**Crie Seus Próprios Agentes de IA: LLMs Gratuitas para Todos!**

**Transcribed by** [**TurboScribe.ai**](https://turboscribe.ai/?ref=docx_export_upsell)**.** [**Go Unlimited**](https://turboscribe.ai/subscribed?ref=docx_export_upsell) **to remove this message.**

Today is the day to make an LLM run 100% locally on your computer. Nice! You will reduce your costs with your agents within the configuration of your crew using the Crew AI with the LLM on your computer. It's now here on the Sandeco channel.

Let's go! Hello, how are you? I'm Professor Sandeco Macedo. I am a professor and researcher at the Institute and Federal University of Goiás. In addition, I am the ambassador of Campus Para Brasil and my goal is to make you use smart agents in your daily life.

Well, this is a question that people always ask me. Sandeco, how do you run an LLM on my own computer without using the APT chat to integrate with my Crew AI? How do I do that? Well, I'm going to show you in this video, which is a free mentoring, about the Crew AI book that I released a few days ago. Volume 1, Crew AI for Beginners.

We have a mentoring group there, the people who are following, right? You can find out more about it in the link in the description below, okay? Very well, the first thing you will do is access this site here, olhama.com, right? And this is where we will download the LLM. Here, I'm going to decrease the zoom and we're going to download it here. You come here in Downloads, click here in Downloads and download it to Windows.

Download it to Windows if you have it on Windows, right? Or for Linux or for Mac. There are for all operating systems. I'm using Windows here because of my studio, right? It's all Windows, okay? It's a teacher's studio, guys.

Teacher's studio, that's right. It's in poverty. Okay? After you download it here, you will come here in your Downloads, right? I'll show you my downloads here.

After you download it, you will have this file here. Just click twice and then that standard Windows installation, right? You come from Next, right? Install, Next, Next, Next, OK. Everything will be wonderful.

Now, the next step, I'm going to do the following. Let's go here in Models, okay? So, here in Models, I'm going to download the following. I'm going to look here at Llama 3.1. You can use others too.

Look how cool. I have Llama 3.1, okay? We're going to use this 8B here. But just to show you that there are other models here.

You can use here, for example, Gemma, Mistral. There's Kiwin 2 here, right? Kiwin 2 here, look. This model of this 0.5B is very small, okay? So, do your tests there.

For everyone, the default for you to download will be the same thing. How am I going to do it? I'm going to go back here on Llama. I'm going to click here on Llama so we can download it, right? So, it's going to open the model library here, okay? And I'm going to choose the library for us.

Llama 3.1 will be here, okay? Instead of copying and pasting here, run this again, click here, open here, click again. It will do the following. It will give you this option here.

Llama 3.1 2.8B. It's this Llama that I'm going to download here on the computer. And look how simple it is. I'm going to copy here.

I'm going to open a Windows window here, right? A command window, okay? Very simple, everyone knows, okay? So, here's the thing. To require Llama on our computer, I'm going to use Llama. So, Llama, Llama, right? Llama, less, less, version.

You see if it's installed correctly. So, when I run here, you'll see that I have version 0.3.6. Okay? To get Llama 3.1, how am I going to do it? I'm going to open Llama here. Pull, okay? Write this command here.

Pull. And then you paste exactly the same name, which you got here. I'm going to increase the screen size a little bit so you can see it better.

You see? The same name, copied and pasted there. You open your command prompt again. And just run it here.

I've already downloaded it. And everything will show up like this. Kind of already downloaded, right? So, you do the following.

You wait for this download to happen. It takes a little while because it's 4.7GB. If I'm not mistaken, it's a little big. Then you'll download it all on your computer and then you can put it to run.

Okay? Very well. You see in this case that I've already downloaded Llama 3.1 here on my computer. It's cute here.

How do I run it? To play with it here in the command prompt itself. Look. I'm going to do the following.

I'm going to come here. I'm going to copy this command here so I can run it. Llama run, the name of our model.

So, I'm going to run it here. I press enter. And you'll see that it's trying to run it here.

I'm going to have a chat with it right here. You'll see. I'm going to talk to this LLM, okay? Very well.

Now it's loaded beautifully here. Let's talk to it. I'm going to say hello here.

Hello. How are you? How are you? There it goes. Hello.

I'm fine, thank you. And you? How is your day today? I'm great. I'm wonderful.

Isn't it? That's great. I'm going to ask it a technical question here. Let's see if it's going to answer me nicely.

I'm going to ask if it understands about network transformers. Talk to me about network transformers. Okay? Talk to me about network transformers.

I'm going to wait for it to explain it to me. Ah, it's a wretch. See? I didn't explain it.

Hey, teacher. You didn't ask a very well done question here. I said, talk to me about network transformers.

It understood that it was the transformers of the cartoon Transformers, Optimus Prime, Bumblebee and everything else. Look at that. It got confused.

It thought that the Transformers here in this case is the cartoon Transformers. I want it to tell me about it like this. Talk to me about network transformers.

Now, yes. Network transformers are a neural network architecture that aims to improve the ability to learn data sequences such as text, voice and image. The Transformer Neural Network was presented in 2017 in the article Atencio e Isola Unidos.

That's exactly it. That's all we know about the Transformer Neural Network. So, it's working.

