Build Small Hackathon

61 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/LIVE/7OTGEJXAILY?SI=WYTTC5DV_6TODXMX
https://www.youtube.com/live/7otgeJXailY?si=wyttC5dv_6TODXMx

SUMMARY

The kickoff event for the "Build Small" hackathon emphasized the importance of developing smaller AI models, encouraging participants to create practical and imaginative applications using models with fewer than 32 billion parameters. The event featured various sponsors who provided resources, guidance, and prizes to support participants in their projects.

- *The hackathon has two main tracks: "Backyard AI" for practical applications and "Thousand Token Wood" for creative projects.*
- *Participants can use models with a maximum of 32 billion parameters, with specific submission requirements including a demo video and social media post.*
- *Over $48,000 in cash prizes and additional model credits are available, with multiple categories for winning.*
- *Sponsors include OpenAI, Nvidia, and others, each offering unique models and resources for participants.*
- *Participants receive credits for using various platforms, including $250 in Modal credits and access to OpenAI's Codex.*
- *The event encourages collaboration and community engagement through Discord support channels.*
- *The hackathon aims to revive interest in smaller, fine-tunable models amidst the trend of large-scale AI models.*

Yeah. >> Okay, we should be live. >> We should be live. All right. Awesome. Hey everyone, I think you guys can see me. If you guys can hear me. Uh those of you who are joining us on YouTube, can you share on the comments like are you able to hear me?

Awesome. Thank you so much. Yeah. Um all right, let's get started. uh we are very very excited to kick off this hackathon. So this hackathon so the goal of this kickoff is to you know talk about the the highle detail of the hackathon and also we will try to go into some low-level details as well. uh we will be joined by our sponsors uh who have made this hackathon possible and uh I'm really really excited for you know all of us to be sharing uh this moment then stage um before we get started um can everyone tell me like where they are joining from which city or which country they are joining from maybe you can wait for one more minute and then I'll just start right Awesome.

So we are having some folks from India. People are joining from UK, London, Ghana. Oh, nice. Brasov, Romania, Canada, India. Amazing. Amazing. Thank you so much. Uh thank you so much for you know giving us time and we are very excited. So yeah let's get started. So um the hackathon uh that we are uh spons that we are you know conducting that that's that has already started it's called build small and the uh we are very very thankful to sponsors who have you know made it possible and very soon you will be listening to them as well and what they are you know what they want to talk about their technology and what they are bringing to the

hackathon and uh thank you so much for everybody joining and thank you to our sponsors. So what exactly is build small? So we thought that right now the whole AI scene is dedicated to or you know dominated by large API inference you know providers and all these large models. So we thought let's take it back to when to the time when it was all models were tinkerable you can fine-tune them you can you know have lots of fun with them.

So we just wanted to bring that whole era back and that's what you know had us started on this uh on this hackathon journey and planning journey and everything else. So how we have planned this hackathon is that we have created two tracks. So these two tracks are uh the first track is uh called backyard AI and uh so backyard AI means uh that we expect you to build something that is uh maybe you know useful to you or useful to somebody that you know and the idea is that you know building something not just for the sake of it but building something that is useful and uh you know maybe makes somebody's life somebody who's dearest to you. their life you know a bit better tiny bit better then the second track that we have is called thousand token wood so it's uh it's uh it's the name itself is whimsical in nature and the way we have named it we are planning to you know excite the community in building uh something that's creative it's a little bit out there and you know it's uh it's it's it's a wacky and and then things like that so we are really excited about these two themes and hope you guys all are too.

And uh so so one example for the backyard AI could be you know a storybook generator that your kid can use and maybe a thousand token wood could be an interactive game. So you know you can use your imagination. These are just two examples. So there are three broad rules to this hackathon and the first one is that we are limiting the model use to 32 billion parameters. So that means that you can use any model or any combination of models but every model should be below 32 billion parameters like for example to the extreme you can use multiple models which are bel below 32 billion parameters that should be fine and uh and uh so if we talk about for example the uh you know kind of models the total model uh total billion total parameter count should be 32 billion and not the active parameters. So just putting these things out there for to you know remove confusion and then uh the submission would be a gradio app and you have to submit it under the hugging face uh org that's the hackathon or and uh uh the SDK can be uh you know grad it can be uh uh it can be docker as well but underneath that it has to be a gradio space and uh what we are also expecting the third thing that is most important about the you know submission is that we are expecting a demo video.

Demo video is important because that will help us you know understand what your app is able to do even if you know we are not able to run it uh uh at the time of evaluation maybe due to due to GPU limit you know limitation or API exhaustion or something else and we want you to create a social post as well. So when you create a so you have to do you have to make sure three things one that your app is using models that are below 32 million parameters and then second that you have to submit it under the org so you have to join the org and the third thing is that you have to submit a video and a social post proof. You can post your video and your social post in a link to the social post in the space readme. Um so there are lots of amazing prizes that we are uh able to offer you thanks to our uh you know generous sponsors. Uh so on offer we have around $48,000 cash prizes and then uh we have 20,000 model credits. Uh and then uh Nvidia has you know uh agreed to sponsor us with two RTX 580s and then we have J GBD Pro for one year and so you have around 29 ways to win all these. So what I mean by that is you have 29 different categories in which you can win these prizes. So there's lot to

