[{ "text": "if you want to chat with your docs if", "start": 0.08, "duration": 3.0 }, { "text": "you want to chat with your text files", "start": 1.56, "duration": 5.64 }, { "text": "your PDFs csvs Excel files anything any", "start": 3.08, "duration": 6.16 }, { "text": "type of document really this is such a", "start": 7.2, "duration": 5.12 }, { "text": "great project private GPT is my most", "start": 9.24, "duration": 6.0 }, { "text": "popular video of all time I made it", "start": 12.32, "duration": 5.6 }, { "text": "months ago and since then the developers", "start": 15.24, "duration": 5.16 }, { "text": "have built a ton of new functionality", "start": 17.92, "duration": 4.84 }, { "text": "and really changed the course of private", "start": 20.4, "duration": 5.56 }, { "text": "GPT completely and so today I'm going to", "start": 22.76, "duration": 5.2 }, { "text": "show you the updated way to install it", "start": 25.96, "duration": 3.44 }, { "text": "I'm going to show you all of the new", "start": 27.96, "duration": 4.279 }, { "text": "features and we have a special guest at", "start": 29.4, "duration": 5.92 }, { "text": "the end so let's go so this is private", "start": 32.239, "duration": 6.121 }, { "text": "GPT it is completely open source you can", "start": 35.32, "duration": 5.88 }, { "text": "run it entirely locally with a local", "start": 38.36, "duration": 5.199 }, { "text": "open- Source model you can also use chat", "start": 41.2, "duration": 5.08 }, { "text": "GPT if you want to everything is super", "start": 43.559, "duration": 6.041 }, { "text": "flexible now and private GPT has really", "start": 46.28, "duration": 5.2 }, { "text": "transitioned into becoming a developer", "start": 49.6, "duration": 3.4 }, { "text": "product so I'm going to show you a", "start": 51.48, "duration": 3.0 }, { "text": "little bit about that but it's still", "start": 53.0, "duration": 3.44 }, { "text": "just as strong for the end user if you", "start": 54.48, "duration": 3.399 }, { "text": "just want to load up your documents and", "start": 56.44, "duration": 3.56 }, { "text": "chat with them this is still one of the", "start": 57.879, "duration": 3.84 }, { "text": "best options out there and so this is", "start": 60.0, "duration": 4.32 }, { "text": "the GitHub page it has nearly 40,000", "start": 61.719, "duration": 5.521 }, { "text": "Stars almost 5 1 half th000 forks and", "start": 64.32, "duration": 6.08 }, { "text": "now they have a super easy to use API", "start": 67.24, "duration": 5.12 }, { "text": "and the way you can think about the API", "start": 70.4, "duration": 4.079 }, { "text": "is it's essentially an extension of the", "start": 72.36, "duration": 5.6 }, { "text": "open AI API and really many projects are", "start": 74.479, "duration": 6.241 }, { "text": "using the open AI API as the standard", "start": 77.96, "duration": 4.24 }, { "text": "and building off of that including", "start": 80.72, "duration": 3.439 }, { "text": "autogen and what that means why that's", "start": 82.2, "duration": 4.76 }, { "text": "so important is it makes private GPT an", "start": 84.159, "duration": 6.201 }, { "text": "easy dropin replacement for chat GPT and", "start": 86.96, "duration": 4.799 }, { "text": "then you get all of this additional", "start": 90.36, "duration": 3.52 }, { "text": "functionality around retrieval augmented", "start": 91.759, "duration": 3.481 }, { "text": "generation so we're going to check out", "start": 93.88, "duration": 2.919 }, { "text": "two things I'm going to show you how to", "start": 95.24, "duration": 3.64 }, { "text": "install the basic user interface and", "start": 96.799, "duration": 3.481 }, { "text": "show you a couple of the settings and", "start": 98.88, "duration": 2.44 }, { "text": "then I'm going to show you around the", "start": 100.28, "duration": 3.24 }, { "text": "API and so the first thing to note is", "start": 101.32, "duration": 4.759 }, { "text": "that the original version of private", "start": 103.52, "duration": 4.279 }, { "text": "GPT", "start": 106.079, "duration": 3.761 }, { "text": "is still active it's called the", "start": 107.799, "duration": 4.161 }, { "text": "primordial version so if you want that", "start": 109.84, "duration": 3.68 }, { "text": "which was launched in May 2023 which is", "start": 111.96, "duration": 3.28 }, { "text": "also the same month