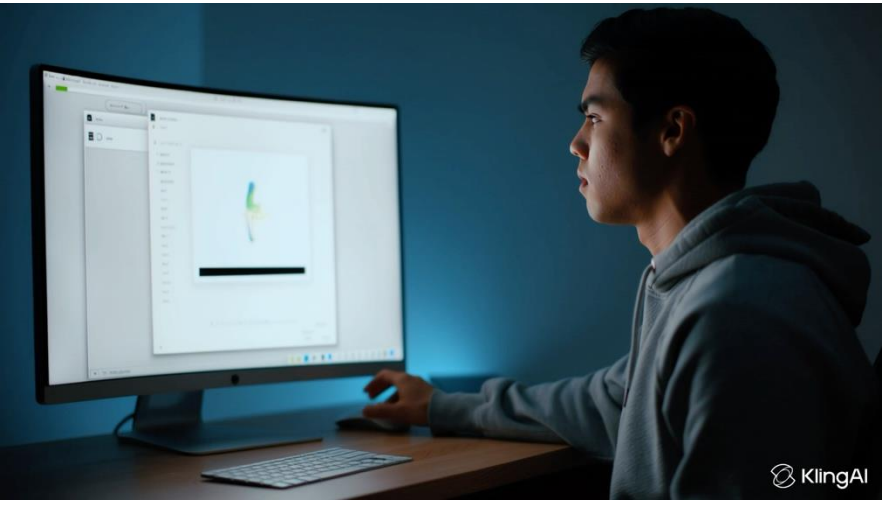
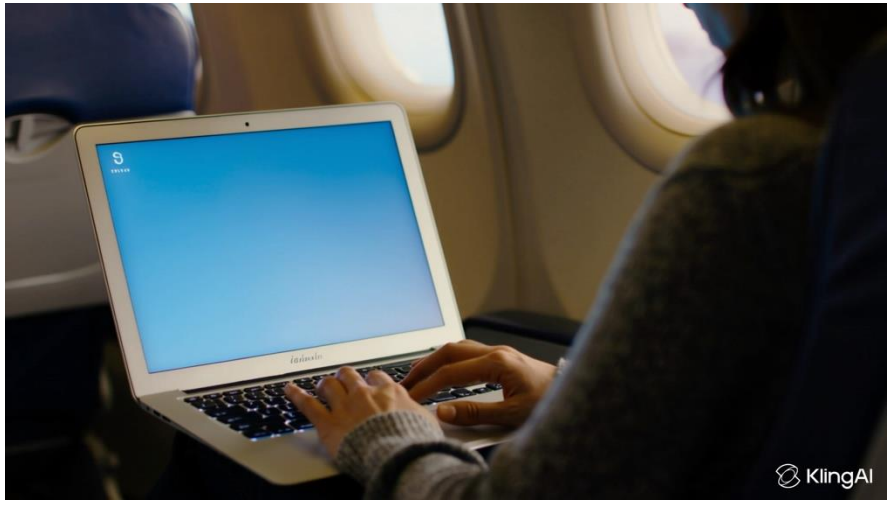
Cloud LLMs incur high cost, privacy risks, and rely on Internet

When you get shower and notice the camera.

When your AI agent relies on cloud APIs for a long time.



When writing your paper on a plane but there is no internet.



What if we run LLMs locally?

Totally Free!

What if it runs locally?

Keep my privacy local!

Run it locally!

Use it everywhere!

