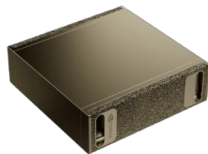


# Inference trillion parameters everywhere

Redesign LLM for everyday device



# LLM is better with more parameters, but it's memory bound



1999\$ + PC cost

2999\$

3600\$

1999\$

RTX5090 GPU

DGX Spark

Mac Studio M4

AMD AI Max+ 395

32 GB VRAM

128 GB UM

128 GB UM

128 GB UM

1792 GB/sec

273 GB/sec

273 GB/sec

256 GB/sec

30B model max

100B model max

100B model max

100B model max

32B model ~20 t/s

32B model ~20 t/s

32B model ~20 t/s

Personal AI PC: up to **32B** model usable

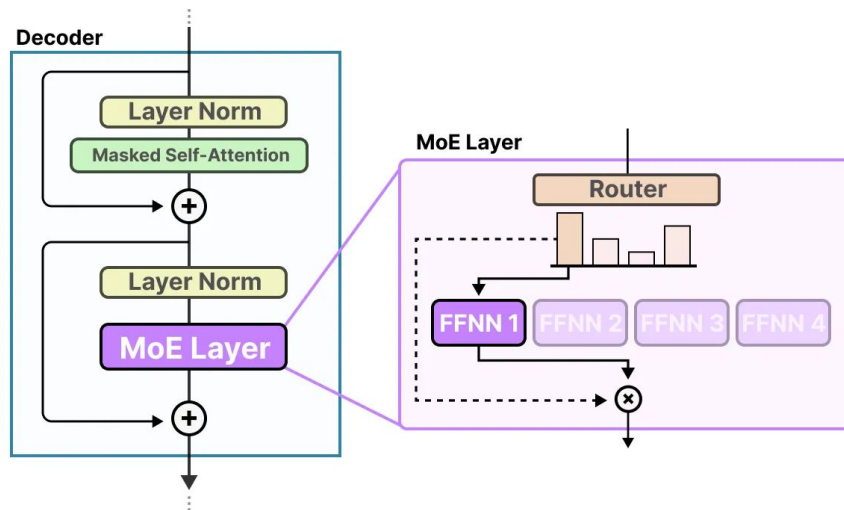


MoE: reduce computing, but not memory

  
671B Parameters



1342 GB in FP16  
671 GB in FP8  
335 GB in FP4



A **router** (also called a **gate network**) is added which is trained to choose which expert to choose for a given **token**.

## Problem:

- ❑ MoE architecture doesn't fit general PC's architecture
- ❑ High bandwidth memory is expensive and limited
- ❑ Distilled models lose task specific ability dramatically
- ❑ Every vendor has their own DSL & runtime

## Solution:

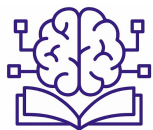
- ❑ Break LLM into smaller pieces
- ❑ Load them into HBM by tasks
- ❑ “Vulkanised”



# Training Process



task specific corpora  
hundreds of categories



source large model  
100B ~ 1T parameters

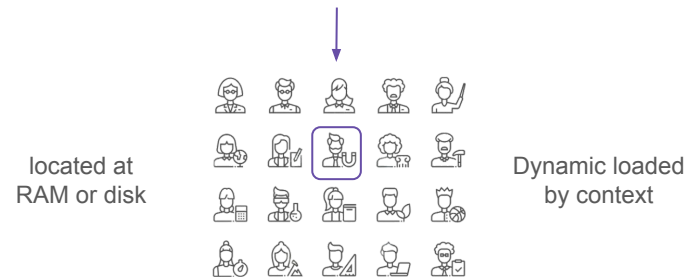


agent based router  
small encoder classifier



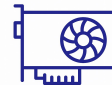
distilled small agent models  
100+ of 1~7B models

