Using RAG to Build a Conversational Model on the Criminal Code of the Republic of Armenia

Authors: Armine Papikyan, Khoren Movsisyan

BS in Data Science

Supervisor: Vahan Petrosyan

CEO at SuperAnnotate

Abstract

Large language models (LLMs) have been the talk of the tech town since the introduction of the transformer architectures. Recently, OpenAI launched GPT agents, which are conversational assistants designed for specific areas. Our paper presents a conversational agent that focuses on Armenian Criminal Law. We have experimented with different models and techniques to adapt the assistant to legal data, finding some methods more effective than others. Our model enables the user to ask questions about laws or cases and get straightforward, clear answers. This project aims to make it easier for people to access and understand legal information, simplifying the process of getting legal help.

Introduction

Artificial intelligence has greatly advanced in recent years, significantly impacting various industries, including the legal sector. The development of transformer architectures, highlighted in the influential paper "Attention is All You Need," [1] has changed how machines understand and generate text that resembles human communication. This breakthrough has led to the creation of advanced language models like ChatGPT, which have not only garnered massive investments but also encouraged the development of more specialized applications.

Among the innovations in this field are the GPT agents introduced by OpenAI, which are tailored to provide expertise in specific domains. In the scope of this project, we have developed a conversational agent that could serve the specific needs of those interacting with Armenian criminal law. The necessity for such a specialized tool is the following: legal professionals, students, and the general public often face hurdles when they need to understand or retrieve information from legal texts, which are typically dense and complex. In order to find information about the legal sector or a specific case, one has to go through hundreds of law articles manually which sometimes are also overwhelming in terms of heavily technical legal terms.

Our project addresses these challenges by creating a user-friendly platform where one can ask questions about Armenian criminal law and receive clear, accurate responses.

The accuracy is backed by the Retrieval Augmented Generation (RAG) [2] system that the assistant is based on. By RAG, it's able to retrieve the document that has closest similarity with the query. The RAG system uses a curated JSON file containing titles and content of law articles as its knowledge base.

We have also curated a list of question-response pairs, incorporating case-specific scenarios into our fine-tuning dataset.

Our testing phase involved deploying various OpenAI and Mistral models, assessing their performance by analyzing how well they handled specially curated questions aimed at evaluating the system. Through rigorous trials and comparative analysis, the RAG system consistently showed superior performance, leading us to select it as our primary model. Given that the RAG system delivered satisfactory results, and further fine-tuning would require significantly more resources without adding substantial value, we decided to adopt RAG as the primary method for deploying the assistant.

For the UI/UX component of the project, we developed 'Chainlit' and 'Streamlit' user interfaces for our conversational model. This interface not only facilitates natural and coherent dialogue flow but also ensures that the sequential logic of conversations is maintained, which is crucial for accurately addressing legal queries.

Data

In the scope of this project, we have used "Criminal Code of the Republic of Armenia" [3] as of 23 May 2018 in English language, and "Criminal Code of the Republic of Armenia" [4] in Armenian. For fine-tuning purposes, we have also collected a curated list of question-response pairs containing practical and case-specific questions.

The data manipulation process involved converting the data into a JSON file, with each ID corresponding to one law article. This approach was adopted to simplify and enhance the intuitiveness of the retrieval process.

Here's a sample of data:


```
{
 "1.": {
   "title": "Criminal legislation of the Republic of Armenia",
    "content": "1. The criminal legislation of the Republic of Armenia consists of this Code. New laws that envisage criminal liability shall be included in the Criminal Code of the Republic of Armenia. 2. The Criminal Code of the Republic of Armenia is based on the Constitution of the Republic of Armenia and on principles and norms of international law."
 },
 "2.": {
   "title": "Tasks of the Criminal Code of the Republic of Armenia",
    "content": "1. Tasks of the Criminal Code of the Republic of Armenia shall be as follows: to protect human and citizens' rights and freedoms from criminal encroachments, rights of legal entities, property, environment, public order and security, constitutional order, peace and safety of humanity, as well as to prevent crimes. 2. For the purpose of implementing these tasks, the Criminal Code of the Republic of Armenia stipulates the ground for criminal liability and the principles of criminal legislation, determines what acts dangerous to the public are deemed to be crimes and defines the types of punishment and other criminal-law enforcement measures for commission thereof."
 },
 "3.": {
   "title": "Ground for criminal liability",
     "content": "The sole ground for criminal liability is criminal offence, i.e. the commission of an act that contains all elements of corpus delicti provided for in the criminal statute."
 },
 "4.": {
   "title": "Principles of criminal legislation",
```

"content": "The Criminal Code of the Republic of Armenia is based on the principles of legality, equality before the law, inevitability of liability, personal liability, fault based liability, individualisation of justice and liability and humanism."
}}

Methodology

For the scope of this project, we have used 3 different models - GPT 3.5 Turbo, GPT 3.5 Turbo 1606 and Mistral's Mixtral-8x7B-Instruct-v0.1. Each of these models' performances were tested on a RAG system.

System Architecture and Data Preparation:

Our project uses several key techniques and libraries, related to retrieval augmented generation, including Langchain, OpenAI's API, and FAISS for vector storage and retrieval. The primary data source for our conversational agent consists of a curated JSON file containing Armenian criminal law articles. Each article is detailed with a title and corresponding content, organized to serve as the knowledge base for the agent.