PRIMA.CPP

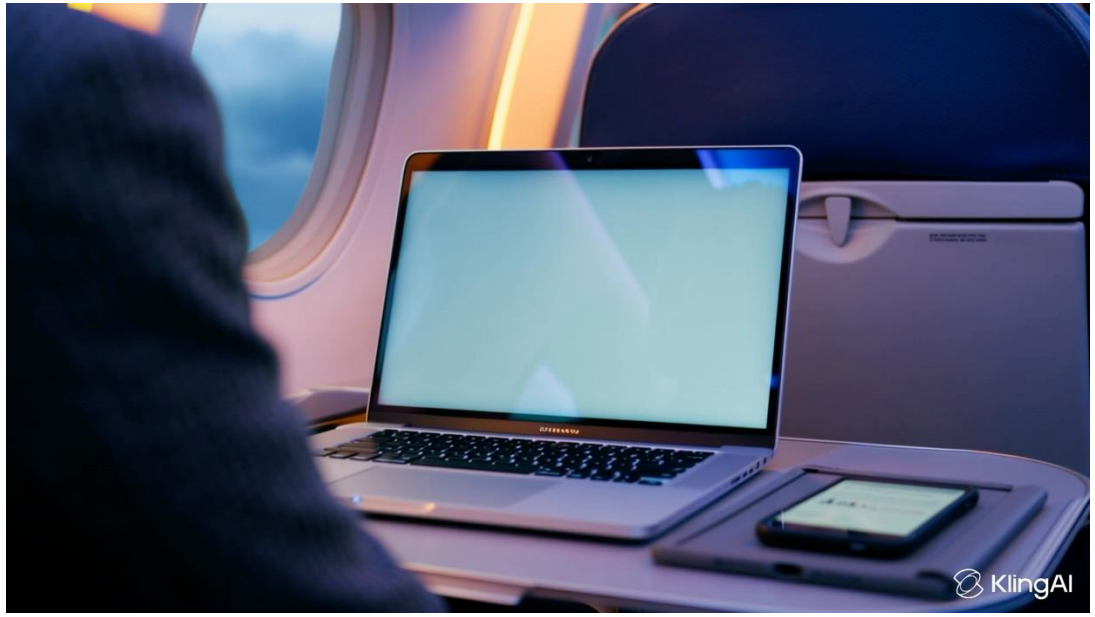
ICLR'26

Fast 30-70B LLM Inference on Heterogeneous and Low-Resource Home Clusters

Zonghang Li, Tao Li, Wenjiao Feng, Rongxing Xiao, Jianshu She, Hong Huang, Mohsen Guizani, Hongfang Yu, Qirong Ho, Wei Xiang, Steve Liu

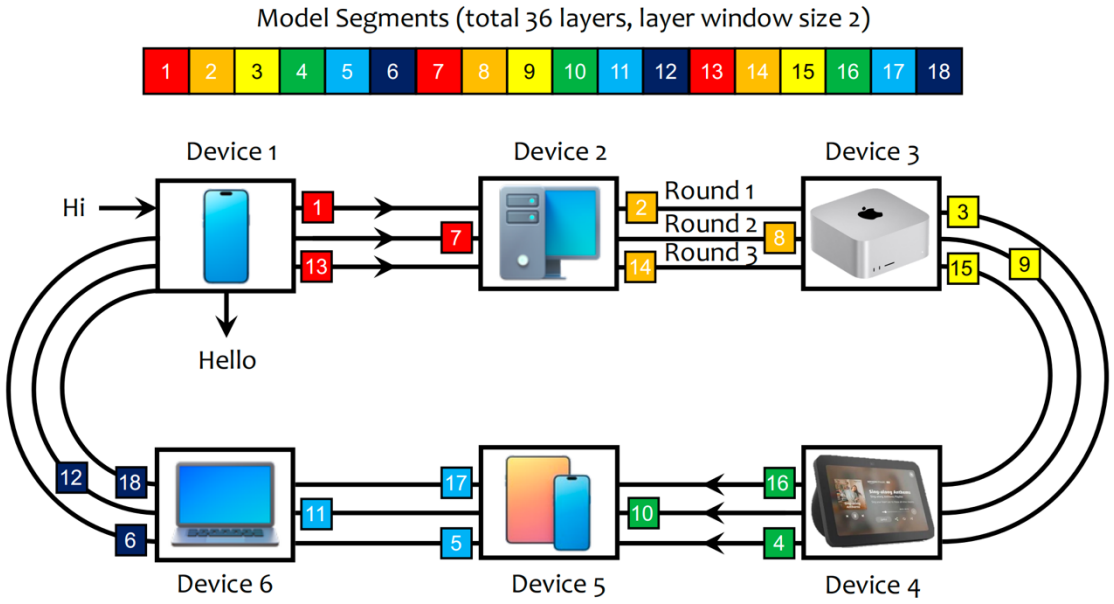
MBZUAI, UESTC, CityU, LTU

Still worried about token fees, privacy, and no internet? Try **PRIMA.CPP.**
Pool your devices on hand to serve 30-70B models locally —— **free, secure, and fast!**

