vLLM

MEETUP · THAILAND

vLLM Bangkok

High-performance LLM serving, distributed inference and the Thai AI ecosystem — talks, panels and networking.

DATE & TIME

Sun 6 Sep · 2:00 – 7:00 PM

Registration opens 1:15 PM · Free admission

VENUE

JAS Green Village

3rd floor, Ramkhamhaeng · Bangkok

AGENDA

1:15 – 2:00	**Registration**
2:00 – 2:20	OPENING **vLLM Roadmap in the Thailand Ecosystem** AGICAFET
2:20 – 2:50	TALK **Introduction to vLLM: High-Performance LLM Serving & Maximizing Performance on Limited GPUs** Tun Jian Tan · Core Contributor, vLLM · Principal AI Engineer, Embedded LLM
3:00 – 3:45	PANEL **AI Adoption in Thailand** — **Dr. Pin Siang Tan** · CTO & Co-Founder, Embedded LLM — **Tuan Luong** · APAC General Manager, Tensormesh — **Dr. Komes** · Principal AI Evangelist, KBTG — **Sutthinee Theppariyapol** · Head of Product, FlowAccount · Moderator
3:45 – 4:00	TALK **Thai and Global Perspective: Connected Point** CK · Fastwork
4:00 – 4:30	TALK **Inside vLLM: Production Best Practices, Model Integration and Roadmap** Wang Tiezhen · Head of Ecosystem, Inferact
4:30 – 5:00	TALK **Scaling Distributed Inference with llm-d** Jing Wen Ng · Specialist Solutions Architect, AI, Asia Pacific, Red Hat
5:00 – 5:30	TALK **Supercharging vLLM with LMCache: KV-Cache Reuse in Practice** Shaw Nguyen · Solution Architect, Tensormesh
5:30 – 5:45	**Break**
5:45 – 6:20	PANEL **vLLM Developers in Thailand** — **Sattaya Singkul** · Specialist Lead, AI Solution Engineer, True Digital Group — **Natthanan Bhukan** · Researcher, Tsinghua University — **Pongsasit Thongpramoon** · Senior Solutions Architect, NVIDIA — **Kanin Kearpimy** · Fellow, Python Software Foundation · Moderator
6:20 – 7:00	**Lucky Draw + Networking**

vLLM Bangkok Partners

Sun 6 Sep · 2:00 – 7:00 PM
JAS Green Village

MEETUP · THAILAND

STRATEGIC PARTNER

embedded

AGICAFET

HOST PARTNER

Jaymart GROUP

TECHNICAL PARTNER

SVOA EXPERT BEYOND IT

msi

GLOBAL COLLABORATION

Red Hat

inferact

tensormesh

COMMUNITY PARTNER

PYCONTH

FLOWACCOUNT

ooca

NATACHAT

In 2025 vLLM Meetup Bangkok: Embedded LLM

Inside vLLM Meetup Bangkok: Why This World-Class AI Project is the GitHub Darling of 2025

Natdhanai Praneenatthavee 4 min read · Nov 30, 2025

2

On November 21, 2025, I had the opportunity to attend the vLLM Meetup Bangkok. While the evening atmosphere was casual and friendly, the content shared on stage was absolutely "blazing hot." This event brought together top-tier minds in LLM Serving and Inference Optimization.

embedded


vLLM ♥ PyTorch
torch.compile
flexattention
vLLM
enhances
PyTorch
torchao
accelerator agnostic
memory allocator
piecewise graph
PyTorch
accelerates
vLLM
AMD
vLLM BANGKOK

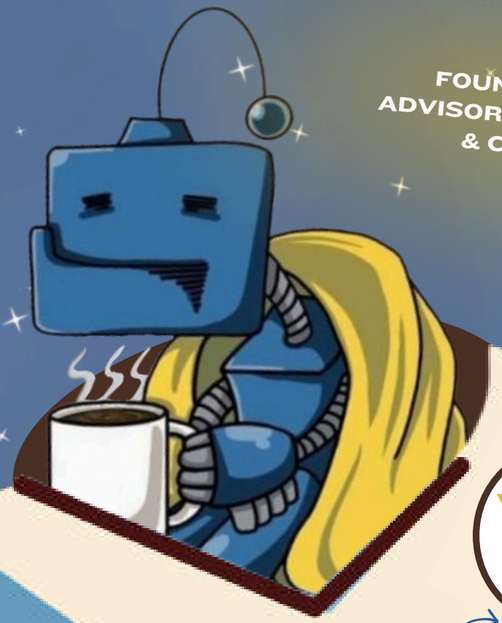

SESSION 2 — 1:50 PM
vLLM: The Universal Backbone of LLM Inference
Scale Without Compromise
With Embedded LLM
High-performance inference for scalable, cost-effective LLM deployment — the infrastructure layer that makes production AI economically viable.
vLLM