that I reviewed it", "start": 113.52, "duration": 3.199 }, { "text": "you can find that here but if you want", "start": 115.24, "duration": 2.919 }, { "text": "the updated version that's what we're", "start": 116.719, "duration": 3.32 }, { "text": "going to be talking about right now", "start": 118.159, "duration": 4.161 }, { "text": "thanks to the sponsor of this video", "start": 120.039, "duration": 4.841 }, { "text": "service now service now enables", "start": 122.32, "duration": 5.28 }, { "text": "businesses to automate a ton of their", "start": 124.88, "duration": 5.48 }, { "text": "processes enabling a more productive and", "start": 127.6, "duration": 6.12 }, { "text": "efficient team and now they offer direct", "start": 130.36, "duration": 5.44 }, { "text": "AI Integrations including Azure open Ai", "start": 133.72, "duration": 3.799 }, { "text": "and service now's own large language", "start": 135.8, "duration": 3.519 }, { "text": "model which allows for an even greater", "start": 137.519, "duration": 4.121 }, { "text": "level of automation thanks to the", "start": 139.319, "duration": 5.2 }, { "text": "generative AI controller and now with", "start": 141.64, "duration": 6.319 }, { "text": "their now assist AI solution you can", "start": 144.519, "duration": 5.921 }, { "text": "layer AI onto every one of your teams", "start": 147.959, "duration": 5.161 }, { "text": "within your business from it a customer", "start": 150.44, "duration": 5.32 }, { "text": "service to HR to developers and just as", "start": 153.12, "duration": 4.199 }, { "text": "an example with now assist for let's say", "start": 155.76, "duration": 4.28 }, { "text": "the customer service team you can", "start": 157.319, "duration": 5.161 }, { "text": "decrease response times summarize cases", "start": 160.04, "duration": 5.0 }, { "text": "gather context more quickly and make all", "start": 162.48, "duration": 5.0 }, { "text": "of your resolution data super consistent", "start": 165.04, "duration": 4.64 }, { "text": "and with now assist for creators you", "start": 167.48, "duration": 4.679 }, { "text": "can actually give them the power of AI to", "start": 169.68, "duration": 4.279 }, { "text": "generate code greatly accelerating the", "start": 172.159, "duration": 3.601 }, { "text": "time to deployment so be sure to check", "start": 173.959, "duration": 3.64 }, { "text": "out service now's intelligent AI", "start": 175.76, "duration": 3.6 }, { "text": "platform to see how it can automate and", "start": 177.599, "duration": 3.321 }, { "text": "improve your business today the link", "start": 179.36, "duration": 3.44 }, { "text": "will be in the the description below and", "start": 180.92, "duration": 4.36 }, { "text": "thanks again to today's sponsor service", "start": 182.8, "duration": 4.719 }, { "text": "now so we switch over to the private GPT", "start": 185.28, "duration": 4.28 }, { "text": "documentation and they really spent a", "start": 187.519, "duration": 4.401 }, { "text": "lot of time on this documentation it is", "start": 189.56, "duration": 4.44 }, { "text": "very thorough and as a developer I", "start": 191.92, "duration": 3.44 }, { "text": "really appreciate that so if we scroll", "start": 194.0, "duration": 2.84 }, { "text": "down we see this quick local", "start": 195.36, "duration": 2.84 }, { "text": "installation steps and that's what I'm", "start": 196.84, "duration": 3.64 }, { "text": "going to be walking you through we're", "start": 198.2, "duration": 4.119 }, { "text": "going to set this up entirely locally", "start": 200.48, "duration": 3.64 }, { "text": "we're not going to use chat GPT at all", "start": 202.319, "duration": 3.041 }, { "text": "so switching over to our terminal the", "start": 204.12, "duration": 3.8 }, { "text": "first thing we're going to do is clone", "start": 205.36, "duration": 4.2 }, { "text": "the repo and before we get started all", "start": 207.92, "duration": 3.12 }, { "text": "of these commands I'm going to put into", "start": 209.56, "duration": 3.16 }, { "text": "a gist I'm going to put them in the", "start": 211.04, "duration": 3.52 }, { "text": "comments below so you don't need to copy", "start": 212.72, "duration": 4.159 }, { "text": "these down as we go you'll find them all", "start": 214.56, "duration": 4.84 }, { "text": "in the gist below so here we go get", "start": 216.879, "duration": 