Document Management:

The DocumentManager class is responsible for handling the loading and management of legal documents. Upon initialization, it accepts a filename which refers to the JSON file of law articles. The documents are then loaded using the TextLoader function from the Langchain community package, which reads and stores the content for subsequent processing.

Due to token limitations with the OpenAI API, it was necessary to divide the data into 10 parts for the chatbot to be able to read the document. As a result, the Chatbot will only be able to answer questions from the part of the Criminal Code that is currently chosen. However, with a higher tier of OpenAI API integration, the limit of tokens is lifted and you can access more tokens at once. With more tokens you can have less divisions.

We also experimented with asynchronous sequencing of documents, but this approach proved too time-consuming, as it required waiting for one document to process before moving to the next, often exceeding the one-minute limit. To save time, we determined that dividing the documents and indicating the approximate location of the law article was the most efficient solution.

Embedding and Indexing:

For embedding the documents, we utilize the OpenAIEmbeddings class, which interfaces with OpenAI's API to convert text data into vector embeddings. These embeddings are crucial for the retrieval process, as they represent the documents in a high-dimensional space. Once embedded, the documents are indexed using a FAISS database, an efficient similarity search and clustering of dense vectors, which facilitates fast and accurate retrieval of information.

Agent Configuration:

The AgentManager class manages the conversational AI aspect of the project. It initializes with a specific model, tailored for criminal law, using OpenAI’s different GPT model. The agent is set up to interact with a retrieval system by incorporating a retriever tool, which uses the FAISS database to fetch relevant document embeddings based on user queries.

Retrieval-Augmented Generation (RAG):

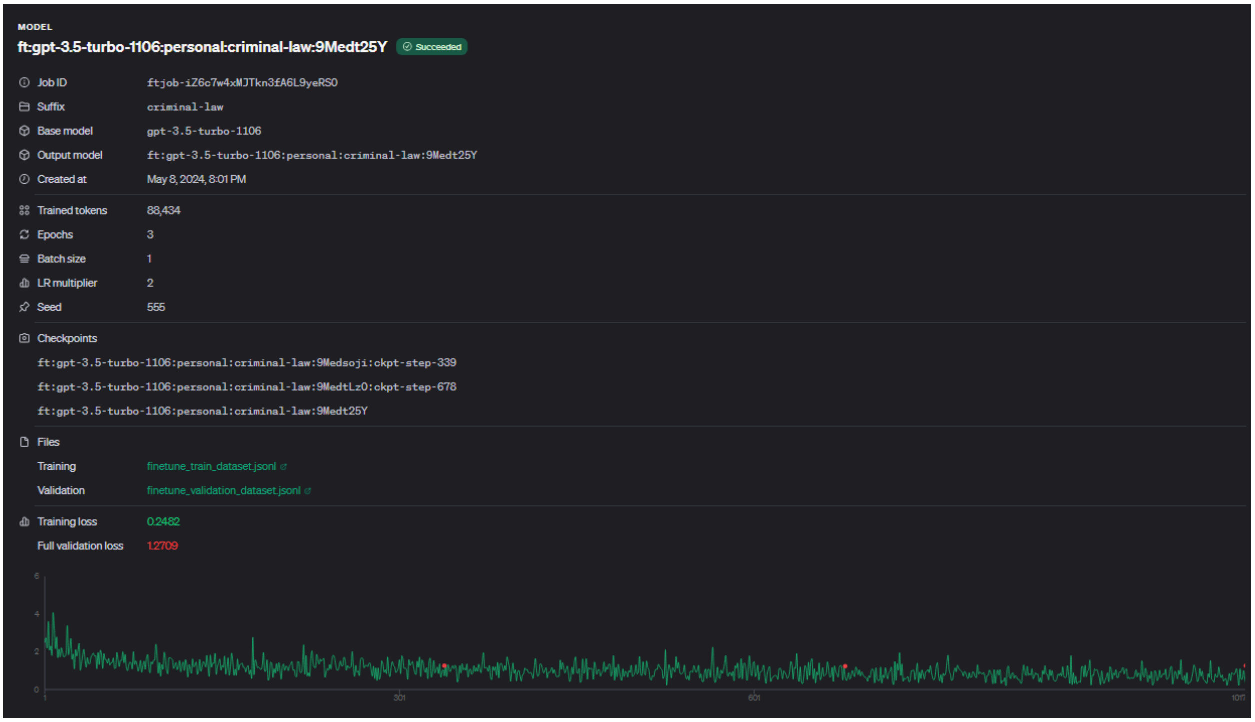
The conversational agent operates on a RAG system, where it dynamically retrieves information from the indexed database as needed to generate responses. This approach combines the benefits of neural language models with the precision of database queries, ensuring that responses are both contextually relevant and factually accurate.

User Interface:

To facilitate interaction with the system, a 'chain lit' user interface is deployed. This interface is designed to maintain the sequential logic of conversations, ensuring that each interaction builds logically on the previous one, which is crucial for complex legal discussions. While the specific code for this interface is not detailed here, it integrates seamlessly with the conversational agent to provide a user-friendly experience.

Fine-tuning component:

We also employed OpenAI's GPT 3.5 model and fine-tuned it using a carefully curated list of question-answer pairs. However, the results, which we will discuss later in this paper, didn’t show much improvement, leading us to favor the RAG system.



Trials on Criminal Code in Armenian Language

In the initial phase of the project, we implemented RAG using the Armenian Criminal Code in the Armenian language. The results were not impressive, which led us to use the translated criminal code. A language model that works effectively in Armenian for criminal law would require significant pre-training on a large corpus of the Armenian language.

