MY WEIRD PROMPTS

Podcast Transcript

EPISODE #110

Building the Ultimate Local AI Inference Server

Published December 27, 2025 • Runtime: 21:13

https://myweirdprompts.com/episode/local-ai-inference-server-guide/

EPISODE SYNOPSIS

Are you struggling to run the latest AI models on your aging hardware? In this deep dive, Herman and Corn break down the technical requirements for building a dedicated local inference server in late 2025. They move beyond simple chatbots to discuss "agentic" code generation—systems that can autonomously debug and test projects—and why these sophisticated tools demand massive amounts of VRAM. From the technical hurdles of the KV cache to a step-by-step shopping list for a dual-RTX 3090 PC build, this episode provides a comprehensive hardware roadmap for developers. They also weigh the pros and cons of Apple’s unified memory architecture versus the raw power of DIY Linux builds, exploring how quantization can help you squeeze more performance out of your budget. If you value privacy and need the speed of local execution, this is the hardware guide you've been waiting for.

DANIEL'S PROMPT

Daniel

I’m interested in running the GLM 4.7 model by Z.ai locally for agentic code generation. My current 12GB VRAM setup isn’t powerful enough to maintain a usable context window. What local hardware and approximate cost would be required to build an inference server that provides a baseline of decent, usable performance for this type of model?

TRANSCRIPT

Corn

Hey everyone, welcome back to My Weird Prompts! I am Corn, and as always, I am here with my brother.

Herman

Herman Poppleberry, at your service. It is great to be here. We have got a really technical, hardware focused prompt today that I have been itching to dive into.

Corn

Yeah, you have been vibrating with excitement since we listened to the audio. This one comes from our housemate Daniel, who lives here with us in Jerusalem. He is looking into running a specific AI model locally, and it sounds like his current computer is just not cutting it.

Herman

It is a classic problem, Corn. The software is moving at light speed, and the hardware we bought two years ago is already feeling like a vintage typewriter. Daniel is asking about the GLM four point seven model from Z dot AI. Specifically, he wants to use it for agentic code generation.

Corn

Okay, slow down already. You lost me at agentic. I know what a sloth is, and I know what code is, but what is an agentic code generator?

Herman

That is a great place to start. Most people think of AI like a chatbot, right? You ask a question, it gives an answer. But an agentic system is more like a digital intern. Instead of just writing a single function, an agentic model can look at your whole project, find a bug, think about how to fix it, try a solution, run the tests to see if it worked, and then try again if it failed. It acts as an agent on your behalf.

Corn

Oh, so it is actually doing the work, not just suggesting the words. That sounds like it would need to keep a lot of information in its head at once.

Herman

Exactly! And that is where Daniel is hitting a wall. He mentioned he has twelve gigabytes of video RAM, or VRAM, and that it is not enough to maintain a usable context window. For those listening who might not know, the context window is basically the short term memory of the AI. If you are writing code, the AI needs to remember the file you wrote ten minutes ago to make sure the new code actually fits.

Corn

And twelve gigabytes is not enough? That sounds like a lot for a normal computer.

Herman

For gaming or video editing, twelve gigabytes is decent. But for these massive new models like GLM four point seven, it is like trying to fit a gallon of water into a thimble. Especially in late twenty twenty five, these models are getting more sophisticated and their memory requirements are ballooning.

Corn

So Daniel wants to build an inference server. He wants a dedicated machine just for running this AI. Where do we even start with that? Is he looking at spending a few hundred bucks or are we talking about a second mortgage?

Herman

Well, it is definitely more than a few hundred bucks, but we can definitely find a sweet spot. To run GLM four point seven with a decent context window, we need to talk about the two main paths: the PC route with dedicated graphics cards, or the Mac route with unified memory.

Corn

I remember Daniel mentioning in the audio that he is not really a Mac guy, but he heard they were good for this. Why is that?