win and there is a lot of opportunity for everything. And uh to make this uh thing work, what we did is we started with these uh hugging face credits uh $20 credits that gives you zero GPU access and that gives you inference provider access and uh so since all of you are joining us in this hackathon or and it has a team status what that means is that everybody by default gets the zero GPU u accessibility. you can create zero GPU spaces and you can use zeroGPU uh for around 40 minutes per day. So I think that's much more than enough that you need for you know experimentation and building and hosting your apps and you can build up to 10 zerog apps. So if somebody has a question that you know how many apps can you submit to the hackathon you can submit as many apps that you want but if you're using zero GPU then you will be limited to 10 zero GPU spaces and then we have uh model has supported us with $250 uh model credits for every uh every participant that has joined us that has registered with us and uh openai has also generously given us you know $100 for first thousand registrations So we have a constant support available on discord and we will be we are planning to do these AM kind of a sessions with our sponsors uh most likely on June 9th, 10th and 11th. So stay tuned for further you know communication.

Um we how we have structured this whole thing is we have these bonus quests and what that means is that you have to tap as many as you can from these six like it has to be off-grid that means that it it doesn't have to be off-grid but if it is off-grid and what off-grid means that you're not using APIs but you're using a model that is running locally for you in the space then that makes you eligible for the badge that is off the grid and similarly you can understand the rest of the badges. So we have different kind of prizes which you know which centered around these badges. So we have that information available on the uh on the landing page. So feel free to read that. I will skip ahead right now and then uh um just one second.

Yeah. Yeah. And um yeah, I think uh yeah, I think we are uh that's it from my side. Uh Hannah, are you around? I can give you the control. All right, folks. Uh I think Hannah will be talking about uh something new that we are building at uh Gradio and we are very excited about it. Let me hand over the Google to Hannah. Uh >> yeah, you can just show the next. >> All right. Yeah, sure.

Yeah. Uh yes. So something we've been working on in the past couple of months which is really exciting is due. Workflow. Um so now we can now build really cool AI pipelines without code. Um and it's all built on top of radio. Um so what you can do is now drag and drop nodes on a canvas to chain together image generators, voice models, summarizers, um etc. which is really exciting. Um there's lots of cool spaghetti UI um tools out there and I think this is like a long time coming that I think we're all really excited about um launching.

Um so you can plug in any hugging face model or data set and you can browse by category um like image, audio, video, uh 3D text and you can also plug in your data sets. Um you can also filter to zero GPU spaces um and also search with semantic search which you can do on the hugging face hub. Um, but like more complex workflows, you can also mix in your own custom Python functions. Um, and once you've built uh any kind of workflow, you can share or hand it off as a Gradio app and save it to the hub.

Um, you can save it to uh it'll be saved as a workflow uh file and you can launch it to like any other Gradio app. It's again it's just another layer on top of radio. Um, and others can use your workflow in your spaces too. Um, so

we're really excited to see what you might build with it. We're very much open to any creative submissions around it. It's it's very much in its betas launch. Um, and we're still working on it. It's got some um little things to work on still. But um, if you do uh have any suggestions or or see any bugs, then please raise any issues. Um, and we'll work on them as quickly as possible. Um, but yeah, so go crazy with it. Let us know what you build with it. Um, and we'll also have some more documentation and tutorials to help get you started with it very soon. Um, so yeah, hopefully you enjoy using that. Um, and it'll be released today.

>> Awesome. Um, thank you so much Hannah for that. Uh, uh, I'll now, uh, start, you know, handing it over to our sponsors and, uh, maybe we can start with, uh, uh, let me see this one second.

Yeah, I'll call uh Steven onto the stage. >> Uh yeah. Hey. Hi, Stephen. How are you doing? >> Hey, very good in you. Thank you. >> Yeah, >> thank you for having me be with us and yeah, stage is all yours. Take it away. Perfect. I was not supposed to go first. So, let me just make sure everything is ready. Um, so you can share everything. That would be this one.

And then I'll be good to head over. By the way, I'm going to intro myself quickly. But I am Stefan Batifo. I am developer relations for Black Forest Labs and we are one of the sponsors for this event. I'm actually very happy about it. Uh I found the event and I thought it was a really cool idea to do something that is, you know, all about small models because lately everything about has been about, you know, really bigger and bigger models. So I'm just going to walk you through our BFL tiny bits than the models that I'm suggesting you're using for this event. Um I also wrote some guides so you can get the best out of the models and then I'll let you have fun with it. So yeah, I as I said I'm Stefan Ode here and for the people that don't know BFL, so Black Friday labs, we are frontier visual AI research lab. Um our models, you know, we focus a lot on open source and open weights. Our models in total have more than 400 million downloads on hugging face. Um and we're a bit more than a year old now. We're like 85 people 90 people. um a lot of researchers and really what we want is always to train your model to push the frontiers of visual intelligence and AI um and that's what we're working on.

if you know stable diffusion. Our founders come from there. Um they also invented latent diffusion back then. Um and since then you know I've been working in the field and then they decided to create black forest labs a bit more than a year ago as I've said but I'm not here to talk about the company but I'm here to talk you about the models. Uh for this one we have flux to client which we released in January now and this one it's an open weight open source models and there are different models for this one. We have two versions. We have the 4 billion version and the 9 billions version. And for the events, you can use both. Um during this talk, I'm going to focus more on the 4B version just because it's really really small and really would allow you to do like many things during um during the hackathons.