# Inference Process



PCIe 10~25 GB/s

≤ 1s latency



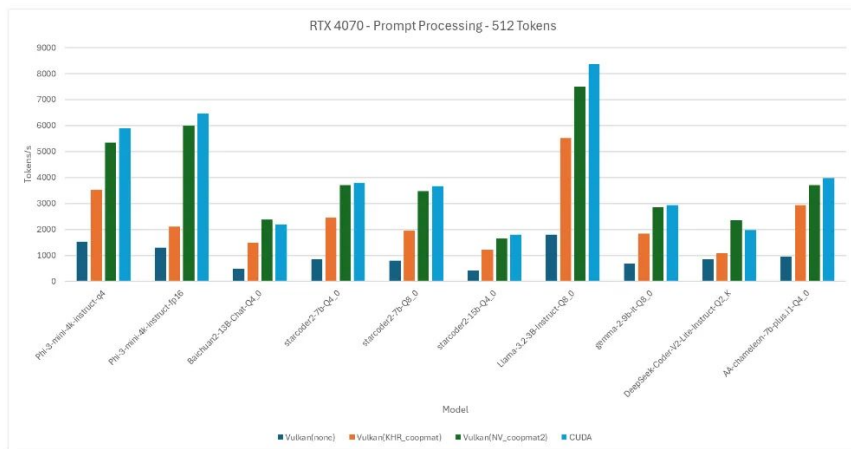
Stay in HBM for the same **context**

# Potential of Vulkan

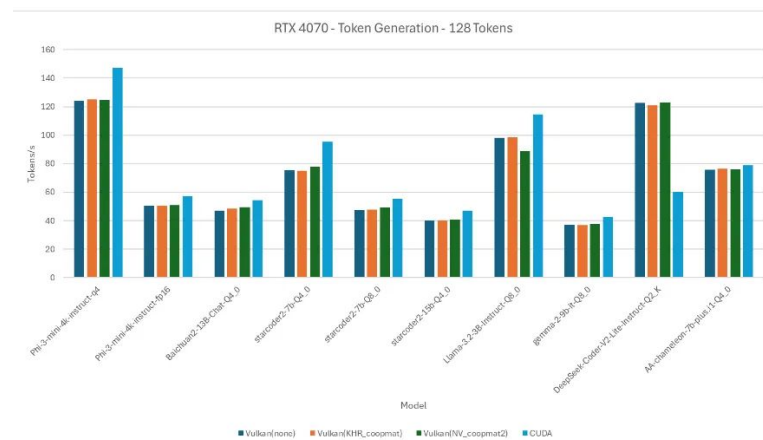
- ❏ Cross platform
- ❏ High performance
- ❏ Example: ggml



Performance



Performance



## Outcomes:

- ❑ Unbound inference at edge side
  - ❑ Much cheaper inference
  - ❑ Makes private large scale inference possible
  - ❑ Cross-platform code base
- ❑ Modularity brings scalability
  - ❑ Fine tuned agents instead of prompt-based
  - ❑ Much lower cost for continual knowledge update
- ❑ Sustainability
  - ❑ Build ecosystem for difference models
  - ❑ provide long term service for deployment system

## Business Model

- ❑ APP based platform for models. On PC and mobiles.
- ❑ Target all customers:
  - ❑ To Business: Privacy
  - ❑ To Individual: Flexibility
- ❑ We provide:
  - ❑ Thousands of small proprietary models varied by sizes and tasks
  - ❑ On demand subscription and download
  - ❑ Automatic router to load model by context
  - ❑ Ecosystem for small models





## Business case: to individual

Aim for high performance, easy access, lower cost, privacy and flexibility.

- ❑ A high school student is preparing for exam, parabase has models trained with latest exam questions.
- ❑ A gaming PC owner want to use his discrete GPU as a personal AI host.
- ❑ A French learner wants lower latency and a more stable connection to an LLM, so they can run a French language model directly on their phone.
- ❑ An author or innovator wants to use an LLM for support but doesn't want to risk their ideas being exposed to the service provider.
- ❑ ...



## Business case: to enterprise

Aim for local inference and fine-tuning for full control, and low cost and risk for continual knowledge updating.

- ❑ A university department wants to train proprietary models tailored to their subject matter for student use.
- ❑ A law firm wants to train a model using financial case data from the past five years.
- ❑ A car company wants its AI assistant to remain accessible even in tunnels.
- ❑ An AI company wants to host its own LLM server to avoid relying on OpenAI
- ❑ ...

## Why now?


### ❖ Private AI environment on its way to launch

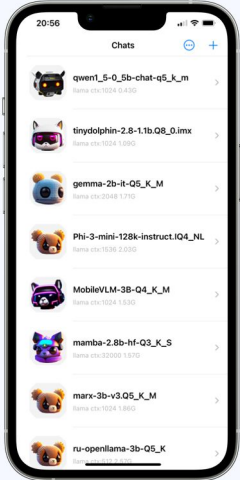
- AI PCs
- Vendor's SDK
- Efficient hardware

### ❖ Dense model shows a good performance

- Llama 1B, 3B, 8B
- Qwen 0.5B, 1.5B, 3B, 7B
- 0.5~16 GB: fits in most HBM.

