GPU Hosting and Open Source AI Will Revolutionize or Kill WordPress

Mark Maunder

On the eve of WordCamp US 2024 we find ourselves in the midst of a revolution. It is perhaps the most profoundly transformative technology revolution our species has experienced in our short history in this Universe.

In fundamental terms, since computers have existed we have been programming them by writing individual functions by hand. We recently discovered how to train functions to solve problems so complex that all the programmers in all the world, working the entirety of their natural lives, would not be able to solve these problems using traditional programming.

But it goes beyond that. We've figured out how to train an artificial brain to have a conversation with us that includes reasoning capability and problem solving. The hilarious part is that it was as simple as creating an AI model that predicts the next part of a conversation, which is a solution so simple it has left swathes of academia bitterly disappointed in both computer science and our own species. Surely our power of reason is more complex, more nuanced, more… special?

We've even figured out how to train an artificial brain to be a world-class programmer, which will ultimately help us to create far more effective artificial brains. As the invention of computers created technology that accelerated innovation in hardware and software exponentially, so innovation in the field of AI will further accelerate the field.

WordPress is what powers the majority of websites that publish content on the World Wide Web – content that until now has mostly been created by humans. Large Language Models are spectacularly good at ingesting content, doing things with it and producing new content by combining, summarizing, reformatting and even reimagining the subject matter with a user-defined fresh take.

In this post I'll discuss the powerful enabler that is agentic AI, how it will transform user experiences, how we are putting this technology to work at Wordfence, the chasm in the market around GPU enabled hosting for WordPress, why it is an important enabler for society, and the incredibly exciting business opportunity that GPU enabled WordPress hosting represents.

Note: If you're new to AI, for clarity in this post I'll refer to AIs as "models" for the most part, which is what an AI is in it's portable and deployable form.

The Future is Agentic

For the past year a concept known as "agentic AI", "function calling" or "tool calling" has been gaining momentum. That is, the ability for a model to call functions that a developer has defined and given the model access to. This takes a model, which is an isolated electronic brain receiving inputs and outputs, and gives it arms, legs, hands and feet. Tool calling gives a model the ability to do stuff at will. Don't worry, they don't call functions directly. They express their intention to call a function with parameters and the developer's code decides if the function is actually called.

Tool calling capability in AI first showed up in the mainstream last year when OpenAI announced some limited capability. Since then we've seen it appear in many other models including Meta's Llama. I use an application called vLLM to host our large language models in a highly scalable way, and vLLM supports tool calling for Hermes and Mistral and I've added that capability for Llama 3.1 which we're using internally – and I'll contribute a pull request at some point. The point is that tool calling – or "agentic" capability in models – is a rapidly growing field and all the bits and pieces you need to do this out of the box haven't been created yet but are rapidly emerging.

UX Will be Transformed

The punchline is that you're going to be able to expose the entire WordPress API to an LLM, make a few recommendations about how to use the API and what to avoid, and the future of publishing with WordPress will be you having a conversation with an AI that will go something like this: (Lets call the AI Charlie)

- Hey Charlie, you up?
- Mark you know I'm always up. How can I help?
- Go read the latest aviation news and let me know if I've written about anything being discussed in the past 2 years.
- Done, and yes a plane nearly landed on a taxiway and you suggested ILS be made mandatory on visual approaches a year ago.
- OK go and see if there is any regulatory progress on that.
- There is. There's a draft bill they're trying to attach to the FAA reauthorization bill.
- Cool, write a post that covers the taxiway near-landing, mention my thoughts on ILS and bring in any relevant data in the draft bill and reference your sources. I also want ILS to be the focus keyword with a density of at least 20 and give me three catchy headline options.
- Done. Here it is.
- Looks good. Use the first headline, publish it, and then be fairly strict about how you moderate comments as they arrive. I only want comments that are adding new data to the story.
- Will do.

- And email the mailing list, but engaged contacts only and it's time for us to clean house so delete all unengaged contacts after you back them up. We'll do the final deletion after a few weeks once we're sure we got it right, so go ahead and schedule that.
- Done.

I've taken a few liberties with the aviation facts in this post – the FAA bill has already been approved by the Senate – but you get the idea. There are a quite a few functions that Charlie is calling in our example including URL fetching, WordPress API functions to read old posts, publish new posts and moderate comments, and API functions from a mailing list provider like MailChimp, Aweber or Hubspot to send email and manage a mailing list.

Agentic Systems are Transforming Security Research

At Wordfence we have an internal agentic system called Murphy that assists with security research. Murphy uses Llama 3.1 405B quantized down to 8 bits, running on an 8 GPU cluster of Nvidia H100 GPUs. Murphy has a range of functions I've built including URL fetching and the ability to accept a ZIP file and scan the entire file for vulnerabilities in code. Llama 3.1 makes agentic calls – or tool calls if you prefer – to other AI systems to evaluate code for vulnerabilities. We're exploring ways to make this capability available to vendors and security researchers so let me know at our booth at WCUS if you're interested in this.

Our user interface for Murphy is Slack and Murphy exists as a slackbot. It turns out a platform that already facilitates conversation, including threading, is ideal for interacting with a system emulating a human.

The biggest limit on what we can do with AI models currently is the size of their context window. Models that are the industry leaders at generating and auditing code have a context window of 128,000 tokens or roughly 80,000 to 100,000 words. Google's Gemini Pro 1.5 supports a context window of up to 2 million tokens and is unique in that respect. Models tokenize words and a word can be one or more tokens depending on the tokenizer used. That context window size limits the amount of code a model can ingest at once and hold in it's electronic brain while evaluating the code for vulnerabilities or performing any kind of analysis.

