it is no surprise that deep seek has

been getting a lot of attention lately

and I suppose in keeping with that theme

for today's video we are going to do a

quick local install and test of the

newly released Janice Pro Models from

Deep seek these are multimodal models

that will allow you to actually ask a

question about an image and get that

question answered as well as actually do

some text to image Generation all from

within the same gradio web interface

which is prettyy d cool we will begin

**Setup And Install**

here just in the GitHub and we can see

that on today's date which is January

27th 2025 Janice pro has been released

an improved version of Janice and for

today's video that's what we're going to

be testing now there are two versions

that have been released there is a

Janice Pro 1 billion parameter model and

a Janice Pro 7 billion parameter model

going to the actual hugging face

repository for either of these where we

can see all of the model files host

we can see that the 7 billion parameter

model is about a 15 GB download while

the 1 billion parameter model is about

like four four and a half total so there

is a big difference in size of these

models and due to that I suppose that

one of the common questions would be how

much vram does either of these utilize

so for the video today we're actually

going to be testing both of these models

just to see one how they perform and two

how vram utilization differs just to

give kind of a benchmark for those at

home who may be interested in actually

seeing if they could run either of these

or perhaps both of these on their local

system it is really quite simple to get

this actually all set up and going and

that's basically what we're going to do

right now if we go down to number three

here or the quick start in the Janice

GitHub repository we can see that

installation of this is quite easy I

like to use cond environments which is

what I'm going to do here and I am

simply going to create a new en

environment I cannot talk and type at

the same time very

well and we can see that they just have

one requirement here which is just

python greater than or equal to 3.8 I am

using Python 3.10 and the Dy flag I have

appended to the end of this command will

just basically go ahead and

automatically say yes to anything cond

may have asked me now that that's done

we can essentially just go ahead and

activate our newly created conduit

environment and once we are in here our

next step is to actually clone this

GitHub repository onto our local machine

so we just type git clone and then paste

the URL of this repository which we can

find when we go here into the code

button and the drop down we find it

right here and once we do this it will

go ahead and clone this onto our local

machine once that's done our next step

is to just go into that folder so CD

Janice and I will just clear my terminal

um routinely just to kind of keep the

amount of old text in their light so

it's easier to read the next thing is

really just to do pip install dpace

period but because we're wanting to

actually do the gradio web interface as

well which is essentially the same thing

you would see if you were to go to the

hugging face page for deep seek and you

see that there are some spaces running

right here and one of them is chat with

Janice Pro 7B so this actual like

website look if you will is the gradio

web interface and we want to actually

clone this and run it on our local

machine so in order to do that we're

going to scroll down to the end of

section three here in the quick start

and we actually do see that there is a

slightly different installation command

here which will include that kind of web

interface and allow us to run that

locally so once I get back down there we

will just copy this and paste it into

our

terminal and that will go ahead and

install the dependencies as well as the

Janice repository in the way that it

will allow it to work with itself and

the models we're going to download and

once that has completed we can

essentially just run python demo app

genpro dopy which will open the web

interface however what I want to do here

is first off try the 1 billion parameter

model so we are actually going to need

to make one change in this script to

actually tell it to download the 1

billion parameter model instead of the 7

billion parameter model in order to do

that we we are going to go into the demo

folder like I said I can't type and talk

the demo folder on our local machine and

find that script which is appor Janis

proo dopy we are going to copy this and

I'm going to edit this in Nano which is

a inter terminal text editor but you can

edit this in vs code or any other IDE

you would

like when we open this file we will be

able to scroll down here and see that

under the model path right now it is

written as the 7 billion parameter model

so when you run this for the first time

it will essentially download the model

files and things from the hugging face

repository which we can see right here

so it will download all these when you

run it for the first time because I want

this to initially start and use the 1

billion parameter model we are just

going to change this to reflect that so

instead of 7B it will be the 1 billion

parameter model the second thing that I

want to make note of is that when you

actually start this it will ALS Al make

a live sharable link for the gradio web

interface being that we are going to

just want to run this locally on our

machine and don't need a actual live

sharable link to give to anyone we can

scroll all the way down here in this

demo script and basically on the last

line here where it says demo. launch and

in parenthesis it says share equals true

we can actually just go ahead and remove

that and in doing that it will just give

us the local URL of this running on our

local host or our local machine so that

