dr. Homoki Péter, 2023-03-19 20:00:00 +0100; demo "law firm chatbot" (https://chatbotdemo.homoki.net using OpenAI API access)

Introduction

For the sake of experimentation and to gain experience, I have spent some time since 15th March creating a basic frontend for the OpenAI API. I aimed to customize it in a way that could give a working demonstration of some of its uses for small law firms.

I have started with the simplest, most informative, although probably *not* the most useful application: a chatbot. I wrote a frontend for the OpenAI GPT 3.5/4 model, using the currently available API and the standard methods provided by OpenAI. I attempted to customize the usage in a way that approximates what a small law firm (such as my own) could theoretically expect from a chatbot such as this one.

This approach forced me to consider both practical issues of use and of possible deontological risks and problems – at least those applicable to Hungarian lawyers.[1] So although this demo chatbot is purely for research purposes, it is as real as it can be.

This blog is written by a lawyer for fellow lawyers, so it is not the rather primitive programming behind the demo that is interesting.[2] For lawyers, it's usually also not relevant who operates the model, what's the exact name of the model is how the technical parts work etc. However, after all these hours of work with the model, I am convinced that it is imperative that a wider range of lawyers investigate these capabilities in more detail.

Even after years of research on this subject, I've found this large language model to have an astonishing range of capabilities. I see plenty of opportunities that are relevant to legal professionals as well, and these are a sign of how models like this could change the way we work, in a shorter timeframe than expected.

Last year's research on this subject has become obsolete in many ways: what the actual capabilities are, what the most promising tools are for specific objectives, what could become reality in just a couple of years time, and that some of the worries of last year are no longer relevant.

So, seeing the capabilities of this foundational model, I am convinced that similar large language models will have a profound effect on how lawyers will work in the future, in all segments of this profession.

That is the reason why, I also believe that it is not a waste of time for any lawyer, no matter how old or experienced, to better understand how these models work, what kind of limitations they should expect, and what some of the current constraints are.

Using the demonstration as an excuse, I would like to share with you some of the experiences I have gained so far working with OpenAI GPT models and provide a little background information. Even if

some of this basic information is of a technical nature, I have only highlighted information that, from a lawyers' perspective, could be relevant for later uses. By using this narrative, I believe we can also lay down some very basic structure for the much needed future discussions in the area of how lawyers will be able to use foundational models.

What is GPT, ChatGPT and the OpenAI API?

GPT is an acronym (generative pre-trained transformers) of a specific family of neural network-based language models, originally created by OpenAI LLC in the ante-diluvial years of 2018.

For the eyes of the uninitiated, language models are just large files used by arcane software to process some software requests, they are critical building blocks in natural language processing software running on computers, to identify the content of texts (classify or extract information elements), to translate or otherwise create new text according to instructions. There are hundreds and thousands of language models, but not all of them are published or available for the public.

Since 2018, OpenAI has released a number of new versions of its GPT model, all being trained on larger and larger texts (corpora) with some changes in architecture. The first version to make the headlines as a possible way “to spread misinformation” was GPT-2, but each new version came with progressively more media coverage and frenzy. The latest, GPT-4, was released 14 March, 2023, and provides very impressive improvements over the previous GPT-3.5 (which was itself, already very impressive).

The media coverage was greatly boosted when OpenAI released a “consumer front-end” for their language model, that was finetuned for a chatbot functionality. This was called “ChatGPT” in 30 November 2022 (relying on the version GPT-3.5). Currently, for a monthly fee of 20 $ (+VAT), users can enjoy the chatbot functionality of GPT-4 under the trade name “ChatGPT Plus”.

Version 3 and beyond of GPT are not downloadable, Microsoft (the biggest investor of OpenAI) has exclusive license to the models since September 23, 2020. Regardless, the language models are all accessible via the web services called application programming interfaces (API) provided by OpenAI, since at least early 2021. So currently, the main methods of accessing these language models is either via the consumer front-end (which are not intended to be served to one’s own clients), via the APIs (which require some front-end themselves) or via another providers building themselves upon these APIs. So there is no on-premise use possible, and all requests have to go through OpenAI and will come from them. There is already a limited access for using *some* of the OpenAI models from Microsoft’s cloud offering, which is important due to widespread regulatory requirements (e.g. in financial industry etc.) in relation to cloud computing solutions.

It’s very important to understand that there are other, fully open and downloadable large language models[3] similar to GPT that are almost as good in many aspects, and there are also language models that are still better at certain tasks than GPT, and also that due to the current setup and limitations, it is simply not possible to carry out certain, very important language related tasks when using GPT.

Nevertheless, for illustration purposes, let's see how this chatbot works, why it is not really the best suited for the job of a law firm chatbot, and in what ways could smaller law firms use other services of the OpenAI API, with what kind of limitations.