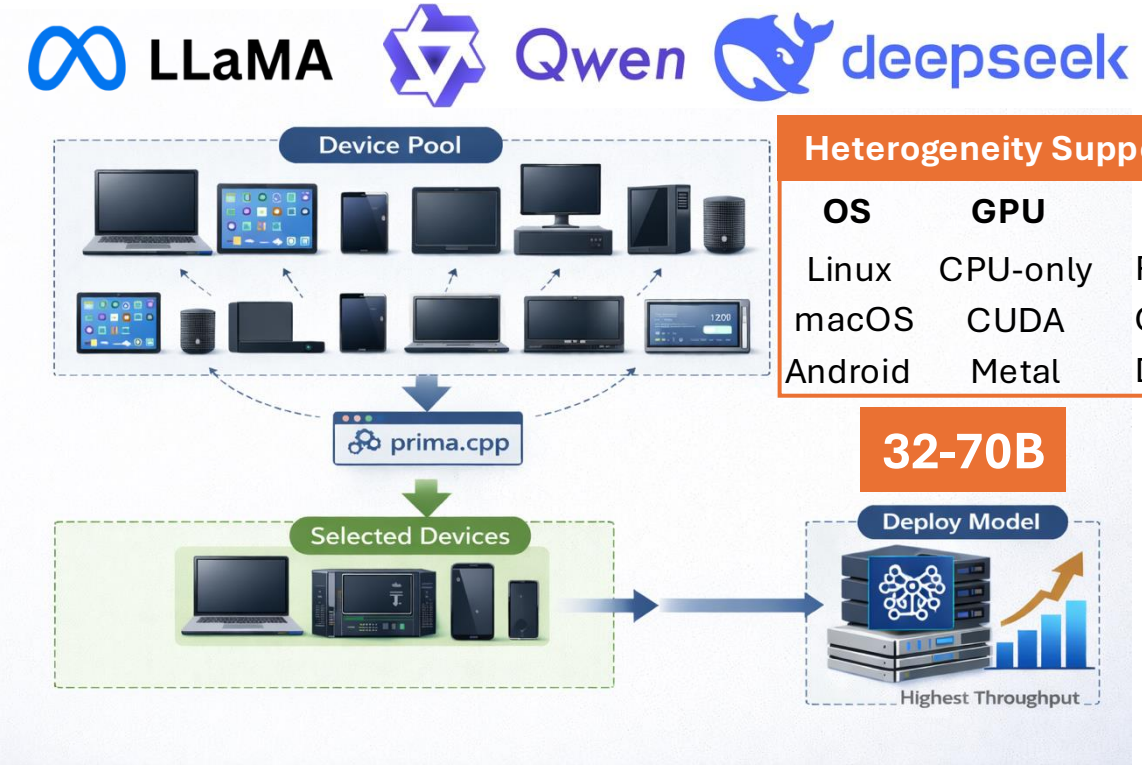


Big Features, Simple Setup

Pipelined-Ring Parallelism (PRP)



