Stack (2)



Malathi Venkatesan

malathilatha199615@gmail.com | +91 8940981260

Saravanan Gandhi

saran91amul@gmail.com | +91 8144669858

Theme: **A Teacher's Narration of the Problem**

Real Problems Come From The Classroom

- Most Solutions Designed Top-Down Miss Daily Realities
- This Theme Centers The Teacher's Lived Experience
- Classroom Issues Emerge Continuously During Lessons

Highlights The Implementation Gap

- Training Shows What To Do But Not How To Do It Live
- Teachers Struggle To Apply Methods In Chaotic Settings
- Theme Exposes Mismatch Between Training And Practice

Focus On Just-In-Time Support

- Help Often Arrives Weeks Later Via Inspections Or Visits
- Delayed Support Misses The Moment The Problem Occurs
- Demand Immediate, Offline Tools That Assist In The Moment

Offline-First Data Flow

A sophisticated zero-latency pipeline that executes high-performance inference entirely within the local browser environment

User Input

Bilingual educational queries are processed locally without making any server calls.

Browser Interception

The system intercepts requests to prevent external data routing for maximum privacy.

Model Retrieval

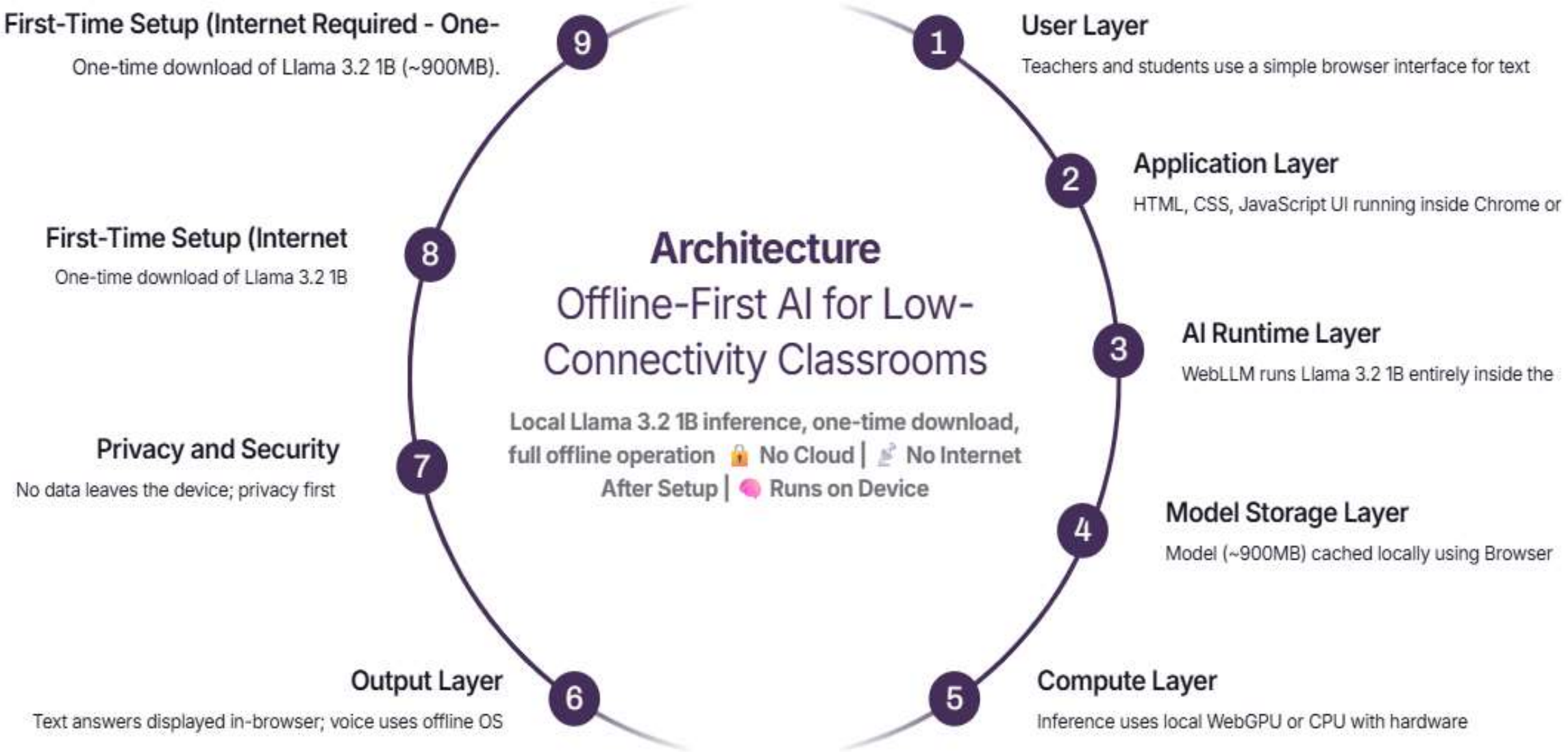
The quantized Llama-3.2 model is pulled from the local Browser Cache API storage.