THE AGENTIC WORKFORCE
IBM
AI Alliance
SCALING MULTI-AGENT ORCHESTRATION & VERIFIABLE AI
Welcome and Introduction
1. AI Application Developer Partner:
Bob the builder
4.Building AI Agent you can trust
through Responsible Frameworks
embedded

Maker to Maker #4

15 FEBRUARY 2026
1.00–7.00 P.M.

CLEVERSE RAMA9



FOUNDER, AI TECH ADVISOR IN GOVERNMENT & CORPERATE

VLLM: THE UNIVERSAL BACKBONE OF LLM INFERENCE

THE PARADIGM SHIFT: FROM NODE GRAPHS TO DATA TABLES FOR AL AGENTS

BEYOND COPILOTS: THE EVOLUTION OF DEVELOPMENT WITH BOB

IBM

AL ENGINEER, GOING FROM THAILAND TO GLOBAL.

GOOGLE'S SECURE AI FRAMEWORK (SAIF)

SOLO TO SCALE: THE END-TO-END BLUEPRINT FOR BUILDING BUSINESS-DRIVEN AI

TRUSTWORTHY AI

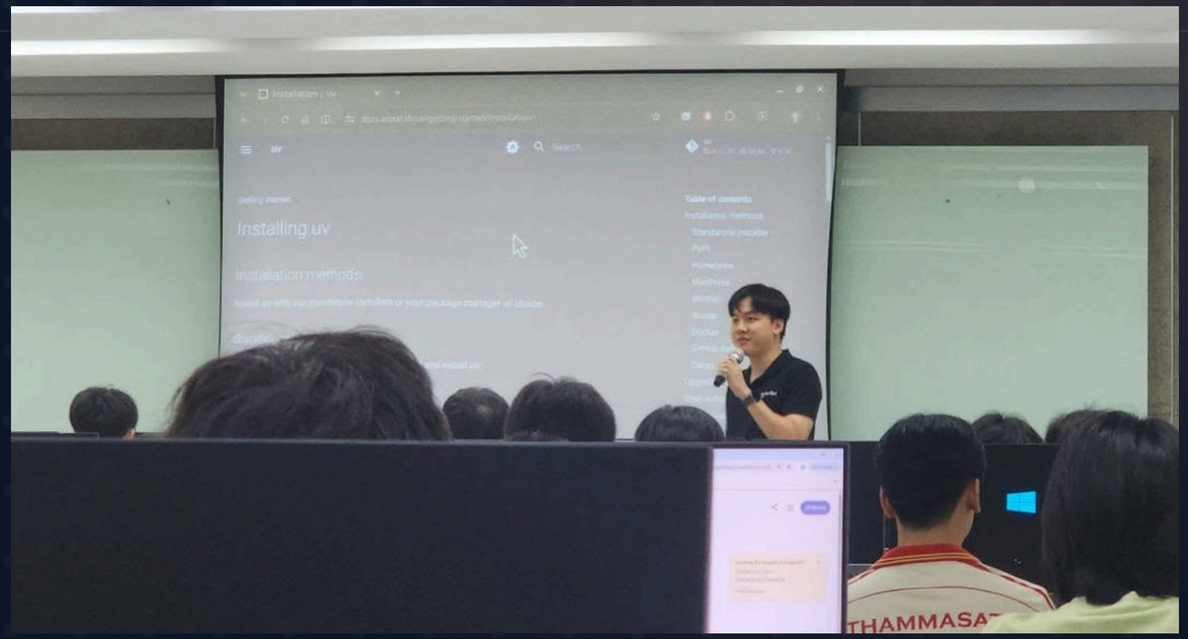
BIOROBOTICS & SMARTCITY AND UK JOURNEY

AI JOURNEY 2026, AGENTIC-AI PROFESSIONAL IN 2026

HOW TO PLAN ORGANIZATIONAL CHANGE WITH BRICKS AGENTIC TRANSFORMATION FRAMEWORK

HACKING SYSTEM PROMPTS: WHAT WE LEARNED FROM LEADING LLMS


Maker to Maker #4
15 FEBRUARY 2026
1.00–6.00 P.M.
CLEVERSE RAMA9
MEET OUR SPEAKERS
LIM HOI PANG.
LLM Solution Architect of Embedded LLM
TITLE: THE PARADIGM SHIFT: FROM NODE GRAPHS TO DATA TABLES FOR AL AGENTS
TAN TUN JIAN
vLLM Core, Maintainer, AI Principal Engineer of Embedded LLM
TITLE: VLLM: THE UNIVERSAL BACKBONE OF LLM INFERENCE


Installing uv
Installation methods


vLLM-Omni
vLLM Recipes
GuideLLM
vLLM
LLM Compressor
vLLM Semantic Router


Z-Image-Turbo Online Generation

Opening vLLM Bangkok

https://vllmbangkok.com/

vLLM Bangkok Day

Event | Agenda | Blog | Community | Sponsor | ภาษาไทย | Registration is full

TH Official event by the vLLM creators

vLLM Bangkok Day

A day of production vLLM — talks from the creators, real Thai deployments, and the people building inference infrastructure in Southeast Asia.

