**Hungarian Journal of Legal Studies** 64 (2023) 3, 435–455
DOI: 10.1556/2052.2023.00475

AKADÉMIAI KIADÓ

# Large language models and their possible uses in law

PÉTER HOMOKI[1] and ZSOLT ZŐDI[2*]

[1] Lawyer, Licensed in Hungary, Hungary

[2] Institute of the Information Society, Ludovika University of Public Service, Budapest, Hungary

## ORIGINAL RESEARCH PAPER

Received: April 13, 2023 • Accepted: July 18, 2023
Published online: April 15, 2024

## ABSTRACT

The paper explores the potential applications of Large Language Models (LLMs) like ChatGPT in the legal field, focusing on how they can enhance access to law. We begin by elucidating the fundamental workings of LLMs and their current and future general applications. The core of our study predicts the utilization of LLMs in various legal domains, especially where tasks like text retrieval, generation, labeling, and classification are prevalent. We argue that tools like ChatGPT could play a pivotal role in these areas. Additionally, we discuss the limitations and customization requirements of LLMs, particularly for legal uses. An experiment conducted by one of the authors, involving a tailored version of GPT for small law firms, serves as a practical example, but building on this, the paper also proposes ways in which LLM-based applications could democratize access to justice, making legal assistance more accessible and efficient for the broader public. This study contributes to the understanding of the intersection between AI technology and legal services, highlighting both the opportunities and challenges in this field.

## KEYWORDS

Large Language Models (LLMs), legal chatbots, access to law, AI in legal services, chatbot limitations, legal information retrieval, ethical considerations in AI, legal text analysis, OpenAI API, legal technology integration

[*] Corresponding author. E-mail: zodi.zsolt@uni-nke.hu

# 1. INTRODUCTION

This study seeks to answer how computer programs similar to ChatGPT (that are based on Large Language Models (LLMs), which can 'understand' human (natural) text and generate a response accordingly), can be used in the field of law, and more specifically, how they can improve access to law. Since these systems have only recently become available to the wider public, it is not our goal to outline a comprehensive vision of the future. What we can undertake in this study is to suggest some possible uses based on the current capabilities and limitations of these LLMs.

The study proceeds as follows. In the first, introductory part, we briefly describe the essence of the operation of LLMs, and existing and prospective general (non-legal specific) uses. In the second part, we 'predict' the uses in law. Within this, we give an account of those legal fields, where text retrieval, generation or labelling or sorting (classification) is the primary goal, and therefore it is very probable that GPT, or similar LLMs can play an important role in the near future. In the third part we list the limitations and customisation needs of ChatGPT or OpenAI's GPT in general, providing an example of an experiment that one of the authors conducted regarding the customized version of GPT for small law firms. Then, partly based on this, we also outline some possible ways in which LLM-based applications can contribute to access to justice for the wider public.

# 2. TECHNICAL BACKGROUND AND GENERAL USE

## 2.1. What are large language models?

Large language models are a very successful set of tools that belong to the Natural Language Processing (NLP) branch of the wide field of artificial intelligence (AI). AI also refers to several other areas, such as machine perception or robotics, but these areas have little direct impact on the world of law, so we will not deal with them. NLP is a field of scientific study and engineering that is related to modelling the use of natural languages, and to solving problems related to the generation or 'understanding' of such text.

Within NLP, language models have been used since the late 1940s[1] as (probabilistic) computational models representing how people interact with their environment through the use of natural language. Such models capture statistical representations of the natural language people use, to answer questions such as, what is the likelihood of specific words appearing after another given word or groups of words? Based on this capability, various algorithms have been used in a wide range of applications, from generating natural looking text to machine translation, summarisations, or sentiment analysis.

While there have been many different approaches to how languages are modelled in computers, in the last ten years, neural network-based language models have become the most successful at this task. The availability of enormous amounts of data and the capabilities of the latest neural network architectures called transformer-based models[2] have made these types of language models the dominant approach within many fields of NLP.

[1] Shannon (1948).

[2] The current generation, including GPT, is based on an architecture called Transformers, see Vaswani et al. (2017).

Neural networks are a concept of computation very different from traditional computers, which belong to the so-called von Neumann architecture. While computers are explicitly programmed with rules to follow, neural networks learn all their ‘programming’ during optimization and training. Whereas traditional computers have separate memory and processing units, neural networks are built on many[3] separate, very similar processing units (called artificial neurons) that are organised in different layers and interconnected into networks in accordance with the tasks they are used for. When they receive feedback regarding their outputs, they change their operational parameters (weights and biases) until these parameters achieve an optimum during their training. This approach makes it possible for neural networks to efficiently perform complex tasks such as NLP that simply cannot be achieved with traditional programs. On the flip side, it also makes them less interpretable than traditional computer programs.

