Noah Kravitz (00:01.265)
Hello, and welcome to the NVIDIA AI Podcast. I'm your host, Noah Kravitz. Right now, the world is watching AI evolve faster than ever before. And that progress isn't just being fueled by technological breakthroughs in scale. It's being fueled by human collaboration. Open source models, open data sets, and shared research are giving developers, enterprises, and governments the building blocks they need to innovate together. NVIDIA has been part of this movement from the very beginning.

contributing open libraries, publishing datasets and research, and most recently, sharing families of open models. Which brings us to today's episode. We're talking about Nemotron, specifically unlocking the secret of Nemotron. On the surface, Nemotron may look like just another open model family, but the real story is how it anchors NVIDIA's strategy for building accelerated infrastructure and driving increased adoption of AI everywhere. Joining us to unpack this open secret

are two of the leaders driving this work forward. Bryan Contanzaro is vice president of applied deep learning research at NVIDIA, and Jonathan Cohen is vice president of applied research at NVIDIA. Bryan and Jonathan are here today to talk Nemotron. can't wait. Gentlemen, welcome to the AI Podcast. Thank you so much for making the time to join us.

Jon (01:20.258)
Thank you for having us.

Bryan Catanzaro (01:22.625)
It's great to be here.

Noah Kravitz (01:22.661)
So let's start at the top, and I'll direct this one to you, Bryan, to get us going if that's all right. What is Neamotron? And as a follow-up, why did NVIDIA decide to build its own family of models when you already work with essentially every major model builder out there?

Bryan Catanzaro (01:41.634)
Nemotron is NVIDIA's open technology for artificial intelligence. Nemotron includes models that we train. It also includes data sets that we release, as well as algorithms and methodologies. our goal with Nemotron is to support the community in building customizable AI that can be integrated deeply and tightly into the beating heart of every business around the world.

Our second goal with Nemotron is to help NVIDIA design systems for deploying and constructing AI. There's a lot of questions about how AI works that touch the various design decisions that go into building NVIDIA's software and hardware systems. And we can answer those questions better because we build Nemotron.

So, you know, ultimately we're excited to open up Nemotron even further and continue to put it out there for the community. We love learning from the community. Nemotron is built in collaboration with the community where we learn a lot from what others are doing in the community and then we try to contribute what we can back. We think that this is a great opportunity for NVIDIA to support the AI industry.

Jon (03:05.98)
Maybe I could also add, know, so Nemotron is the models that we release are a family of large English models, text and multimodal. And we've kind of settled on I think there's like

Noah Kravitz (03:16.411)
John, I apologize for interrupting you. I don't know if something moved, but you sound a little further away from the microphone than you were.

Jon (03:30.008)
closer. I changed nothing.

Bryan Catanzaro (03:31.692)
It sounds better now. It does sound better now.

Noah Kravitz (03:32.741)
Yeah, it does, right? Okay.

Jon (03:34.84)
That's where I moved in like three inches. That's really weird.

Bryan Catanzaro (03:37.27)
It sounded like you were in another country. It was really far.

Noah Kravitz (03:46.171)
Yeah, no, it was clear just kind of far away, but definitely better now. So thank you. Sorry about that.

Jon (03:49.026)

We're, okay. What was I gonna say? yeah, sorry. I'll just start over. Yeah, so Nemotron is a collection of large language models and it's probably worth saying, so they're text models and multimodal LLMs and we've kind of settled on like three sizes or we think of them as weight classes. So we have smaller models that we call nano models. We have...

medium-sized models we call super models and then we have the largest frontier size models which we call ultra. So Nemotron collectively refers to everything Bryan said and then this family of models that were

Noah Kravitz (04:30.713)
And so how does Nemotron fit into NVIDIA's broader AI strategy? Because from what I understand, it's not just, and I say just, the models are huge, but not just the models, but it's kind of a cornerstone for growing the ecosystem.

Jon (04:47.468)
Yeah, well, you know, if you think of NVIDIA as a, as an accelerated computing platform company, and you asked the question, well, what does an accelerated computing platform mean in this age of AI? So it includes chips, it includes networking, it includes software stack, but it also includes the models. And, you know, when we think about what is a platform today, a platform is all of those components. And if you're building AI applications and you care about
the quality of the models, but you also care about the performance. Like Bryan mentioned, one of the reasons that we train Nemotron models is so we can learn. We are pushing the limits ourselves so that we can learn and make sure that our platform is the best. But it also means we can do co-design. can cooperatively design the model architecture, the software stack, and the hardware across the whole, all of the hardware components all together. And we've been doing that. And that gives us opportunities to make things more efficient, lower latency, higher throughput.

more energy efficient by improving things across that entire stack up into the model architecture. so Nemotron is a really important part of that strategy as a accelerated computing platform company where our success comes from this full stack co-design and optimization.

