## Retrieval Augmented Generation for Legal Journals

Sheeba Moghal, Shingai Nindi, Powell Sheagren

### Introduction

In a now infamous filing last June, lawyers representing a man who claimed to have suffered physical damages during a plane flight submitted a claim in part generated by AI to court. The claim used pre-existing cases to show precedent and all together seemed to have a coherent argument. There was only one problem, the judge and opposition couldn’t find the cases referenced and, even after the lawyers followed up with ChatGPT, they were unable to prove their existence or validity. They claimed to be using the AI tool as a search engine and it cost them the case and some hefty fines for negligence. In this project, we hope to provide an alternative to this legal malpractice by using a retrieval augmented generation (RAG) model to build on a pre-existing LLM such that accurate and correct information can be found through searches.

Law is just one of the many fields in which LLM ‘hallucations’ can cause tangible harm and lead to negative outcomes for users. The LLMs never claimed to be able to be experts on all topics but they do offer a framework for others to build specialized systems to retrieve specific information rather than simply predict the next word in a series. Since so much of legal work is research into precedent, we thought it would be an apt place to apply RAG so that a future model can save lawyers time while not costing them a drop in accuracy.

First, we will discuss the data that we sourced to complete this task, then we will discuss the methodology behind our models and the tools we used. Afterwards, we will show the results of our models and test their validity. Finally, we will return to the initial problem of researching legal cases and show the efficacy of the method to complete that task.

### Methodology

#### Data Sources

For our project, we were lucky that there are quite a few pre-existing repositories of legal filings and documents. This is mostly due to the fact that lawyers need to be able to access the records as easily as possible to ensure they can find relevant sources and we are lucky that this ease of access has led to a few important corpus’ which we could have used. We ended up primarily choosing the Caselaw Access Project (case.law) which is a website for a team which is hoping to make all published US court decisions available and public. We specifically used the Alaska cases to limit the amount of data we’d need to use and to offer a specific case. The initial dataset was modified from their original structure to fit our models, allowing them to be called upon and represented in a way that will improve user experience in the final project. While we did consider using multiple sources of data at the same time, it became clear that the merging of the data would be more tiring although we are not ruling it out for future more refined models with increased computational power.

#### Programming

All of the programming was done in python with traditional data wrangling and cleaning packages like pandas and numpy. When working specifically on the model, experiment one utilized BERT and more specifically colBERT as a tool to find document similarities in a variety of ways which will be explained later. Another one of our models used FAISS in a similar way and we used huggingface as a package amongst other packages used for language transformation.

### Experiments

We started by running three different approaches for the retrieval augmented generation to compare and contrast validity and results: ColBERT, FAISS, and TF-IDF. After some work, we decided to scrap the TF-IDF approach and focused on the other two. I will now break this section description into two to discuss the methods and processes we used for each.

#### ColBERT:

We began by importing data from case.law and standardizing it for efficient querying and retrieval in the future. This mostly involved managing the metadata so that formatting was in an ideal format and documented properly. The next step involved determining what a query would look like, what questions to ask, and what format the responses would be in. We decided to create multiple fields to look up different aspects such as Name, Legal Abbreviation, Date, Jurisdiction, and a custom option. The query was then preprocessed by cleaning the text and pulling out specific relevant information. Next we created a system that queried the data based on the input through matching dates, names, or other information referenced.

Given this initial prompt, we tokenize and embed it using HuggingFace’s bert-base-uncased which also includes a context step and relevancy fine tuning. This transforms the query into a dense vector representation which is fed into a ColBERT (Contextualized Late Interaction over BERT) model which computes cosine similarity with other documents to ensure relevance and semantic alignment. With this dense retrieval system in place, we needed to add a generative aspect to it to return useful information not just a document number. In order to do this we used BERT traditionally, conditioning on the documents found and using it for an encoder-decoder transformer architecture which creates a summary with the decoding output. This information is then passed back out of the system, providing discrete information about the document and then prompting for a summarization which is generated as mentioned. A few options are also shown in order to show not just the closest but the cluster that are nearest to give more options.

We built and trained this model on a local machine which slowed down the processing but once everything had run we were satisfied with the results. The ColBERT Similarity measures were able to find documents that were relevant to the queries and the evaluation metrics (f-score, precision, and recall) were all high for at least one of the documents retrieved. There were a few limiting factors though. The data was inconsistent at times and the model could not account for abbreviations; also the general purpose embeddings may miss domain specific information. Additionally, we were working on a system with a fixed token limit and scale and so further work might need to be done to advance the model by expanding those values.