Neural networks[4] work in such a way that they are not explicitly programmed, they are just optimised during training so that the given neural architecture performs the given task effectively by using the results of the training, called the parameters. The optimised ‘model’ contains both the parameters (weights and biases) and the architecture itself.[5] This model is used for a specific task, such as text generation or classification, where a program library provides some input (a numerical representation of a text[6]) to the neural network, and the neural network generates outputs (also called as ‘predictions’) as quickly as possible.

Neural networks have been built according to many different architectures, and as we have mentioned above, being able to train these networks on the huge amount of data that recently became available has been a real game-changer in terms of their capabilities. One of the major reasons for this breakthrough is the ability to effectively train these models on a vast corpus acquired from the internet without the need for humans to manually create enormous amounts of training datasets for language understanding tasks (e.g., for a given input, what would be the correct output?). Creating such human ‘golden sets’ for all the different possible tasks of a neural network would be prohibitively expensive. Instead, automated training methods are used with the huge datasets, and these automated methods can substitute manual creation of golden sets. For example, masking a certain word in a text and predicting what that masked word could be is one such method – this method is often referred to as a type of unsupervised training of the language models.

Since 2017, the most performant language models have been created in two phases: first, a pre-trained model is created using unsupervised training methods. This is the most resource-intensive part of the training and is thus very expensive; nobody really wants to repeat it unless necessary (e.g. to get better results for a given language).[7] However, for most models,

[3] With current large language models, these numbers are in the magnitude of several hundreds of millions of artificial neurons and more.

[4] Of course, both the architecture of neural networks and the libraries calling these networks from traditional computers still have to be engineered to the specific task.

[5] This is a language model, but the word ‘model’ here also refers to a specific instance of the neural network as a computing application, containing one or more files invoked by the computer (e.g. a text representation of the architecture and a .h5 file for the parameters).

[6] Called embeddings.

[7] Nemeskey (2021).

the pre-trained phase is not directly usable for real-life NLP tasks (for example, it can guess a masked word very well, but that is hardly a practical use case). That's why a separate fine-tuning phase is needed after the pre-training. Fine-tuning is much less resource intensive; only a few hundred examples are often sufficient, but usually, a different fine-tuning is needed for different tasks, such as multi-label classification of text,[8] extractive question answering,[9] or text generation.

Since 2017, these pre-trained models have improved the state-of-the-art results in many benchmarks and outperformed other models that might have been trained from the ground up for a given task (with e.g. task-specific neural network architectures).[10] These pre-trained models have shown the power of large, generic language models that are trained on a huge corpus, and have also clearly shown that the larger the training data is, the better the results will be, and the more performant the model will become. Also, these pre-trained models have clearly demonstrated that they can be used to retrieve not only linguistic but also common sense and world knowledge.[11]

Hardly two years later, another revolutionary change took place:[12] some of these large language models became even more versatile in terms of their capabilities, requiring no further fine-tuning to achieve impressive results on benchmarks. When published, GPT-3[13] was unique in being able to be adapted to many ('downstream') tasks without fine-tuning, and when provided with a few in-context examples, GPT-3 was able to generalize to unseen cases without further fine-tuning. The model attained a versatile *meta-learning ability*, so that even without expensive fine-tuning, it can effectively recognize the patterns necessary for giving the correct answer.

While fine-tuning is still able to improve their performance, these models are capable of achieving excellent results in a large number of fields with minimal or no fine-tuning at all. Instead of the relatively resource intensive fine-tuning on hundreds of examples, now it is sufficient to give a dozen or even fewer examples (in few-shot learning). Even better, different tasks can be achieved by using different human-engineered (and human readable) prompts. For example, a classification can be carried out with a simple prompt such as the following: '*Which of these choices best describes the following document?: "[Class A]", "[Class B]", "[Class C]"*', while a text can be changed from first person to third person by giving the simple instruction '*Convert this from first-person to third person (gender female)*'. Astoundingly, these prompts can lead to language models performing complex tasks, such as translating code[14] into human languages or the other way around, or translating between human languages.

[8] For example, labelling all sentences in the judgements that are related to competition law.

[9] Based on a longer text, providing short answers to questions asked by the users during inference time, such as what kind of terminations are included in the contract, who the contracting parties are, etc.

