

# PriceSenz

## Large Language Model Based Talent Match Making Engine

Team 3: Umar Ali-Salaam, Ann Biju, Whitney Humecky, Philmore Kuong

Mentor: Dr. Pankaj Choudhary

Company Sponsor: PriceSenz - Muhammed Shaphy, Hazim Kuniyil, Bijith Moopen

### Abstract:

It has been a long applied technique to filter applicant resumes with keywords to narrow down the number of candidates a hiring manager would need to review. This is a reasonable, yet flawed, method. A qualified candidate may not only use *just* the right keywords to overcome getting filtered out. This is a loss for both the candidate and the hiring organization. Large Language Models have experienced major improvements in capability and quality in recent years. An LLM has the potential to apply context to the prompt, and return an answer that doesn't depend on specific words, but the specific meanings of many words. An LLM powered resume and job search tool can break through the limitations of exact word matching, and provide a more equitable, practical, and valuable experience to all parties involved.

## TABLE OF CONTENTS

1. [INTRODUCTION](#)
2. [RESOURCES](#)
3. [KEY ROLES](#)
4. [COMMUNICATION PLAN](#)
5. [TIMETABLE](#)
6. [IMPLEMENTATION DETAILS](#)
7. [ISSUES AND LESSONS LEARNED](#)
8. [EVALUATION](#)
9. [ETHICS DISCUSSION](#)
10. [CONCLUSION](#)
11. [FUTURE WORK](#)
12. [CONTACT INFORMATION](#)
13. [APPENDIX](#)
14. [SIGNATURES](#)

## 1. INTRODUCTION

What do talent acquisition professionals do when they have a thousand applicants for one position? Historically, they've used keyword filters to reduce applicant resumes to only those which mention specific skills and experience they are looking for. The problem with this is that there are many ways to say the same thing. It is not practical to tailor these keyword filters by hand to account for all of the possible ways someone could describe a candidate's experience, research experience, hard skills, etc. This is why Generative AI tools need to be developed with talent acquisition in mind. People who need jobs, and companies who need employees, need better solutions.

PriceSenz, a talent solution provider, is looking to develop a Generative AI powered candidate to job match tool. The tool is able to connect qualified candidates' resumes to job descriptions. The product in development by UTDDiscovery Team 3, as subsequently described, will be for internal use by PriceSenz talent acquisition specialists, who are currently unable to review every resume in their database.

## 2. RESOURCES

### Software

#### Google Colab

Colab is like Google Drive for writing code. It allows users to share updates to scripts in real time, and run the script in cloud computing space. The team used Google Colab to share development progress as a team and collaborate on debugging, different approaches to the goal, etc.

#### HuggingFace

HuggingFace houses AI models, discussion and collaboration spaces for users, datasets and cloud computational resources. The team utilized an embedding model, BAAI/bge-small-en, an open-source Flag Embedding that can map any text into a

low-dimensional dense vector. These vectors are used in a similarity search against the database to retrieve an initial amount of resumes for the LLM.

### LLaMA - Large Language Model by Meta AI

Large Language Model developed by Meta. LLaMA is trainable via fine tuning or a method called retrieval. LLaMA can be communicated with by an API (application programming interface) in a chatbot style of query and response. The model used here is llama-2-13b-chat, which is a 13 billion parameter fine-tuned model for chat completion. This model is a middle ground between LLaMA 2's 70b (the most accurate) and 7b (the fastest model) chat model.

### Pinecone

Pinecone, an information retrieval system, provides vector database storage for LLM models to access remotely for Retrieval Augmented Generation. Retrieval Augmented Generation is essentially introducing outside domain-specific knowledge to an LLM. In this use case, resumes from Pinecone are introduced to the LLM so it has a knowledge base that it can work with. Pinecone stores the resume data the model accesses, reads, and returns to the user in alignment with the request. The project utilizes the data from Pinecone in conjunction with LLaMA to create the NLP solution.

### Replicate

Replicate provides cloud space for open-source machine learning models. Anyone can push new models or train existing models on Replicate's servers. The team accesses a specific iteration of LLaMA, llama-2-13b-chat, through Replicate.

## Hardware

### Dell personal computers

Used to access resources described above. No on-device development or deployment is expected or recorded.

## Experts

Sponsor Hazim Kuniyil:

PriceSenz Mentor, assists when the team was slowed down by learning new tools via self-study such as LangChain implementation and component connections.

Faculty Advisor Dr. Pankaj Choudhary:

University of Texas at Dallas, Professor - Mathematical Sciences and Associate Dean of Graduate Studies. Dr. Choudhary assisted with technical knowledge as needed and professional development of team members and relationship between the sponsor and the team.