is more of a preferential thing but

personally I always like to just make

sure it doesn't generate that sharable

live link if you will once that's done

the only other thing we need to do is

make sure that we go back out of the

demo folder because you want to run the

actual command here from just the Janice

folder and no sub folder from within it

and as we do this now we will see that

it will spin up a gradio web interface

for us that will be using the 1 billion

parameter model now I have already gone

ahead and tested these on my machine so

it is not going to download the model

files for me because it has already done

that however if you're following along

and running this for the first time on

your machine when you run this it will

go ahead and download the actual model

files which will take a little bit of

time depending upon your network speed

remember the 1 billion parameter model

is probably like a 4 and 1 half gigabyte

download and the 7B was like 15 to 16

and when we do this we are going to see

**1B Testing**

some things but one of those things will

be a clickable link that will actually

bring us to the web interface running on

our local machine we can now see that we

have a local URL to click on which will

go ahead and open what we saw in that

hugging face spaces except this is all

running on our local machine and we can

now go ahead and actually test Janice 1B

or Janice Pro 1B now to begin I think

what I'm actually going to do is start

with the text to image generation and

then I will actually save the output

from that and then use that for the

multimodal understanding I never really

like to test things with the default

examples they give because I like to I

suppose add my own personal touch into

it and then test it on like unseen data

if you will or something of that sort

now prompting is very important because

the more descriptive it is is correlated

to generally how well the image comes

out as they say here but I prompting so

we're just going to

do a sports car driving on the Las Vegas

Strip now I am using OBS as well so when

we look at the vram utilization of the

card right here we want to keep in mind

that OBS and screen recording is going

to add uh one or two gigabyte overhead

to the actual usage of the card so with

that said it just generated 5 images of

perhaps questionable quality and we can

see that it did seem to top around 14 GB

of video RAM usage um and here are the

images I'm not really going to be

analyzing the results and things like

that this is more just kind of a trial

and just to see how it works and how it

looks and things of that sort but I do

want to say that my prompt is pretty

poor and they say that more detail can

produce a better image so I suppose we

can just go ahead and use one of their

demo prompts but instead I think instead

of intrinsically designed I write

intrinsically designed sports car so

we'll add some personalization into this

prompt and we'll once again go and

generate the images and we'll see if

they are of better quality with that

intricate

prompt okay and that I I'm seeing 17

gigs of video memory used right there

which is kind of a

lot all right

these are definitely of much better

quality and I think the next thing we're

going to do now

is sure so we have saved that image to

the desktop now and we are going to

scroll back up all the way to the

multimodal understanding which to put it

in plain English just basically means

you can show it a photo and then ask it

about the photo and it will be able to

actually respond to your query which we

will do now what is this

a stylized

eye

okay what is this be specific as

possible

please much better this is an artistic

depiction of an eye the eyes is

intricately detailed striking Blue Iris

unfortunately so that's good but my

sports car thing didn't really seem to

make any effect here but that is okay

for the one B this is pretty much what

we have and I'm basically just going to

jump straight into this and go ahead and

now try this with the 7B model and as I

do that the first thing I'm going to do

is just close this so I'm going to do

control C which will close the server

and I'm going to clear out of there so

if we actually tried to do anything else

in here we would see that it is no

longer being served because we shut down

that um script if you will I am going to

now need to once again modify the demo

script that we did before because I want

to now point it to the 7even billion

parameter model so I'm just going to use

Nano again to quickly edit this and we

scroll down once more to the model path

here and this time instead of the 1B

model we are going to change that to

7B and again I do actually have these

already downloaded on the system because

it just makes a video faster to go as

well if I don't have to sit and wait for

big files but remember that you only

want to run this from the actual Janice

folder and no subfolders within it or

else it just wouldn't work so we will

now do python demo and this will

**7B Testing**

actually go ahead and run this with the

7 billion parameter model as opposed to

the 1 billion parameter model so we'll

get to kind of see the difference in

them and what they say and things of

that sort we can now see that it has

once again loaded except this time we

are going to be able to use the 7

billion parameter version of Janice Pro

which we have here now there obviously

is no visual difference when you open

the gradio web interface however I do

actually want to give it that same image

that was generated with the 1B and we'll

ask it the same question which is what

is this and let's just see how its

output differs with a simple question

like that so before I believe it had

just had a stylized eye and we can see

that with the same basic question this

is a detailed illustration of an eye

okay so this is almost more akin to when

