

Forum:[Hack Forums](#)

Thread: [I built my own "Neuro-Sama"](#)

oRequiem - August 11, 2026, 10:08 pm

For anyone that doesn't know who Neuro-Sama is, she's basically an AI VTuber that has the ability to have memory, as far as I know = learns on the spot, can sing (I have an idea of how vedal987 does that), can talk with a human-like cadence (not too much by choice), can read and reply to chat via voice/text, can train to play games, etc.

I built my own pretty close to how Neuro-Sama is, but Neuro-Sama is quite advanced since vedal987 was probably the first to do something like this, and has the gpu power to do more. Mine has memory, personality (develops over time), thinking about mood changes depending on the conversation, gave her the ability to read my twitch chat, I also created a discord bot so she can join a vc and/or reply to people that mention her, and a filter system so she doesn't randomly say anything bad in my bannedwords.txt file. I would like to eventually fine-tune an LLM (her brain) for vision (ability to train her to play games and see what's on my screen).

I haven't really bothered working on the frontend of the project because of localhost issues (basically can't connect to a temporary web UI I made when my AI is "running"). It would be nice to make a clean/minimalistic web UI for ease instead of just focusing on the terminals being opened.

Curious to see if anyone has any more ideas I can work on with the project.

UPDATE: I did end up giving the AI an emotion and mood system. I am currently using Ollama for the LLM, but I might end up switching to llama.cpp since I've noticed that it seems to be better overall compared to Ollama.

UPDATE 2: I officially went to running her from Ollama to llama.cpp; Apparently, the performance and speed difference is noticeable, but I think I will stick with llama.cpp

UPDATE 3: Just gave the AI a "Hybrid Brain." I'm only using free API's for now since I'm trying to keep this 100% free; my local LLM is purely the fallback right now. I sadly can't implement a vision model using an API key since it gets rate-limited way too fast, and I simply don't want to invest any money in this project right now. 3 brains right now, 1 for deep thinking, 1 for local fallback, and 1 as her "main" brain.

UPDATE 4: I finally routed her TTS Python Audio perfectly in OBS Studio. It was a little tricky since I use my main stream mic as "Voicemeeter Banana" so using CABLE-Input was a bit tricky to split Python Audio in my OBS Audio Sources, but found a way to make sure my AI didn't talk through my mic's streamline.

UPDATE 5: Finally did a 2nd test running my AI Vtuber on stream today. Its responses were perfectly fast and accurate, had A LOT lower fps due to Call of Duty taking up way more VRAM than it should, especially when I'm running whispers locally, the way I'm routing my AI's Python TTS audio made it so it speaks in-game so people in the lobby heard replies when it replied to my twitch chat. Still no vision yet, but that's in the works. Still trying to figure out what single player game to test first for my AI to learn to play.

UPDATE 6: For giving her vision and to play games, I'm currently working on an organized

structure and blueprint to start implementing. It looks like I'm going to have to invest some money so I don't get rate-limited almost instantly. Now for the Discord bot version of her, I'm currently working on implementing her hybrid brain so she can talk in voice channels freely, just like how I can talk to her when running her with my mic unmuted. Before, I had to make it so temporary .mp3/.wav files would go to my temp folder; she reads, responds (Python audio and local LLM would crash), then responds (never responds in voice due to LLM crash), then the audio files auto-delete. With her new hybrid brain, I can make it so that part isn't local unless it's needed as a fallback (which wouldn't work anyway, so it's 100% relying on the API's, hence the money investment). Vision will take longer to implement, so I wanna at least get this Discord bot version of her somewhat closer to being finished, and this will be a pretty good upgrade.

UPDATE 7: So 2 of my cloud-based models were deprecated at the same time. I spent 7 hours straight last night rewriting a large portion of my codebase, which is triggering me. For some reason, after my cloud-based models were deprecated, my AI couldn't detect any of my API keys even though they were 100% active. I finally got 2 of them to work (another new API), and it's somewhat back to normal again. For some reason, 1 of the API's stopped working, but the other started working again randomly. It's a pretty odd issue I ran into. Currently spending hours figuring out how to get her Discord bot to receive voice channel audio and respond via voice. Seems like nothing is really working; I've tried about everything that I can think of, including audio routing to a virtual audio cable. Right now, she only responds in the voice channel when someone types to her, not when they speak.

UPDATE 8: I've been attempting to invest some money in the project. One of the services I use for my API keys has really bad payment methods (all require a bank), which is annoying. There is crypto, but it requires a wallet extension that is overly confusing to me. I will provide more updates whenever I get some money invested in the API. I did test out Google's API, but their UI is a convoluted mess, and I'd prefer staying away from it.