So, as I've said, we have two sizes and both of them can do text to image as you can see on the bottom left here and then they can also both do editing. So, you can edit an image, you can do, you know, you can start with the texture image, get an image and then you're not happy with it or you want to change the background or you want to change anything in it, you can just edit it directly. And then you can also have like different styles as well. You know, we can see a bit of a glimpse here where we have like more anime style, more artistic, more, you

know, like the person that we have with a broccoli. Uh, you can basically decide everything you want. Then what I want you to pay attention to is that we also have different versions of those models. So we have the base models, which is the one you want to use if you're going to train Lauras. So if you're going to fine tune the models, this is the one you want to use because they are the best one for it. Those are the ones you can shape the best uh when you train them.

And then we have the distilled version which is a way faster version which is the one you should use when you want to then you know make an inference and check your lauras and see if it's working and deploy it to production. And yeah why it's so nice is because four billion models four billion parameters sorry for model is very very small nowadays. Um so you can basically run it everywhere. You can run it on consumer hardware. You can run it on the zero zero GPU from hugging face. Uh I'll show you right after, but I have a space where you can play with it and the models are hosted there directly.

I also for this I wrote a guide actually um teaching you and telling you how to fine-tune flux 2. Um I invite you to look at it. Um it's here. You can find it in the blog. Uh we will share the link as well very likely. I will share the link after my talk but it's really telling you how to fine-tune the models you know how to fine-tune something you know how to build the data set how to fine-tune a text to image Laura and then also how to fine-tune an image editing Laura because you can also do those um which you can see here for example which is like an ugly data set um so here on the left you have the input picture and then I add a Laura on top of this and so then I get this output then uh from it.

So really have a look at it. Uh you will get to know how to use it. I'm also writing you know the different tools you may use uh to do loras. Then you will have to create a data set as well if you don't have one. Again it's written. You don't need that many images. It's between like 15 to 40 depending on what you're doing. Like 15 to 20 if you want to teach style and a bit more if you want to do an editing.

Um again you know read the docs. um and read the guide and you'll be able to do everything. It's also fairly fast. We can like you know it can be done in an hour on an air 490 4090 which is a small GPU. If you have a bigger one you know it will be faster obviously and you can do different loras. So this one is a style Laura which is you know you teach this style like this pixel art sprite style and on top of this you know you can do the editing Laura that I've shown before. Um and again this is just an example you know you can do many different style. Um you can see here as well you know it's really learning the editing and not learning the subject.

This Laura has actually been trained on animals and it was mostly like the data set was mostly animals and then we made them ugly but then it still works you know for like a city or something like this. We have a space uh which you find the link the link here and I'm just going to show you because I'm going to run out of time quite soon. This is what it looks like. Um, you can, you know, do like a text to image generation and then you can also do something like an image edit. So, if I show you, uh, actually, let me use my webcam.

It's working. It's not working. Uh, but it's fine. I can do it here. So, you can see here I did it before. You can just use my webcam, take a picture, and then transform it. Um and when you do then you know it will run on the

zero GPU and then you'll be happy and you'll have everything. You can also load your own loras. This is something that I have. I also wrote down on how to train the Laura there as well. So you know to make sure that everything that you need to know is written and yeah that should be everything. Have fun for the hackathon.

I'll be hanging out on Discord as well. You can find basically all the links here. guides, the model, you know, the starter space as well. Have a look at our docs and then if you want to train something, have a look at the AI toolkit, which is likely the one uh that is the most convenient for you. That' be it. Thank you. >> Thanks a lot, uh Stephen. That was amazing. Um, so, uh, Black Forest Lab has very generously supported us, uh, with a cash prize pool of $5,000 and, uh, they have, uh, created a nice guide and I'll link it on the landing page and, you know, feel free to use it and it's a very nice starting guide, getting started guide for, you know, building with the client and, uh, gradu, you know, submitting something to the hackathon. Uh, all right. Thanks. Uh I'll next move to uh uh the open BNB team. We have Zong with us. I'll add him to the stage.

Hey Zong, how are you doing? >> Hi UV. Hi everyone. Uh my name is Chong and I'm from the OpenBM team and I will quickly introduce the team and also show you guys the uh eligible models for this competition. And I will share my screen right now. Yeah. So uh first let me quickly introduce uh our team. Uh so we are open BMBB uh which is a open source community co-ounded by the THU NLP and modelbest and we are dedicated to advance the development of the LM models on devices and hardware and we have the open source uh mini CPM family which is a lightweight high performance ondevice LMS and spanning from language multimodel and edge devel uh deployment and we have a comprehensive tool chain for model compression, quantization and inference optimization. So these enable those LM to run efficiently on the res uh resource constraint hardware such as smartphone, IoT devices and robots. And a little bit more about our community is that uh we are founded by the uh THU NLP which is Chinua NLP lab uh which is China's uh earliest research group for NLP and large language models and we have over 200 papers at top conferences and for uh uh 44,000 citations and we are also sponsored by the modelbas which is a worldleading team in efficient ondevice models and we have customers in law industries, car industries and IoT industries and our open source community has uh 139K GitHub stars and over 32 a million total downloads and uh so for the mini CPM model family uh we have models in the text area which is the mini CPM series and we have uh the mini CPM M O and V which are the omni and vision models. And we also have like a TTS model called the walk CPM.