A well-known, publicly available multilingual model is mGPT. When posing a question to it on Huggingface, it provides arbitrary responses without even properly ending the sentence. This was a hint that we would need massive computational resources to better train it in the language.

The RAG with the GPT-3.5 model was more promising but still not sufficient to continue the project with it. Here are some of the responses that we received from the RAG-optimized GPT-3.5 model.

1. *Ինչպե՞ս է հոգեկան առողջությունը դիտարկվում անձի քրեական պատասխանատվությունը որոշելիս*

Հոգեկան առողջությունը դիտարկվում է անձի քրեական պատասխանատվությունը որոշելիս, ըստ Հոդված 22-ի, սահմանափակ մեղսունակությունը կատարելիս հոգեկան առողջությամբ չի գիտակցել իր արարքի հակաիրավականությունը կամ ամբողջությամբ չի կարողացել ղեկավարել իր արարքը, ենթակայ է քրեական պատասխանատվության: Հոդված 23-ում նշված է, մեղքը կատարելիս արարքը կատարել է մեղավորությամբ, *եթե այն կատարողը գիտակցել է իր արարքի հակաիրավականությունը կամ թեև չի գիտակցել: Հոդված 24-ում նշված է, դիտավորությունը և անզգուշությունը կարող է դրսևորվել դիտավորությամբ կամ անզգուշությամբ:*

As can be observed, the grammar, syntax, and semantics are inconsistent. Although the model seems to attempt to fetch and retrieve the answer, hinting that the retriever itself does a decent job and that the embedding works normally, the incorrect grammar flow leads us to the problem of an insufficiently pre-trained model.

2. *Ի՞նչ է նշանակում «սահմանափակ քրեական պատասխանատվություն» տվյալ օրենքով:*

«Սահմանափակ քրեական պատասխանատվություն»-ը նշանակում է այն դեպքերը, երբ անձը կատարում է մեղավորություն, որը հանցանքը կատարելիս հոգեկան առողջության խնդիրների հետևանքով ամբողջությամբ չի գիտակցել իր արարքի հակաիրավականությունը կամ ամբողջությամբ չի կարողացել ղեկավարել իր արարքը:

3. *Սահմանափակ քրեական պատասխանատվությունը կարո՞ղ է պատիժը նշանակելիս մեղմացուցիչ գործոն լինել:*

Այո, սահմանափակ քրեական պատասխանատվությունը կարո՞ղ է պատիժը նշանակելիս մեղմացուցիչ գործոն լինել, այո՛:

Although if reading carefully the meaning can be derived from the messages, the language problems are still very prominent.

RAG vs. SFT

After deciding to proceed with the translated criminal laws, our initial idea was to compare two well-known techniques: Retrieval-Augmented Generation (RAG) and Supervised Fine-tuning (SFT), to see which would better align data with our language model. We hypothesized that RAG would be more effective for such projects.

Retrieval-Augmented Generation (RAG): RAG has shown promise by retrieving relevant knowledge from external databases, enhancing the accuracy and credibility of the generated content. This is particularly useful for tasks where accuracy of response is important, such as fact-checking, or document retrieval, which we have in our case.

Feature Comparison	RAG	Fine-tuning
Knowledge Updates	Directly updates the retrieval knowledge base, ensuring information remains current without the need for frequent retraining, suitable for dynamic data environments.	Stores static data, requiring retraining for knowledge and data updates.
External Knowledge	Proficient in utilizing external resources, particularly suitable for documents or other structured/unstructured databases.	Can be applied to align the externally learned knowledge from pretraining with large language models, but may be less practical for frequently changing data sources.
Data Processing	Requires minimal data processing and handling.	Relies on constructing high-quality datasets, and limited datasets may not yield significant performance improvements.
Model Customization	Focuses on information retrieval and integrating external knowledge but may not fully customize model behavior or writing style.	Allows adjustments of LLM behavior, writing style, or specific domain knowledge based on specific tones or terms.
Interpretability	Answers can be traced back to specific data sources, providing higher interpretability and traceability.	Like a black box, not always clear why the model reacts a certain way, with relatively lower interpretability.
Computational Resources	Requires computational resources to support retrieval strategies and technologies related to databases. External data source integration and updates need to be maintained.	Preparation and curation of high-quality training datasets, definition of fine-tuning objectives, and provision of corresponding computational resources are necessary.
Latency Requirements	Involves data retrieval, potentially leading to higher latency.	LLM after fine-tuning can respond without retrieval, resulting in lower latency.
Reducing Hallucinations	Inherently less prone to hallucinations as each answer is grounded in retrieved evidence.	Can help reduce hallucinations by training the model based on specific domain data but may still exhibit hallucinations when faced with unfamiliar input.
Ethical and Privacy Issues	Ethical and privacy concerns arise from storing and retrieving text from external databases.	Ethical and privacy concerns may arise due to sensitive content in the training data.

SFT with 0-shot prompting: SFT aligns the model with specific behaviors, styles, or preferences, making it useful for general domain familiarity. However, it tends to make more assumptions and can hallucinate based on its training (since it answers or "guesses" questions based on data it's been trained on), whereas RAG minimizes such risks by retrieving data from external sources.

