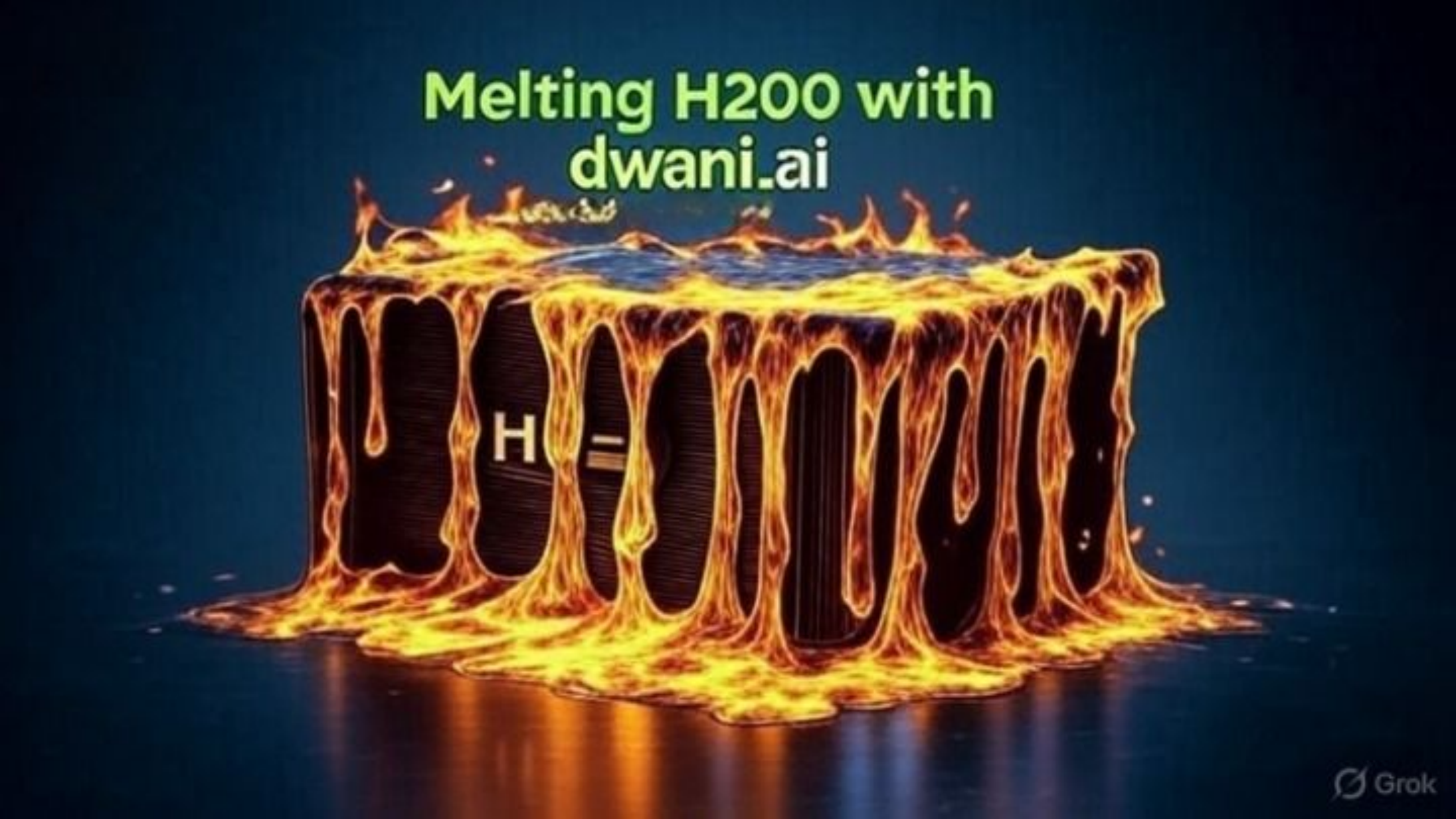

Melting H200 with
dwani.ai
H
Grok

TEAM

Dual Challenge: Dwani.ai & Benchmark Blitz

Dwani.ai Goal

Serve 10-10,000 LLM inference requests concurrently.

Benchmark Blitz

Run Qwen3-0.6B for 5 min, maximize GPU core usage.

