**Document Fulfillment AIML-20250714\_153021-Meeting Recording**  
**July 14, 2025, 3:00PM**  
**20m 42s**

**Gantla, Sandeep Reddy 0:32**  
Hi, Tim. Good morning.

**Mize, Joseph J 0:39**  
Good morning, everyone.

**K K, Ganesh 0:45**  
Hello everyone. Good morning.

1:19  
Everyone.

**Vautrinot, Kevin**  
Hey everybody.

**Smithee, Adrian 1:26**  
Alright, well Gwin is out today, so I think our main agenda item was to demo the login credential setup. Sandeep, Ganesh, Ajay, are those are those working and ready?

**Gantla, Sandeep Reddy 1:42**  
Yes, uh is is this still everyone or uh is anyone joining the meeting? She wanted to come before we start.

**Smithee, Adrian 1:50**

**Smithee, Adrian 2:26**  
Yep.  
Yep, I can see it.

**Mize, Joseph J 2:29**  
I can see it.

**Gantla, Sandeep Reddy 2:29**  
Um.

**Vautrinot, Kevin 2:30**  
Yeah, we can see it.

**Gantla, Sandeep Reddy 2:32**  
OK sure. So this will be our home page of the screen. We need to register first here. So I already had an account here so I will jump into the login first. So.

**Mize, Joseph J 2:51**  
Now that’s pointing to a local host. Is that something that we’re gonna do in DNS to resolve with something else?

**Gantla, Sandeep Reddy 3:00**  
So as of now it’s my local system. We gonna deploy it in the server. We need some you know permissions and authentications to deploy it in the server. So we are working for them and as well as Michael is working on to you know, procure the server from you know and.  
Register the servers. So once we get all those details, we can deploy it in the servers. So once we get in login as of now if you look at the uh stats, we have no stats because we don’t have any data. So that’s why we are getting the stats as zero. If you look at the documents we have zero document. This is a empty project so I’m uh demonstrating from beginning.

**Gantla, Sandeep Reddy 3:40**  
So once you get into the projects we have now added a feature called select a project. So this is kind of a section we are dividing every project into sub group. So if you have if you are attending multiple meetings right if in one day if you are phone attending 4 meetings and you.  
We wanted to track all four meetings. We need to group them four meetings. So here what we did is we need to create a new project. If you create a project like say print, we are working something like print migration. So I’m demonstrating in the same way.  
So this is a new project. Once we click the new project we will see this print migration is added. So we need to upload the documents of the meeting print migration. Once we click on this it will be showing. We can choose the files which we wanted to store in the print migration project. So if you look.  
OK I’m as of now I’m storing 17 print migration projects so these all 17. So our application will also support the batch processing where it will the document will store in the vector database as well as the metadata will be converting into SQL database. So as of now selected.  
So once I click on the process files, we’ll support the batch processing. So each of these will be parallely working and converting into the vector database and it will store it in the vector database and the metadata of the data like whatever the file name and the title of the.  
Extract the data from the file names and whatever the file name is that we also extract the title name and we will store all this in the SQLite database where we have the all the information. So once we stack like convert this all the metadata and we’ll as the vector database. So once we store in the.  
Vector embedding we can see those in the Faiss vector database. So let just give a couple of seconds while we uploading.  
Once we upload the files, we get a notification. If there is any fails of the uploading, we will get a notification. If there’s is everything is successful, we will get a successful message saying out of 17 files how many files we have uploaded.

**Mize, Joseph J 6:33**  
And this is local and those files aren’t very large. It’s um.

**Gantla, Sandeep Reddy 6:42**  
Files are not much but when it comes to vector embeddings right we are using an open AI embeddings APIs. So this we need to send each request to. If you look at here we have said we have the text we have is length of the text is 2000. So similar way every chunk is converting into vector embeddings. So that means from our text. We are converting into vector numerical format database so it will take. We are requesting for every document we are requesting an API so that will take some time to call the API and get the data, convert those into vector embeddings and also we need to store it in our local vector in database. It takes some time.  
So yeah, now we can see. As of now we can see we have total documents of 17. We get the stats how how we have 97 vector indexes. So that means we have total 97 chunks. In a couple of seconds we have are able to divide it into 97 chunks.  
So this is how it works in the vector embedding. So let’s come to the you know and cannot. I also wanted to add one more project so it will be easy for you. So let me.