In this research, we implemented RAG on GPT 3.5 Turbo 1106 and Mixtral-8x7B-Instruct-v0.1 models. The best performing model was GPT, since our analysis showed that this model offers the best balance between token usage and performance. For benchmarking, we curated a list of 40 questions to test the plain GPT 3.5 model, the RAG-optimized GPT 3.5 Turbo 1106, and Mixtral-8x7B-Instruct-v0.1. We also tested a few questions with GPT 4 on RAG, which yielded promising results but required more resources and had stricter token limitations, leading us to continue with the RAG-optimized GPT 3.5 Turbo 1106 model.

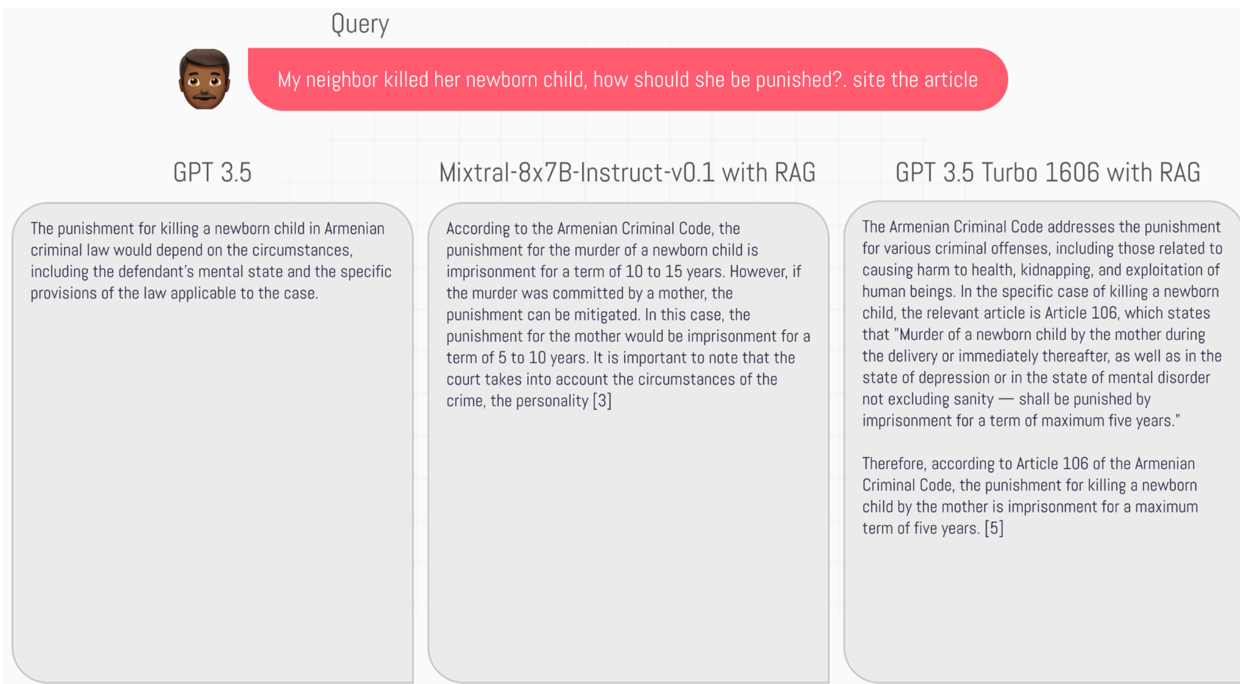
Evaluation and benchmarking

We have curated a list of 40 questions to ask to test the models and evaluate their performances. According to our analysis, the GPT 3.5 Turbo 1606 model optimized with RAG provided more detailed, accurate, and in-depth answers. The difference between this model's performance and the usual GPT 3.5 with RAG is not dramatic, but the 1606 model still gives better results.

Throughout our analysis, we concentrated more on comparing GPT 3.5 Turbo 1606 with Mixtral-8x7B-Instruct-v0.1 model. In general, GPT performed much better than Mistral's model.

When asked about Armenian Criminal Code to the plain GPT 3.5 model with no RAG optimization, it answers questions very generically. The usual answers that it gives to questions about crimes is that the action can be subject to imprisonment or fines, depending on the circumstances of the case.

Here are a few specific examples of how different models respond to some law-related questions. At the end of each answer by GPT's and Mistral's models, there's a 0-5 rating of the answer to then use in evaluation. The rating is based on the accuracy of the answer, the accuracy of document retrieval, and how detailed the answer is.



Query

what is insanity and which article talks about it?

GPT 3.5

Insanity, in legal terms, refers to a mental state where a person is unable to distinguish right from wrong or understand the nature and consequences of their actions. In Armenian law, provisions related to insanity may be found in various articles of the Criminal Code, particularly those concerning diminished responsibility or incapacity due to mental illness. However, specific articles would need to be referenced from the Armenian legal documents to provide precise information.

Mixtral-8x7B-Instruct-v0.1 with RAG

Insanity is a mental disorder or condition that makes a person unable to understand or control their actions. In the Armenian Criminal Code, insanity is discussed in Article 21, which states that a person who commits a crime while in a state of insanity cannot be held criminally responsible. The article also outlines the procedures for determining whether a person was insane at the time of the crime. [4]

GPT 3.5 Turbo 1606 with RAG

Insanity in the Armenian Criminal Code is addressed in Article 25. It states that a person shall not be subject to criminal liability if they were in a state of insanity while committing an act dangerous to the public. This includes situations where the person, due to chronic mental illness, temporary mental impairment, dementia, or other diseased condition, could not realize the dangerous nature of their action or control it. [5]

Query

My friend told me that the Red Cross Emblem I have sticked on my car is against law. Is he correct and why?