Problem

Current script underutilizes cores; RTX 5090 falsely outperforms H200.

TensorRT-LLM: The Powerhouse

Framework

TensorRT-LLM for max throughput, low latency, multi-GPU scaling.

Dwani.ai Integration

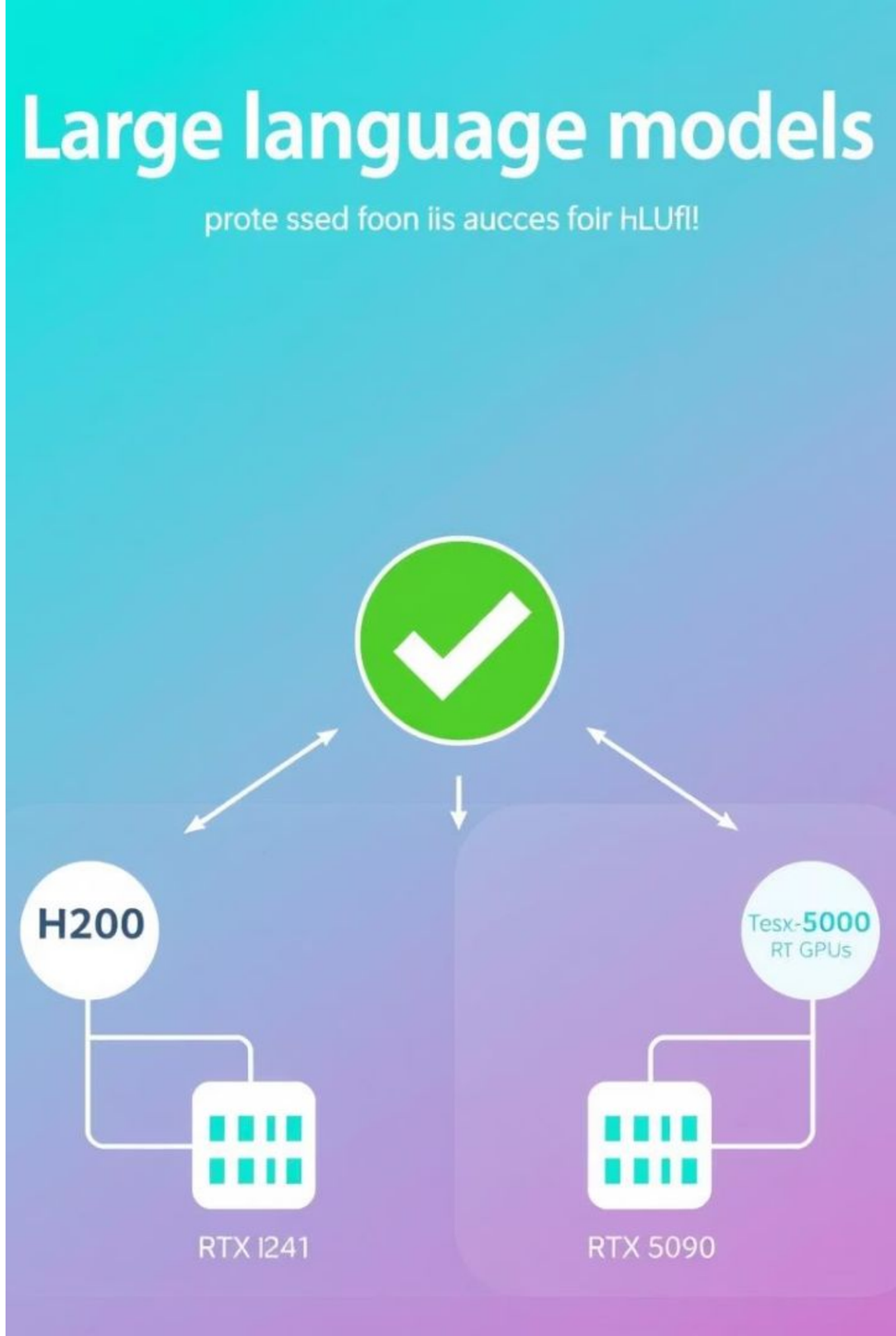
Load all models (Gemma 27B, Qwen 30B, etc.) with FP8 quantization.

H200 fits 76.4 GB, RTX 5090 time-shares.