[10] This is called the pre-training and fine-tuning paradigm; see Liu et al. (2021).

[11] Petroni et al. (2019).

[12] Liu et al. (2021).

[13] See Brown et al. (2020). Although, even GPT-2 had similar flexibility and meta-learning capabilities, it was considerably less effective, see also Radford et al. (2019).

[14] See, for example, the specific Codex model released by OpenAI prior to GPT-3.5 and 4 (that has since been superseded by these).

While this revolution was already accessible to specialists back in 2020 (and to programmers in 2021), the visibility of this exploded into public awareness only with ChatGPT in late 2022. This approach works only with the largest of the transformer-based generalist models that rely on prompts (instructions, completions, demonstrations) as described above. Even these generic language models need further research and fine-tuning to suitably follow instructions and to avoid harmful responses.[15]

Before turning to the use of OpenAI products and their APIs (application programming interfaces), we have to underline that these capabilities are not specific to OpenAI,[16] although, at the time of writing this study, they have a clear technical lead ahead of other companies, and an unquestionable lead in marketing.

This study is not a technical paper about what kind of LLMs perform better in NLP tasks, nor should it be understood that the types of models that we discuss (e.g. autoregressive or unidirectional language models or those using prompt-based methods) are inherently better than other models (e.g. such as the bidirectional model called ‘BERT’). We just want to provide some specific examples of the use of LLMs in the field of law that either already work in practice or could at least work in theory. In this study, we are interested in exploring LLMs that are relatively easier to implement for a larger user base, because this perspective is a priority for our research.

## 2.2. What is GPT?

Since 2018, OpenAI has released a number of new versions of its autoregressive[17] type of language model called GPT (Generative Pre-trained Transformer), 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.

With the help of the GPT-3 model, it was possible to run various labelling, text generation, text completion, summary, translation and generally dialogic tasks by means of text task definitions (prompts) as mentioned in the previous point, but this could only be done via application programming interfaces (APIs), so most of the world simply did not take notice of this (or any other models outside the workshop of OpenAI with similar capabilities).[18] The performance of the GPT model was further enhanced when OpenAI introduced GPT-3.5 on 28 November 2022.

Two days later, ChatGPT was introduced as well, serving as a user interface primarily for consumers to access a fine-tuned version of the GPT-3.5 model. The media coverage was significantly boosted by the release of this ‘consumer front-end’ for the language model, fine-tuned for chatbot functionality. To achieve the impressive performance of GPT-3.5,

[15] See the example of what work had to be done for the language models of OpenAI to understand and follow instruction prompts (‘InstructGPT’) at Ouyang et al. (2022).

[16] Wei et al. (2022).

[17] Here, autoregression means a training approach applied to a deep learning architecture, where the training is done by learning to predict the next token in a sequence where the previous tokens are given.

[18] One notable exception was the GitHub Copilot released for the public on 29 June 2021; see link1.

a considerable amount of human feedback was provided in a reinforcement learning method[19] to make the answers (chat completions) as close to what humans would expect as possible.[20]

The latest generation, GPT-4, was made accessible on 14 March 2023, offering substantial improvements over the previous GPT-3.5. Since March, users can enjoy the chatbot functionality of GPT-4 as well, under the trade name 'ChatGPT Plus', for a monthly fee of 20 USD (+VAT).[21]

Originally, chat functions within NLP were only used as a special field to enable convincing conversations, reduce the costs of the prohibitively expensive call centers or customer service support lines by answering simple questions online, or asking frequently needed clarification questions (e.g., before submitting a ticket to a human operator, or making reservations, etc.) Question and answer uses were made possible by matching dialogues with the most probable intentions, holding conversations, and extracting relevant information from previous statements. However, with the capabilities of ChatGPT it became clear that chat functions can also serve as an excellent interface for all the different varieties of tasks that LLMs are capable of, and understandably, human users prefer this approach as long as it remains reliable.

ChatGPT is a system whose operation is not transparent for end-users. Inputs provided by the users (even by paying users) may be used for further training of the model, and the language model (currently GPT-3.5 and GPT-4) is fine-tuned for chat discussions, with a user interface provided for chatbot use.

Even though ChatGPT uses very capable models, it is important to be aware that ChatGPT is not suitable for professional use other than research or replacing web searches (provided that no confidential information is used in the search queries). ChatGPT is a consumer product, and business users will have to rely on the application programming interface[22] made available by OpenAI. This programming interface uses the same model as ChatGPT but with some crucial differences: a) data provided via these APIs is not used for training (as stated by OpenAI in the terms of use at the time of writing[23]); and b) there is a very low fee to be paid based on the length of the text submitted and received.