GPT 3.5

Your friend might be correct depending on the context. The Red Cross Emblem is protected under international law and is meant to signify humanitarian aid and protection. However, unauthorized use of the emblem, especially in contexts unrelated to the Red Cross or humanitarian activities, can be prohibited by law. It's essential to adhere to regulations regarding the use of such symbols to avoid legal repercussions.

Mixtral-8x7B-Instruct-v0.1 with RAG

Yes, he is correct. According to the Armenian Criminal Code, the use of the Red Cross Emblem is prohibited unless it is used by the International Committee of the Red Cross, the International Federation of Red Cross and Red Crescent Societies, or the national Red Cross or Red Crescent Society. Violation of this provision is punishable by a fine or imprisonment. The reason for this prohibition is to protect the symbol's neutrality and ensure [5]

GPT 3.5 Turbo 1606 with RAG

According to the Armenian Criminal Code, making use of the protective emblem of the Red Cross in a manner that is contrary to international treaties is illegal. This includes using the Red Cross emblem in a way that is not authorized during military operations. The offense is punishable by a fine or imprisonment for a maximum of three years. [5]

Query

I read from news report that someone engaged a child in robbery, how should they be punished?

GPT 3.5

Engaging a child in robbery in Armenian criminal law could lead to penalties under various articles of the Criminal Code related to both robbery and exploitation of minors. The punishment would depend on the specific circumstances of the case, including the age of the child and the level of involvement of the perpetrator.

Mixtral-8x7B-Instruct-v0.1 with RAG

According to the Armenian Criminal Code, engaging a child in robbery is punishable by imprisonment for a term of 5 to 10 years. Additionally, the person who committed the crime will also be banned from holding certain positions or engaging in certain activities for a term of up to 5 years. [3]

GPT 3.5 Turbo 1606 with RAG

Engaging a child in the commission of a criminal offense, such as robbery, is punishable by imprisonment for a term of one to five years. If the act is committed by a parent, teacher, or another person in charge of the upbringing of the child, the punishment can be imprisonment for a maximum of five years, with or without deprivation of the right to hold certain positions or to engage in certain activities for a term of maximum three years. If the act is accompanied by violence, the punishment can be imprisonment for a term of three to seven years, with or without deprivation of certain rights. If the act is committed in an organized group or in the commission of a grave offense, the punishment can be imprisonment for a term of five to ten years, with or without deprivation of certain rights. [article 165] [5]

Query

I know that my sister is planning to poison her husband. What happens if I dont report about it?

GPT 3.5

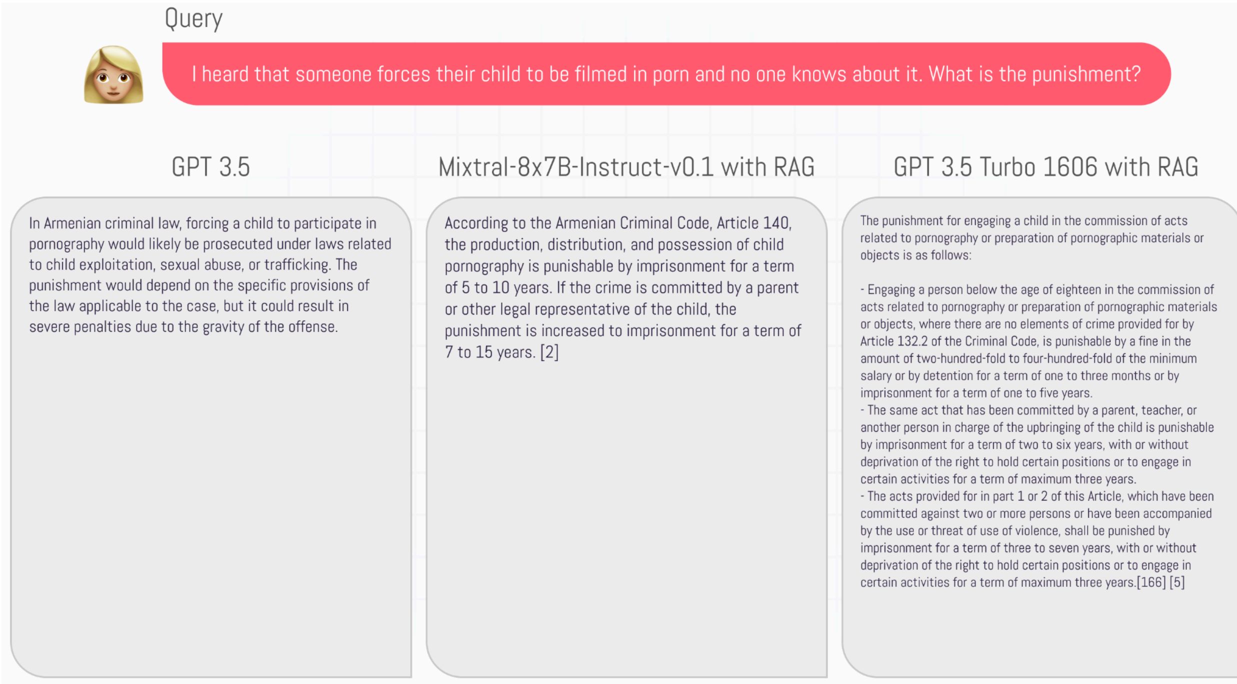
Failure to report a planned crime, such as poisoning, in Armenian law could potentially lead to legal consequences depending on the specific circumstances and provisions outlined in the Criminal Code. This could vary from being considered an accomplice to the crime to facing charges for failure to prevent a serious offense.