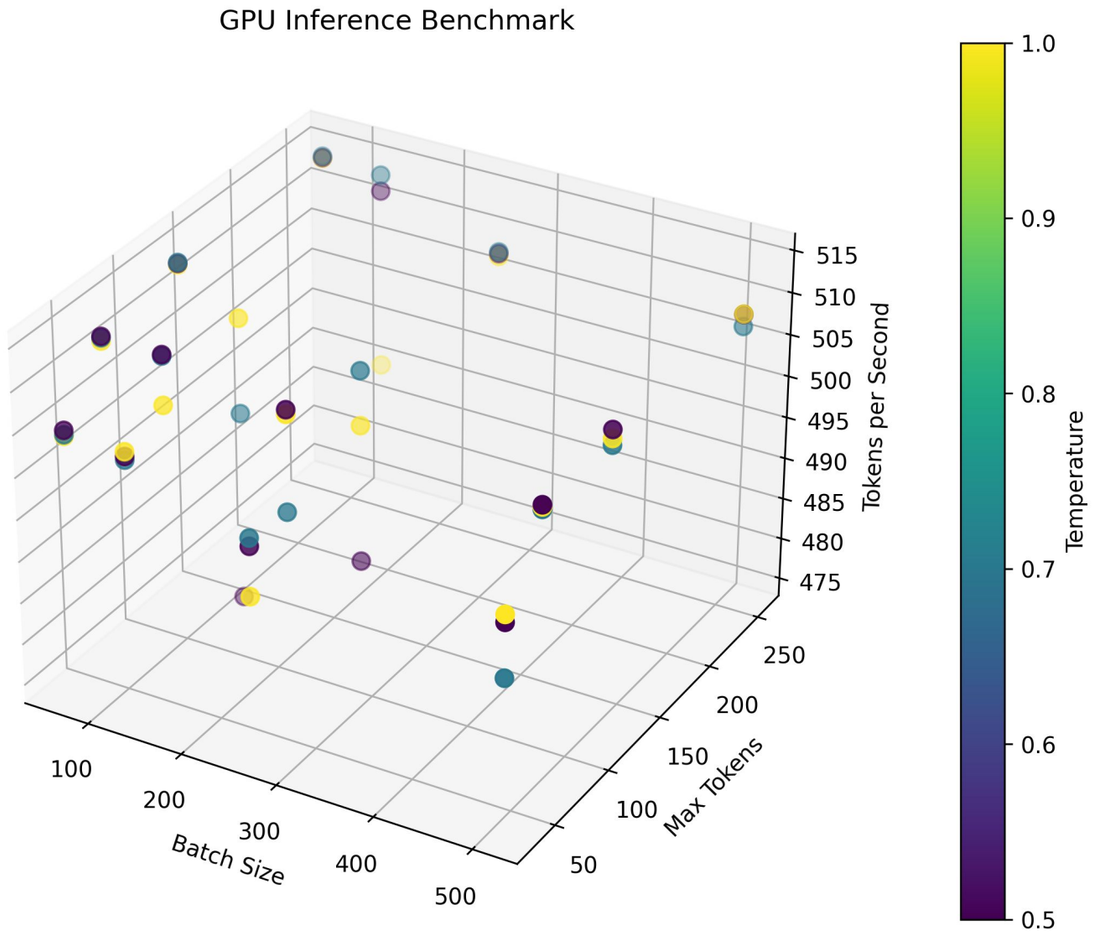
Benchmark Strategy

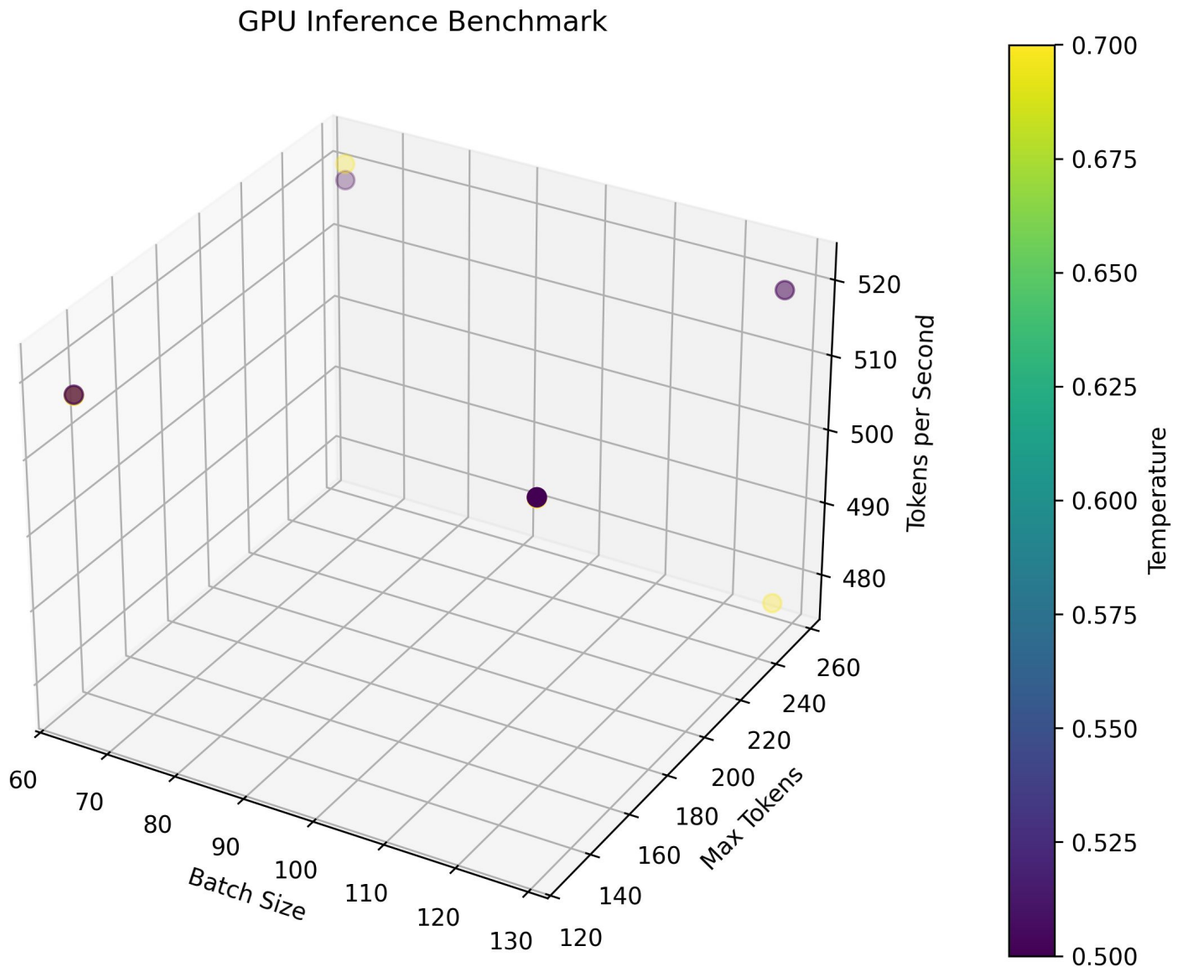
Qwen3-0.6B, 5-min run, dynamic batching.

Metrics: Throughput (tokens/s), latency, utilization, temp, power.



Results That Win





GPU Brrr Squad: Unstoppable

Dwani.ai Solution

Scales to 10,000 requests, melts H200 with efficiency.

Benchmark Fix

Max core usage, H200 beats RTX 5090.

Innovation

TensorRT-LLM + profiling for performance & health.

Team

Experts in GPU optimization, inference, benchmarking.

Join the GPU Brrr Revolution

Resources

Provided: RTX 5090 (United Compute Cloud).
Needed: H200 access (United Compute team).
Software: TensorRT-LLM, Nsight, nvidia-smi.