5.64 }, { "text": "clone and then the URL and it's IM", "start": 219.4, "duration": 4.479 }, { "text": "Martinez SL privat GPT and then hit", "start": 222.519, "duration": 3.44 }, { "text": "enter once you have that cloned we're", "start": 223.879, "duration": 4.0 }, { "text": "going to CD into that new directory CD", "start": 225.959, "duration": 4.761 }, { "text": "private GPT now in the documentation", "start": 227.879, "duration": 4.241 }, { "text": "they use Pi M but I'm a big fan of cond", "start": 230.72, "duration": 3.799 }, { "text": "so that's what we're going to be using", "start": 232.12, "duration": 4.56 }, { "text": "today and cond allows you to isolate", "start": 234.519, "duration": 4.241 }, { "text": "your python environments making module", "start": 236.68, "duration": 5.839 }, { "text": "management that much easier so we're", "start": 238.76, "duration": 6.72 }, { "text": "going to type con create DN private GPT", "start": 242.519, "duration": 4.521 }, { "text": "python equals 3.11 and then hit enter", "start": 245.48, "duration": 2.92 }, { "text": "and I already have an environment named", "start": 247.04, "duration": 3.16 }, { "text": "to that so I'm going to go ahead and", "start": 248.4, "duration": 3.24 }, { "text": "remove it and create this new one but", "start": 250.2, "duration": 2.959 }, { "text": "you probably won't come across this", "start": 251.64, "duration": 3.519 }, { "text": "warning all right then we hit enter to", "start": 253.159, "duration": 4.401 }, { "text": "proceed all right from there we're going", "start": 255.159, "duration": 4.281 }, { "text": "to grab this Command right here cond to", "start": 257.56, "duration": 3.199 }, { "text": "activate private GPT we're going to", "start": 259.44, "duration": 3.319 }, { "text": "paste it and that's how we're going to", "start": 260.759, "duration": 3.761 }, { "text": "activate our environment hit enter now", "start": 262.759, "duration": 3.321 }, { "text": "you know the environment is activated", "start": 264.52, "duration": 3.399 }, { "text": "because it says so right there next", "start": 266.08, "duration": 3.839 }, { "text": "we're going to use poetry to install the", "start": 267.919, "duration": 3.84 }, { "text": "UI and the local version and if you", "start": 269.919, "duration": 3.921 }, { "text": "don't have poetry installed you can use", "start": 271.759, "duration": 3.601 }, { "text": "Brew to install it and of course I'm", "start": 273.84, "duration": 3.32 }, { "text": "installing this on a Mac but the", "start": 275.36, "duration": 4.24 }, { "text": "installation process should be quite", "start": 277.16, "duration": 4.24 }, { "text": "similar on a PC I don't believe Brew is", "start": 279.6, "duration": 4.64 }, { "text": "available on the PC but you can just", "start": 281.4, "duration": 5.079 }, { "text": "Google how to install poetry on a PC so", "start": 284.24, "duration": 3.6 }, { "text": "here we go Brew install poetry and I", "start": 286.479, "duration": 3.521 }, { "text": "already have it so I'm not going to do", "start": 289.0, "duration": 3.44 }, { "text": "that next we're going to do

what we said", "start": 287.84, "duration": 6.0 }, { "text": "poetry install d-wi UI comma local hit", "start": 290.0, "duration": 5.52 }, { "text": "enter and that is going to handle all of", "start": 293.84, "duration": 3.68 }, { "text": "the installations for us it's really", "start": 295.52, "duration": 3.679 }, { "text": "really nice and easy all right there we", "start": 297.52, "duration": 3.399 }, { "text": "go everything's installed it looks like", "start": 299.199, "duration": 3.921 }, { "text": "it got installed perfectly we have one", "start": 300.919, "duration": 3.72 }, { "text": "little warning right here but I'm going", "start": 303.12, "duration": 3.359 }, { "text": "to ignore that for now next we're going", "start": 304.639, "duration": 4.0 }, { "text": "to use poetry to run this script and", "start": 306.479, "duration": 3.921 }, { "text": "it's the setup script and one important", "start": 308.639, "duration": 3.481 }, { "text": "thing to note is a lot of the settings", "start": 310.4, "duration": 4.0 }, { "text": "that we use to customize private GPT are", "start": 312.12, "duration": 4.16 }, { "text": "found in this setup script so if you", "start": 314.4, "duration": 3.76 }, { "text": "want