And for agents we have agent CPM, ultra uh rack and the chat dev. And for tools we have tools in evaluation, training and data. So for this event any model under 32 billion parameters is eligible for this competition and we have some recommended model for each category. Uh so first is for text uh the text model we have recently released the mini 51B model uh which is a lightweight and fast text model perfect for local first apps and this model is only 1B parameter and it ranked uh number one on the artificial analysis intelligence index uh for tiny models and it's great for like building a local personal assistant a study tutor or a like email helper.

And for our release, we also showed a desktop pet for this model. And something small and great for the desktop also works for this model. And we have an older model which is the mini CPAM 4.18B. And this is a stronger text reasoning and problem solving uh model uh for apps that needs like deep thinking. And it's great for like reasoning assistant uh study s and uh planning assistant. And also we have the uh the TTS models which

is a end to end voice generation uh with TTS voice cloning and creative audio. Uh so you can clone like a person's voice with this model and generate speech. So for example, you can clone like a famous person's voice with this model and build an app with it.

And uh we have recommended model for vision and multimodel as well. Uh so we have the mean CPMV 4.6. Uh this is a 1.3b parameter model. Uh and it's better for uh image understanding OCR and uh document assistant and video understanding. So you can build something like a receipt uh bill parser homework image tutor and shop manual assistant. And we also have the mean CPM O4.5 which is a omni model. It's the world first full duplex uh omnimodel uh model. So it can continually see listen and speak naturally. So for example, when it's talking, you can interrupt the model within that conversation and it it can respond uh instantaneously uh to your speech.

And a little bit more about the project ideas and price distribution. Uh so we are sponsoring uh $10,000 in total uh for for two tracks together. And the first track is Spark uh backyard AI. Uh so for this one, you build something that solve a real world problem for someone and small practical uh apps wins this one. And we have a total prize pool of $5,000. Uh so the first uh the winner will get 2500 and the second place will get 1,500 and third will get a,000. And the other track is the,000 token wood.

And you can build something playful, strange, uh delightful, and AI native with this uh track. And we will also offer $5,000 for this pool. And it's the same price distribution for first, second, and third. Yeah. And for this event, we are also offering free APIs on our end. And here are the API address and the authorization token. And we have the API for Min CPM 4.18B and the min CPM v4.5 and the minpm v4.6 six the two version the reasoning and the non-reasoning version and uh on the right side is a sample usage of how you can utilize this API and uh for model deployment uh you can use vam or llama CPP and for vlam it's best for GPU serving and has high throughput and open AI compatible APIs and for llama CP it's best for local first quant uh quantitized model and laptop and edge devices. And below are some sample code that you can use.

And here's also a quick start for transformer. Uh for example, like you can deploy the uh MCPM v4.6 image as well as the mean CPM uh 51B uh both like using the transformer uh architecture. And this is a sample template of how to build a gradual app. And for the quick start, you can pick like a concrete use case. And you can choose a matching uh mcpm model for this case. And you can build a minimal gradial app and deploy on the hacking face.

And uh you can submit your uh space link. And uh to eligible for the reward uh you have to use the mcpm model as a central part for the app and thank you very much and uh hope everyone uh get started on your mini CPM journey and build something fun together. Yeah. Thank you. >> Thank you so much Zong. And uh just for everybody's knowledge uh so open BMBB has sponsored us with uh $10,000 US and that means that you know uh there are separate tracks for OpenBM and uh to be eligible to winning uh in this category you will have to build with uh mini CPM models. Uh more details are available on the landing page and u yeah uh thank you so much Zong. I'll uh move over to uh VB. Thank you Zong. Okay, thank you.

>> Let me add VB quickly. >> Hey, hey, >> hey, VB. How's it going? How you doing? >> Pretty good. I'm going to try and share my screen. Um, >> cool. Um, so I guess I have like about um 5 to 10 minutes. First of all, thank you so much to the Hugging Face team for um putting this together. I'm quite excited about um about this like about the hackathon and what everyone builds. Um quick intro. I'm VB. I'm part of the uh developer experience team uh here at OpenAI. I predominantly work on uh Codeex as well as our OpenAI API platform. Um and uh the reason why I'm here is because um OpenAI is supporting this um hackathon and you um to build with codeex. And so as part of this we would be supplying um about a thousand um codeex credits that you can utilize to build during the during the hackathon as well as beyond the hackathon as well.