Today, the largest language models are certainly capable of creating text of a quality that is difficult to distinguish from human-generated text. Based on the samples provided, linguistically

[19] Reinforcement learning is a type of machine learning that is not referred to as either supervised or unsupervised learning. The learning agent is able to process some input from its environment as a feedback, and thus updates the connections within the neural network; see Russell and Norvig (2016) 830.

[20] This Reinforcement Learning from Human Feedback (RHLF) was first researched by OpenAI with InstructGPT; see Ouyang et al. (2022); and see also footnote 15 on how the InstructGPT was essential in the creation of ChatGPT as well.

[21] Even though the chat function in the search engine Bing is supposed to be using GPT-4, currently, in practice the answers are much more restricted and cautious in Bing compared to what is available on ChatGPTPlus or via the APIs.

[22] Using specific libraries provided by OpenAI, very simple computer programs invoke this API, for example, defining a prompt and providing some optional examples, then sending the 'openai.ChatCompletion.create' message.

[23] See 3.(c) of Terms of Use at link2. (at the time of access): 'Use of Content to Improve Services. We do not use Content that you provide to or receive from our API ("API Content") to develop or improve our Services. We may use Content from Services other than our API ("Non-API Content") to help develop and improve our Services. …'

correct complex transformations can be carried out on some sample texts, be they contractual or judgment provisions. However, this does not mean that the use of these models does not have fundamental application limitations, which, for example, the publishers of the GPT model also describe over several pages,[24] and which are applicable to ChatGPT as well.

Translating these deficiencies into practical terms from the legal point of view, in a way that allows us to draw conclusions affecting a wide range of society, is not an easy task. This will only be possible through extensive mapping and experimentation of individual applications. However, such experimentation and research are essential, as the training of future professionals must be based on these revealed and abstracted limitations, and the tasks of legal work must be adapted to such characteristics.

It's crucial to understand that there are other, fully open and downloadable large language models similar to GPT that are almost as good in many respects. There are also language models that perform even better at certain tasks than GPT. Due to the current setup and limitations (e.g., GPT is not available for download and can only be used via the API provided), it is simply not possible to carry out certain essential language-related tasks when using GPT.

Nonetheless, the availability of prompt-based, few shot learner large language models capable of carrying out various NLP tasks with just a human-readable change of the prompt, and by simply providing instructions, will not only change the economics of the legal profession and the way we carry out these tasks, but in the long term, it may also change law itself (and most other professions without a dominant physical element). Some authors from reputable universities have even dared to call the GPT model a first, early manifestation of artificial general intelligence (AGI).[25]

So even if GPT-4 is not a 'strong AI', and even if it will never be able to conclude the never-ending disputes about how to achieve artificial general intelligence (and also how to define human intelligence and humanity or how to differentiate our roles from those of other types of intelligence), it is a clear sign that there are unexpectedly simple mechanisms working behind our complex language capabilities, including our legal thinking. Probably, a lot more of these activities can (and should) be automated than we are currently comfortable with.

## 3. PRACTICAL USES OF LLMS IN LAW

The purpose of this section is to present the most important uses of large language models (LLMs) in law, where we mean LLMs with capabilities at least as advanced as that of GPT-3. In some parts, more specific examples are useful both as illustrations and as support for our claims; there, we might refer to the capabilities of GPT-3.5 or GPT-4 (even if we just use 'GPT' in general). The list below is far from complete. Additionally, we note that so far, only blog posts and non-scientific analyses have been published on the operation and legal uses of GPT. No systematic, scientific investigation has yet been published, so we have had to treat the cited sources with criticism.

[24] OpenAI (2023).

[25] Bubeck et al. (2023).

## 3.1. Text retrieval and legal information for the public

The earliest legal activity to be computerized was text retrieval. Searching in large bodies of text (e.g. a massive legal case database) with simpler tools (e.g., full-text search engines, or with the help of other content-oriented tags and indexes) has existed since the 1950s. However, text retrieval for legal use became truly effective when *semantic search* appeared. Semantic search is a collective term that refers to methods that use, in addition to plain text, either the searcher's intention or the representation of the deeper relationships of the sources (texts, images, other digital files) to find and output the results. The machine may recognize deeper connections (e.g., by recognizing the context described above or by recognizing image elements), or they may be produced beforehand through human effort (e.g. by labelling). The best example of an advanced semantic search engine is Google itself, which tries to figure out the searcher's intent based on various factors, including their geographic location and search history. It also determines the 'deeper meaning' with the help of millions of parameters, including how many other pages refer to the given webpage or text.