Somali - August 13, 2026, 7:05 am

(Aug 11, 2026 10:08 PM)oRequiem Wrote: [\[CITE\]](#) For anyone that doesn't know who Neuro-Sama is, she's basically an AI VTuber that has the ability to have memory, as far as I know = learns on the spot, can sing (I have an idea of how vedal987 does that), can talk with a human-like cadence (not too much by choice), can read and reply to chat via voice/text, can train to play games, etc.

I built my own pretty close to how Neuro-Sama is, but Neuro-Sama is quite advanced since vedal987 was probably the first to do something like this, and has the gpu power to do more. Mine has memory, personality (develops over time), thinking about mood changes depending on the conversation, gave her the ability to read my twitch chat, I also created a discord bot so she can join a vc and/or reply to people that mention her, and a filter system so she doesn't randomly say anything bad in my bannedwords.txt file. I would like to eventually fine-tune an LLM (her brain) for vision (ability to train her to play games and see what's on my screen).

I haven't really bothered working on the frontend of the project because of localhost issues (basically can't connect to a temporary web UI I made when my AI is "running"). It would be nice to make a clean/minimalistic web UI for ease instead of just focusing on the terminals being opened.

Curious to see if anyone has any more ideas I can work on with the project.

Everything you do seems amazing just one note.

When you add "bannedwords.txt" it looks harmless, but it actually makes your AI much dumber, since it constantly has to remember and avoid those words.

So if you need that, make sure to apply it as a filter on the final response not as part of its memory. And honestly, you'd be better off focusing on a physical device rather than a website AI model. Your project will get noticed by everyone once it can walk, talk, and see your chats, bots, etc. all at the

same time. Imagine saying "find the person who said this" and it does it, then says "here you go, I put that person on your desktop as user.txt with his messages from this year.

That way it's synced with everything and becomes an actual real entity that can talk.

oRequiem - August 13, 2026, 8:51 am

(Aug 13, 2026 07:05 AM)Somali Wrote: [\[CITE\]](#) Everything you do seems amazing just one note.

When you add "bannedwords.txt" it looks harmless, but it actually makes your AI much dumber, since it constantly has to remember and avoid those words.

So if you need that, make sure to apply it as a filter on the final response not as part of its memory.

And honestly, you'd be better off focusing on a physical device rather than a website AI model. Your project will get noticed by everyone once it can walk, talk, and see your chats, bots, etc. all at the same time. Imagine saying "find the person who said this" and it does it, then says "here you go, I put that person on your desktop as user.txt with his messages from this year.

That way it's synced with everything and becomes an actual real entity that can talk.

Well it's technically not part of her memory, she just checks the bannedwords.txt file for words that filter out so if she says a sentence with one of those words, she will just say "filtered," I do have a fallback so if the file corrupts for some reason, she will go to her default in the code itself.

CL1B - August 20, 2026, 4:35 pm

What a cool read man, cool stuff. I just watched the short video and while it doesn't show much in action; in context with this thread it's pretty cool man. Goodluck!

oRequiem - August 20, 2026, 4:57 pm

(Aug 20, 2026 04:35 PM)MarlboroMan Wrote: [\[CITE\]](#) What a cool read man, cool stuff. I just watched the short video and while it doesn't show much in action; in context with this thread it's pretty cool man. Goodluck!

The test was specifically how well the game performed and if she responded at all/how fast her responses were with all of that running.

Basically was running this:

- Main Terminal (Her startup)
- Whispers Locally (Her TTS)
- VTube Studio + Her VTuber Model (Autonomous Movement - Still working on tuning it)
- Black Ops 7 (COD takes a ton of VRAM by default)
- 2 Terminals to Bridge my Twitch Chat to give her the ability to read
- OBS Studio
- StreamLabels
- Chatterino
- Google Chrome (YouTube for music, X while it's Idle, and Twitch Dashboard)
- Discord just sitting there idle

I have a ton of examples on Discord, just haven't published them in on X or YouTube.

I got her to hear music and dance with her VTuber Model so far.

Mr.Cat - August 22, 2026, 4:35 am

It would be interesting to see this in action on a youtube video or something

oRequiem - August 22, 2026, 11:27 am

(Aug 22, 2026 04:35 AM)Mr.Cat Wrote: [\[CITE\]](#) It would be interesting to see this in action on a youtube video or something

Here's a test I did a while ago

<https://www.youtube.com/shorts/JN-34PgYsgg>

oRequiem - *October 12, 2026, 4:59 pm*

Sucks, but I haven't worked on this in a few weeks.

Last time was my last edit, will get back into it at some point though. Very fun project, just been busy.