to customize anything we can do", "start": 316.28, "duration": 3.359 }, { "text": "that so let's take a look at the", "start": 318.16, "duration": 3.759 }, { "text": "customizations now and if we go to the", "start": 319.639, "duration": 5.081 }, { "text": "settings. yo file this is where we can", "start": 321.919, "duration": 4.84 }, { "text": "actually change the different settings", "start": 324.72, "duration": 3.56 }, { "text": "here for the local model we're going to", "start": 326.759, "duration": 4.521 }, { "text": "be downloading the BLS mistl 7B instruct", "start": 328.28, "duration": 5.639 }, { "text": "model but the documentation also says", "start": 331.28, "duration": 4.639 }, { "text": "that llama 2 works really well so you", "start": 333.919, "duration": 3.801 }, { "text": "can try either of those models and yeah", "start": 335.919, "duration": 3.241 }, { "text": "because those are Cutting Edge open", "start": 337.72, "duration": 3.24 }, { "text": "source models so if you wanted to change", "start": 339.16, "duration": 3.159 }, { "text": "it if you wanted to experiment with", "start": 340.96, "duration": 3.12 }, { "text": "other models this is where you would do", "start": 342.319, "duration": 4.72 }, { "text": "so you can also use Amazon sag maker and", "start": 344.08, "duration": 5.04 }, { "text": "so if you wanted to host your model at", "start": 347.039, "duration": 3.841 }, { "text": "Amazon sag maker this is where you would", "start": 349.12, "duration": 3.519 }, { "text": "enter the endpoint name right here and", "start": 350.88, "duration": 3.84 }, { "text": "if you wanted to use open AI you can do", "start": 352.639, "duration": 3.721 }, { "text": "that right here as well but we're going", "start": 354.72, "duration": 3.479 }, { "text": "to stick with all of the standard", "start": 356.36, "duration": 3.88 }, { "text": "settings for this setup so so switching", "start": 358.199, "duration": 3.801 }, { "text": "back to our terminal we're going to run", "start": 360.24, "duration": 5.36 }, { "text": "poetry run Python scripts SLS setup hit", "start": 362.0, "duration": 5.4 }, { "text": "enter and this may take a little while", "start": 365.6, "duration": 2.76 }, { "text": "because it's actually going to be", "start": 367.4, "duration": 2.4 }, { "text": "downloading the models we need the", "start": 368.36, "duration": 3.279 }, { "text": "embedding model as well as the large", "start": 369.8, "duration": 3.72 }, { "text": "language model and just a reminder the", "start": 371.639, "duration": 3.601 }, { "text": "embedding model is the model that", "start": 373.52, "duration": 5.079 }, { "text": "converts text into Vector storage and", "start": 375.24, "duration": 4.6 }, { "text": "here you can see we're downloading the", "start": 378.599, "duration": 3.521 }, { "text": "mistal instruct model which is about 4", "start": 379.84, "duration": 4.919 }, { "text": "GB a little bit over 4 GB and you know", "start": 382.12, "duration": 4.68 }, { "text": "mistol is one of my favorite models", "start": 384.759, "duration": 4.84 }, { "text": "because it's small it performs extremely", "start": 386.8, "duration": 5.92 }, { "text": "well and it runs easily on my machine", "start": 389.599, "duration": 4.921 }, { "text": "okay that's it that only took a couple", "start": 392.72, "duration": 3.64 }, { "text": "minutes so that's awesome and as a", "start": 394.52, "duration": 5.84 }, { "text": "reminder private GPT is using", "start": 396.36, "duration": 4.0 }, { "text": "llama.el which means that you have to", "start": 400.479, "duration": 5.041 }, { "text": "use GG UF format and any model that you", "start": 402.8, "duration": 4.32 }, { "text": "actually want to test out which is fine", "start": 405.52, "duration": 3.56 }, { "text": "because that's an awesome format and by", "start": 407.12, "duration": 4.24 }, { "text": "default it's using chroma DB as the", "start": 409.08, "duration": 4.48 }, { "text": "local Vector storage all right next we", "start": 411.36, "duration": 4.88 }, { "text": "have to set a few values and this is", "start": 413.56, "duration": 4.479 }, { "text": "specific to a Mac now if you're on a", "start": 416.24, "duration": 3.2 }, { "text": "Windows machine check out the", "start": 418.039, "duration": 3.401 }, { "text": "documentation they talk about what to do", "start": 