Herman

It is all about how the memory is shared. In a PC, your system has regular RAM and your graphics card has VRAM. They are separate. If the AI model is twenty gigabytes, and your graphics card only has twelve, the whole thing slows down to a crawl because it has to keep moving data back and forth. But a Mac with Apple Silicon has unified memory. The processor and the graphics cores all pull from the same pool. So if you buy a Mac with one hundred twenty eight gigabytes of RAM, the AI can use almost all of it.

Corn

That sounds like a massive advantage. But if he wants to stick to a PC build, what is the baseline for decent performance?

Herman

For a PC build in December twenty twenty five, the gold standard for local AI is still the NVIDIA RTX series. If twelve gigabytes is not enough, the next logical step up is twenty four gigabytes. You can find that on the older RTX thirty ninety or the forty ninety. But here is the kicker, Corn. For agentic code generation, even twenty four gigabytes might be tight if you want a really long context window.

Corn

Wait, so if he buys a top of the line graphics card, he might still be limited?

Herman

Possibly. See, the model itself takes up space, and the context window, the memory of what you have been doing, also takes up space. As the conversation gets longer, the memory usage grows. This is what we call the KV cache. To have a baseline of usable performance, Daniel is probably looking at a multi GPU setup.

Corn

Two graphics cards in one box? That sounds like something out of a sci-fi movie. Is that even allowed?

Herman

Oh, it is more than allowed, it is encouraged in the AI community! You can take two used RTX thirty ninety cards, which each have twenty four gigabytes, and link them up. Suddenly, you have forty eight gigabytes of VRAM. That is enough to run GLM four point seven at a high precision with a very healthy context window.

Corn

Okay, let's talk numbers. If Daniel goes out today and tries to build this two card monster, what is the damage to his wallet?

Herman

If he goes the used route, which is very common for these builds, he can probably find thirty ninety cards for about seven hundred to eight hundred dollars each. So that is sixteen hundred for the GPUs. Then you need a beefy power supply because those cards are thirsty, a motherboard that can actually fit two giant cards, and a case with enough fans to keep it from melting. You are probably looking at a total of around twenty five hundred to three thousand dollars for a solid, home-built inference server.

Corn

Three thousand dollars just to run a coding assistant? Man, I think I will just stick to eating leaves and hanging from trees. That is a lot of money!

Herman

It is, but think about the value. If you are a developer, and this agentic tool saves you ten hours of work a week, it pays for itself in a couple of months. Plus, you have total privacy. None of your code is being sent to a cloud server owned by a giant corporation. For some people, that privacy is worth every penny.

Corn

That is a fair point. I guess I forget how much people value their secrets. But before we get deeper into the Mac versus PC debate and the specific specs, let's take a quick break for our sponsors. Larry: Are you tired of your computer being slower than a turtle in a tar pit? Is your VRAM so low it can barely remember its own name? You need the RAM-Slammer Five Thousand! The RAM-Slammer is the world's first external memory expansion that plugs directly into your wall outlet. Using patented lightning-capture technology, it converts raw electricity into digital thoughts, giving your computer an infinite context window! Does it work? My cousin says it does! Is it safe? Define safe! The RAM-Slammer Five Thousand comes in three colors: Carbonized, Scorched, and Void. Warning: do not use near water, pets, or people you care about. May cause localized gravitational anomalies. RAM-Slammer Five Thousand! BUY NOW!

Corn

...Alright, thanks Larry. I think I will pass on the localized gravitational anomalies today. Herman, back to reality please.

Herman

Yeah, let's definitely stay away from plugging our memory into the wall outlet. So, we were talking about the cost of a PC build. About three thousand dollars for a dual thirty ninety setup. But we should also mention the forty ninety. A single forty ninety is incredibly fast, but it still only has twenty four gigabytes of VRAM. In late twenty twenty five, we are starting to see the fifty series cards, and everyone is hoping for more VRAM, but NVIDIA likes to be stingy with it.

Corn

So if Daniel really wants to go pro, and he wants to avoid the headaches of building a PC and dealing with driver issues, is the Mac actually a better deal?