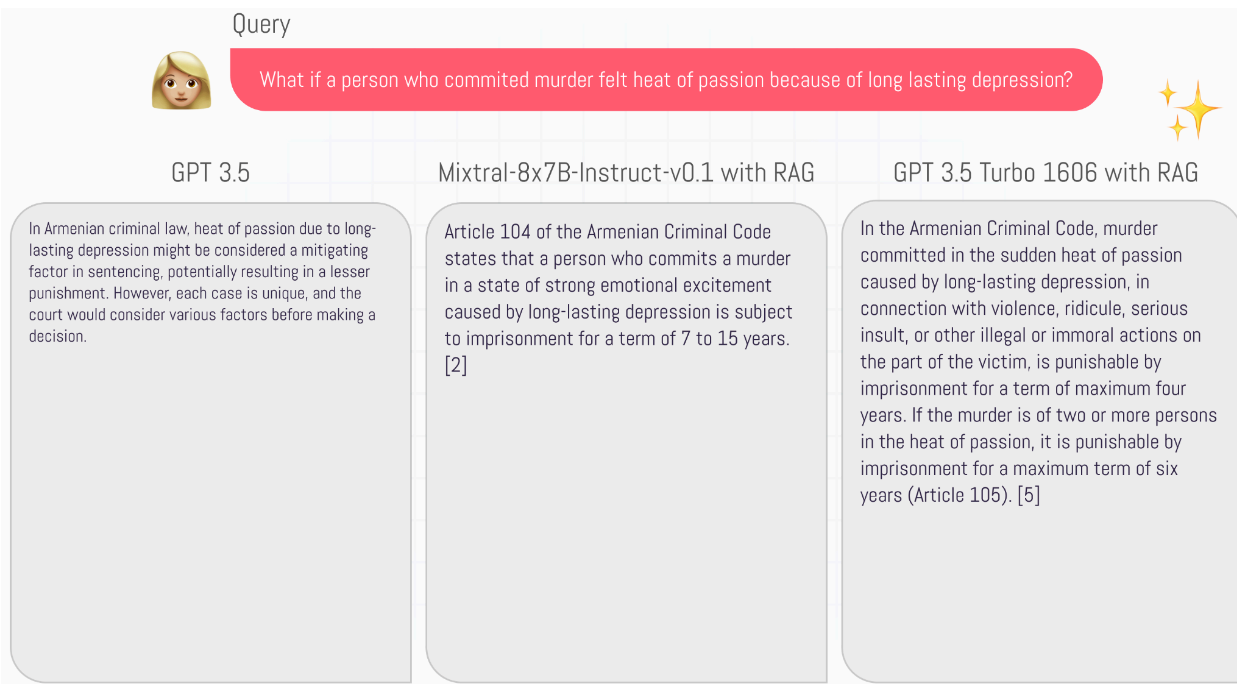
Mixtral-8x7B-Instruct-v0.1 with RAG

Article 38. Failure to Report a Crime If a person knows about the preparation or commission of a crime and fails to report it, they will be punished with a fine or imprisonment. [4]

GPT 3.5 Turbo 1606 with RAG

If you know that your sister is planning to poison her husband and you don't report it, you may be subject to punishment under the Armenian Criminal Code for "Concealment of a Crime." The specific consequences and penalties for this offense are not provided in the information retrieved. However, it is generally considered a serious offense to conceal a crime, especially one as severe as poisoning, and could result in legal consequences. It is important to report any criminal activity to the authorities to prevent harm and ensure justice. [4]





Here's a short comparison of the two models:

GPT 3.5 Turbo 1606 with RAG: Almost excellent quality in retrieval, embedding and quality of LLM response. One small minus of this model can be the lack of Chain-of-Thought reasoning – if a question is very complex and requires going from

Document 1 to Document 2, then another document may fall short in connecting lines. We refer to this issue later in the article.

Mixtral-8x7B-Instruct-v0.1 with RAG: Average quality in embedding, lack of quality in retrieval(mentioning article), average quality of LLM response. The model falls short in retrieving the correct document; however, when finding the document, it adequately shares the content of the law article and can connect well with the given question. The model is good for generic questions but lacks performance in complicated cases.