WebGPU Inference

Hardware-accelerated processing occurs directly on the local GPU for near-zero latency.

Local Output

Text-to-speech or UI rendering is performed on-device for the final user experience.


First-Time Setup (Internet Required - One-
One-time download of Llama 3.2 1B (~900MB).
9
1
User Layer
Teachers and students use a simple browser interface for text
2
Application Layer
HTML, CSS, JavaScript UI running inside Chrome or
First-Time Setup (Internet
One-time download of Llama 3.2 1B
8
Architecture
Offline-First AI for Low-Connectivity Classrooms
Local Llama 3.2 1B inference, one-time download, full offline operation 🔒 No Cloud | 📡 No Internet After Setup | 🧠 Runs on Device
3
AI Runtime Layer
WebLLM runs Llama 3.2 1B entirely inside the
Privacy and Security
No data leaves the device; privacy first
7
4
Model Storage Layer
Model (~900MB) cached locally using Browser
Output Layer
Text answers displayed in-browser; voice uses offline OS
6
5
Compute Layer
Inference uses local WebGPU or CPU with hardware

One-Time Download for **Forever Offline** Use

Download once; runs offline for low-connectivity schools

Model downloaded once (~900MB)

Single download size: approximately 900MB

Stored securely in browser cache

Local storage in the browser cache for safety

Works without Internet

Full AI functionality offline after download

Ideal for rural and low-connectivity schools

Designed for environments with limited internet

The Technical Stack

Leveraging cutting-edge web technologies to turn standard hardware into high-performance AI workstations

WebLLM (Apache TVM) serves as the core **Inference Engine** for local model execution.

WebGPU API enables direct **Hardware Access** for accelerated graphics and compute processing.

Vanilla HTML5/CSS3 provides a **Zero Bloat** user interface for maximum performance on old hardware.

Llama-3.2-1B-Instruct model provides high-quality bilingual educational assistance.

Future Strategic Roadmap

The evolution of PathShala AI from desktop stability to mobile accessibility and curriculum-specific fine-tuning

Phase 1

Current Stability

Stable desktop offline AI architecture running on Windows and Mac browsers via Chrome.

Phase 2

Expanding availability to Android and iOS via specialized Chrome WebGPU support.

Mobile Optimization

Phase 3

Voice Integration

Integrating Whisper-tiny (Wasm) for voice-to-text accessibility for low-literacy users.

Phase 4

Domain-specific fine-tuning on NCERT National Curriculum data for regional relevance.

Curriculum Specificity

Democratized AI for All

By combining edge computing with local language models, PathShala AI is building a more equitable educational landscape where intelligence is no longer restricted by internet access.
Ensuring quality education reaches every corner of the world, regardless of connectivity.