WHEN

Sunday, 6 September 2026 · 14:00-19:00

WHERE

JAS Green Village Ramkhamhaeng

Registration is full

All 250 places taken View the event page ↗

Get directions ↗

vLLM in Thailand

The companies, labs and universities running vLLM here. If your team is on this list, you are not doing it alone.

Add your team → 5 organisations

AGICAFET

Bangkok

Organiser of vLLM Bangkok Day. Builds and serves Thai large language models, including the ThaiLLM series.

> Serving Thai-language models in production, including NVFP4-quantised checkpoints.

agicafet.com ↗

Embedded LLM

Strategic partner of vLLM Bangkok Day and a core contributor to vLLM. Works on inference optimisation across GPU platforms, parallelism strategy for large MoE and multimodal models, and private on-premise serving.

> Upstream vLLM contributions (DFlash, KV-cache compression and offloading, prefix caching, chunked prefill, sleep mode) and production serving on limited GPUs.

embeddedllm.com ↗

Red Hat

Enterprise open source. A major contributor to vLLM and to open model serving on Kubernetes.

> Upstream vLLM contribution and enterprise inference platforms.

redhat.com ↗

Tensormesh

Inference infrastructure — making LLM serving faster and cheaper at scale.

> Inference optimisation and KV cache systems built around vLLM.

tensormesh.ai ↗

Inferact

Partner of vLLM Bangkok Day, working on AI inference.