LLMs can bring a real breakthrough in the search of legal texts. However, search and retrieval of legal texts is not as easy (yet) as simply asking a natural language question or feeding the LLMs with all the legal texts, expecting them to 'remember' them, and then asking questions. The problem lies more in the first part of the process.

While LLMs can answer questions they were pre-trained on[26] - and a surprisingly large number of legal texts have been included in pre-training[27] - this approach is not very practical for legal uses. First, the pre-training was carried out based on data prior to September 2021 and is not updated periodically, neither with laws nor with court cases. Secondly, in legal work, having access to the widest possible scope of relevant public documents (including local regulations and court cases) is often critical, and reviewing non-public documents, such as contracts, also forms an essential part of legal work.

These large legal texts cannot be simply fed into the model because the maximum length of input text for LLMs is rather limited. For example, for BERT, this is as small as 512 tokens,[28] and even for the latest GPT-4, this peaks at 32,768 tokens. That is why, even with large language models, one has to use a staged information retrieval: a first step uses a fast, but coarse retrieval method, and a second stage ranks the possible set of answers and either presents it to the user as-is, in its raw form, or answers the question from the input provided by this retrieved document snippet and a chat completion function.

[26] Petroni et al. (2019).

[27] With ChatGPT, we can identify that pre-training included many acts of the Hungarian Parliament before September 2021, and also some government decrees etc. Of course, the scope of documents included in the pre-training are probably larger for legally relevant English language documents as well (e.g. 92% of the 570 GB corpus used for pre-training GPT-3 was in the English language, and 'only' 0.08% in the Hungarian language, which is relatively high, considering that the second largest corpus of French and the third of German are also only around 1.8 and 1.7%, respectively); see link3.

[28] The conversion between text and token depends on both the language and the embedding methods used in a language model. For illustration, Article I of the Treaty on European Union is 677 characters in English and *142* tokens in the embedding used by GPT-3, while it is 630 characters in Hungarian, but *362* tokens.

The question of providing a better neural network based solution to the first step is not trivial, and is subject to numerous research[29] projects. However, there are already some solutions that work to some degree and rely on the capabilities of large language models. Being statistical models, all language models have to use numerical representations of the text, called embeddings. Different embedding methods have different costs in terms of use and provide different performance. Large language models use ‘contextualised dense vector embeddings’. This means the following: while standard, term-based information retrieval methods (such as TF-IDF) rely on the frequency of occurrence of specific words in a text (or documents in a corpus), and retrieve information based on such frequencies in the keywords of the question, neural retrieval methods rely on a neural network based transformation of both the question (query) and the documents to be searched. These learned transformations are called embeddings, and they are able to capture the semantic similarity between words. ‘Contextualised’ embeddings like those used by LLMs can also capture context-dependent meanings of the words. This enables a much richer understanding of words and sentences, both in the question (the information retrieval query) and within the documents to be searched through.

Thus, LLMs can help even with the first stage of information retrieval by converting the documents, document parts – or even just some automatically generated summaries of larger documents – into contextualised dense vector embeddings and storing them in a fast database. When a question is formulated for the search in a natural language query, LLMs are used to transform the query into an embedding, and the vector database can be efficiently searched for the closest, most similar embeddings, which also means the documents (parts or summaries) that are closest in *semantic meaning* to the question.

A second stage can be used to review, rank or score the several ‘possibly relevant’ document parts and retrieve only the most relevant, or to feed the LLM both the retrieved short document part and the query as a prompt (input). This latter version could be used to either directly answer the question in natural language or to extract the relevant part of information from the given document part (e.g. the date of termination).

While such solutions may even be integrated into ChatGPT,[30] the most important aspect from the point of view of access to justice is that it can revolutionise the provision of legal information to both professionals and laypeople.[31]

Relying on legal texts and expressions in their context, LLMs could, in theory, be made able to answer questions asked by laypeople in non-legal language, and to formulate the answers in non-legal language. In addition, instead of simply repeating the text extracted from legal sources, it can respond precisely to the question asked and reformulate legal information into practical steps. These models can produce all this in continuous operation (24/7), practically immediately, with minimal costs and without any necessary social interaction with humans.

However, there are certain limitations and customisation needs for the system, which we will address in section 4 below.