## Other players: LLM farm

 **LLM Farm** [Docs](#) [News](#) [GitHub](#)



## LLM locally on iOS and MacOS

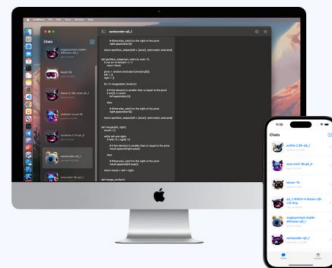
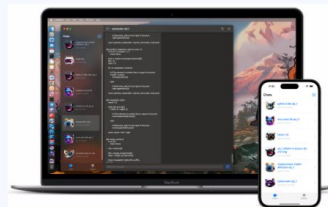
Absolutely free, open source and private.

[Download Latest From TestFlight →](#)

[Download Stable From AppStore →](#)

## Absolutely free

LLM Farm provides all features absolutely free of charge! No hidden fees, subscriptions or feature limitations - all features are available for use at no additional cost.



## Open Source

The core is a Swift library based on llama.cpp, ggml and other open source projects that allows you to perform various inferences. A class hierarchy has been developed that allows you to add your own inference.

## Other players: Edgerunner

### Dynamically switching between SLMs.

**Intelligent routing of requests:** EdgeRunner will intelligently switch between models, dynamically routing requests to the best task-specific model for the use case.

**Power efficiency:** Saves RAM and power, increasing efficiency and performance.

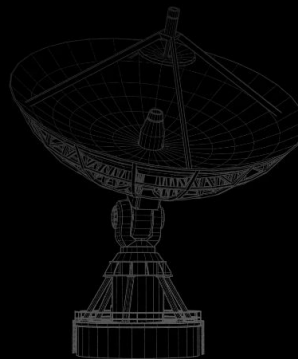
**Streamlined accessibility:** Makes Generative AI simple and boring, enabling widespread adoption.

**Ever-evolving standard:** Becomes the enterprise standard for leveraging multiple models at once, continuously leveraging the latest models, future proofing intelligence.



### Our Mission.

To build Generative AI for the edge that is safe, secure, and transparent



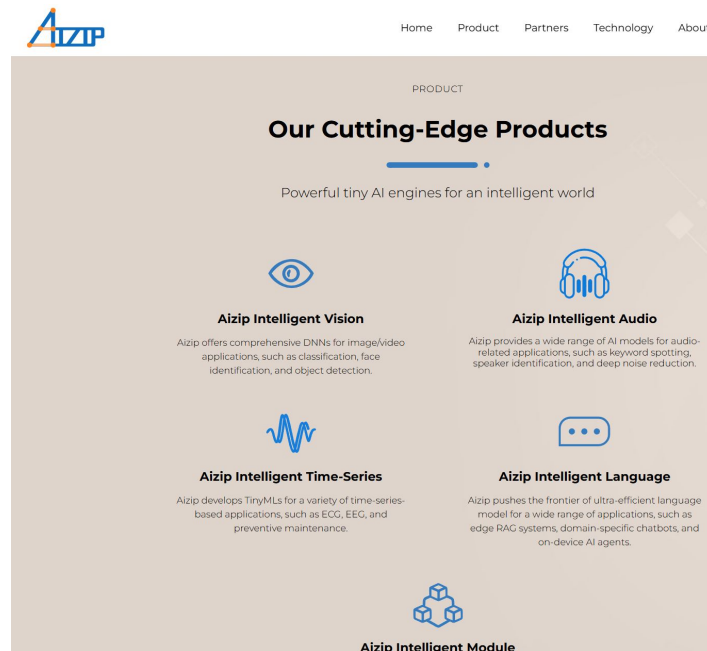
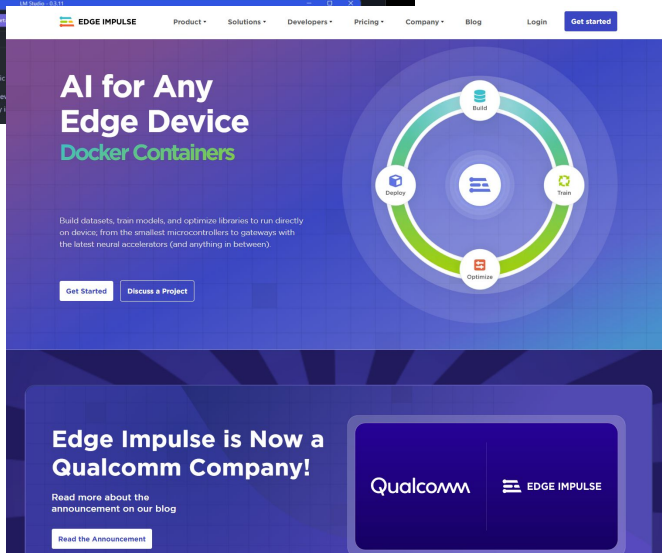
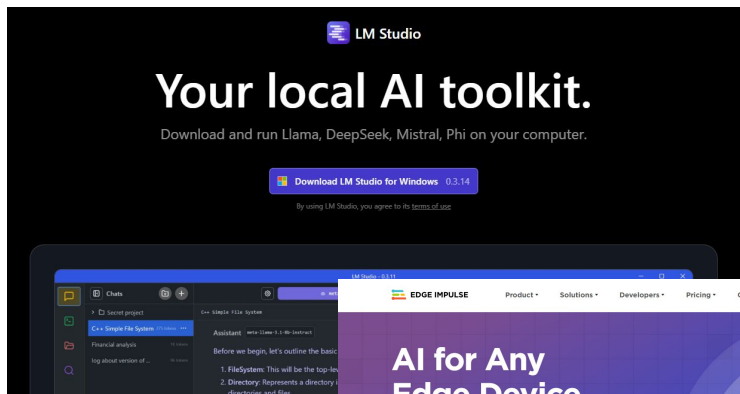
Our founding team has deep experience in developing SOTA open source foundation models, optimizing models for edge use cases, and building technology that solves real-world challenges.