**Mize, Joseph J 7:50**  
And I I know Michael wanted to do this and have it automatic from the meetings and then say to a location and then load into here. I notice that you’re doing in a text file. Can teams save that as a text file or can it take it from a doc?  
Uh, a word box.

**Gantla, Sandeep Reddy 8:09**  
It will support both. You can see it supports docs, text and PDF files. You might we can you you if we wanted to store the information right? Whatever the information from the meetings and you wanted to add some more information for the meetings and you can open it docs, text or PDF files you need to up.  
Whatever information you want, you need to know include in the those documents. If you upload those documents you can start working on like if we can ask questions on top of the data. So it’s the automation process. We haven’t started

**Mize, Joseph J 8:58**  
And now is the different. I know docs are a lot larger than text files. Is that gonna make a difference in the time it’s going to take to?

**Gantla, Sandeep Reddy 9:09**  
It’s it’s in the local machine, right? So our local machines are not highly configured machines. So in the servers it will be a little more faster. As of now I have uploaded like 17 in parallel.

**Dubba, Jeevan R 9:21**  
So if you see for the next time daily right when we start using this folder, so it may not take that much time because we are updating only the yesterday meeting document. not all at a time right? It will be stored if all the 17 document will be stored again tomorrow you’ll be uploading only the one document which happened today.

**Gantla, Sandeep Reddy 9:21**  
Yes.

**Mize, Joseph J 9:39**  
Correct. But my question was these are text files are only a couple of kilobytes. it looks like compared to a doc file which may be a couple megs. So I was wondering I’m having a brain fart right now. Sorry when it changes from the one file type to the other file type, is it going to take longer?

**Dubba, Jeevan R 9:40**  
Yeah.

**Dubba, Jeevan R 10:01**  
Yeah, currently we are extracting the meeting notes from the M365, right? So we are copying into the document only and we can go with the document like whatever the document currently we are pushing, we can go in a similar way, hot text, anything will work.  
If you push the different file also it will work format that is notable.

**Vautrinot, Kevin 10:24**  
What about performance though?

**Dubba, Jeevan R 10:26**  
Yeah, it’s worked. It will work.

**Mize, Joseph J 10:26**  
Yeah, that’s what I’m worried about, ‘cause the doc files are gonna be 100 times the size of these text files.  
So even if it’s one, it’s still gonna.

**Gantla, Sandeep Reddy 10:35**  
Just wanted to clear things out. So here it’s it’s not on the basis of the file size. So once you upload the file size it will store it in the local storage where the server is deployed, right? So once it’s stored it in the storage then we read the.  
We have multiple libraries we use for the text file. We directly open and we can read it. For the document we use libraries to the document and the PDF we use by PDF library. So once we open those documents, whatever the text we have right, we extract the text and we convert those text into vector.  
Embeddings. We are using the open AI uh uh APIs where we are not connected directly connecting to the open AI. So we are connecting for um with the API which is United AI Studio is um is um.  
Giving us the APIs and that APIs will be direct to Open AI. So then we get Azure Open AI subscription and we get the response from the Open AI to United AI Studio and United AI Studio to our thing. So it will take some time because that is the IT is built on top of APIs.  
So every API request it needs some time. So we are that’s why it is taking time to

convert the documents. If you look at for the second time I only uploaded 3 documents. So there is information whatever the information in the three documents it is converting into word to vector. So whatever the words we have we were converting it to vectors. So it takes some.

**Mize, Joseph J 11:52**  
Yeah.

**Gantla, Sandeep Reddy 12:11**  
Time to convert and store it in the local database. It’s a vector database which is a fire database which is we are what we are using here and once we store the all the information we can start retrieving the information from the database. I think this makes more clear.

**Mize, Joseph J 12:30**  
I’m sorry for all the questions. I have one more. Usually when you’re doing an API to open AI, you have to use tokens. But if this is local, are those tokens going to be required or is that something that’s bypassed by being local?

**Gantla, Sandeep Reddy 12:32**  
Mhm.  
Mhm.

**Mize, Joseph J 12:48**  
Are we still using tokens?

**Gantla, Sandeep Reddy 12:49**  
No, we we yeah we are we are consuming the tokens from the open AI. So even though it’s in the local or it’s server we are using, we are consuming Azure Open AI’s tokens which will be added to our subscription.

**Mize, Joseph J 13:06**  
OK.

**Gantla, Sandeep Reddy 13:07**  
Yeah. Can we go ahead?

**Mize, Joseph J 13:09**  
And.

**Gantla, Sandeep Reddy 13:12**  
Sorry, can we go ahead with the demo or you have any any more questions?

**Mize, Joseph J 13:19**  
I guess we can go ahead. The only other question is what version are we gonna be using? Like for open AI I never mind. Go ahead. Sorry.