[29] See Luan et al. (2021); Ma et al. (2021).

[30] See link4; link5.

[31] For example, see the possibly largely computer generated blogpost, Rogan (2023).

## 3.2. Text generation and document assembly

Another form of use, which was computerized and utilized by legal professionals early on, was text generation, more specifically, document assembly. Document assembly systems usually consist of two different modules: one focusing on the authoring of the templates (defining relevant text parts, their relationships, sources of information to be included, and the business logic that defines the document creation process). The other module is the interview module, where end users input all necessary information specific to the instance of the document being created. The system fills out variables, combines the text elements according to users' instructions, and prepares a relatively accurate document.[32]

Since GPT was created specifically to produce text, it's no surprise that ChatGPT can write almost perfect legal documents at first glance, as noted by Jack Shepherd in his blog about ChatGPT. At the same time, considering that LLMs do not understand the law in the same way people do but put one statistically appropriate word after another, it is normal that such documents contain a number of quite primitive errors.[33] As Shepherd notes, since it does not understand the context, it very rarely asks clarifying questions before providing some results. For example, it never asks about the governing law and thus sometimes produces sentences that make no sense overall. His conclusion is that 'at least right now, the use case for that version of ChatGPT that he was using is less *drafting contracts* and more *producing first drafts of contracts*'.

This blog post was about using version GPT-3.5 on a chat tool that is not intended for professional use. How could legal professionals make good use of LLMs in terms of text generation and document assembly?

Similar to text retrieval, LLMs can be used in a multi-staged approach as part of a more complex system. If we just focus on the text generation of larger documents such as contracts, a possible approach would be to define and engineer three different steps.

The central part of the document assembly solution would be an approved clause bank that would operate similarly to the first stage of the text retrieval 'text base' as described above. The clause bank is to be built from generalised provisions of text that can be reused in as many contexts as possible, while:

a) still retaining clear references to the specific roles of entities appearing in the clause (including both subjects such as parties or objects such as properties, movables, rights, etc.), and also
b) storing pertinent metadata about the given provisions (such as the governing law or jurisdiction where the given clause may be used, how much and in what way the provision benefits certain contracting parties, or any other information that would be relevant in a specific context and for inclusion in a given document).

Here, LLMs' job is merely to facilitate a search based on content (represented in embeddings).

In the process of training and personalizing large language models, the most challenging aspect is not creating the clause bank but rather defining the 'table of contents'. This 'table of contents creator' step begins by asking the user (referred to as the 'interviewee') about the contract's specific needs and then determining which clauses should be included in the

[32] E.g. see Susskind (2010) 101; Homoki (2022a, 2022b).

[33] Shepherd (2023).

document based on those requirements. This is the stage where further fine-tuning of the given LLM is necessary, and which also makes it essential to address the issue of the scope of the document assembly system.

Currently, contracts are very dissimilar in nature: standardisation in language is more of an exception than the rule, even within a given jurisdiction and language. The wider the possible set of requirements are, the more likely that such a document assembly system will be using inappropriate or dangerous clauses, and the more thorough the legal review process will have to be after assembly.

It does not seem realistic to achieve a refined and balanced document assembly system for an entire legal system. Also, LLMs are probably not appropriate to define the actual problems of a consumer based on the direct instructions given by that consumer (cf. 4.4).

However, it is realistic to create a manageable sized 'table of contents creator' for a given company, even a large one, or for a given law firm or public notary's office, who serve a well-defined, standardised range of customers. Instead of consumer-facing solutions, it is more appropriate for these document assembly solutions to interface with professionals.

With such a restricted scope in mind, the fine-tuning necessary for this 'table of contents creator' could be much simpler in theory: it would need a couple of hundred text pairs composed of 'stated requirements'–'necessary headings'. From the users' point of view, a separate user interface would make sense to restrict the most important requirements to a tree of most common choices, with some extra place for customised, individual instructions. This interface would be responsible for creating the 'stated requirements' serving as the input to the table of contents creator LLM. Of course, the possible set of common choices and the 'necessary headings' are best created from the existing contract corpora of the given company, but this will be discussed in the next section.

The final step, which is the easiest to accomplish, is to make the necessary linguistic and text adjustments to each separate clause retrieved from the clause bank based on the 'necessary headings' output of the second stage (e.g. changing declensions, conjugations, the number of parties, terms). These tasks are trivial for LLMs, while this was problematic for non-LLM based document assembly systems, especially for non-English speakers.[34]

In addition to the subject of document assembly, LLMs (especially GPT) can be used in various ways for text creation and writing assistance. They are excellent tools for spellchecking and for stylistic recommendations, as well as verifying the citation formats.