Bryan Catanzaro (06:06.392)
thing I wanted to add to that is that these days there are new things that are part of accelerated computing that maybe people haven't considered. So for example, data sets that we use for pre-training and post-training models have a dramatic effect on how quickly the model converges. In fact, comparing different revisions of our Nemotron pre-training set, we've accelerated pre-training by a factor of 4x.

just by having a smarter pre-training data set, which means that you can actually train a much smarter model with the same amount of compute.

Noah Kravitz (06:44.763)
Yeah. Bryan, if I could interject real quick, what makes the data set, one data set, better, more optimized to help the model converge faster than another?

Bryan Catanzaro (06:58.222)
Well, what we're trying to do with LLMs is build something intelligent that can help us solve problems. It can answer questions. it can reason. It turns out that if you just take all of the texts that humankind or computers have ever produced on the internet and train an LLM on it, that's kind of where the community started many years ago. But it turns out that's not the most intelligent way of building AI because a lot of that text isn't adding very much intelligence. And so.

every organization that builds LLMs spends an enormous amount of effort and compute in understanding their dataset, refining it, rephrasing it using synthetic data generation. And the effort that we put into, these datasets has an enormous impact on how quickly the models train. And also on the overall strength of the model once it's trained. And so these days, I believe that the datasets that we release as part of Nemotron

are an important part of NVIDIA's accelerated computing efforts because it's not really possible to think about how fast a system is for training. If you're training on a data set that's not very smart, it's going to take an enormous amount of compute to get to the same amount of intelligence as if you were training on a data set that was much more polished. so, you know, that's kind of the genius of accelerated computing is that we...

Noah Kravitz (08:09.467)
Right, right, right.

Bryan Catanzaro (08:23.544)
try to understand the problem from first principles and we try to optimize the entire stack end to end. And these days it seems pretty important that NVIDIA's accelerated computing platform includes Nemotron.

Jon (08:41.024)
I can give another example that's interesting at inference time. So reasoning, the way these models reason is they generate thinking tokens, right? You ask it a question and then it generates a lot of tokens as it thinks through the answer. And there's very clear examples

where you can generate a lot of tokens and not actually make a lot of progress towards the answer, or you can be more efficient, generate fewer tokens and make more progress. And again, from the same perspective of accelerated computing, you don't really care

Noah Kravitz (08:43.665)
please.

Jon (09:10.148)
Did it generate 10,000 tokens? Well, you do care. If you can generate the same quality answer in 2,000 tokens instead of 10,000 tokens, that's a 5x speedup. And so that's also part of the accelerated computing story. So all of these are opportunities we have to make things faster.

Noah Kravitz (09:23.462)
Right.

Bryan Catanzaro (09:27.232)
Exactly. Accelerated computing has never just been about how many arithmetic operations per second you can perform. It's really about what capabilities do provide.

Jon (09:37.688)
Yeah. Yeah. And I think the key to NVIDIA's historic success is as a company, we've always focused very much and had deep expertise on the actual end applications people care about. Whether it was computer graphics or high performance computing or deep learning or now modern AI, it's really thinking about what's the end goal and how do you build a platform that gets you that goal with the least amount of time you have to wait, lowest latency, highest throughput.

Noah Kravitz (09:49.073)
Mm-hmm.

Noah Kravitz (10:04.807)
You talked some about the openness, the collaboration and co-design that's such an important piece of this. Open source, a big part of Nemotron. What does it mean to call Nemotron one of the most open artificial AI developed... Sorry, I'm going to restate that. fumbled. What does it mean, and maybe Bryan, I'll ask you this first, but either of you, what does it mean to call Nemotron one of the most open AI development efforts that we've ever seen?