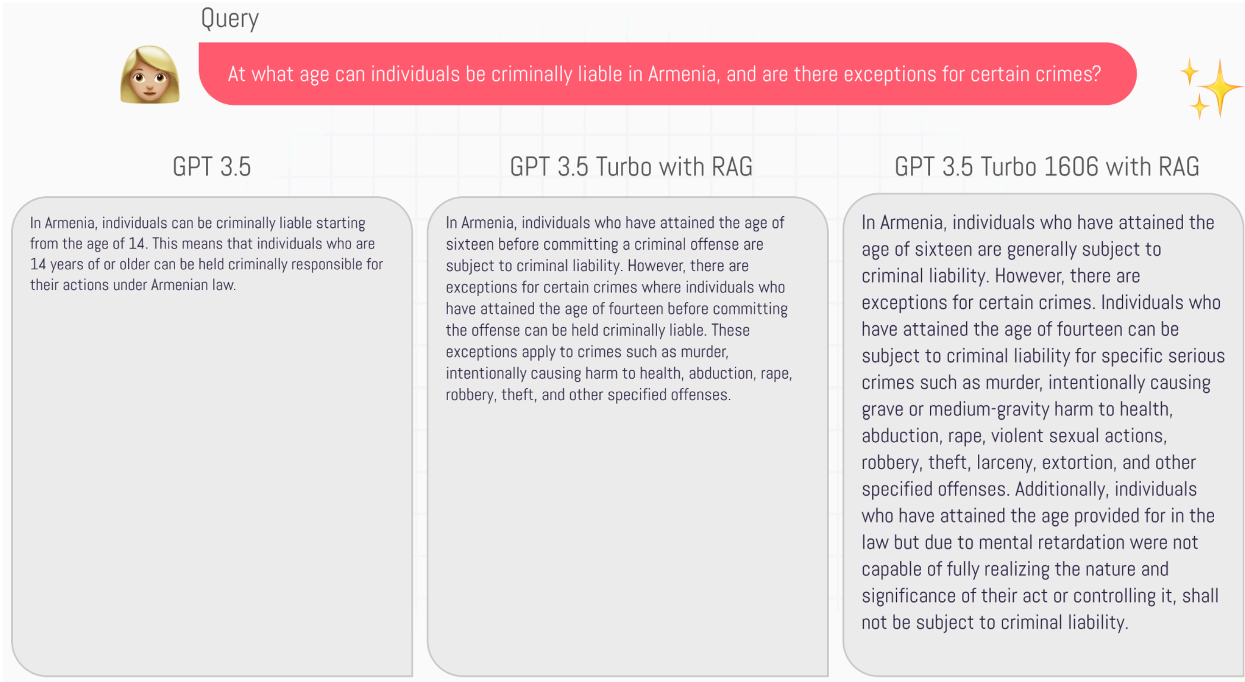
GPT 3.5: Doesn't give wrong answers, but the answers provided by the model are too generic and, as such, hard to compare against those of the models with RAG.

One general observation about temperature: when increasing the temperature, we noticed that the model becomes more "lively" or friendly and also gives more detailed and long responses compared to when the temperature is 0. However, it's worth mentioning that adding too much temperature may result in hallucinations.

In our dataset of the questions with the responses of GPT's and Mistral's models, we have rated each response based on how good it matched with the actual response of the question and the actual law article that it refers to. For this, we have carefully curated the questions and did deep analyses of the criminal law documents. Here's the benchmark with the corresponding evaluations of each model.

Benchmark (Higher is better)	GPT 3.5	mistralai/Mixtral 8x7B-Instruct v0.1 (with RAG)	GPT 3.5 Turbo 1606 (with RAG)
Grade (As graded by us)	2,5* The answers provided by the model are too generic and, as such, hard to compare against those of the models with RAG.	2,9	4.7

Initially, we aimed to use fine-tuning to better align the model with criminal law and cover a broader range of use cases. However, the fine-tuning dataset, which included about 400 pairs of practical and case-specific questions, did not significantly outperform RAG and even struggled with direct questions from its training dataset. For this next example, the fine-tuned model gave a more concise and short answer – "Individuals in Armenia can be held criminally liable from the age of sixteen, subject to certain exceptions for those aged fourteen to sixteen."



There are two potential improvements we believe could enhance the fine-tuned model's performance:

1. Expanding the dataset with a larger and more diverse set of question-response pairs.
2. Involving legal professionals or data trainers to enrich the dataset with practical and varied use cases, and then training the model on this enhanced dataset.

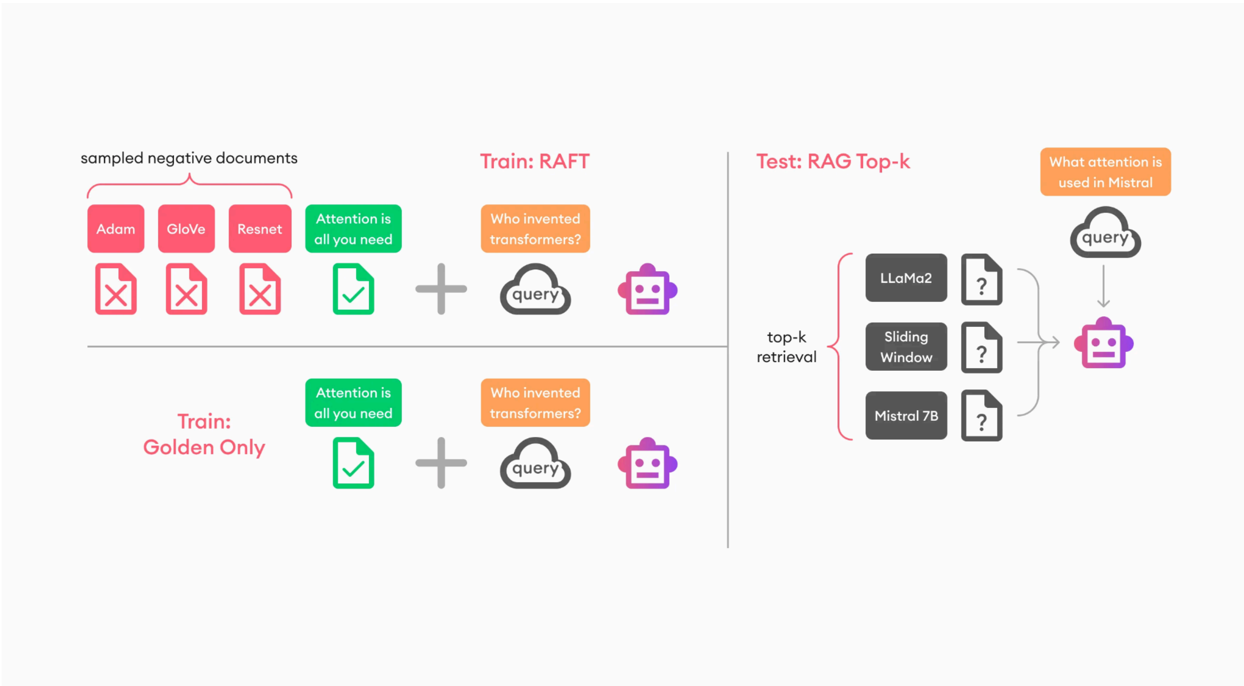
Given the need for a model that excels in fact-checking and information retrieval, we concluded that RAG is the optimal approach for this project.