### 3. KEY ROLES

“Lead” - delegate responsibility, define subtasks of the larger goal, manage progress, resources, timeline, etc for section.

Umar Ali-Salaam: Weekly reports, retrieval lead.

Ann Biju: Minutes, research lead.

Whitney Humecky: Meeting Organizer, documentation of progress.

Philmore Kuong: Prompt engineering lead.

### 4. COMMUNICATION PLAN

The team communicated via direct message group chat and in person at biweekly working sessions. Working Sessions were held 5:00 to 7:00 pm on Tuesdays and Thursdays in a regular meeting location and through Microsoft Teams from 3:00 to 4:30 pm on Tuesdays. Absences from team or mentor meetings were communicated in advance.

Company Sponsor and Faculty Advisor were updated on progress, each Thursday at 1:00 pm in a recurring Microsoft Teams meeting. When Team 3 needed assistance on any portion of the expectations or tools, the team contacted Hazim Kuniyil of PriceSenz. Kuniyil provided both his resources and his time to assist the team.

## 5. TIMETABLE

8/31-9/14: Research and learning

Studied resources suggested by PriceSenz on necessary tools for project completion.

9/14-10/12: Work on retrieval

Ran a successful prompt and response algorithm that trains the LLM with feedback.

Should incorporate LangChain - a python package that can be used to retain context for the LLM to consult when responding to a prompt.

10/12-11/9: Language generation and prompt engineering

The chatbot (local, program internal chat interface before API implementation for testing purposes) developed returned readable resumes that are appropriate to the job description. System prompts (the base prompt for the LLM to know what its objective is overall) were refined to produce high quality outcomes.

11/9-12/7: Testing and refinement

All components are tied together and complete, the LLM is able to access the Pinecone data, and retain context so the user can provide feedback on its response - LangChain, the user is able to utilize the end-product to prompt the LLM and have the desired number of resumes that fit the desired qualifications in plain English.

## 6. IMPLEMENTATION DETAILS

The model essentially follows a 3 step process Resumes Retrieval, Map Reduce, and a Gen. AI

## procedure

1. Resume Retrieval:
  - 1.1. Embed a job description using HuggingFace SentenceTransformer model: BAAI/bge-small-en
  - 1.2. Query for data with a similarity search from Pinecone using the vector from step 1.1
  - 1.3. Process the extracted resumes to remove excess symbols, extract names, and parse into LangChain's Document type
2. Map Reduce
  - 2.1. Split the Document of resumes into smaller chunks. Due to token limits with LLMs, RecursiveCharacterTextSplitter, from LangChain, creates smaller chunks that are within the token limit without having to sacrifice context (losing words), to Map. Smaller chunk sizes lead to a longer processing time but allows for larger outputs and while larger chunk sizes do the contrary.
  - 2.2. Input the desired prompt for mapping, in other words, the information wanted from each resume
  - 2.3. Input the desired prompt for reducing, this is how the LLM will rank each resume. The prompt can be adjusted to rank each resume by any metric desired.
3. Generative AI
  - 3.1. By feeding the ranking into the LLM which also includes the mapped information from 2.2, the user can then ask the LLM about information from the Map Reduced resumes. Some of the capabilities demonstrated within this model include:
    - 3.1.1. Retrieval Augmented Generation
      - 3.1.1.1. The model can gather accurate information from the knowledge provided to it. If asked specific questions about the resumes, the LLM will answer correctly.
    - 3.1.2. Chain-of-Thought Prompting
      - 3.1.2.1. Our LLM can follow logic and breakdown steps of reasoning to output an accurate response that is inline with those steps of reasoning. For example, if we ask "If 3 out

of 4 people ranked 1st, 3rd, and 4th, where does the fourth person rank?” our LLM would respond with “2nd”.

3.1.3. Conversational Memory

3.1.3.1. The model we created is also able to remember previous conversations with it. We can ask our model about questions or comments we made to it previously and return the appropriate response