Demo law firm chatbot using GPT-3.5 and GPT-4

Limitations of customization by examples and prompts

The current demo chatbot uses the engine called GPT-3.5, but that's just purely for reasons of economy: answering via the GPT-4 costs 15 times as much as GPT-3.5.[4] Thanks to the OpenAI API, it is easy to customize to some extent how the chatbot works, what kind of answers it gives, and most importantly, what kind of responses it should refrain from giving.

As you can see from the source code, besides the mandatory branding of the front end to the law firm (which is a very basic web application in this case), this customization is made via question and answer examples and prompt instructions. The examples are made of pairs of questions and answers, while the prompt instructions are effectively fed to the model before the user can input their own questions (providing some built-in "bias" based on which to give answers).

These customizations tell the chatbot what kind of persona they should play (an assistant, a receptionist or a lawyer etc.), how they should act and also, what kind of information they should definitely serve.

In these customization texts, using plain English and Hungarian language, I've tried to include some of the most basic deontology rules applicable to law firms (such as no answers that could be understood as comparative advertising etc.), while at the same time, providing the absolutely necessary information about the law firm "marketed", such as contact data and area of expertise etc.

The latter is vital, because most of these sophisticated language models tend to "hallucinate" and for the moment, they are not doing an internet search on their own. For example, I gave the model explicitly the phone number of my office, but not the physical address. During a test, I've asked the chatbot for the contact details of the law firm in general (not just the phone number), and the completion included a very precise and existing physical address – but that was not my law firm's.

However, in terms of size, there are very strict limitations which also affects how much customization we can do. For GPT-3.5, there is a strict limit of 4096 tokens, that includes both "prompt" (the question) and "completion" (the answer). Also, the prompt size includes our examples and prompt instructions, as well as the chatbot user's actual question.

Language models have to turn characters, words and sentences into tokens before they can process them. The size of a sentence in tokens depends a lot on the language and the words used, and there is no hard rule.[5] For my case, mixing English and Hungarian text (trying to create a multi-language chatbot), this means that half of the tokens that can be used is already taken up by the customization

texts, and the remaining half should include both the actual user question and the answer from the chatbot as well. This is a very serious limitation.

So even if a lot more customization would be useful, and a lot more information about deontology rules or the firm could be inserted, there is simply not enough place for that. For example, I have tried to include some references to the core principles of lawyers in the EU, so that the chatbot could respond in a way that reflects these values, but that would have made the demo useless due to very short answers.

The good news is that this is not a theoretical limitation, and even with GPT-4, you can already use about eight times as much tokens in total.

What can lawyers use such a chatbot for?

So, for what purposes can we use such a chatbot? Here, we use the term "chatbot" in the strict sense: the demo front-end that relays the end users' questions to the language model hosted by OpenAI, with some minor customizations.

We can use a chatbot like this to provide information about our firm in a slightly more entertaining way than what can be achieved on a plain website. Additionally, we can provide this information simultaneously on other channels, such on a Telegram or Viber chatbot etc. Essentially, it's just advertising and marketing.

This can give a relative edge to the law firm, at least until most other law firms have the same tool. The extra entertainment value comes from chatbot's ability to pretend to be a lawyer, allowing users to ask the chatbot about legal issues without the need to explicitly define all questions and answers, as was necessary with earlier generations of chatbots.[6] Of course, to do so, the terms of use must clarify that this is not legal advice and should not be used for any real purposes.[7] It's important to differentiate between this entertainment value and the legal advice actually given by the law firm (and not by the chatbot).

Even the current terms of the OpenAI API usage policy clearly state that these models should not be used for providing legal services without a qualified person's review[8]. This means that, due to these usage policies of OpenAI, this model may not be used in consumer-facing front ends. That is, unless a reckless lawyer takes responsibility in advance, giving their blank approval no matter what answers the chatbot provides to any legal issues asked. This may satisfy the requirement of the OpenAI usage policy, but would otherwise be manifestly unethical.

At least in its current state, this chatbot is not best suited for all typical chatbot cases. It might give users incorrect answers regarding contact details or the firm's area of expertise. It is not ideal for booking appointments with lawyers. Even if GPT excels at interpreting the intent of potential clients, and could technically be capable of checking a calendar for free time slots, it is currently much easier and more reliable to do so via a dedicated application (with the possibility of connecting to payment services to give weight to the time slots booked).

While this particular demo chatbot can only be used for client-facing purposes, the processing capabilities of the OpenAI API (including GPT completion uses) extend beyond this simplistic chatbot functionality. The salient feature of the model is its ability to converse fluently in many languages, but rather (since GPT-3.5) its capacity to give astonishingly accurate answers to very complex questions, as long as the question does not relate to facts beyond September 2021.

Let's get a quick overview of this based on current experiences, and I believe we will have to return to these issues later for more in-depth discussions.