Um we'll also have a bunch of prizes. Um I would let Yuvie cover that with the end. Um, perfect. So, this is just like I put together like four or five sites just to just give like a quick primer on Codex in case you haven't used it and just some of the um some of the recent stuff that we have um added and shipped as part of Codex. So, Codex is our um solution and our um one singular way of uh you know using a a software agent or a coding agent. um built on top of all of our uh state-of-the-art models like GPD 5.3 Codex, GPD 5.4, GPD 5.5 um which is then like integrated with within the Codex app, IDE, CLI. So which means that you can use the same coding agent across different softwares and um it works across the tools that you use um which means that Codex can interact with GitHub, with Figma, with you know notion, uh Google Docs, whatever it may be. um you can connect Codex to. So wherever you work, what that means tangibly for you for your hackathon is as you are developing models um you know tools, demos, apps, grad servers, MCP servers and so on and so forth, you can um use Codex as the backbone for um orchestrating all of these things and actually building the code that would power all of these tools, right? And um um in fact hugging face has its own plug-in um which you can also directly use from Codex as well. Um of course just a very brief um history sort of lesson on the models itself. The Codex app is as good as uh the models that power it. Um and you know in the past sort of six months we've seen like an incredible um almost like a flywheel um being put together between research as well as product. We went from GPD 5.1 Codeex Max um which was our first model trained on truly running long longunning tasks all through compaction all the way now to GPD 5.5. We've had about six seven models that we've released specifically focused on codecs and coding. Um and we believe GPD 5.5 is the smartest and the fastest model for um real world work. Um it's also a model which um has learned to interact with tools that you use. for example, you can use it for computer use, you can use it for browser use. Um, and whilst it is like significantly more powerful, it is it is also quite a bit token efficient, which means that you don't really uh spend as much um you know in API compute costs or just just just generally it doesn't take as much to be able to do the same thing that you were doing before with any other coding agent. Um this is just a very quick graph comparing the jump that we saw from GPD 5.4 to GPD 5.5. Um though the darker line is GPD 5.4 the lighter one is GPD 5.5. And um you can see that the dots represent reasoning levels. So for the same reasoning level um you can see that how token efficient GPD 5.5 is. It gets better score on terminal bench 2.0. Terminal bench 2.0 is um is a benchmark which um which indicates how good a model is um it's it's like a good proxy of how good a model is on day-to-day tasks um as well as coding tasks right and uh you can see that for the same reasoning effort the model is much more token efficient right um at the same time it is it scores higher on terminal bench as well and so you can see like almost like a clear sort of win between GPD 5.4 4 and GP 5.5. Um, and that's something which we're quite excited about. Um, and so on. Um, bringing it back to um, to Codex itself, Codex as a solution um, has become significantly better in the um, in the last like couple of months. Codex now has an inapp browser which means if you're developing some sort of dashboards uh apps, servers or anything uh really that requires

some sort of UI element CEX can automatically um look at that in the inapp browser as well as interact with that annotate that play around with that and so on and so forth.

You can also use image generation directly within uh within codecs. You can connect to remote SSH connections. So I saw that model is also here. So in case like you're using like a model VM, you can connect um directly via SSH to the model VM and um you know use um codeex app to then orchestrate stuff or anything that you're doing on the VM itself. Um and um and yeah, something that I'm quite excited about these days is slash goal. Um it's um it's now a generally available feature which allows you to let Codex pretty much rip through from a prompt all the way to an outcome. Um and if you give like a like a prompt saying that hey like you know use goal mode and I I don't know train a model train a two billion parameter model on so and so uh strategy. um it's quite likely that um goal mode will figure out what's the best way to go about it, how to set up your environment and everything that comes with it and then at the end of like couple hours actually give you a fine-tuned model. So um if it's something that you want to play around with, I would strongly recommend playing around with it. Um you can do the same thing everything that you can do on the Codex app, you can do the same thing from Codex mobile. Um so especially during the hackathon times um not everyone can dedicate 24 hours um day in and day out on the hackathon itself. So in case you want to like remotely check in on all your uh on the progress of how your um build is going, you can do that directly from the Codex um mobile app as well. Um then there's like a small bit about Codex plugins. Um plugins allow you to connect to all of these different uh disperate sources. Um, as I mentioned before, could be GitHub, could be Figma, could be hugging face. Um, and pretty much anything that you can think of. Um, use plugins are available directly within the Codex app. So, just head over to plugins and figure out which one uh works best for you or just ask Codex to create one for you. Um, with that, I would take a pause. Um, have a have a great hacking experience and enjoy the hackathon.

Yeah, thank you so much V for your time and OpenAI for the sponsorship. Um, just before you go, I'll quickly add uh uh few things about the sponsorship. So there is uh there is a cash award that's being sponsored by OpenAI which is uh you know $10,000. So we have three prizes in that category, first, second, and third. And the uh the most interesting thing about this sponsorship is that you know uh the evaluations will be done by codeex. So ideally what uh codex will be checking is that there have to be uh you know codex attributed commits in your uh in your in your project and how you'll submit is uh make sure that you create a app on the uh hackathon or like everything else but in the readme of your app uh you have to mention the GitHub repo public GitHub repo and in that repo there have to be uh you know this codeex attributed commits. So in uh the codeex you know will be evaluating based on that and if you have a holistic usage of codecs like you use it to fine tune your model or you used it to you know create agents and whatnot then that will be ranked a bit higher and then you know you'll be more eligible to win in this uh category. So yeah it's it's very interesting and a very new thing. Um yeah thank you so much V for your time again and uh yeah let's uh yeah I I'll move to the next uh >> thanks a lot. Thank you.