**Dubba, Jeevan R 13:32**  
TPT 4.0.

**Gantla, Sandeep Reddy 13:32**  
OK.

**Mize, Joseph J 13:34**  
Are we gonna use the mini or the turbo or?

**Dubba, Jeevan R 13:36**  
Yeah.

**Gantla, Sandeep Reddy 13:36**  
Mhm.

**Dubba, Jeevan R 13:39**  
Normal one, the 128 cable. Yeah, 12.

**Mize, Joseph J 13:40**  
This is standard. Okay.

**Gantla, Sandeep Reddy 13:45**  
And we are also using the text embedding uh large model from the open AI. It is the API where we convert the mods to vectors.

**Mize, Joseph J 14:02**  
Thank you.

**Gantla, Sandeep Reddy 14:02**  
Yeah. OK, now we have two projects and we have 20 documents in it and we can adding. If you want to add more projects, you can create the projects and we can start asking questions on top of it. So for this if you wanted to ask any questions related to the document, whatever.  
The information we have right. So if you wanted to get all the information for date wise we want to get a summary we can go ahead and choose the provide all meeting summary in the full like full arranged by date so we can ask this question so it will be going to get the same we can ask in any format. It’s not a it’s it’s just sample format you can ask it.

In a different way also it will understand the data and intention of the user and it will be generating the answers. So we can also ask anything related to meetings. So as of now if we ask this um.  
Response we will get it day by day it will arrange according to the day. So from from first date on the day we have the transactional print meeting and we have. So similarly we get all the summary of the date. So each day whatever happens it will be getting the answers and the details. So we can ask any date we can say June 9th or the 5th questions we will also have the follow up questions where you can ask anything you wanted to know more about a meeting on happened on the 9th or June 5th, June 8th any date you can start asking.  
Like, can you summarize?  
Summarize the meeting.

9th So it will extract the information from here like we’re asking about the June 9th, right? So we’ll we’ll extracting the information from the June 9th as the 9th and it will look for the documents which has the date with the 9th and June and it look for the meeting for the Zoom and.  
Like June 9th and whatever the meetings happened on the June 9th it will it will give you a summary. So as of now we have on June 9th we have only one meeting. We got a summary extracted of that one meeting. So this is a summary of that meeting. We can ask anything if there is in the documents if you have any information.  
And whatever the information you have in the documents, whatever the way you are organized and whatever the information it is stored, we can get the information from this.  
You can start asking anything related to how it um any follow up questions. If you want you can ask uh.  
If you have anything from this thing you wanted to know from uh 9th meeting what um Michael has discussed because we haven’t stored it. Um the documents what we have stored right? We don’t have anything like a person’s name it’s like what Michael is discussed what um.  
G1 it has been discussed all these things if we able to you know store it in the document. So then we can ask questions like like related to the person. We also support the person wise we know we will also extract the information from the documents while we are converting into vector databases.  
If there is any information about the persons then we will extract those person’s information and also we store those information in the SQL database. So that in that case we are we are storing the person’s information. If you ask any questions related to the person’s information then the SQL light will.  
Give the um information like there is the document has um the person’s what the person is talking about. Then the retrieval mechanism what we have developed it will combine the.  
SQL as well as the semantic search as well as the keyword search. So the keyword search from the SQL Lite database and the similar semantic search will happen from the fires vector database. We are by combining both of them we will generate the output response. So and the SQL will.

in the vector databases. So we it retrieves the all the change related to the Michael and then it will gather and whatever the change it is written right it will give it to the LM model and the LM model will read the information whatever available in the chains and out of that we will generate the response like this is the user’s intention like user’s intent to wanted to know about this Michael what he’s speaking on the meeting like 9th or.  
Uh last like whatever the question we we was asking right? So uh based on that we will get the this LLM will be summarized whole the thing and give the response. This is how the architecture and the flow works. So as of now I don’t have any clear diagram or something where I can show it and explain.  
Uh, yeah, that’s all from the architecture prospective.

**Mize, Joseph J 19:23**  
Can I ask a question?

**Gantla, Sandeep Reddy 19:24**  
Sure, sure, sure. Go ahead.

**Mize, Joseph J 19:26**  
I noticed in the beginning you started this out as a project. Can you use those project as resources? And when you create your prompt usually in there you can specify resources, so specify additional projects as resources.  
So it can span projects as well or within this.