For current commercial providers of such existing plugins and Word add-ins, it will probably be quite difficult to remain relevant in the coming years. The reason is that these LLMs are very versatile, and their functionality can be changed by simply giving different prompts and examples, so with minimal programming, a single LLM-based plugin can cover what previously only several different plugins provided. Also, considering that most legal professionals use standard commercial office applications as their everyday tools, it is rather likely that the suppliers of these applications will provide some or most of these plugin functionalities out-of-the-box for a wide range of professionals in exchange for a subscription fee,[35] thus displacing the market for existing plugin providers.

[34] See Homoki (2022a) 22.

[35] On 16th March 2023, Microsoft announced Microsoft 365 Copilot; see Stallbaumer (2023).

## 3.3. Analysis of law: classification, text extraction and language inference tasks (e-discovery, e-due-diligence, legal analysis)

Tasks related to the natural language understanding branch of NLP comprise the third major area to discuss. This field includes classification of text segments (from the level of tokens to several documents), extracting information from text (such as dates, entities) and determining the relationship between two pieces of texts through 'natural language inference' (for example, is one sentence supporting or refuting another, are these arguments defined in relation to the same point of law, is there a contradiction between this conclusion and this statement? etc.). Let's take a look at each of these subfields in a little more detail.[36]

Automatic classification is a very old branch of NLP. The aim is for the machine to sort through a large amount of text and categorise parts of the text into predefined categories based on certain rules (referred to as labelling the text segments with different categories). The rules can be very simple (e.g. occurrence of words or groups of words in certain texts) or more complicated, such as the semantic content of the text (for example, is this an employment contract with a salary above 500,000 EUR that is no longer in effect?). Such a classification is used by many legal IT systems, and we highlight two of them here, the so-called e-discovery and e-due diligence systems.

E-discovery (also known as electronic disclosure in some jurisdictions) refers to finding a large number of documents related to legal proceedings such as litigation, audits, investigations, etc., where the information sought is in electronic format. The importance of e-discovery varies from jurisdiction to jurisdiction, depending on the conditions and likelihood of a court ordering such disclosure and the possible consequences of not complying fully with the request. In the United States' jurisdictions, market demand for computerised support in e-discovery was strong enough for this field to grow into a major product segment. E-discovery works by using technology to help find relevant information about a case. It is the process of preserving, collecting and analysing electronic data in response to a discovery request in a legal proceeding. This is partly information retrieval, but also a problem to be solved with the help of natural language understanding, mostly through classification.

Another typical purpose of document sorting is legal due diligence, where the aim is to find certain signs of risk within large amounts of legal documents or to find specific types of documents that have to be examined by lawyers or by automatons in more detail. Due diligence activities are usually carried out in relation to certain events, such as preparing for the sale or acquisition of a business (to determine the risks and soundness of the acquisition, or the purchase price), or as part of a wider audit activity (to find, for example, non-compliance). A typical task, for example, is to search for contracts among many thousands that contain different liability or termination rules than is usual (cluster analysis or analysis of outliers, both as unsupervised classification), or find those contracts that are subject to mandatory arbitration.

As seen above, classifications can be made according to unsupervised machine learning methods (cluster analysis), but also according to very specific criteria, which was usually based on supervised learning. In this regard, LLMs can drastically simplify the costs of classification

[36] Homoki (2022a) 24–27.

and enable users to discover new ways to classify document provisions, without having to fine-tune separately for each different classification task.[37]

Of course, this depends on the type of classification and the content to be classified. One has to be aware of the aforementioned token limits, but with GPT-4, even shorter contracts (e.g. employment contracts) may be fed into a single prompt together with the instructions. In most cases however, it is not the best approach to feed the complete contract, so the relevant provisions will have to be extracted first. If the relevant contract part is still too large, it may be split into multiple parts and sent to the LLM piece by piece. However, in such cases one has to be careful not to split the text in such a way that some relevant context is lost for the classification task (e.g. cross-references between termination rights in the contract). Another valid approach here is the same as mentioned in the part on information retrieval (retrieving the relevant provisions of the contracts from a database by way of the similarity of embeddings and representations of the texts).

The outstanding ‘few shot’ and ‘zero shot’ learning capabilities of GPT and other LLMs make it possible to use these LLMs in such a way that only a good prompt is defined for several simultaneous classification tasks; then the same prompt is fed with each separate provision of all the contracts, piece by piece.