#### FAISS

Similarly, for this model we imported data from case.law and then embedded, tokenized it to ensure a dense representation. Again most of the embedding was done on the metadata rather than the text of the file itself as we are looking for topline similarities. For the next step, we indexed all of the data using the Facebook AI Similarity Search (FAISS) which created a model that we’d end up using for the retrieval aspect of the problem. For this model, we used a very similar query structure with a Name, Legal Abbreviation, Date, Jurisdiction, and custom option for people to look up topics and cases. This was to create more similarities between the two models which would allow for an ease of comparison.

Again in a similar way, we prompted for a query in one of the categories which was also tokenized and embedded in a similar way to the metadata itself. With this representation, we passed it into the FAISS indexing which has its own methods of determining similarity and then printed out the cases which were closest to the prompt in addition to a brief snippet of the case itself for reference. This information is then passed back to the user and there are again multiple options shown to give access to other documents in the set.

After constructing and training the model as mentioned above, we were successfully able to get a prototype running which produced the top 5 documents nearest to the query. We were satisfied with its performance although there is room for slight improvements in similar places to the limitations placed upon our first model.

### Discussion:

With our two models established, the next step was to evaluate the difference between the two and compare results on the same prompts. For accuracy metrics we used precision, recall, f1-scores, and normalized discounted cumulative gain (nDCG) and evaluated the results on the first few best options rather than just the first. Here is a table of results for two of the prompts we used, for reference the k is the ranking of similarity of the prompts

**ColBERT - “What about Hillyer”**

| Query | k | Precision | Recall | F1-Score | nDCG | Match Type |
| --- | --- | --- | --- | --- | --- | --- |
| What about Hillyer | 1 | 1 | 1 | 1 | 1 | Partial |
| What about Hillyer | 3 | .33 | 1 | .5 | 1 | Partial |
| What about Hillyer | 5 | .2 | 1 | .33 | 1 | Partial |

**FAISS - “What about Hillyer”**

| Query | k | Precision | Recall | F1-Score | nDCG | Match Type |
| --- | --- | --- | --- | --- | --- | --- |
| What about Hillyer | 1 | 1 | 1 | 1 | 1 | Partial |
| What about Hillyer | 3 | .33 | 1 | .5 | 1 | Partial |
| What about Hillyer | 5 | .2 | 1 | .33 | 1 | Partial |

**ColBERT - “Case on McIntosh”**

| Query | k | Precision | Recall | F1-Score | nDCG | Match Type |
| --- | --- | --- | --- | --- | --- | --- |
| Case on McIntosh | 1 | 1 | .5 | .67 | 1 | Partial |
| Case on McIntosh | 3 | .67 | 1 | .8 | 1 | Partial |
| Case on McIntosh | 5 | .4 | 1 | .57 | 1 | Partial |

**FAISS - “Case on McIntosh”**

| Query | k | Precision | Recall | F1-Score | nDCG | Match Type |
| --- | --- | --- | --- | --- | --- | --- |
| Case on McIntosh | 1 | 1 | .33 | .5 | 1 | Partial |
| Case on McIntosh | 3 | .33 | .33 | .33 | .8 | Partial |
| Case on McIntosh | 5 | .2 | .33 | .25 | .7 | Partial |

To summarize, the two models were both successful but on non-specific prompts the ColBERT Model outperformed the FAISS one on all metrics which means its what we recommend be used for the problem going forward. Additionally, it should be mentioned that when new data, or in our situation cases, are added then the FAISS model would need to be re-run to index everything while the ColBERT model can scale without added computation which is another reason to recommend it.

In the future, we do hope to create the potential to scale out such systems to other state’s legal filings as that was a limiting factor in our data. Also the BART model used to summarize lost some context of the documents due to its window so the model could be improved with a larger context. We hope to add functionality for more increased queries instead of the current options system and integrate it into a larger legal workflow so that it is easily accessible by lawyers who need the tool. There are all implications for the future but we are happy with our model as a proof of concept.

### Conclusion

As we have shown, AI can be effectively used to expedite the research process of law work especially when it comes to citing proper sources and finding ones relevant to the current case. We discussed the problem which we hoped to address, walked through our methodology and the systems that we used, and then described the two models we built before discussing their successes and failures. We found that both models succeeded well in pulling relevant information from the documents and, apart from stylistic differences, the end results looked very similar.

To return to the initial new story, our tool would really only help access sources which were not hallucinated by the large language model, the rest of the malpractice is not addressed but future works which combine retrieval augmented generation with other full texts of legal filing could fill in this gap. As machine learning techniques move forward, it's important to make sure that mechanics are put in place at each step to ensure quality of output and accuracy of the information. If not, AI companies may be in place to be sued themselves and the lawyers will be sure not to use ChatGPT for research that time.