**Gantla, Sandeep Reddy 19:49**  
Yeah that’s what yes. Yeah we we we divide everything project we store in a metadata what we in the SQL when we are storing right we store the informations

**Gantla, Sandeep Reddy 19:49** *(continued)*  
…like meeting ID, the project. In this context it’s a project project ID and the chunk ID and the file.  
Name all this information, we store it in a structured format so that whenever the users ask questions related to that, we will we will retrieve the information based on the intention. So we have the participants.  
In the schema we have the participants where we store the whoever the participant. If there is any mention in the document right participants, then we extract those information and we store it in the participants list and the date range if you’re not only from the June 9th.  
We can also ask a related question too from meetings from the last month like give me a.  
Um, meetings summaries.

**Mize, Joseph J 20:51**  
Correct. But these these are all from meetings that we input into this project. I’m saying if you have another project with another set of meetings, can you pull those as a resource into here as well?

**Gantla, Sandeep Reddy 20:58**  
Mm.  
So that that will be that will be the next step. As of now we worked on this. So in that case what we can do is we can we can give an option like like iterate or some other symbol. When you put it you will you will get all the list of the projects. So if you want if you get get the print if like if in our case we have 3 projects I think.

**Mize, Joseph J 21:15**  
OK.

**Gantla, Sandeep Reddy 21:30**  
One is default project, print and transaction. If you want to ask in only particular summaries or meetings, some context from particular meeting so you can choose. As of now we have the print migration right? So when you choose at the rate and you will see the list of folders like projects.  
So you need to choose this project and start asking questions. So once you start asking questions and it will it will go to the project and whatever the metadata we store in the SQL database. It will only look for the that project’s information only we will it does not go to the other project’s information and put it in this our project. So it will it it’s complete.  
Isolated from each project to project as well as the user to user. So Michael, you’re storing the documents in your username. So Michael cannot able to access that information because Michael has to upload those information. So each person has his isolated environment so they can start storing whatever the files.  
They want and they can uh ask questions from the this prospect like um the user in

whatever the user wanted to know about the documents you he can start asking questions.  
And we also give an option to file particular file also not only the folder, we also give in particular file to file. If you wanted to only know information from particular files like 123 like 3 files. If you want to know from these three files I wanted to know something something I wanted trying to search something from these three.  
Meetings and you can start tagging those three and start asking only from those files. So it will it will save us you know context and the tokens from the LLM models as well as the you know embedding models. So that’s where the another future if it it you if the team want us to build.  
This future in this environment, so we take some time to build it. but it it is a you know. we can make that thing also.  
I think I’m clear for the question that we have been raised.  
Any question Sandeep?”

**Mize, Joseph J 23:47**  
Uh, I already asked the question. I put the hand down. Thank you.

**Vautrinot, Kevin 23:49**  
OK.

**Gantla, Sandeep Reddy 23:50**  
Mm-hmm. Yes, Kevin, anything you want?

**Vautrinot, Kevin 23:52**  
So I have a lot of like I want to see what happens when you type certain things in. I guess one question I have is like where are we in the process here? Do we want us to like send you prompts or do I like know this is on a local machine. Is there a way that we can get access to it and play around with it? Like what? Where are we in the?

**Dubba, Jeevan R 24:14**  
So let me explain like step by step to everyone. Currently we’re building the model and it is in a local system as a next step site. First I think Mike has already created the Aid ID. So once we get Aid 10 created then we’ll be creating the request on the web server to host this.  
The link in the web server so that everyone can able to get an access to this link, everyone can able to give the prompts, we can able to see that meeting do it as a next steps. Gantla has already sent the details to the Mike on the how to procure the web server and we have no like we need to get the updates from the Mike whether he has raised on that.  
Once we get the web server then the same tool can be hosted in the URL so we can able to keep we can able to access the same thing what something is doing.

**Vautrinot, Kevin 25:04**  
OK, so you don’t need us to give you a thumbs up that says Yep. I we think this works. Let’s go to the next step. The next step is OK, let’s get this hosted in a web server and then we can play around with it, right?

**Dubba, Jeevan R 25:17**  
Yes, yes we need to get the web server so that once you start using this tool and if you have any ideas or thoughts or any changes required then we can able to do from our end. But as of now the tool look like this and good.