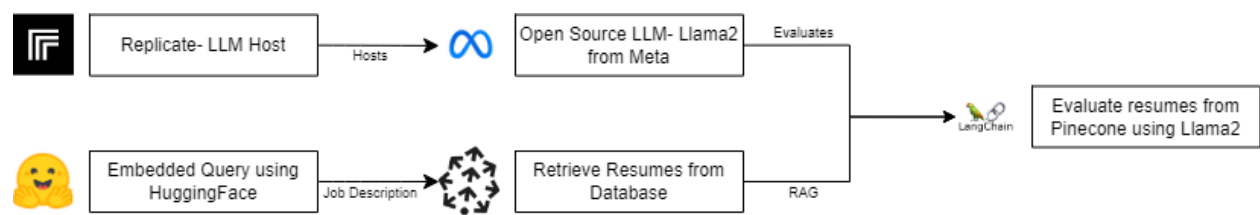


Figure 6.1: Data Flow Diagram

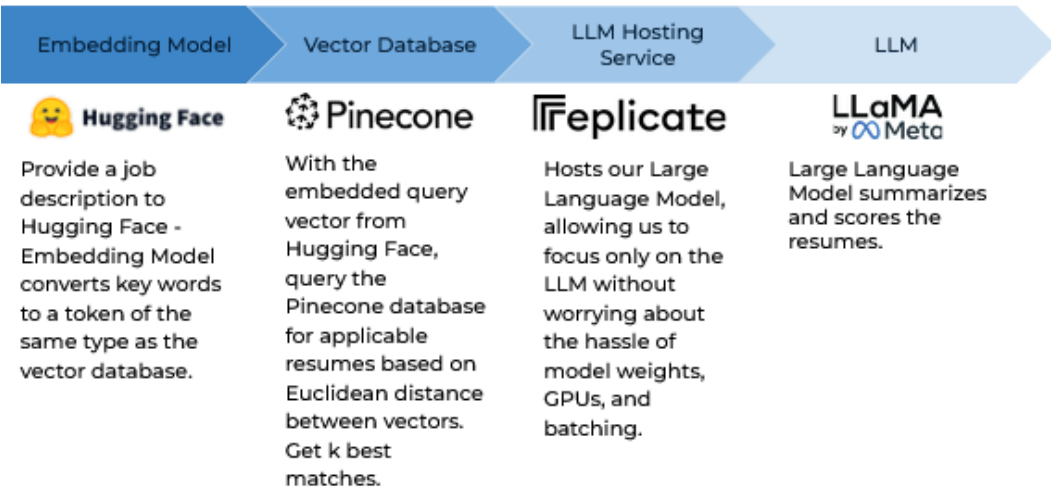


Figure 6.2: How the Model Works

7. ISSUES AND LESSONS LEARNED

Data input character length: Use MapReduce summarization technique to contextualize bigger sentence chunks. Even using MapReduce, we would situationally get a “too many tokens” error.



To remedy this, we utilized RecursiveCharacterTextSplitter which essentially splits our document containing all the resumes we queried from Pinecone into chunk sizes of our choosing. Chunk overlap indicates if you want any overlap between chunks, in this case we did since maintaining an overlap within chunks helps the LLM keep context.

```
text_split = RecursiveCharacterTextSplitter.from_tiktoken_encoder(  
    # the smaller chunk size the longer it takes to run  
    chunk_size = 300, chunk_overlap = 2)  
docs = text_split.split_documents(docs)
```

LLM Hallucinations: Our first approach to hallucinations (when the LLM outputs information that it has made up since it does not actually know the answer), was to lower our LLM's temperature. However even after setting the temperature to the lowest possible (0.01) our LLM was still giving us hallucinated results. We then began to try different ways of reducing hallucinations such as changing our system prompts. System prompts are templates containing instructions for how the LLM should act and respond to user inputs. We explicitly instructed our LLM to only use the information we provided to produce its output while also providing it the information within the prompt so it would know where the information is. By doing so, we were able to eliminate the problem of hallucinations as all the outputs from our LLM after doing this were accurate to the original resumes.

Lack of supervised training data: Improvised with job descriptions available online and evaluated response per team's judgment.

## 8. EVALUATION

The project's success is defined by PriceSenz's customers, both those looking for a job and those hiring, are able to be more efficiently matched to jobs than current methods. The team doesn't have access to the user analytics data of PriceSenz, so it's up to PriceSenz to relay the project's success to the team. With that being said, user analytics such as placement rates (the rate that suggested employees are hired), time to hire (how long it takes for a company to find

and hire an employee), and comparative analysis (compare stats to previous methods) are integral in determining the project's success. In terms of our own evaluation of success with respect to this project is the accuracy of our model. Within the times we evaluated our model's accuracy, such as, testing the successfulness of Retrieval Augmented Generation, Chain-of-thought prompting, and conversational memory, we were able to get results that directly matched the resume which the output came from. This suggests that we were able to successfully integrate RAG within an LLM.

## 9. **ETHICS DISCUSSION**

There were a few ethical dilemmas discussed amongst the team before and during the development of the chatbot. Which includes: The potential of bias in the model's training and tuning that causes a pattern of exclusion of certain types of candidates from being matched to jobs. The team also had to consider the potential of chatbots similar to this one that can reduce the amount of recruiting jobs in the job market for people working in Human Resources. Another dilemma is transparency; if a candidate is unaware as to why their resume will or won't be selected it can cause their job search to be hindered or challenging.