419.44, "duration": 4.64 }, { "text": "specific to a PC but for the Mac this is", "start": 421.44, "duration": 3.68 }, { "text": "what we're going to be doing and", "start": 424.08, "duration": 2.72 }, { "text": "switching over to the documentation if", "start": 425.12, "duration": 3.919 }, { "text": "you have an envidia GPU here it is this", "start": 426.8, "duration": 4.239 }, { "text": "is what you look for Windows Nvidia GPU", "start": 429.039, "duration": 3.361 }, { "text": "support and then you follow these", "start": 431.039, "duration": 3.44 }, { "text": "instructions and this is the main code", "start": 432.4, "duration": 3.519 }, { "text": "that you're going to be running that is", "start": 434.479, "duration": 3.881 }, { "text": "specific to Windows but since we're on a", "start": 435.919, "duration": 4.761 }, { "text": "Mac here's what we're going to do cmake", "start": 438.36, "duration": 4.0 }, { "text": "args equals and then we're going to say", "start": 440.68, "duration": 4.04 }, { "text": "llama metal on pip install Force", "start": 442.36, "duration": 5.119 }, { "text": "reinstall no cache llama CPP Python and", "start": 444.72, "duration": 4.319 }, { "text": "then hit enter okay it looks like we", "start": 447.479, "duration": 3.56 }, { "text": "actually got some errors tree of", "start": 449.039, "duration": 5.081 }, { "text": "thoughts AER chat streamlit pedals I", "start": 451.039, "duration": 5.081 }, { "text": "don't think these are related to the", "start": 454.12, "duration": 3.68 }, { "text": "project though yeah and looking through", "start": 456.12, "duration": 3.44 }, { "text": "the code base they have no mention of", "start": 457.8, "duration": 4.2 }, { "text": "Trio thoughts AER chat streamlit pedals", "start": 459.56, "duration": 4.199 }, { "text": "so I think this is related to my local", "start": 462.0, "duration": 3.36 }, { "text": "machine these are all projects that I've", "start": 463.759, "duration": 3.28 }, { "text": "try to play around with and now they're", "start": 465.36, "duration": 3.279 }, { "text": "just incompatible so I'm just going to", "start": 467.039, "duration": 3.12 }, { "text": "ignore that I think it's fine you", "start": 468.639, "duration": 3.881 }, { "text": "probably won't see this next we need to", "start": 470.159, "duration": 4.32 }, { "text": "set this variable", "start": 472.52, "duration": 5.76 }, { "text": "pgp profiles equals local make run now", "start": 474.479, "duration": 5.801 }, { "text": "this is a really important step to", "start": 478.28, "duration": 4.039 }, { "text": "follow and I think a lot of people ski", "start": 480.28, "duration": 3.96 }, { "text": "this step so make sure to run this hit", "start": 482.319, "duration": 3.801 }, { "text": "enter okay and I think that's it now", "start": 484.24, "duration": 4.12 }, { "text": "it's all loaded up let's give it a try", "start": 486.12, "duration": 4.32 }, { "text": "there it is private GPT and it uses", "start": 488.36, "duration": 4.48 }, { "text": "gradio for the UI but of course now that", "start": 490.44, "duration": 4.36 }, { "text": "it's a more developer focused product", "start": 492.84, "duration": 4.44 }, { "text": "the point is you can add it to any UI", "start": 494.8, "duration": 4.16 }, { "text": "that you want so let's experiment let's", "start": 497.28, "duration": 3.599 }, { "text": "see if this works so if we look up here", "start": 498.96, "duration": 3.72 }, { "text": "in the top left we see mode we have", "start": 500.879, "duration": 4.16 }, { "text": "query documents now that is the standard", "start": 502.68, "duration": 4.44 }, { "text": "chat with your docs setting then we have", "start": 505.039, "duration": 4.481 }, { "text": "llm chat and that means you just want to", "start": 507.12, "duration": 4.84 }, { "text": "do standard chatting with an llm and it", "start": 509.52, "duration": 4.68 }, { "text": "won't actually do retrieval and then", "start": 511.96, "duration": 4.759 }, { "text": "context chunks is interesting because", "start": 514.2, "duration": 4.04 }, { "text": "that is just what you're getting from", "start": 516.719, "duration": 3.2 }, { "text": "the vector database so if you actually", "start": 518.24, "duration": 3.08 }, { "text": "want to see the data going back and", "start": 519.919, "duration": 3.441 }, { "text": "forth from the vector