**Vautrinot, Kevin 25:24**  
OK.  
OK.

**Dubba, Jeevan R 25:34**  
Say initially we thought of automating from the direct teams to this portal, but there is a human intervention required for one step. Suppose this tool is already having 30 plus documents till today. Tomorrow we have another meeting and that meeting needs to be uploaded. So what you need to do like once?

**Dubba, Jeevan R 25:54**  
When the meeting is completed, just you need to put the two prompts. The prompts initially we have given to the summarized meeting and the six will get the content from the copilot on the 365 of that meeting and we need to paste that file into this full UI.  
So that up to date all the meetings will be there in one place.

**Vautrinot, Kevin 26:19**  
That makes sense.

**Dubba, Jeevan R 26:19**  
Yeah, as the next steps to get access everyone so that once you started using you’ll be having a lot of questions, lot of changes needed As for you then we can able to do that changes recorded.  
OK, that’s all from our end today. So Sandeep, what you’re going to work for tomorrow or the next thing in the tool?

**Gantla, Sandeep Reddy 26:43**  
Jeevan, for deployment purpose we need to, you know, work with the, you know, deployment team. So they we need to work with them and we need to get the access for the deployment servers and we need to deploy it in the servers. So we need to set up an environment in the deployment.

**Dubba, Jeevan R 26:57**  
Yeah, for that. Uh, for that.

**Gantla, Sandeep Reddy 27:00**  
Mhm.

**Dubba, Jeevan R 27:01**  
Yeah, for that work we need to get access to the web server. So let’s

connect with Mike tomorrow, like what is the update on the web server. Once we have the web server, then we can connect with the deployment team and we can make you able to host this UI.

**Gantla, Sandeep Reddy 27:03**  
We.  
Mhm.  
Sure. Jeevan.

**Smithee, Adrian 27:15**  
Are you able to tell based on this this HCP console what stage it’s at? Trying to look in here and see this is there’s a request I think.

**Dubba, Jeevan R 27:26**  
This is the AID ID. I think we sent out the details Ganesh to procure the web server with this AID ID and I don’t know like whether he has requested. This is not the tool where we need to request the web server. There is a separate.

**Smithee, Adrian 27:42**  
OK, OK, got it, got it.

**Dubba, Jeevan R 27:43**  
This is only for the eight ID, yeah.  
And do have the data ID with us, Ganesh or Sandeep?

**K K, Ganesh 27:51**  
No, Jeevan actually. I think it should be in the recording. So Michael actually created it.

**Dubba, Jeevan R 27:57**  
Yeah. can you share Adrian again the screen at least or type that 880 in the chat so that it will be useful for social.

**Smithee, Adrian 28:04**  
Yeah, so I’ll put the link right in the chat.

**Dubba, Jeevan R 28:06**  
Um.. not link. A link will not open us like only the eight ID. If you see the number eight ID, yeah a AID under score some number is there. Just put that in the chat.

**K K, Ganesh 28:07**  
In the recording no other way.

**Gantla, Sandeep Reddy 28:07**  
Michael already see you.

**Smithee, Adrian 28:10**  
Oh, OK, OK. You just want the number. Yep, you bet.

**Gantla, Sandeep Reddy 28:18**  
Even already Michael has been shared that information.

**Dubba, Jeevan R 28:21**  
Yeah, yeah I need this. Yeah, OK.

**Mize, Joseph J 28:24**  
Now, is this installed just a custom install or is it like a Kerberos or a docker or?

**Dubba, Jeevan R 28:32**  
No, currently we are requesting the web server to host this link. As of now we are not requesting anything in any additional info.

**Mize, Joseph J 28:43**  
Yeah, you said you’re requesting it to host a link, but it has to have a server, it server.

**Mize, Joseph J 28:46**  
it the install. Is that like just a regular Linux install or is it a docker? A Kerberos? What type of?

**Dubba, Jeevan R 28:46**  
Yes.  
Yeah.

**Mize, Joseph J 28:57**  
The VMS or uh, hosting environment.

**Gantla, Sandeep Reddy 29:02**  
So it might be a Windows, yeah.

**Dubba, Jeevan R 29:03**  
So yeah, with Windows, yeah.

**Mize, Joseph J 29:06**  
It’s Windows OK.

**Dubba, Jeevan R 29:07**  
Yeah.  
OK, that’s all from my end and let’s discuss tomorrow on the next steps.

**Gantla, Sandeep Reddy 29:22**  
So any questions from anyone? If not, we can drop.

**Smithee, Adrian 29:31**  
Alright, thanks everyone.

**Dubba, Jeevan R 29:31**  
OK.