Automatic Device Selection

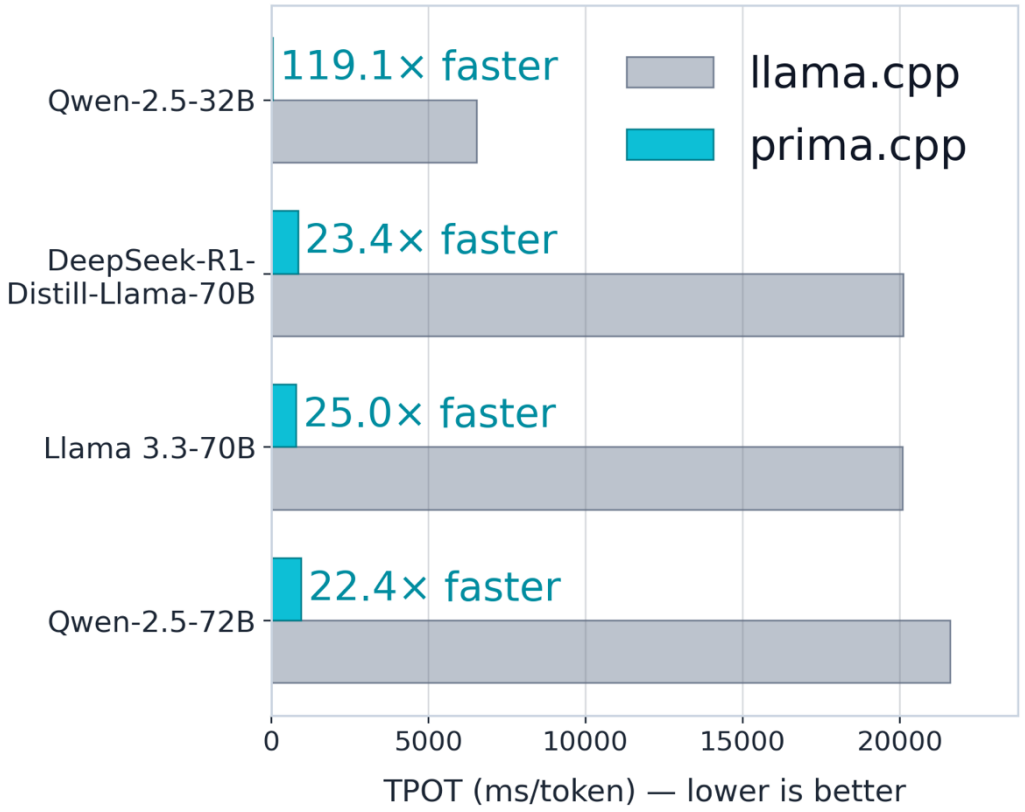


Don’t worry about heterogeneity, prima.cpp helps you allocate layers.

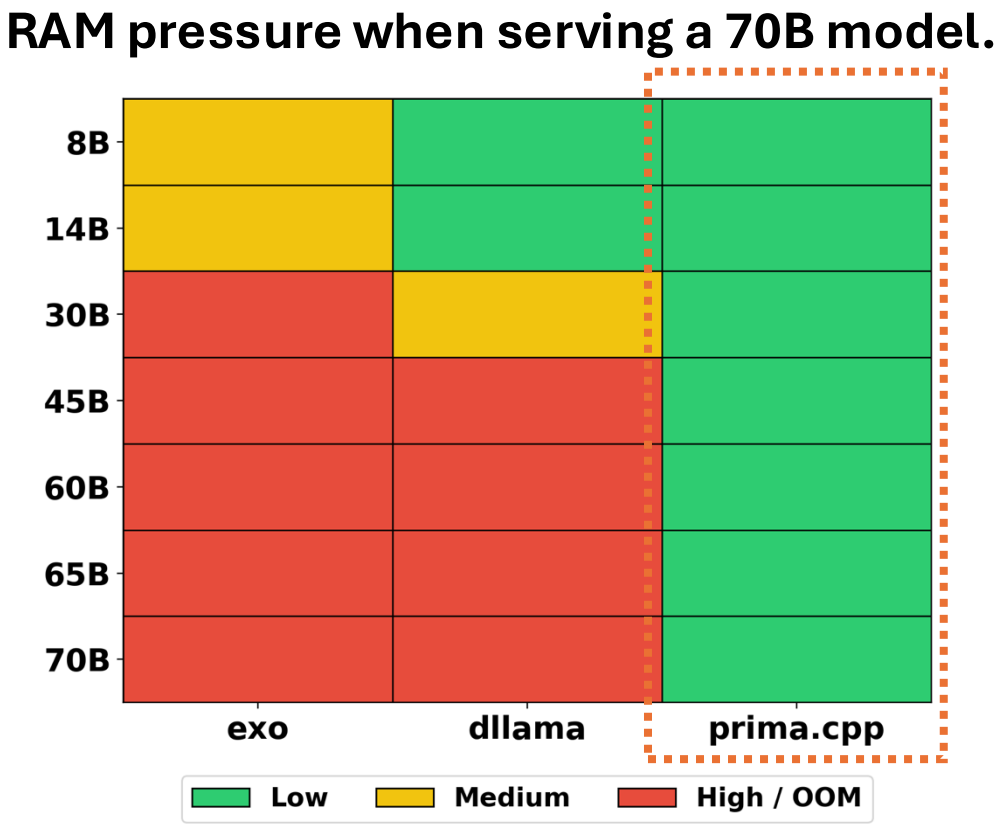
Collect as many devices as you can, prima.cpp helps you select devices!

Fast, Low RAM Pressure, no OOM

Up to 119x Faster



Low RAM Pressure, No OOM



Thank you!

GitHub

Lizonghang/prima.cpp

gitee

zonghang-li/prima.cpp