It's a beauty here. The Llama 3.1 on my computer. Ok? To connect the Llama with the Cru AI, I'll need to install the LangChain Llama.

It's simple to do. Too easy. Look how it is.

To install, you'll do the following. Just like this. To install the LangChain Llama here, you'll do it like this.

Installing, right? If you're going to use the pipe, just give the command pipe install, right? LangChain Llama. Now, if you want to install using poetry, using poetry, you'll use the command poetry add LangChain Llama. Ok? Very simple.

The first thing I'm going to do here is import. I'm going to import my code here. I'm going to import the OS library and then I'm going to import the Llama LangChain library so we can use the Llama here, ok? Too beautiful.

The account is too beautiful. Look. And now, what am I going to do? I'm going to add the Cru AI libraries down here.

For us to create the agent, the task. It's going to be a simple agent. I'm just going to ask him to build me a love letter.

Very simple. Ok? The next step here is this. There's a problem.

There's a problem. Right? When you're going to connect Cru AI with Llama, you have to take into account that you're connecting with the GPT, ok? Take into account. So, he needs you to activate a key from OpenAI here, but it's weird.

Putting this NA here, this NA here, right? It's like a mock, right? A hidden thing. It's a lie thing to deceive the tool itself to be able to work. Okay? Then the next step, what is it? I'm going to create the connection with Llama here.

And then I do the following. I'm going to call who? I'm going to call the chat library, Llama. Okay? Very well.

Open and close parentheses here. The next step is to call the model. So, who is my model here? The model has to be exactly the model that I installed, which is the model here.

Llama 3.1 2.8b. Right? So, this is the model we're going to use. Because it was the model I looked for to install here. Now, a very important thing.

When we use Llama, it will be installed on a local server on our machine, right? So, Crue AI will see Llama through this server. So, we're going to put here http:// as if it were connecting a site, connecting an IP address. So, I'm going to put here localhost.

Okay? Localhost, which means a connection with the machine itself. And then I'm going to connect with the 11.4.3.4 port. Okay? So, this is the address and the connection port with Llama.

Once that's done, everything is ready. Now, we just need to put the agents. So, I'm going to put the agents here.

I'm going to create the agents. I'm going to put it here. Now, it's very important.

My agent is a love letter writer. So, I'm going to put his LLM here, being exactly Llama's LLM. Don't forget to do this.

This is very important. The next step is to define the agent's task. You already know how to do this, based on our classes.

I put the task here, which is to write a love letter to the agent. But look at the detail here. I'm going to ask him to write this love letter in a file called amor.md. This amor.md is the love letter that I want to write to a certain person, which the agent is going to write.

The next step is to create our crew. Basically, I'm going to create a crew here, where I'm going to put my agent, the agent-writer, the agent's task, which is the writing of love letters, and the basic sequential process. The next step will be the following.

I'm going to execute it by putting the recipient and the people who are going to receive this love letter. So, let's go. Here I'm going to express the feelings of the recipients, who is the person who is going to receive this love letter.

So, the person who is going to receive it is going to be, the recipient's name is going to be Maria. And the feelings that I want it to contain in the love letter is going to be exactly this here. Eternal love, deep affection, admiration.

This comma doesn't exist, guys. Very well. I corrected it in Portuguese.

The next step is to activate the kick-off, to activate it by executing our crew. So, I'm going to put here our recipient, the recipient receives the recipient, and the feelings receive the feelings here. So, everything is very beautiful, everything is ready.

And the next step that I'm going to do, I'm going to make an impression of the results down here. So, the impression of the results will come out. Let's run this whole thing here to make it work.

It's going to be too beautiful. The Llama 3.1 along with your, I pressed chat, with your crew.ai running everything locally. Nothing is coming out to the web here.

Let's run it. Now, I'm going to say, there's a little error here, what is it? Results. Ah, the variable name is wrong.

Now it worked. Now it worked. While I'm running here, I'm going to run for you, I'm going to say the advantages of using this type here of LLM running locally.

The first thing is the cost. It drops absurdly, that is, you have no cost, except for the cost of running your machine. Right? Second place is that you have a data protection, right? That is, the data you have, for example, if you have sensitive data, in here, using this local LLM, it's all protected.

Okay? So, it's a very big advantage that you have when you run this LLM along with your crew.ai on your local machine. It's running the agent here, it's running the agent. Its output will be exactly the love letter.

Okay? Let's see. Very well. Look here, when it finished, I had the love letter printed.

It says here, My dear Maria, when I think of you, my heart beats fast, my soul lights up, invaded by a feeling of peace and such. That whole thing he's saying here. Another thing I want to do here, I'm going to reload here, you can see that the file appeared here, amor.md. I'm going to open it for you and look at my letter here.

What a beautiful thing. My dear Maria, when I think of you, my heart beats fast, my soul lights up, invaded by a feeling of peace and such. Okay? I hope you enjoyed this video.

Don't forget to subscribe to the channel. Here in the description, there is a link to a WhatsApp group where we will give you support on questions about crew.ai and also about the crew.ai book that we released. Okay? I hope you enjoyed this video.

A big hug for you and see you next time.

**Transcribed by** [**TurboScribe.ai**](https://turboscribe.ai/?ref=docx_export_upsell)**.** [**Go Unlimited**](https://turboscribe.ai/subscribed?ref=docx_export_upsell) **to remove this message.**