Yeah, I'll add Shashank from Nvidia. Uh, thank you Shashank for joining us. >> Stage is yours. Um, >> hey, can you see my screen? >> Yeah, I can see you. >> Thank you. >> Perfect. Hey everyone, uh, this is Shashankma from Nvidia. I'm a developer advocate and product research manager. Uh, for the build small hackathon, we wanted to give you a quick map of uh, the various, you know, Nvidia Neotron models that we have. available

that fit the bill. Uh under 32 billion size uh parameters, very efficient for for local as well as edge AI across uh spanning across modalities with you know language, text um sorry text uh audio, video embedding models for document parsing and uh document extraction uh speech. So we have we have the whole deal uh big package uh with Nvidia Neotron.

So, Neimotron models are designed for uh you know being to hit this you know tough trifecta of being efficient, open and intelligent. All these models are openly available on hogging face as well as you know you'll see many of them have uh you know quantized versions NVFP4 checkpoints available as well which make them really efficient for you know the latest uh GPUs that we have that can leverage the hardware native engines. Um so they basically just rip tokens. Um so in short like starting with the general purpose uh language models this is for reasoning chat as well as tool use. Uh the these are great for building classic assistant chatbot coding helpers uh rag applications u a being part of longrunning agent workflows that are efficient. Uh so we have two models in this category. The neatron 3 nano 30 billion parameters total and three billion active leveraging mixture of experts as you can probably tell. And in addition to that there are some u uh you know architectural optimizations that were done to make them really really fast for a given unit of intelligence.

Um in addition to that we have a much smaller edge model available as well. It's called Neotron 3 nano4B and these uh this is for you know RTX and Jetson devices useful for local assistant you know building games and powering you know non-play characters NPCs uh reasoning and tool use agents as well. So these are these fit the bill under 32 billion parameters. uh great for this hackathon uh in terms of general purpose language models. Um next we also have uh you know more advanced models. We have the multimodal uh variant of nano. It's called a neatron 3 nano omni. It's the same 30 uh billion parameter size 3 billion active. It's built on top of the previous nano model that uh we we I shared in the previous slide. And this is omni understanding. That means it can understand audio, image, text, video um and and the output is text. So it can reason over all these different modalities. Great for document intelligence tasks, GUI agents and it won a bunch of benchmarks um you know all across the board um you know when it was released uh about a month or so ago.

And then uh if you want to go deeper into math and code we have a specific fine tune of uh the Neatron Nano. Uh it's called Neimatron Cascade and this this is gold medal winning performance on math and code. great for you know if you're going to if you want to go niche into that domain and also for tool integrated reasoning tasks as well. Now moving on to other uh you know modalities on the on the speech side of things. We have the Neotron 3 ASR the automatic speech recognition models. Uh this is great for building voice agents uh live captioning uh multilingual transcription building meeting audio apps accessibility tools what have you.

Um in addition to that we have this whole category of Neimatron speech models. Um Neimatron 3.5 ASR is one of our latest additions. uh but then we have others as well including uh TTS uh speaker diorization um and and you know streaming variants of of of this model and so if you can go into hugging face and then search for neatron speech collection you'll see a bunch of these uh as well finally for uh retrieval augmented generation for vision and document uh intelligence we've got something called this neatron parse uh this is specific this is a tiny model specifically designed uh it's purposeuilt for document extraction. So if you have this task of you know extracting uh text, tables, uh markdown, bounding boxes from PDFs, PowerPoints, forms, reports, screenshots,

uh what have you, this model does really well at that. It's it's it's a tiny model. It's like less than 1 billion parameters. Um and so it kind of is purpose-built for this uh workflow. Fits right in the center of this rag um as as sort of an intermediate model. uh then that that flows into a larger model for uh you know final reasoning uh like the nano and uh in terms of the uh you know vision and document intelligence and for rag use cases we've got some embedding models as well that fit the bill uh with Neimotron there's the Neotron coal embed uh vision language uh embedding model it has 4 billion and 8 billion sizes and this is great for you know difficult retrieval over dense pages tables charts uh and then There's a smaller variant of that, right? It's called it's a post train version of Llama. It's called Llama Neimatron embed VL 1 billion. And this is for lightweight multimodal retrieval augmented generation as well.

And finally, if you want to make your app safer, we've got some content safety models for uh multimodal content moderation. Uh this this helps you provide, you know, input and output safety checks, uh policy enforcement. Uh so this is like a great model that the that comes at the maybe the end of your agent to ensure that everything that comes in or goes out in fact right is safe. So we've got uh models across uh you know language uh modality, text modality and then speech uh embedding models uh content safety models and um and finally so we we this is an exciting announcement. We released Neatron 3 ultra yesterday. So now this model is not for this hackathon but this is something that we wanted to share out uh with you all. This is an efficient large uh you know frontier class accuracy open model uh for longunning agents high-end reasoning and orchestration and so I would encourage you to check this out you know outside of this hackathon uh even it has uh you know some of the best best uh throughput for for for you know frontier class intelligence especially compared to other frontier class open models and uh you know bookmark these pages take a screenshot uh we've got some a bunch of resources on how to get started with Neatron models. There's a whole Neatron repository and if you take uh if you know scan these QR codes you'll you'll get dropped right into the GitHub repository. You can learn more about Neatron Altra as well. Um here's a screenshot of the Neatron Neatron GitHub as well. We've got uh you know usage cookbooks for deploying these models uh using these models in various kinds of use cases as well as training these models in case you want to fine-tune these further. Uh so we've got all of those resources available as well. Uh that's all folks.