Herman

It might be. If he looks at a Mac Studio with an M four Ultra chip, he could configure that with one hundred ninety two gigabytes of unified memory. That machine would be an absolute beast for local AI. It could run models much larger than GLM four point seven, and it could handle context windows that would make a PC cry.

Corn

But Daniel said he is not a Mac guy. Is the software support there?

Herman

It has gotten much better. Tools like Ollama, LM Studio, and especially MLX, which is Apple's own machine learning framework, have made running models on Mac incredibly easy. It is often a one-click setup. On a PC, you are often dealing with Linux, CUDA drivers, and complex Python environments. It is a lot of fun if you are a nerd like me, but it is a lot of work.

Corn

Okay, so let's weigh the options for him. Option A is the DIY PC with two used graphics cards for three thousand dollars. Option B is the Mac Studio, which probably costs... what, five or six thousand?

Herman

Yeah, a high-spec Mac Studio is going to be in that five to seven thousand dollar range. It is a significant jump in price, but you get a lot of memory and a very small, quiet machine. Those dual GPU PCs are loud and they put out a lot of heat. It is basically a space heater that also happens to write code.

Corn

Well, we do live in Jerusalem, and it gets cold in the winter. Maybe the space heater is a feature, not a bug!

Herman

Ha! True. But there is a middle ground. Daniel could look at a single GPU build with an RTX forty ninety and use something called quantization.

Corn

Quantization. That sounds like another one of your fancy words. Break it down for the sloth in the room.

Herman

Think of it like a high-quality photo versus a JPEG. A high-quality photo has all the detail, but the file is huge. A JPEG throws away some of the information that your eyes don't really notice to make the file smaller. Quantization does that to an AI model. You can take a large model and squeeze it down from sixteen bits to four bits. It loses a tiny bit of intelligence, but it takes up a quarter of the space.

Corn

So he could run a bigger model on his twelve gigabyte card if he just squeezed it enough?

Herman

To an extent, yes. But Daniel's problem is the context window. Even with quantization, the memory needed for the conversation itself, the KV cache, does not shrink nearly as much. For agentic code generation, you really need that raw VRAM. If he wants a baseline of decent, usable performance, I would say twenty four gigabytes is the absolute minimum, and forty eight gigabytes is where it actually starts to feel good.

Corn

So, if he is building this today, December twenty seventh, twenty twenty five, what are the specific parts he should put on his shopping list to get that forty eight gigabytes?

Herman

Okay, here is the recipe for a solid mid-range inference server. First, a motherboard with at least two PCIe slots that are spaced far apart. These cards are thick! Look for something like an ASUS ProWS series. Second, two used RTX thirty ninety cards. You want the thirty ninety because it has twenty four gigabytes and it is much cheaper than the forty ninety. Third, a sixteen hundred watt power supply. You do not want to skimp here. If those cards both spike at once, a cheap power supply will literally pop.

Corn

Pop? Like a balloon?

Herman

More like a small explosion with a side of blue smoke. Avoid that. Fourth, at least sixty four gigabytes of system RAM. Even though the AI lives on the GPU, the system still needs room to breathe. And finally, a case with excellent airflow. I am talking six or seven fans.

Corn

And the cost for all that together?

Herman

If he is savvy with used parts, he can get that done for about twenty eight hundred dollars. If he buys everything brand new, and maybe goes for the newer cards, he is looking at four thousand plus.

Corn

That is a lot of money, but I guess it is an investment in his craft. You mentioned something earlier about the context window math. How do you actually calculate how much memory you need for a certain number of words?

Herman

It is a bit technical, but the rule of thumb is that for every thousand tokens, which is about seven hundred fifty words, you need a certain amount of VRAM for the cache. At sixteen-bit precision, it is about one gigabyte of VRAM for every eight thousand tokens. So if Daniel wants a thirty-two thousand token context window, which is about the size of a small book, he needs four gigabytes just for the memory of the conversation, on top of the size of the model itself.

Corn

So if the model is twenty gigabytes, and the context is four gigabytes, he is already at twenty four. He is maxed out on a single card.