A possible roadmap for further research

This blog post is not an advertisement for OpenAI, so it's not my intention to list all the possible uses of the OpenAI API. I aim to convince fellow lawyers that it is technically very easy to connect to these endpoints (or use similar services from other language models from other providers or open-source models). This does not require significant money and effort, and if someone has these connections built into versatile applications, they can greatly improve their law firm's capabilities and even save money currently paid to more suppliers. For lawyers who use a large number of different IT products, these APIs could also serve as a way to reduce the number of required products and the costs of integrating them.

Of course, in the case of proprietary language models such as GPT, this also comes with a price, because lawyers will have to rely more and more on the provider of the large language model, and they may change their pricing or their terms of use at any time.

For now, let's take a look at the further areas of use for the latest GPT models.

Besides the demo chatbot, the same chat completion API calls can also be used for translations from one language to another – just the built-in prompts for this purpose will be different. Of course in relation to translation purposes, we have to remain mindful of the appropriate token size limits. We can also send prompts (instructions and texts) about correcting the style of the text, checking typos instead of using a spell checker, or simply changing the nouns, declensions, or conjugations in the text according to some rules.

With a handful of appropriate examples in the customizations, we can also convince the OpenAI API to accurately classify diverse client data (if we have the authorization to send them), and using the answer from the API, provide our own software with specific instructions, like which other software to call or which parameters to use when calling different software. For example, by sending the header information, the email text, and previously feeding some simple taxonomy, the OpenAI API could return the suggested filing locations of the emails. We can also use the chat completion API to tag emails that seem urgent or otherwise require the attention of a partner.

These uses have little to do with the usual chat completion functionality. However, the capabilities of these later GPT models are such that they can be accurately used for such purposes as well. Even OpenAI suggests that previously used "text completion" tasks (where the focus is on generating

longer text from a prompt) should now use chat completion API calls instead, because GPT-3.5 is much faster and more powerful.

So these chat completion APIs could also be used for text generation, even to create whole contracts or drafts. However, the documents lawyers need to draft usually have to comply with a large number of requirements that could be client-specific, project-specific, or even specific to the drafting lawyer. How can this be achieved with the GPT models?

To make the GPT models better suited to specific applications of the legal profession, providing more detailed prompts before each conversation session is not the best approach. The activity of building on top of the powerful large language models is called “finetuning” in this industry. Even just a few examples can greatly help the model in better understanding the tasks at hand and providing more accurate answers. Also, this finetuning tends to be technically neither costly, nor very complex, so the main value in this phase is the expertise of the domain specialists who work out the dozens or hundred correct question-answer pairs (or appropriate classifications etc.)

While more open language models make it possible to finetune their pretrained models in many ways, with the OpenAI API, this is rather restricted. Currently, the latest model that can be finetuned is GPT-3 (so no finetuning for 3.5 and 4), but it is a very simple and straightforward process, once someone has the set of questions and optimal answers, and it is much cheaper to do than providing QA examples before a question, like we did with the demo chatbot.

This is the appropriate approach for contract generation where the set of requirements (and the set of appropriate clauses etc.) is a lot longer that could fit into a prompt. But this approach could also be used to provide more accurate and predictable answers in very specific fields, without hallucinations, but without the need for finding every possible questions that a user may ask. This could include questions of local law, processing large knowledge bases etc.

So, there a lot of ways how large language models can be effectively utilized by legal professionals, from document drafting and contract generation, to question answering, and how finetuning could help.

But at the same time, we should not forget about some other important issues that will also need our attention, like how the training of a law student or a lawyer should adapt to using these large language models, which could become excellent teachers.

However, as a prerequisite, lawyers would be needed to evaluate the domain-specific accuracy of the answers provided. This could start with creating domain specific benchmarks (separately at the national and EU-level) for some major areas of law, to more accurately assess how the chat completion question answering capabilities correlate with these. We must determine the strengths and weaknesses of these chat completions in the legal applications, because no one else will be able to answer that in our stead.

---

1. The Hungarian ethical rules explicitly include the full CCBE Code of Conduct. ↩

2. No programming related information is included in this blog post. If anyone would like to continue to experiment, please see the GitHub page for a complete source code. Please do not approach me for any technical related help, I have neither time, inclination or expertise to do so. ⮌
3. Such as those based on the LLaMA released by Meta, or some versions of BERT etc. ⮌
4. The demo chatbot works with GPT-4 as well, ⮌
5. There is a general rough estimate of 4 characters per token, but that is just for English text. ⮌
6. E.g. in DialogFlow-based NLP chatbots or even more primitve, keyword-based chatbots. ⮌
7. Terms of use saying: "*Do not use this chatbot for any real purposes, including for trying to get legal advice or legal services. Use this chatbot only to see for yourself why or why not this very popular model provided by OpenAI can or cannot be used for such purposes.*" ⮌
8. https://openai.com/policies/usage-policies: "*Unauthorized practice of law or offering tailored legal advice without a qualified person's review: OpenAI's models are not fine-tuned to provide legal advice. You should not rely on our models as a sole source of legal advice*" ⮌