Thank you so much Shashank. This was amazing. Um thank you for you know for laying down like which models are eligible and which are not. And uh uh for everybody's information we have this uh on uh we are putting this up on landing page as well. So Nvidia is sponsoring uh two RTX 580s and uh so one will be given and charged or evaluated by Nvidia team and uh you have to build with the neatron models that are showcased in this uh in this uh session and uh for the second one you have to build with the same neatron models but then they will be evaluated based on the community like if you have the maximum number of likes and then you know uh interaction and everything else then you stand to win that RTX50. 580. So yeah, thank you so much Sashank uh for giving us time and for joining the hackathon.

Thank you. >> Thank you. >> I'll thank you. I'll quickly add uh Felicia from Model and uh >> Hello. >> Thank you. Hey. Hi Felicia. How are you doing? >> Good. How are you? Thank you for having us. >> Thank you. Thank you for joining. Yeah. Stage is all yours. >> Awesome. Um hi everyone. Um really excited to support um the build small hackathon um here at Modal. Um I'm Felicia. I work on developer community here and modal is an AI infrastructure platform um that developers use to run a variety of AI workloads. Um you can run

inference um train models um run batch processing um and utilize sandboxes to build coding agents. Um all of this um sort of uses modal's primitive of um sort of like Python function um that you can use um and sort of like um and access CPU GPU compute um at scale. Um and so um kind of diving into different like use cases and ways you can use modal. Um you can run um VLM inference in 200 lines of code. Um you can um do supervised fine-tuning in 300 lines of code. Um use sort of modal volumes to keep your data um have your data in one place. Um use open source libraries and serverless infra to make um parallel hyperparameter sweeps trivial. Um and then um recently we've been seeing a lot of activity around um coding agents. Um here's an example of sort of like running open code in a modal sandbox. Um but also recently um we uh released an integration with um the OpenAI's agent SDK. And so um our very cool engineer Eric built um this demo that um runs um the agent SDK with modal sandboxes to do parameter golf. um sort of in this theme of like yeah building small models. Um, and for this hackathon, um, like Eveie mentioned, we're, um, giving folks, uh, $250 in modal credits. Um, this should be enough to get you guys building, um, with a variety of open source models, um, run training if you're interested, sandboxes. Um, and if you have questions about how to build with modal, um, you can reach me at, um, felic_modal on the hugging face discord. Um, but also highly recommend joining the modal Slack channel. Um, where you can just speak directly to our engineers and get technical support there. Um, yeah, and that's it. Um, really excited to see what you guys build. Um, I'm sure UV will dive into more of like the prizes and details there.

Thank you so much, Felicia, for uh, and Model, you know, for joining and generously sponsoring every participant with a $250 uh, Modal credits. I think that's that's that's a lot. That's uh that can actually help you know anybody wants if they want to fine-tune a model or if they want to host a model because we have these categories and these badges like if you fine-tune a model you stand to get more chance to win a prize and along with that model is also sponsoring a prize which is a model category. So you if you are using model and you have to specify that in your space read me so you become eligible to win around $20,000 model credits and that's uh that's uh again an amazing prize. Um so yeah we look forward to the uh you know hackathon what what everybody builds with their uh most loved platform. Thank you.

>> Thank you. >> All right I'll move uh I'll invite uh Nikita from Jet Brains uh to the stage. Hey Nikita. How are you doing? >> I'm doing well. Thank you. >> Thank you for joining. >> Give a quick quick quick talk about about Milan 2. Uh this is the model that we released just this Monday. By the way, wanted to say that build small is such an amazing hack idea. When our team saw it, we decided to immediately jump on it because like if we weren't sponsoring, we would probably participate ourselves. Such a cool idea.

Thank you, Hugging Face. Thank you, Graio, for this opportunity for everyone. Uh so uh the only thing that I'm going to present today is our LLM called Milum 2 very fresh model 12 billion parameter mixture of expert model uh which is which is quite good on benchmarks uh it's permissively licensed optim somewhat optimized for code uh and but you can use it for whatever you want coding is probably a little bit better and everything else. And when the model really shines is high throughput mode when you have a lot of requests in parallel, it really works really fast.

So it would be nice to if you could check this out. Uh you can build a lot of things to with it. Here are some ideas

that we um advertise from our team. Um, we share the llama CBP weights. If you want to have Malik support, we haven't merged it yet, but please tag us on Discord. If you want to, we'll try to fast forward it somehow. Uh, on the QR code link, you can find the guide how to use the model, how to deploy it, uh, and everything like that. Uh we have two versions one is thinking one instruct the thinking one actually reasons instruct one doesn't produce reasoning traces and thus is blazingly fast and I think you can figure out how to use it essentially you can fine-tune it you can do Laura please break the model if everything works it's fine if something doesn't work let us know we'll try to help you out uh but it should be more or less straightforward for you. I really hope that everyone can can play around with it. So, uh I actually invited the guys from my team to help you out on the Discord here. The team actual engineers behind the model, the ones who were working on pre-training, post training, they know everything about it. So, and can help you out as well. We will be available on Discord. haven't gotten everybody's discord handles yet. Here is the one that I could were able to instruct mine and even the narrow ones. So, but we please talk to us and I hope you'll have a great time and you'll have some feedback for us to improve the model next time. We'll try to be quick and maybe train one as train another one as we as the hackathon goes. I don't know.