So right now we build workarounds like processing one file at a time or finding clever ways to combine files for processing. A larger context window would be transformative for not just security, but for all fields. A very large context window means you can feed "Save the Cat", Robert McKee's "Story" and Joseph Campbell's "Hero with a Thousand Faces" into an LLM in a single query and ask it to write the best screenplay ever written. Well… it'll probably just write Star Wars Episode IV.

If you're having a conversation with a model and, lets say the model is providing companionship for you along with assisting you with various tasks, the context window is what limits the amount of conversation history – or relationship history – that you can have. Every time you prompt a model you have to provide the entire conversation thus far, or if you're thinking long-term, the relationship thus far. Context window is a model's memory about you, the world and itself. Larger context windows are critical if we are to have models that have current knowledge and up-to-date memory.

The future is large AI models with huge context windows calling functions and many of those functions will access other models which may in turn call their own functions to help fulfill not just individual tasks, but whatever your generalized goal is. That may be harvesting a healthy apple crop this fall and sharing your journey and learnings with others, or it may be positioning yourself as the preeminent source for breaking aviation news.

Where is GPU Enabled WordPress?

In the WordPress community we enable publishers. We democratize publishing – as the WordPress tagline goes. How is it possible that when I search for GPU enabled WordPress hosting, there is none available? How can this possibly be? The only explanation is that publishers who are beginning to be users of AI are using closed source APIs from OpenAI and other providers. As a proponent of open source, and someone who has witnessed the positive impact of open source on the development of the World Wide Web over the past 30 years, we need to seize this opportunity.

WordPress is open source. WordPress stands for the democratization of publishing. Why should we democratize publishing? Because if we only share approved thoughts and ideas, we won't have any new ideas and it will create a concentration of power among the few who have the ability to decide which ideas are allowed to be shared. For example, if you can't criticize people in power, they will probably stay in power even if they are doing very bad things. And if you can't criticize old ideas and freely introduce new ideas, you will have an ossified society that never evolves and is stuck with antiquated values.

Just like WordPress democratized publishing, open source AI will democratize computation. But if the most powerful and capable models are kept behind lock and key and we are provided limited, monitored and metered access to them, most of us will be at a severe disadvantage to the wealthy and powerful who have unfettered access to the newest most powerful models. The best writing, the most advanced physics, the newest engineering tools and techniques and the newest trading tools and business applications will all be created behind closed doors using powerful closed source AIs.

For a brief period in the early history of the Web the dominant web server was closed source and produced by Netscape. I built my first web application on Netscape's Commerce Server around 1995. Here's Netscape's server page in 1996 and here is a snapshot of Apache's home page around the same time right as Apache started to become the dominant web server. If Web serving remained closed source, it would have choked out innovation before it got started, and we would never have seen the technology revolution, investment, product innovation and new wealth creation that we saw in the late 1990s extending deep into this century.

The Great Concentration of Power Has Started

If you need to navigate insurmountable regulatory hurdles to develop or host AI, it ensures that only wealthy and powerful companies who have large legal teams and budgets have the resources to develop and deploy AI. The wealthy and the powerful in the AI space are doing their very best to raise the drawbridge of their castles and construct regulatory moats that no one else can cross without great effort and expense.

We are at a point of inflection. Or as the cool kids would say, we're at a "moment". Now is when we decide if our society will be able to train and host our own models on our own machines which we can share with others, that they can further fine-tune and share back to us – or if AI will become outlawed and our only access will be to older approved models via monitored and metered access points.

That is why it is critically important for the WordPress community to act now, act fast, and to support and invest in open source AI. In the WordPress space we should be downloading open source models, modifying them and contributing them back to the community and developing open source code around those models and contributing that code to the community. By creating a vibrant and productive open source AI movement within WordPress, we add our weight to the grass roots movement behind open source AI and ensure that this critical computational resource remains democratized and accessible to all.

Open Source AI Hosting is a Massive Opportunity

The good news is that open source AI presents a massive business opportunity for the most profitable segment of the WordPress market which is hosting. I put the global WordPress hosting market at somewhere between 10 and 30 billion dollars a year. Not a single host has launched GPU hosting targeted at the WordPress market or as part of a WordPress hosting package. The future of WordPress is AI with local open source models running on WordPress websites.

Companies like Lambda Labs, RunPod and others can't buy GPUs from Nvidia fast enough to satisfy their customers. WordPress hosting companies are missing the biggest opportunity the hosting space has ever seen. Nvidia's market cap today is $2.8 trillion. Two years ago they were just over $300 billion. OpenAI has seen the fastest growing user-base in history. We've never seen growth and innovation in technology at this pace before.

The Universe just delivered WordPress hosting a fresh business model that gives their customers unbelievable value and a huge competitive advantage as publishers. WordPress hosting providers already have customers in the millions, and those customers are already sold on AI. Are a handful of AI hosting startups going to take it all?

Is a new Python based CMS going to emerge alongside a fresh hosting brand offering self-hosted open source AI models? Or is the WordPress community going to rise to the challenge and contribute our open source army to the movement supporting open source AI?

See you at the Wordfence booth in Portland at WordCamp US 2024 this week.

Mark Maunder – CTO @ Defiant Inc.