Bryan Catanzaro (10:35.352)

Well, we really think that it's important for AI to be trusted and widely deployed. And in order for that to happen, we think it's important that enterprises have the option to understand the data sets and the technologies behind AI and fine tune them for their own problems and then integrate them very tightly into the...

software and systems that they use to solve problems for their markets. We think it's AI is not a one size fits all solution. And we've seen in the past, you know, many instances of when open platform technologies really allow different industries to differentiate different solutions for the problems they face. You know, for example, the internet as an open technology had really different implications for different industries like healthcare versus retail.

You know, the way that those organizations use the internet to change the work that they do was quite different. But the fact that the internet was an open technology allowed many, companies, many industries to think about solving their problems in a new way using the internet. And when we think about AI, it seems obvious that enterprises need that ability as well. You know, the world's

most important and valuable data always has the most sensitivity about it. And so we think it's important to support enterprises as they learn how to deploy AI, that they can do it in a way that respects their work, their privacy, the important ways that they go about problem solving in sort of a unique way for their business. And so we think that it's really important that there exists

an open foundation for organizations around the world to build and deploy AI. And Nemotron is how we're contributing to that.

Noah Kravitz (12:41.135)
And so, yeah, down please.

Jon (12:41.152)
I can just add one thought to that. From the perspective of accelerated computing, if you think about, you know, we come up with some way to make a chip faster, how does the world consume the benefits of that, you know, acceleration? Well, in the case of a chip, you buy a chip and you get the benefits. But what if we come up with a technique that makes models more efficient at thinking or a dataset mix that saves you time and training? How does the rest of the world

Receive the benefits of that. Like in what form do you package it? I think the answer, the only answer is we have to teach everyone what we did by sharing it through open source, know, open, open weight models, sharing the data sets, explaining how they work, sharing the algorithm. So I think it's natural that open source is a delivery mechanism for the technology that's going into our platform.

Noah Kravitz (13:35.409)
So from a little bit of a hypothetical, I'm an IT leader or a business leader at an organization, and I'm hearing what you guys are saying, and we want to do this. And we have specialized needs in our industry, and we have troves of our data that kind of represent company intelligence and our special way of doing things that has brought us success. And we're ready to embrace the AI age and transform. We could use Nemotron, and I'm going to walk through this.

point me when I get this wrong. We could use Nemotron to take an open source model and we could customize it, train it on our company data and the rest of industry data things to help it understand what we do and the problems we're trying to solve in our industry. Nemotron could help. We could add reasoning capabilities and that sort of thing to it. And then we have a kind of, and I don't want to misuse the term sovereign. And if you guys want to talk about sovereign AI.

But we would then have our own sort of customized, adapted to our business, our industry, the way we do things, our data, and it's ours because we took an open model, we trained it. And so we don't have to worry about the sensitive data being out in some commercial model somewhere or what have you, because it's our model now.

Jon (14:54.902)
I think that's one aspect. Yeah, that's one. I mean, there's many aspects. So, so, so for example, if you say, you know, NVIDIA trains a model, a Neimotron model, and it's great. But since you've disclosed all your training data and look at your training data, for whatever reason, we have some policies for this data we can't use. And we can say, that's fine. Everything you need to reproduce what we did is there. You can train your own model, excluding that data. Or you say, well, I like the data, but the mix is wrong. I don't know. I'm a sovereign project.

Noah Kravitz (14:55.975)
Close?

Jon (15:25.0)
And it really needs to be very good at speaking this language and understanding this culture. And that data wasn't as represented in your training set as I want it to be.

Everything that we did is transparent. And so you can make these modifications yourself. I mean, that's one aspect.

Noah Kravitz (15:38.161)
Fantastic, yeah, right, right.

NVIDIA has released datasets, recipes, alignment techniques alongside the models. So along these same lines of building trust and transparency, why is all of that important? Why is this full level of transparency important for the end users to be able to customize and deploy safely?

Bryan, you wanna start with this one?

Bryan Catanzaro (16:08.824)
Well, I think ultimately, if you don't know what's in a technology, it's harder to trust it. And every business has different ways of thinking about the problems they're solving. They have different problems. And I think it's important as we get more sophisticated about deploying AI and we integrate it more tightly into business problems around the world for businesses to be able to.

to inspect how was this AI built and therefore I can build trust that it's going to help my business solve problems. I think also the integration is a really important point as well. So with Nemotron models, there's a really broad spectrum of integration. You can run it locally on a machine without any internet. You could also run it through an API in the cloud.

and everything in between, you can deal with your business's sensitive data using the same data management and security protocols that your business already has. And I think for a lot of applications of AI, that level of customizability and introspection is going to be essential.