## 10. **CONCLUSION**

This project has the potential to significantly reduce work-hours per employee as it would be able to automate the sifting-through resume process which is currently being done manually. For example, by reducing the time it takes to find the top 10 resumes for a specific job or vice versa, recruiters would be able to manually refine that list leading to an overall better quality in candidates or simply move forward with the model's recommendations. Furthermore, this product could be hosted to the sponsor's clients as a product that could also have a positive impact on their recruiting process. Therefore the project can have an impact on both the internal-facing and customer-facing side of PriceSenz's business.

## 11. FUTURE WORK

Develop applicant-to-job counterpart tool: Similar to the existing capabilities of the product Team 3 developed, but in a reversed flow of data - an applicant to job tool would be able to evaluate a resume and compare it against all open job postings and recommend which few to apply to.

Build out tool for chat-bot like ease to use: Further develop the lang-chain component so that feedback can be provided from the back end user to access job data.

Integrate into existing tools for talent acquisition specialists of PriceSenz: Build API layer so the tool can be integrated into a website or apps with an easy to use user-interface.

Test different LLaMA versions. LLaMA has other fine-tuned models such as 13B which is known to outperform GPT-3, perhaps different models have better performance compared to ours which may be significant for accuracy and processing time.

## 12. CONTACT INFORMATION

### **Umar Ali-Salaam:**

ualisalaam@gmail.com : <https://www.linkedin.com/in/umar-ali-salaam/>

Soon to be Data Science B.S. graduate, who uses their strong background in Graphic Design and Data Science to create meaningful digital user experiences. Currently leading a team of web developers at Fuel the Future, Umar excels as a UX Designer, bringing a wide skill set to ensure excellence in both user-centered design and business goals.

### **Ann Biju:**

annbiju7@gmail.com : <https://www.linkedin.com/in/annbiju/>

Ann is a soon to be Data Science B.S graduate, who hopes to use the practical and theoretical knowledge she's gained thus far to continue contributing to the world of Data as well as exploring the fields of Artificial Intelligence/Machine Learning.

**Whitney Humecky:**

whumecky@gmail.com : <https://www.linkedin.com/in/whit-humecky>

UT Dallas Data Science Senior and Oncor Electric Delivery Data Analyst, Whitney Humecky is motivated by how ethical data can power more efficient business operations and a higher quality customer experience.

**Philmore Koung:**

philmore.koung@gmail.com : <https://www.linkedin.com/in/philmorekoung>

Graduating this Fall with a B.S. in Data Science, he enjoys utilizing his skills to create impactful real-world solutions. Philmore will be pursuing a Master's Degree in Mathematics with a concentration in Data Science in Spring 2024.

## 13. APPENDIX

Langchain:

[https://python.langchain.com/docs/get\\_started/introduction.html](https://python.langchain.com/docs/get_started/introduction.html)

RAG:

[What is retrieval-augmented generation? | IBM Research Blog](#)

[Retrieval | !\[\]\(830769b31eeeaca920791081939ff8ba\_img.jpg\) !\[\]\(198f559926258ddfad814817bda0ffbc\_img.jpg\) Langchain](#)

Chain of Thoughts:

[Tree of Thought \(ToT\) example | !\[\]\(8bba887393ca45b761e5cb49e755e762\_img.jpg\) !\[\]\(b898b980f2d860cdb0237afbc3664529\_img.jpg\) Langchain](#)

Graph Retrieval:

[Getting Started with Neo4j - Developer Guides](#)

HuggingFace:

<https://huggingface.co/BAAI/bge-base-en>

LLaMa:


<https://replicate.com/blog/all-the-llamas>

MapReduce:


[https://python.langchain.com/docs/use\\_cases/summarization](https://python.langchain.com/docs/use_cases/summarization)

## 14. SIGNATURES

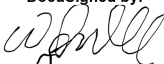
Electronic Signatures (Adobe/DocuSign) are accepted/encouraged.

DocuSigned by:  
  
92517530976343F...

*Umar Ali-Salaam, Team Member*

DocuSigned by:  
  
095094D00055400...

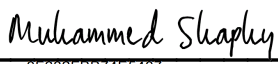
*Ann Biju, Team Member*

DocuSigned by:  
  
8868A64G76704F1...

*Whitney Humecky, Team Member*

DocuSigned by:  
  
7D19B7CF4074453...

*Philmore Kuong, Team Member*

DocuSigned by:  
  
0F682FBB74E5407...

*Muhammed Shaphy, Company Mentor*

DocuSigned by:  
  
502414AF76074B5...

*Pankaj Choudhary, Ph.D., Faculty Advisor*