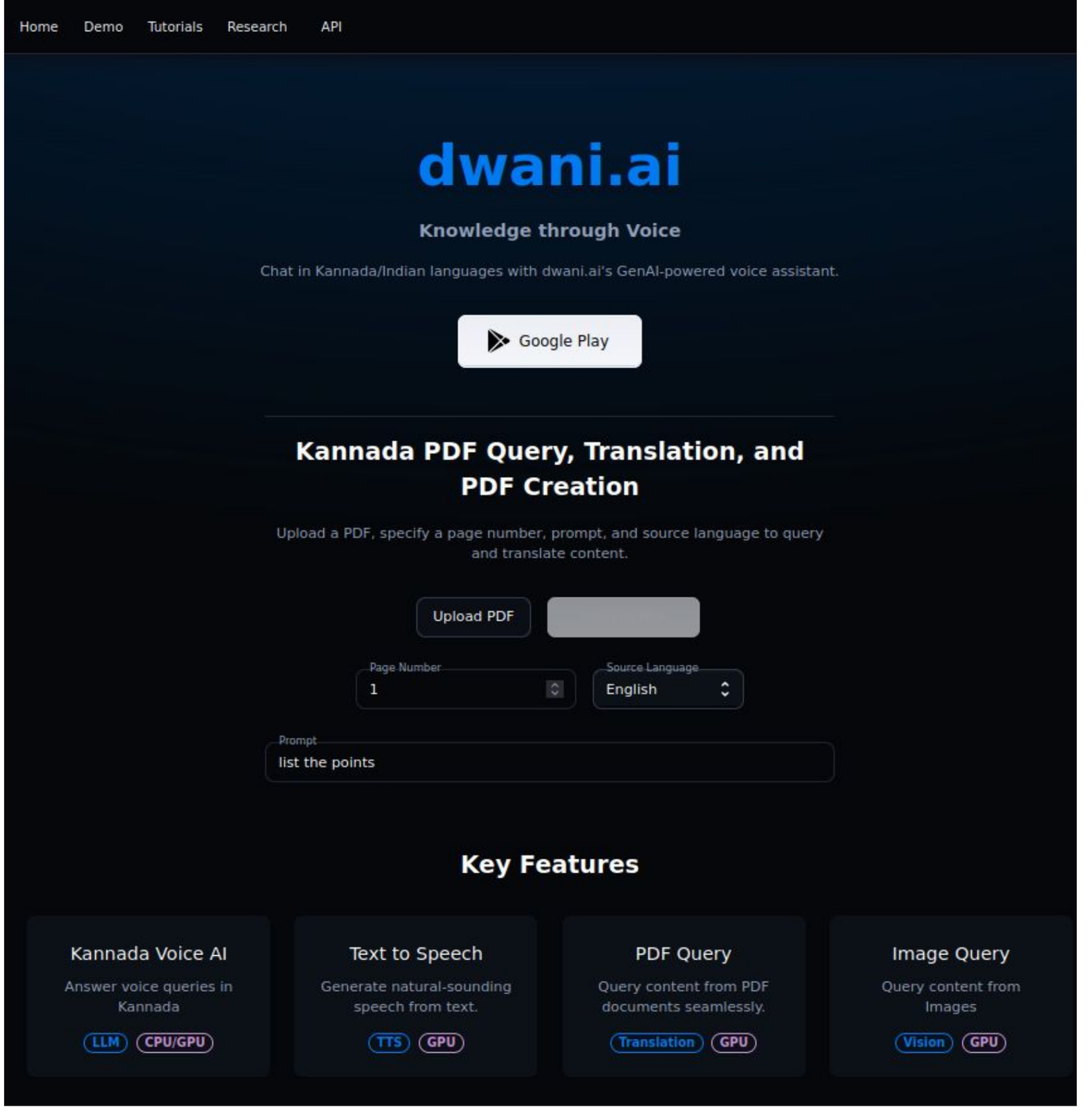
Why Us?

Deliver for dwani.ai, win Benchmark Blitz, melt the H200!

GitHub Link:

Let’s Win

Berlin Hardcore AI Hackathon 2025.


Home
Demo
Tutorials
Research
API
dwani.ai
Knowledge through Voice
Chat in Kannada/Indian languages with dwani.ai's GenAI-powered voice assistant.
Google Play
Kannada PDF Query, Translation, and PDF Creation
Upload a PDF, specify a page number, prompt, and source language to query and translate content.
Upload PDF
Page Number
1
Source Language
English
Prompt
list the points
Key Features
Kannada Voice AI
Answer voice queries in Kannada
LLM
CPU/GPU
Text to Speech
Generate natural-sounding speech from text.
TTS
GPU
PDF Query
Query content from PDF documents seamlessly.
Translation
GPU
Image Query
Query content from Images
Vision
GPU

Android App