I also want to say that I think there's a real big benefit to open technologies in the sense that they tend to develop faster. So, NVIDIA believes that helping AI grow creates opportunity for us. And we think that one of the best ways of helping AI grow is to contribute in an open way to the community. I think when you consider a technology that's being developed kind of

independently by a few different organizations, but they're not able to share very much about what they're doing, there's obviously going to be a lot of reinvention that has to happen and the progress is going to be slower. And so if we are able as a community to come together, you know, to contribute ideas, data, models to each other and learn from each other, I think that we'll progress faster.

Noah Kravitz (18:04.507)
Mm-hmm.

Bryan Catanzaro (18:25.922)
And we've seen that over the past couple of years as various organizations have been contributing to the open technologies for AI. It's really helped the community move forward. like, for example, OpenAI just released GPT-OSS. That was a fantastic thing for the field. Alibaba has been doing some great work with QN models, obviously Meta's.

Noah Kravitz (18:45.745)
Mm-hmm.

Noah Kravitz (18:50.385)
right?

Bryan Catanzaro (18:51.342)
family of llama technologies has been extraordinarily helpful to the field to help the field grow and develop. at NVIDIA, we know that when AI grows, it's opportunity for everyone. It's opportunity for businesses that they can solve new problems and it's opportunity for us because we work with every business that's building AI.

Jon (19:14.912)
Yeah, I mean, a good example of that playing out is our own research groups often will use, like if you have some idea for a way to improve a model, we often will just take one of the existing open weight models, not necessarily Neemotron, that sort of gives you the best vehicle for trying out your idea, improve it in some way, and publish a paper, release the result, right? So like we are building on all the work from these other organizations that release open weight models all the time as well.

Noah Kravitz (19:30.95)
Right.

Noah Kravitz (19:41.479)
And that's, you this is no news to you guys or probably many listeners of the show, but that same sentiment has been echoed so many times over, I mean, over the past couple of years in particular, by guests we've had from all industries and walks of research and life. And you know that the more we're collaborating, the faster we move as a whole. Our guests today

are Bryan Cottonzaro and Jonathan Cohen. They're both from NVIDIA. Bryan is vice president of applied deep learning research.

while Jonathan Serfs is vice president of applied research. And they're here talking to us about NVIDIA Nemotron, family of open models and open technology. We've been talking about the importance of open, open technologies to the AI community in general, to NVIDIA, the learning that goes into informing really the whole stack, the hardware, the models, the software, the connectivity, networking, everything. And the datasets, as Bryan was talking about,

and how it all really comes together to make things advance faster and more efficiently, sort of broadly speaking. Bryan, Neuatron has been a huge effort within NVIDIA, it still is, with many teams working together to bring this to life. The advanced research, I'm sorry, let me rephrase that. Neuatron has been a huge effort at NVIDIA with many teams working together, they still are, to bring this to life, from advanced research to commercially licensed models and data sets now.

Can you guys talk about the pipeline from research to production models, what that's like, what it's been like for Neumetron?

Bryan Catanzaro (21:18.094)
Well, it is a huge effort and it takes a lot of people with different talents coming together to build Neemotron. We've organized the project around basically the different stages of development that a model has to go through, pre-training, post-training alignment and so forth, as well as different functional areas, like for example, long context recall or image understanding.

Noah Kravitz (21:46.919)
Right, right.

Bryan Catanzaro (21:47.4)
And so within each of these areas, we have multiple teams working together, some of which are very researchy, very theoretical, and others are very engineering focused and then a whole spectrum in between. I would say it's a great honor to be part of a project where people are coming together to build something like this. It's also a big challenge, you know, trying to get

Jon (22:09.675)
It's so funny.

Noah Kravitz (22:13.575)
Sure.

Bryan Catanzaro (22:15.146)
So many brilliant minds pointed in the same direction. I think that's one of the central challenges facing every AI development effort around the industry these days is how do we work together to build one amazing thing as opposed to building a hundred small things. And that's really something that's been inspiring to watch come together.

Noah Kravitz (22:30.341)
Yeah.

Jon (22:36.3)
Yeah, if you compare it with like a large scale software effort, know, there's this like famous observation called Conway's law, which is the, the communication patterns that are observed within a piece of software tend to mirror the organ, the communication patterns of the organization, organizational structure that build that software. and training a model is like, I mean, Conway's laws is definitely an issue, but it's just a very different endeavor.

