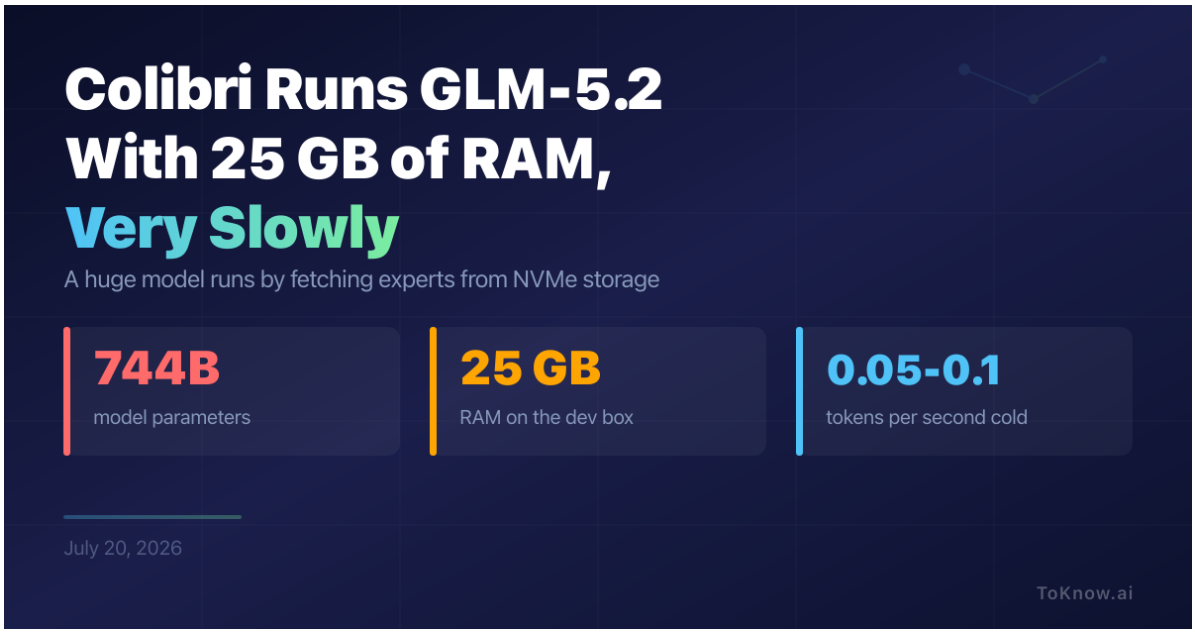
Colibri Runs GLM-5.2 With 25 GB of RAM, Very Slowly

Kabui, Charles

2026-07-20

Read at ToKnow.ai



Colibri is a small C runtime that can run the 744-billion-parameter GLM-5.2 model on a machine with 25 GB of RAM. GLM-5.2 uses a mixture-of-experts design, so it calls only a fraction of its specialist sections for each token. Colibri keeps the model's shared parts in memory and fetches the selected experts from NVMe storage as needed. Its original WSL2 test used 12 CPU cores, a roughly 370 GB four-bit model file, 9.9 GB of permanently resident

weights, and about 20 GB of memory at peak. Startup took about 30 seconds. Cold generation read roughly 11 GB from disk for every token and managed only 0.05 to 0.1 tokens per second.

That speed means one token every 10 to 20 seconds, so a 100-token answer could take 17 to 33 minutes and read about 1.1 TB. The streaming path is read-only, which limits SSD wear, but sustained reads can heat cheaper drives and low memory can trigger wear-producing swap writes. This proves feasibility, but it is not practical for everyday laptop chat. On a different machine with six RTX 5090 GPUs, keeping every expert in GPU memory and RAM removed disk waits and reached 6.84 tokens per second over a 256-token run.

Read More: GLM-5.2's model size, licence, and one-million-token context

Sources:

- Colibri repository and architecture
- Colibri benchmark tables
- Six RTX 5090 experiment log
- GLM-5.2 model card and MIT licence

Disclaimer: *For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation.* ***Read more:*** */terms-of-service*