database select", "start": 521.32, "duration": 3.92 }, { "text": "context chunks so let's switch over to", "start": 523.36, "duration": 3.44 }, { "text": "query documents and we're going to", "start": 525.24, "duration": 3.44 }, { "text": "upload a file I'm going to select this", "start": 526.8, "duration": 4.719 }, { "text": "file which is the autogen research paper", "start": 528.68, "duration": 4.36 }, { "text": "so now we're uploading it it's", "start": 531.519, "duration": 2.88 }, { "text": "processing it which means it's", "start": 533.04, "duration": 2.919 }, { "text": "converting it into a vector database", "start": 534.399, "duration": 3.161 }, { "text": "using the embeddings model and then", "start": 535.959, "duration": 3.281 }, { "text": "we'll be able to use it now as I me", "start": 537.56, "duration": 4.24 }, { "text": "mentioned private GPT is now fully", "start": 539.24, "duration": 4.36 }, { "text": "customizable which means you can set the", "start": 541.8, "duration": 3.76 }, { "text": "chunk size you have a bunch of other", "start": 543.6, "duration": 3.239 }, { "text": "settings that you can play around with", "start": 545.56, "duration": 3.08 }, { "text": "to make sure that you're getting the", "start": 546.839, "duration": 3.841 }, { "text": "best results for your use case there we", "start": 548.64, "duration": 4.8 }, { "text": "go we have it working ingested file now", "start": 550.68, "duration": 4.64 }, { "text": "let's try asking a question okay so", "start": 553.44, "duration": 3.92 }, { "text": "summarize the autogen research paper and", "start": 555.32, "duration": 3.88 }, { "text": "there we go we have a decent summary of", "start": 557.36, "duration": 4.08 }, { "text": "the autogen research paper now again", "start": 559.2, "duration": 3.84 }, { "text": "this is running

completely locally on my", "start": 561.44, "duration": 3.28 }, { "text": "own machine I bet if I tried other", "start": 563.04, "duration": 3.64 }, { "text": "models we might get better performance", "start": 564.72, "duration": 4.04 }, { "text": "and even if we used an open AI model we", "start": 566.68, "duration": 4.12 }, { "text": "might get even better performance now if", "start": 568.76, "duration": 3.92 }, { "text": "we switch over to context chunks let's", "start": 570.8, "duration": 4.279 }, { "text": "see what happens let's do retry and it's", "start": 572.68, "duration": 4.24 }, { "text": "instant and we can look through all the", "start": 575.079, "duration": 4.041 }, { "text": "returning data from the vector database", "start": 576.92, "duration": 4.0 }, { "text": "and of course if we switch over to llm", "start": 579.12, "duration": 4.6 }, { "text": "chat I can just say hello and it's just", "start": 580.92, "duration": 4.68 }, { "text": "like chatting with the mistal model", "start": 583.72, "duration": 4.16 }, { "text": "hello how can I assist you today tell me", "start": 585.6, "duration": 4.76 }, { "text": "a joke why don't scientists trust Adams", "start": 587.88, "duration": 4.72 }, { "text": "because they make up everything so yeah", "start": 590.36, "duration": 4.039 }, { "text": "that's it that is the basic setup for", "start": 592.6, "duration": 3.96 }, { "text": "private GPT and so let's do one more", "start": 594.399, "duration": 3.721 }, { "text": "test I'm going to try uploading the", "start": 596.56, "duration": 3.399 }, { "text": "first book of Harry Potter so we click", "start": 598.12, "duration": 4.24 }, { "text": "upload a file I have it in PDF format it", "start": 599.959, "duration": 4.361 }, { "text": "might be easier to convert it over to a", "start": 602.36, "duration": 4.56 }, { "text": "txt file but let's test it out with PDF", "start": 604.32, "duration": 4.0 }, { "text": "and if we switch over to the terminal we", "start": 606.92, "duration": 3.0 }, { "text": "can actually see the logs and it says", "start": 608.32, "duration": 3.28 }, { "text": "generating embeddings right now so we", "start": 609.92, "duration": 4.0 }, { "text": "can see it working as it goes okay we", "start": 611.6, "duration": 4.16 }, { "text": "can see it's done now let's ask it a", "start": 613.92, "duration": 4.2 }, { "text": "question who is Harry Potter Harry", "start": 615.76, "duration": 4.079 }, { "text": "Potter is a fictional character and the", "start": 618.12, "duration": 3.36 }, { "text": "protagonist of the Harry Potter series", "start": 619.839, "duration": 3.721 }, { "text": "by JK Rowling he is a young boy with", "start": 621.48, "duration": 3.919 }, { "text": "magical abilities who attends hogwart's", "start": 623.56, "duration": 4.16 }, { "text": "School of Witchcraft in magical studies", "start": 625.399, "duration": 4.721 }, { "text": "so likely the model already had that", "start": 627.72, "duration": 4.6 }, { "text": "information but let's try a different", "start": 630.12, "duration": 3.839 }, { "text": "query to make sure that it didn't", "start": 632.32, "duration": 3.28 }, { "text": "already have that information in its", "start": 633.959, "duration": 3.721 }, { "text": "model what is the title of the first", "start": 635.6, "duration": 4.08 }, { "text": "chapter of the first Harry Potter book", "start": 637.68, "duration": 3.68 }, { "text": "the title of the first chapter of the", "start": 639.68, "duration": 3.68 }, { "text": "first Harry Potter book is the boy who", "start": 641.36, "duration": 4.279 }, { "text": "lived and that's correct and if we don't", "start": 643.36, "duration": 4.0 }, { "text": "clear it it will remember our", "start": 645.639, "duration": 4.081 }, { "text": "conversation so we don't have to specify", "start": 647.36, "duration": 4.159 }, { "text": "if we want to keep asking questions now", "start": 649.72, "duration": 3.6 }, { "text": "let's talk a little bit about the API I", "start": 651.519, "duration": 3.56 }, { "text": "switched over to the private GPT", "start": 653.32, "duration": 3.24 }, { "text": "documentation and there's a couple", "start": 655.079, "duration": 3.481 }, { "text": "things that I want to show you first you", "start": 656.56, "duration": 3.92 }, { "text": "can have different settings which is", "start": 658.56, "duration": 3.519 }, { "text": "really nice you can have a version that", "start": 660.48, "duration": 3.76 }, { "text": "runs completely locally you can also", "start": 662.079, "duration": 3.681 }, { "text": "have another version that tests a", "start": 664.24, "duration": 3.159 }, { "text": "different model locally and you can have", "start": 665.76, "duration": 4.48 }, { "text": "another profile that uses an open AI API", "start": 667.399, "duration": 5.481 }, { "text": "so right here our first API endpoint is", "start": 670.24, "duration": 5.399 }, { "text": "ingest and this is a post endpoint and", "start": 672.88, "duration": 5.36 }, { "text": "with that you provide a file and you can", "start": 675.639, "duration": 4.081 }, { "text": "also get a list of the ingested", "start": 678.24, "duration": 3.32 }, { "text": "documents just like that and this is the", "start": 679.72, "duration": 3.679 }, { "text": "completions endpoint and this is the", "start": 681.56, "duration": 4.2 }, { "text": "same exact type of endpoint as open ai's", "start": 683.399, "duration": 4.68 }, { "text": "API and as I said we have a special", "start": 685.76, "duration": 4.04 }, { "text": "guest I'd like to welcome Ivonne", "start": 688.079, "duration": 4.081 }, { "text": "Martinez who is the original developer", "start": 689.8, "duration": 4.64 }, { "text": "of private GPT and also leads the", "start": 692.16, "duration": 4.119 }, { "text": "project today and I have two questions", "start": 694.44, "duration": 4.6 }, { "text": "for him one what inspires you to build", "start": 696.279, "duration": 5.041 }, { "text": "private GPT at first and two what are", "start": 699.04, "duration": 3.599 }, { "text": "some of the coolest features that are", "start": 701.32, "duration": 2.959 }, { "text": "coming up soon when I started playing", "start": 702.639, "duration": 4.481 }, { "text": "around with chbt open Ai apis and llms", "start": 704.279, "duration": 5.281 }, { "text": "in general it became super clear to me", "start": 707.12, "duration": 4.44 }, { "text": "that uh this was a huge opportunity for", "start": 709.56, "duration": 4.36 }, { "text": "the Enterprise ecosystem but when I went", "start": 711.56, "duration": 4.8 }, { "text": "out and asked all the cdos of different", "start": 713.92, "duration": 4.52 }, { "text": "startups uh if they were using this", "start": 716.36, "duration": 3.919 }, { "text": "technology they all said no and the", "start": 718.44, "duration": 4.04 }, { "text": "reason was privacy concerned so at