We advocate for the deployment of Generative AI locally and privately, responding to data gravity and privacy concerns. Our strategy involves developing a suite of domain-specific Small Language Models (SLMs) that operate synergistically. This catalyzes the start of collective intelligence, where multiple SLMs interact and collaborate to solve problems.

Generative AI will transform all legacy unstructured data into relevant intelligence that you can interact with through natural language like a human conversation. In the future, you will be able to speak to your data and it will speak back to you.

We believe technology this transformative should be under your control, living on your device, and isolated from external networks to ensure a private and hyper-personalized experience.

## Other players: to customer local inference tool chain



## Other players: Prompt based agents

### GPTs

Discover and create custom versions of ChatGPT that combine instructions, extra knowledge, and any combination of skills.

精选推荐

Writing

Productivity

Research & Analysis


Education

Lifestyle

Programming


#### Featured

Curated top picks from this week




**SciSpace**

Do hours worth of research in minutes. Instantly access 287M+ papers, analyze papers at lightnin...  
By scispace.com




**Wolfram**

Access computation, math, curated knowledge & real-time data from Wolfram|Alpha and Wolfram...  
By Nathaniel Bauer



**Video GPT by VEED - Instant & Free AI Vid...**

AI video maker powered by VideoGPT. Generate and edit videos with text prompts. Type a...  
By veed.io



**Canva**

Effortlessly design anything: presentations, logos, social media posts and more.  
By canva.com

## Other players: Enterprise demand for private inference



[Home](#) [About Us](#) [FAQs](#) [Careers](#) [Blog](#)

### Distilling AI solutions for enterprise

We enable enterprises to build smaller, more focused AI models with greater performance at a fraction of the cost.

[Contact Us](#)



[Products](#) [Solutions](#) [Developer](#)

EDGE AI PLATFORM

### Ultra-fast, secure edge AI

The enterprise AI platform that takes the guesswork out of edge AI, empowering every developer with precision model to hardware pairings so you can bring AI projects to market faster.

[Get Started →](#)

[Learn More →](#)



[Products](#) [▼](#)

### Full stack enterprise AI.

Simple, secure AI tools built to deliver cost-efficient business AI that just works. From intelligent model routing to enterprise AI agents.

[Book a demo](#)





## Related Research

- **“Beyond distillation: Task-level mixture-of-experts for efficient inference”**, Google AI. 2021
  - ◆ “Determines routing based on task IDs, ensuring that different tasks are routed to distinct experts, thereby effectively minimizing interference between tasks. Compared to token-level routing strategies, this approach only requires loading a subset of experts relevant to the current task during inference, rather than loading all experts of the entire model. This results in reduced communication costs between devices and lower memory usage. “
    - “A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications”. Siyuan Mu and Sen Lin, 2025
- **“Modular deep learning”**, Google Deepmind. 2023
- **“CoServe: Efficient Collaboration-of-Experts (CoE) Model Inference with Limited Memory”**, Beihang University, ASPLOS '25

## Expense:

? for software development.

Fine tuning each small model < ?\$. ?\$ GPU cost in total if rent.

## Timeline:

Model training: 6~? months.

App Development: 6~? months.

parabase.ai  Active

Renewal Date : 4/22/2027 [Renew Now](#)

domain name is secured:)

Parabase 