It's not like I build a module and you build a module and we have a nice clean interface. Somehow all of these things have to get combined together. know, image, Bryan's example, image understanding and long context recall somehow all of it combined together into a single training recipe and a single dataset mix. and so the modularity is, I think less than in software engineering. so the

Noah Kravitz (23:26.982)
Yeah.

Jon (23:30.198)
This idea that you can just decompose it and have lots of teams with sort of clean interfaces between them doesn't really work as well. And so I think there's a real struggle in scaling up an effort like this to a very large team to do something really big.

Noah Kravitz (23:41.699)
Is there a new paradigm emerging of organizing? Yeah.

Jon (23:44.376)
It's interesting question. I wonder, you over the next five, 10 years, there'll be some new law named after someone and some sort of new, you know, management principle here. It's an interesting thing that we've certainly been thinking about. But it does present these

challenges. I think one of the most important principles that we've kind of settled on is you just need a lot of internal openness and transparency. You have to solicit ideas. a lot of people across the company and outside of the company working on all these problems.

You have to solicit all these ideas and you have to encourage them all to work together. That's the only way forward. And so that just takes a very like mature culture and, you know, good leadership and ego lists, you know, operation and everyone being really motivated by the, at the end of the day by the work.

Noah Kravitz (24:19.718)
Yeah.

Bryan Catanzaro (24:35.818)
say also that one of the amazing things about AI is that it's such a general technology that it really changes the way that we do AI. It used to be like 20 years ago when I was a grad student that it was common for people to build state of the art models in computer vision.

on their own, like one graduate student on their own could build a model that's that was state of the art in some important area of computer vision. And, you know, that's kind of how we were trained as PhD students is like, go be brilliant on your own. Well, with with modern AI, the best results come from using industrial scale equipment and, you know, general models that can then be taught how to solve important problems.

Noah Kravitz (25:15.27)
Mm-hmm.

Bryan Catanzaro (25:32.826)
But that requires working together. so one of the first things that AI has changed is the development of AI itself and organizations that can figure out how to collaborate and work together succeed. And, you know, that's one of the reasons also that we really believe in Nemotron as an open project is because we've seen how openness internally has made it possible for us to solve whole classes of new problems with AI. We believe that

as Nemotron and other open efforts come together, bring together more ideas and more force to bear on the development of AI that the results will be stronger.

Noah Kravitz (26:15.707)
Jonathan, NVIDIA has a history of building end-to-end products. Self-driving comes to mind, gaming, course, super pods. But then disaggregating them for the world to use. Does Nemotron follow that same pattern in your mind, and if so, how?

Jon (26:31.798)
Yeah, I think so. think when we talk about that, and Jensen talks about this a lot, what we mean is our solution, the things ultimately that we build are very complicated integrated systems with many layers and many components. And on the one hand, we need to build the whole thing ourselves, because it doesn't work unless you build the whole thing yourself. So we need to train a whole model at the end of the day. It doesn't make sense for us to release, I don't know, a way to make

a reasoning recipe without actually training a model to do reasoning. You know, like you have to do these things and put the whole thing together. But at the same time, I think it's very important that we put, put all of the components into the ecosystem and allow people to consume the parts that they want and not consume the parts that they don't want. So this is how our hardware is. You know, we, we design data center scale computers at this point, but we don't sell it as a single data center. We,

Noah Kravitz (27:05.969)
Sure.

Jon (27:30.412)
design the whole thing, we build the whole thing, then we chop it up into pieces and we sell it through normal sales channels and people, our customers are free to take the parts they want, replace, it's truly an ecosystem. If you don't like the way, you don't like our CPU, use a different CPU. You don't like the storage, use a different storage. You don't like this networking, use a different networking. And we're open and interoperable with all these things. And it's a tremendous engineering challenge to work that way. But I think it's why...

we've been so successful is because it allows us to harness the power of like the entire computing industry because we're not really locking anyone out at all. We're including everybody. And so when we think about large language models, I guess we're thinking in the same way. So we're going to develop techniques and anyone is free to take them. Other companies that train large language models for a living are free to take anything we built. They probably won't take all of it, but they're free to take anything. They want to take the software. That's great. They want to take some of our data sets.