Herman

Exactly! And thirty-two thousand tokens is actually not that much for a coding agent. If you have ten different code files open, you can hit that limit in five minutes. That is why the forty-eight gigabyte setup is so much better. It gives him room to breathe. He could go up to a sixty-four thousand or even a one hundred twenty-eight thousand token context window. That is where the real magic happens. That is when the AI can actually understand the whole project.

Corn

I am starting to see why he is frustrated with his twelve gigabyte card. It is like trying to write a novel on a sticky note.

Herman

That is a perfect analogy, Corn. It is exactly like that. He has this incredibly smart brain in GLM four point seven, but it has the short term memory of a goldfish because of his hardware.

Corn

So, we have talked about the PC build and the Mac Studio. Is there any other weird hardware out there? What about those enterprise cards you see on eBay?

Herman

Oh, the NVIDIA A-series or the old Tesla cards? They are interesting. You can sometimes find an NVIDIA A-sixty-thousand with forty eight gigabytes of VRAM on a single card. They are designed for data centers, so they don't have fans. You have to rig up your own cooling system. It is a bit of a hack, but for a dedicated inference server, it can be a great way to get a lot of memory without the complexity of dual GPUs. They are still expensive though, usually around three to four thousand dollars just for the card.

Corn

Man, there is no cheap way out of this, is there?

Herman

Not if you want "decent, usable performance" as Daniel put it. You can run these models on your CPU, using your regular system RAM, but it is painfully slow. It is like watching someone type with one finger. For coding, you want that instant feedback. You want the AI to suggest the next line before you even finish thinking of it. For that, you need the speed of a GPU.

Corn

Okay, so let's summarize the advice for Daniel. If he wants the most bang for his buck, go for the dual used RTX thirty ninety PC build. It will cost him around three thousand dollars and give him forty eight gigabytes of VRAM, which is plenty for GLM four point seven.

Herman

Right. And if he wants the easiest, most reliable experience and he is willing to pay a premium, go for a Mac Studio with an M four Ultra and at least one hundred twenty eight gigabytes of RAM. That is the "buy it once and forget about it" option.

Corn

And if he is really on a budget?

Herman

If he absolutely cannot spend more than a thousand dollars, his best bet is to find a single used RTX thirty ninety for seven hundred bucks, put it in his current machine if the power supply can handle it, and use heavy quantization. He will be limited to a smaller context window, but it will be a massive upgrade from his current twelve gigabyte setup.

Corn

That sounds like a solid plan. I hope that helps Daniel out. It is cool that he is doing this right here in the house. Maybe once he gets it running, he can program a robot to bring me more snacks so I don't have to climb down from my branch.

Herman

Knowing Daniel, he would probably program the robot to make you do your own chores, Corn.

Corn

Hey, a sloth can dream! This has been a really interesting deep dive, Herman. I actually feel like I understand why VRAM matters now. It is not just about how fast the computer is, it is about how much it can hold in its head at once.

Herman

Exactly. In the world of AI, memory is just as important as speed. Maybe even more so for these agentic tasks where context is everything.

Corn

Well, I think we have covered the bases. Daniel, we hope your new inference server build goes smoothly. Let us know which route you choose!

Herman

And if you need help installing those twenty-four fans to keep the thirty-nineties cool, you know where to find me. I will be the one wearing the thermal goggles.

Corn

Thanks for listening to My Weird Prompts. If you have a prompt you want us to tackle, head over to my weird prompts dot com and send it our way through the contact form. We love hearing from you.

Herman

And you can find all our previous episodes on Spotify or wherever you get your podcasts. We have got a full RSS feed on the website too for you subscribers.

Corn

This has been My Weird Prompts. I am Corn.

Herman

And I am Herman Poppleberry.

Corn

See you next time!

Herman

Goodbye everyone! Give your GPUs a hug for me! Larry: WAIT! Before you go, do you have too much money and not enough mystery in your life? Buy my Mystery Box! It is a box! It is a mystery! It might contain a vintage GPU, or it might contain a very angry hornet. Only one way to find out! Mystery Box! BUY NOW!