I had specified like please be specific

and we can just basically do that now

and see how the 7B model does with okay

so it doesn't seem to have actually been

any more specific than it just was with

the what is this so again this is not

necessarily like a strict academic

testing of this it's just kind of a way

to get a feel for it and see how to

install it and run it and things like

that we can see right now we're using

around 18.4 gigabyt of video memory and

keep in mind OBS is probably accounting

for about 1 and a half gigabytes of that

so make of that what you will I suppose

now I will just try my sports car prompt

again and right here during image

generation is where I'll get a little

concerned if I might o or out of memory

which is basically when your video card

apparatus has no more available video

memory to perform a task you get that

error fortunately

well getting close to the Limit it does

seem that if you have a 24 gbyte card

you should be able to be able to go

ahead and run the 7B of this without too

much issue and again these image

Generations are not the best however my

prompt was again somewhat lackluster so

as we did before let's just go ahead and

put the long demo prompt in but instead

of I we will just type sports car

and we'll see what

happens you can see the images are

taking a bit longer to actually generate

when the prompt is far more detailed

than my lackluster single sentence

prompt so that also speaks to the actual

quality of the outputed image being

heavily dependent upon the intricacy of

the prompt and we can see that this time

while it did still capture the stylized

eye it does also have an automotive

esque theme to it which I would say is

rather nice so I'm just going to save

this once more and we will just

overwrite the original one that I had

done that with and I suppose as like a

kind of like to wrap this up we'll just

pipe this back into it this is almost

like a meta like generation and

understanding weird Loop sort of thing

and we'll ask what this

is a car okay um what is this

and we get better if we ask it a more

intricate question please be specific

I'm going to just ask it like something

totally unrelated to the image as well

and see if see if it references and

responds to this

also side note can you tell

me about

the

purpose of a piston

in a cylinder I guess that's slightly

Automotive related but let's just see

what it

says and obviously it is taking longer

to answer okay so it does correctly

respond to

the weird kind of sub question that I

had asked it very cool suppose uh the

final thing I want to do is just let's

see how high okay so that's locked to

not go over one but even if we put it at

one it will likely have a more

interesting response um 0.1 is

relatively low I think for a temperature

setting

but okay so it was rather similar and I

wonder

if no because I can't show that on

YouTube all right um my curiosity was

something around um whether or not it

has certain restrictions on types of

images it will generate but with that

said I basically just wanted to kind of

cash in on the deepsea hike deep seep

deep seek hype I'm kidding but uh I

wanted to just kind of test this and

show how it was and something I always

get curious about when I see models and

things like that is how much video RAM

do they need to run and can I even run

them so for those who may find

themselves a little confused and be like

well this is only like a 1B model or

like a 7B model like I can run those on

like a Raspberry Pi with

oama these are not really quantized like

those are so even though the parameters

of these models seem small they are

still relatively performance intensive

to actually run and that is just kind of

a nature of I suppose some of the design

and architecture of these models

especially Vision models in general and

ones that can do image generation and

multimodal sort of things like that so I

did I don't know if I said this already

in the video but I did try just doing

the tiny tiniest one along with some

optimizations on the jetsen or Nano

super AIG and un Al it was just not at

all going to be something that would be

possible very likely so um yeah make of

that what you will but for my Jets and

folks out there I do apologize I had to

do a strictly uh desktop PC kind of

focused video other than that um I

**Closing Thoughts**

suppose it is rather interesting there's

definitely some more technical things

and stuff like that here I noticed that

with all of the news around Jeep seek

and things like that there has

definitely been sort of

a um

I don't know how to put this lightly but

a lot of folks have been critiquing like

the data implications of like running

these models and you're giving your data

over to them and things like that I do

want to just point out for the sake of

my videos whenever I am running

something even though like Hey we're

running this locally and offline a lot

of times your computer is still

connected to a network even if you're

running an offline llm I just would like

to make note that all of the computers I

use to film my videos are not burner

devices but they don't have any

sensitive information of mine and in

addition to that they're not actually

able to access other devices on my

network that are like personal devices

and things like that so regardless of

your stance on like data privacy and

things like that I think it is just

important and pertinent to mention that

you want to take some precaution and

what devices you experiment on suppose

with that out of the way that really is

going to wrap this I just kind of wanted

to do something with it and test it and

see how much vram it used and it was

good to I suppose put my liquid cooling

system to the test um since the disaster

if you will all right that's going to

conclude it if you have any questions

leave the comment let me know and thank

you for watching