That's great. They want to take the software and the data sets and some of the training recipes, but modify them. That's great. They want to take the finished models. That's great. So, so in that sense, think philosophically, that's, that's absolutely how we think about products, how we think about hardware, how we think about software. And it's, now how we also think about foundation models.

Bryan Catanzaro (28:45.634)
And I think that's one of the things that makes NVIDIA unique as a big tech company is that although we do full stack and integration, we don't dictate to our customers how that technology is going to be deployed or used. We know that it's not a one size fits all problem. Right. And so we're, we're happy to support companies of all shapes and sizes in every industry.

Jon (29:04.736)
or even assembled.

Bryan Catanzaro (29:14.126)
develop and deploy AI. And because NVIDIA has this orientation, the supportive orientation, where we understand that it's not one size fits all, that actually is the secret to why we able to collaborate with all of these companies. And we want to do that with AI technology as well.

Noah Kravitz (29:30.619)
Right, right, right.

Noah Kravitz (29:35.911)
Brent, kind of switching gears a little bit, but still talking along technical lines, can you share any exciting technical breakthroughs that came about during the Neumetron development process and what they might mean going forward, specifically in terms of efficiency and deployment, but really take it as broad as you like?

Bryan Catanzaro (29:58.019)
Yeah, well, NVIDIA is thinking about AI from an accelerated computing perspective. And we have a belief that the faster we can make a model, the smarter it's going to be. And this follows just because clearly, if we're able to think quicker, then we can get more thoughts in the same amount of time that should help us solve problems. So we're bringing this perspective of accelerated computing to AI in kind of a unique way.

Noah Kravitz (30:16.593)
More thoughts, yeah.

Bryan Catanzaro (30:26.51)
A couple things just from the past few months that we've demonstrated that I'm really excited about. One is we released a model we call it Nemotron Nano V2. It is a hybrid state space model. So it's not a pure transformer model, but it uses this other technology for

reasoning over sequences called a state space model that has some pretty big efficiency benefits. You know, on the same hardware,

compared with other models of the same intelligence, we're about six to 20 times faster. And we're pretty excited about the capabilities of this model, but it's just the beginning. We have really ambitious plans to continue evolving the architectures behind Nemotron, as well as the systems that are used to build and deploy it.

Noah Kravitz (31:08.849)
Sure.

Bryan Catanzaro (31:18.894)
Another thing that we were able to show recently is we trained a nemotron model using four bit floating point arithmetic, and we're able to get world-class results, which is really exciting because using only four bits per parameter of the neural network can be dramatically more energy efficient than using other representations. And we know that the development of AI is

Noah Kravitz (31:42.02)
Excellent.

Bryan Catanzaro (31:48.313)
going to be constrained by the efficiency with which we can train it and deploy it. so showing people new algorithms that are more efficient then is going to help push the industry forward. it's not enough to say, hey, I've got the system. It's really fast at low precision arithmetic if no one understands how to use it. So Nemotron is our way of demonstrating to the community, like, hey, you can take advantage of this amazing

low precision hardware to train a world-class model if you follow this algorithm. So that's the number.

Noah Kravitz (32:21.191)
Right, right.

Jon (32:21.962)
And technically, it's amazing that four bits is enough. Like, if you just think about how little resolution that is, the fact that that works is pretty cool.

Noah Kravitz (32:28.935)
Can you, is it?

Noah Kravitz (32:32.935)
Am I, I'm willing to show my ignorance here, but I don't know if I'm opening up a deep can of worms, but can you unpack maybe for folks listening who like myself don't quite get the full implications of that? Forbid, forbid, sorry, let me rephrase. Actually, no, I'm just gonna stop there. For folks who don't, so maybe can you rephrase for folks who might be listening and myself included who don't fully get the ramifications.

of what doing 4-bit arithmetic and these results really mean.

Bryan Catanzaro (33:07.918)
Well, one fun analogy from my childhood comes from video games. don't know if you remember the eight bit Nintendo system. And then there was the 16 bit Nintendo system and it was like, wow, there's so many more colors with the 16 bit Nintendo. It's like, wow, the, you know, look at that smooth gradient. Right. So if you, if you only have eight bits, you can represent 256 numbers with 16 bits, you can represent about 65,000 numbers.

Noah Kravitz (33:14.779)
That's, mean, yes, of course.