Future area of development: RAFT

When we found out that fine-tuning itself did not yield the desired results because of the reasons mentioned above, we came up with another idea that has emerged very recently and already gained a lot of attention from the AI and LLM community – retrieval augmented fine-tuning (RAFT). There are two main RAFT techniques that if used in the scope of this project with more professionals involved, can significantly boost the model performance.

One potential RAFT technique involves fine-tuning individual components of the RAG system, such as the embedding model, the retriever, the reranker, and the language model itself. A practical approach is to evaluate each component to identify which one is defective or could benefit most from enhancement, and then specifically fine-tune that component. This targeted fine-tuning could significantly boost the overall performance of the model.

Another idea we believe can be invaluable for the scope of this project is derived from the recent paper from UC Berkeley – "RAFT: Adapting Language Model to Domain Specific RAG" [5], which studies how to combine supervised fine-tuning (SFT) with retrieval augmented generation (RAG). With retrieval augmented fine-tuning, we train the model to take a question and documents (even the distracting ones) and come up with an answer that follows a logical thought process. RAFT has proven to be better than just supervised fine-tuning, whether RAG is used or not.



In RAFT, we create training data that includes a question (Q), some documents (Dk), and a corresponding chain-of-thought answer (A*) that's based on information from one of the documents (D*). We distinguish between two types of documents: the 'oracle' documents (D*) that have the information needed for the answer, and 'distractor' documents (Di) that don't help with the answer. Some of the training involves having the right document along with distractions, while other times, we only include distractor documents to encourage the model to rely on its memory rather than just the documents provided.

The language model is then trained using a standard technique to generate answers based on the given documents and questions. This training process enhances the model's ability to perform well in domain-specific scenarios where it needs to use relevant documents to generate answers.

Transparent and unbiased language model for complicated court cases with RAFT;-

In our law assistant project, RAFT can be a useful technique, particularly effective in scenarios requiring chain-of-thought reasoning. This is crucial for addressing questions that are not directly covered in law articles but instead require a sequence of logical deductions from multiple documents (e.g., moving from document D1 to D2, then to D3). The chain-of-thought (CoT) reasoning component of the RAFT system can significantly improve the model's capability to construct answers based on a logical progression, rather than relying solely on the specific content of a single document.

This reliance on the model's internal memory and its ability to link information across various texts can have invaluable implications for even various complex court cases. With its CoT reasoning, the model can answer twisted and complicated questions and resolve cases that would be complicated for the human mind and require much more time to resolve and link the events together. Another important aspect of RAFT is its unbiased nature and transparency, which are paramount in the legal field.

Conclusion

In conclusion, this capstone project has demonstrated the practicality of deploying a specialized conversational agent for Armenian Criminal Law. By employing AI technologies such as Retrieval-Augmented Generation (RAG) and fine-tuning methods, we have developed a system that enhances the accessibility and comprehension of legal information. Despite encountering challenges with initial model performance and language inconsistencies, our analysis indicates that RAG, given its ability to leverage a curated knowledge base, is more effective for this application than other methods like Supervised Fine-Tuning (SFT).

The project has shown that while additional fine-tuning yields only marginal improvements, the primary RAG configuration sufficiently meets the needs for accuracy and relevance in legal consultations. Moving forward, exploring Retrieval-Augmented Fine-Tuning (RAFT) could potentially refine this integration of retrieval mechanisms with fine-tuning processes, aiming for improved precision and adaptability in legal assistance applications.

This work contributes to ongoing efforts in merging AI with legal services, supporting the notion that legal information can be made more accessible and understandable through technological interventions. Future work will likely build on these findings, further enhancing the capabilities of AI-driven legal tools.

References

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). *Attention is all you need*. arXiv.org. https://arxiv.org/abs/1706.03762

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023, December 18). *Retrieval-Augmented Generation for Large Language Models: A survey*. arXiv.org. https://arxiv.org/abs/2312.10997

https://www.venice.coe.int/webforms/documents/default.aspx?pdffile=CDL-REF(2021)022-e

https://www.arlis.am/DocumentView.aspx?docID=153080

Zhang, T., Patil, S. G., Jain, N., Shen, S., Zaharia, M., Stoica, I., & Gonzalez, J. E. (2024, March 15). *RAFT: Adapting Language Model to Domain Specific RAG*. arXiv.org. https://arxiv.org/abs/2403.10131

https://platform.openai.com/docs/guides/fine-tuning

https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

https://huggingface.co/ai-forever/mGPT-armenian

https://twitter.com/rohanpaul_ai/status/1759649968955269170