the", "start": 720.279, "duration": 4.321 }, { "text": "same time I realized privacy was a huge", "start": 722.48, "duration": 4.64 }, { "text": "problem I was very active in the open", "start": 724.6, "duration": 4.44 }, { "text": "source community and I knew about", "start": 727.12, "duration": 4.839 }, { "text": "projects like L chain Lama index prb", "start": 729.04, "duration": 4.64 }, { "text": "open source Vector database and then at", "start": 731.959, "duration": 4.161 }, { "text": "some point nomic released GPT for all", "start": 733.68, "duration": 4.959 }, { "text": "these smaller llns that could run on a", "start": 736.12, "duration": 5.0 }, { "text": "CPU of a normal computer and I said okay", "start": 738.639, "duration": 4.361 }, { "text": "maybe all of these can be put together", "start": 741.12, "duration": 4.079 }, { "text": "and that's how private gbt was born I", "start": 743.0, "duration": 5.0 }, { "text": "created a very simple uh CH gbt like", "start": 745.199, "duration": 4.601 }, { "text": "experience where you could chat with", "start": 748.0, "duration": 3.639 }, { "text": "your documents but the important part", "start": 749.8, "duration": 4.2 }, { "text": "was that you did it fully locally so you", "start": 751.639, "duration": 3.921 }, { "text": "could even run it without an internet", "start": 754.0, "duration": 3.72 }, { "text": "connection we are working on a bunch of", "start": 755.56, "duration": 4.12 }, { "text": "things first of all we are adding more", "start": 757.72, "duration": 3.919 }, { "text": "tools to the API we're going to be", "start": 759.68, "duration": 4.0 }, { "text": "adding more data sources like access to", "start": 761.639, "duration": 3.401 }, { "text": "the internet like connection to", "start": 763.68, "duration": 3.279 }, { "text": "databases and you can expect some high", "start": 765.04, "duration": 4.239 }, { "text": "level tools or apis like summarization", "start": 766.959, "duration": 4.081 }, { "text": "or data extraction coming in the in the", "start": 769.279, "duration": 3.401 }, { "text": "next weeks and months then the second", "start": 771.04, "duration": 4.159 }, { "text": "part a standard way of observing what is", "start": 772.68, "duration": 5.599 }, { "text": "going on within the pipelines and also", "start": 775.199, "duration": 5.041 }, { "text": "uh running evaluation to make sure the", "start": 778.279, "duration": 4.281 }, { "text": "accuracy is high enough for your your", "start": 780.24, "duration": 4.44 }, { "text": "production setup and the last bit is on", "start": 782.56, "duration": 3.839 }, { "text": "the setups themselves because you can", "start": 784.68, "duration": 4.88 }, { "text": "set up private GPT in very uh different", "start": 786.399, "duration": 4.841 }, { "text": "ways you can set it up fully local you", "start": 789.56, "duration": 3.44 }, { "text": "can set it up as a single instance in a", "start": 791.24, "duration": 4.039 }, { "text": "gcp for example or you can have it like", "start": 793.0, "duration": 4.72 }, { "text": "in a in a distributed way where an", "start": 795.279, "duration": 4.881 }, { "text": "instance is hosting private GPT API but", "start": 797.72, "duration": 4.0 }, { "text": "then you have the llm running on", "start": 800.16, "duration": 3.239 }, { "text": "sagemaker for example and the vector", "start": 801.72, "duration": 3.0 }, { "text": "database somewhere else so we're going", "start": 803.399, "duration": 2.68 }, { "text": "to be sharing with the community", "start": 804.72, "duration": 3.479 }, { "text": "different setup uh possibilities because", "start": 806.079, "duration": 4.401 }, { "text": "that's where it comes very very useful", "start": 808.199, "duration": 4.241 }, { "text": "because the whole idea of PR gbd is that", "start": 810.48, "duration": 4.479 }, { "text": "is being used in production so I hope", "start": 812.44, "duration": 4.24 }, { "text": "this feels as exciting as it feels for", "start": 814.959, "duration": 3.801 }, { "text": "us all right thanks for joining us Ivon", "start": 816.68, "duration": 3.399 }, { "text": "and if you like this video please", "start": 818.76, "duration": 3.4 }, { "text": "consider giving a like And subscribe and", "start": 820.079, "duration": 5.2 }, { "text": "I'll see you in the next one", "start": 822.16, "duration": 3.119 }]