Thai Local vLLM Developer First Group

Community & Organizational

AGICAFET X Embedded LLM X Thammasat University

Every month:vLLM Virtual Sharing

Every 6 month: vLLM Thailand Conference

Goal

AI Infrastructure

Inference Optimization

for Thai people, ASEAN and Organization

If your organization

Want to Build together

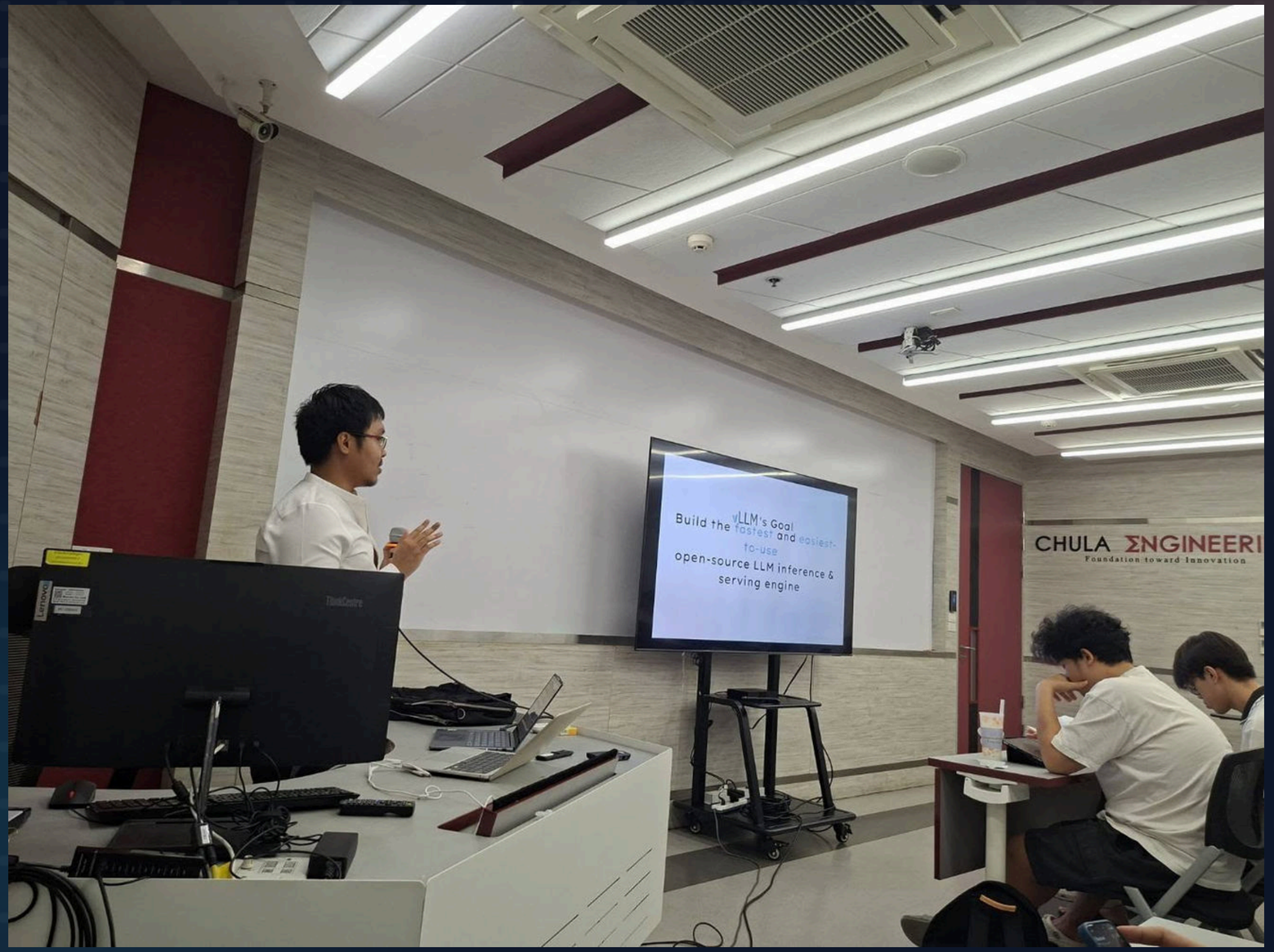
Line: jobfitlin

Emai: natdhanai.pra@agicafet.com

Web: https://vllmbangkok.com/

vLLM Inference Virtual Sharing: more than 600 attends

HSSC 2026 High school from Indonesia



Teching Chula ISE about Inference optimization vLLM

PREMIZE

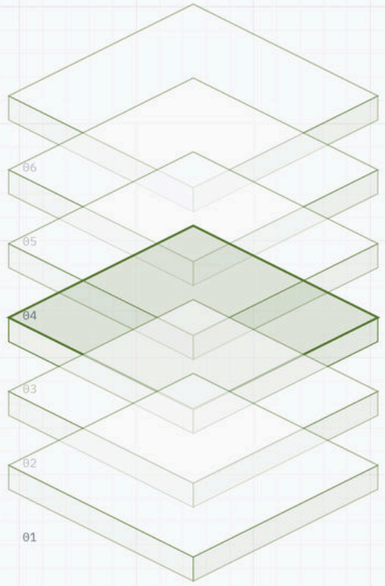
EN / TH

Sign in

01 THE LAYERS

Every layer, engineered by experts.

What feels effortless on the surface is a full AI infrastructure underneath — memory budgets, model serving, retrieval, governance, and experience, designed together as one system on one machine.



06 **Experience**

A clean Thai-first app, customer chat channels, and an API — easy from day one.

05 **Governance**

Policies, human approvals, and a full audit trail around every sensitive action.

04 **Knowledge**

Every page is indexed by meaning; every answer points back to its exact source page.

03 **Intelligence**

A 35-billion-parameter model with a 262,144-token context, beside Thai OCR for scanned documents.

02 **Serving**

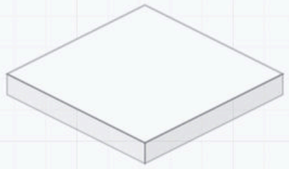
An inference layer tuned for this exact machine — every gigabyte of memory budgeted and managed automatically.

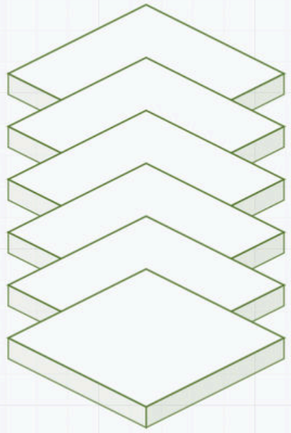
01 **Hardware**

NVIDIA DGX Spark: the GB10 Grace-Blackwell superchip with 128 GB of unified memory.

PREMIZE

EFFICIENT LOCAL LLM





https://premize.agicafet.com/

DGX Spark only DIY

excellent hardware — everything else is homework

- Partitioning 128 GB of unified memory (trial, error, OOM)
- Model serving, quantization, and tuning
- Thai OCR pipeline for scanned documents
- RAG with trustworthy page-level citations
- Governance, approvals, and audit
- Dashboards, metering, and monitoring

DGX Spark + PreMize READY

engineered, measured, ready on day one

- ✓ Every gigabyte of unified memory budgeted
- ✓ Serving tuned for the GB10 — with an auto sleep/wake steward
- ✓ Thai OCR built in, page by page
- ✓ Cited answers out of the box — “Ref page N”
- ✓ Policy → approval → audit, from day one
- ✓ Live dashboards, per-user metering, open API

Thankyou

agicafet.com

agicafet

AGIcafet Co., Ltd.

ตื่นมาโค้ดpython by AGICAFET