Bryan Catanzaro (33:35.757)
With four bits, you can represent 16. So it's a very, very small amount of options to pick from. Like if you're gonna draw a picture using four bit numbers, it's actually gonna be pretty hard to make it look smooth. And of course, what we're doing with our four bit training hardware and software isn't as straightforward as just using exactly.

one of 16 numbers for every parameter in the neural net. They actually come in blocks. The blocks have scaling factors attached to them in hierarchical ways. that's all accelerated by software and hardware that we've built in transformer engine and in our Blackwell GPU generation. And so it's kind of amazing that we're able to take this raw material that's very coarse and rather small.

and we're able to make it flexible enough to train a world-class neural network. And then, of course,

Jon (34:37.176)
But at some level, I always like to think of this as like, you can have any number you want as long as it's one of these 16. And somehow, you know, it still works. It is pretty miraculous.

Noah Kravitz (34:48.295)

Amazing. As we wrap up the conversation, but look ahead to the future of Nemotron, what can developers and enterprises expect next? You've talked about it a little bit, some of the things coming through the pipeline and that you're working on, but what can devs and enterprises expect from Nemotron? And perhaps more importantly, how can they start to engage with Nemotron right now?

Jon (35:14.072)
Well, I can just say, you should expect us to train some big models. We've trained recently some smaller models. We'll be training some bigger models. You can expect us to incorporate more multimodal technology. From NVIDIA, we have some of the world's best, well, I guess the world's best open-way speech recognition models at this point. That technology hasn't really been incorporated into Neamotron, and we're working towards adding audio and these kinds of capabilities. So I think there's a lot of...

which is really cool technology we're working on, really bringing all of the best technology across NVIDIA and concentrating in Nemotron. I think that's something people can look forward to. I know, Bryan, what you would say.

Bryan Catanzaro (35:57.315)
Well, yeah, I would also reinforce how important reasoning is to nemotron. It's been a core part of nemotron development for the last year, and we were super proud that we were able, for example, to take nemotron reasoning and add it to Meta's llama family. We know that there's a lot more work to do to make reasoning even stronger, and we're really excited to do that.

Noah Kravitz (36:21.799)
I can't let a podcast go without making one of these kinds of jokes, but I'm working on making my own reasoning stronger every day. So I can only imagine the effort y'all are undertaking. absolutely. Bryan, John, this has been great. Really informative conversation, but just to hear the two of you talk about Nemotron from the inside out, just a treat. For listeners who want to know more,

Where is the best starting point? Somewhere on the NVIDIA website, social media, where's the place to go?

Noah Kravitz (36:57.179)
I caught you off guard. I'm sorry, we always do this. We need to put this in our template and it's my fault for not doing it. I'm so sorry. We always ask at the end for like a, know, yeah. Yes. Okay, great. So I can rephrase and just, and I'll tee it up with that. Okay.

Jon (36:57.206)
Bryan Catanzaro (36:58.872)
We do.

Jon (37:01.919)
actually

Jon (37:07.586)
Where should we go? Well, I guess I'm gonna sit and download the models from Huggingface. Okay, so let me see that.

Okay.

Jon (37:19.532)
Or I can just talk.

Noah Kravitz (37:19.815)
So for folks who are listening and want to get started with Nemotron, the models are available now.

Jon (37:29.496)
Yeah, so our models are available on Hugging Face. You can download them. You can also experience all of them on build.NVIDIA.com and download them there as well. There's also a number of blog posts and papers. I know, Bryan, what's the best way for people to find all of that?

Noah Kravitz (37:32.999)
Perfect.

Noah Kravitz (37:40.569)
Excellent.

Bryan Catanzaro (37:48.652)
do have a landing page on NVIDIA.com for Neumotron and we're busy filling it out right now, gathering all of the Neumotron content together in one place. So I would go there.

Noah Kravitz (38:00.005)
Excellent. And work in progress, I'm sure, the content, like the technology itself, evolves and evolves. Again, John, Bryan, both of you, I know tremendous amount on your plate with Neematron and everything else, so we appreciate the hour to come on and help shout from

the rooftops, tell the world about all the fantastic work you and your teams have been doing. Congratulations and all the best going forward. As you said, not just inside of NVIDIA, but collaborating with the community and working too.

Raise all the boats together.

Jon (38:32.898)
Thanks for having me.

Bryan Catanzaro (38:33.39)
Thanks Noah.