We'll see. Yeah, this is it for my side. Thank you. >> Thank you so much. Uh thank you so much Nikita. That's amazing. Uh so Malum 2 is the model that uh Jet Brains have recently released. It's a code completion model and uh we have uh Jet Brains is sponsoring us for a cash price of $5,000 US and that's very generous of them and uh you know they are uh the support is available from the team uh on Discord. So if you guys are planning to build with mem 2, I think then that could be a good choice because you have the support and you know you have the whole uh whole ecosystem around it. So yeah, thanks uh Nikita. Thanks again for the sponsorship and for joining us. I'll uh I'll add uh next I'll add uh Julian from Cohair Labs. Uh hey Julian, how you doing?

>> Hi. Yeah, good. Thank you. Thanks so much for having us. Um great. Yes. So, I'm going to talk about the uh models that Kahir and Kahir Labs have sponsored as part of this this hackathon. We're we're super excited to see what you all build. Um so, let me share my screen. Uh okay, here we go.

Right. So uh we have uh two models that we've built that are uh that we want you guys to to look at. They're under the two billion limit. Um I guess just a bit of background on Kahir. So we're an antifi enterprise focused uh model lab. Uh and Kahir Labs is our research arm. Uh and we'll all be in the discord. So uh we'll I'll put our handles at the end. Um so first off the first model is a a kahir transcribe uh which is a model I worked on. Uh it's a two billion parameter model. Uh it's super fast. It is optimized specifically to to run at low latency. Um and it's it's one of the best ASR models out there. If you look at the Hugging Face's own leaderboard, you you see we have some nice benchmark results. Uh we support 14 languages and uh we have pretty broad ecosystem support for this model. Uh so we would uh be very interested to see what you build with it. Um we we're uh we we we feel like if you were to be um fine-tuning and you felt pretty ambitious, this encoder because we have pushed a bunch of the parameters there is one of the smartest uh transcription encoders out there. So there's potentially some interesting things you could do there to extend it to other tasks.

Uh the second model or I should say family of models are a 3.3 billion set of multilingual LLMs. And here the language support is is really really broad. So this these are really great choice if you have um languages that

maybe aren't um considered by more of the um the the the more the the the commonly used LLM. Um and there's five that you'll need to know about. So the first one is base that is the pre-trained model. Um then you have the global which is the best overall. Um it covers all of the 70 languages. Uh and that's probably the default to try first. But then we have three other variants too. So we've got earth, fire, and water. Earth is best for West Asian and African languages. Fire is best for South Asian languages and water is best for European and Asia-Pacific languages.

Uh we have quantizations available and it can run uh on your phone and in your browser. Um it's got a bunch of strengths this model. Um, one of the a few of the key ones are summarization, translation, um, anything that has a cross-lingual um, uh, t part element to the task. So, if you need to understand multiple languages in the same prompt, uh, these are really really strong choices. Uh, and we have a um guide uh, on our blog, so please check that out. We'll post the slides in the discord so you can uh get all of this information after the fact and we will be there in the discord. So myself I I worked on Kir transcribe and Alejandra and Sarro who worked on tiny uh and we're really excited to engage with you guys hear what you hear what you're planning to build um and answer any questions that you have. Thanks so much. Passing back to you Yubie.

Thank you so much Juliana and that was amazing. So uh just uh you know just few things about Koh. So they have generously sponsored us with with uh 5k US dollar sponsorship and that's uh that's get added to the general price pool. So we have a great support present in the discord uh with uh you know the teams that Julian mentioned and you can use tiny they have created a great guide you can use uh transcribe model as well.

So and you know you can fine-tune you can use uh if you end if you end up using doing fine-tuning using model or you know if you end up using quantization using llama CPP you tend to you know get all these badges in like I explained in my uh in my initial uh you know uh session that these are badges and you know more and more badges that you collect then you know you stand a chance to win big. So yeah thank you so much Julian and thank you everyone for you know joining us. Um let me just uh wrap it up.

Uh let's share my screen. Uh just one second. Yeah. Um that's about it and uh just start building. I know you guys might be having some you know questions around um some of the credits that we have uh we have distributed and uh the you know joining deadlines and registration deadlines and everything. So all the questions are there already on the landing page and they are already being discussed in discord. So you know feel free to search on these two locations and uh the homepage for this event is the uh is the hackathon org and you have to be a member of this orc if you want to participate for all these amazing prizes which is you know around $48,000 cash 20,000 model credit and you know judge GBD bro for one year 580s and whatnot. So make sure that you are a member of the organization because very soon we might be you know closing the organization as well the registrations is already closed. So make sure that you are a member of this organization and I think that's about it. Um thanks everybody for joining and uh yeah um see you see you at the hackathon.

Take care. Bye.