Besides classification, the same LLMs may be used to extract relevant information from a very large document set, such as finding contracts above a certain value threshold. The only challenging part in this task is to segment the text in such a way that values necessary for the calculation of the threshold should preferably stay in the same text segment.

Although language inference tools have also been a subject of research in law for over thirty years, they are not yet widely used in legal practice. These tools could be used to reveal the hidden structure of arguments in larger documents such as briefs, verify whether certain claims are supported by law or any of the evidence disclosed, or whether new statements of the plaintiff are in contradiction with previous statements etc. The technical way to do this with GPT would be exactly the same as for classification and extraction, i.e. feeding the statements to be tested against each other in the same prompt and with prompt instructions on the type of relationship between the two sentences that will have to be examined.

Inference tools could benefit other areas as well, such as contract negotiations or legislation procedures by, for example, enriching the automatic summaries of differences between versions, or assisting in the provision of automated explanations for changes, etc.

## 3.4. LLMs as enablers in law – outside direct legal operations

We have listed a number of new kinds of uses for LLMs in traditional areas of NLP, and these uses all relate to how legal professionals directly work with text (for example, drafting or analysing text etc.). Besides these, we expect that some of the most interesting changes will result from those uses of LLMs that enable further scaling of the work of humans or extend the possible use of other tools. We call these ‘catalyst’ uses of LLMs.

Perhaps the most important such use is the *training of humans*, more specifically, legal professionals. Similar to the problems with self-driving cars, LLMs may turn out to be not reliable enough in many critical areas for direct consumer uses, or even in just augmenting

[37] As mentioned above, fine-tuning is also an option for GPT, which could further increase the reliability and performance of these neural classifiers.

critical work of legal professionals. Even in such an unlikely case, even based on the current capabilities of LLMs, the conversational skills of current GPT generations are already able to help train new generations of lawyers at a much lower cost, with a more customised experience, and at a much greater depth than is currently possible in law schools and universities using traditional methods.

With the help of LLMs, training and testing materials for humans can be turned into more practical and realistic exercises (which is an important subject for legal uses), which can be rolled out at scale. These tools can also enable a human to supervise more students at once than is currently possible.

At the same time, it takes an enormous amount of preparation to do this. Such preparation includes a thorough audit of what a specific type of LLM is capable of (without fine-tuning), identifying in which legal fields that particular LLM is not reliable, by fine-tuning in what kind of areas these ‘hallucinations’ could be significantly decreased, and how the methods described above (such as fine-tuning or connecting it to knowledge bases) will affect the reliability in general. There is also no other way to find out whether the LLM will be able to handle (‘understand’) multi-layered, higher level concepts of law as well, and how these abilities are affected by further fine-tuning or by merely relying on connections to databases by vector embeddings.

Only such experiments can show us where and how to use LLMs in training and which fields should be trained by humans.

Another catalyst could be the facilitation of the operation of *knowledge management* systems, by making the capture of individual knowledge easier, requiring less human intervention and supervision. While all organisations today would benefit from the systematic recording of knowledge relevant to their operations and from the easy retrieval of such information, only the best funded and managed organisations have the capabilities to do so. These large, well-funded organisations have dedicated personnel (e.g. librarians, professional support lawyers, quality specialists etc.) to ensure that the processes of operations are well documented and kept up to date (by way of, for example, the quality or information management systems in place). Even for many such organisations, knowledge management may fail to cover every important operational aspect. The most difficult elements of knowledge management are separating relevant knowledge worth recording, recording the knowledge in a way that can be reused outside the original context, and also collecting sufficient metadata about this piece of knowledge to enable its later retrieval. In this area, LLMs are able to help human-staffed organisations to achieve their potentials.

The third area of catalyst is the possible role of *LLMs as middleware between different IT systems* and AI solutions. Even today, GPT is able to act not only as a conversational agent for humans, but as a technical interface between different agents (including other GPT calls by other companies), provided that it has either been pre-trained on such information, been explicitly given such information in the prompt, or is able to retrieve such definitions via third party API calls and act accordingly. Of course, only time can tell how reliable these interconnections will be in the longer run, but the advantage of this approach is its greater flexibility and resilience when certain changes occur in the defined APIs.[38]

[38] There are already some trials available in this area; e.g. Langchain (link6) or other similar experiments, such as Nakajima (2023).

The fourth and last area of a possible catalyst is that of an *enabler of training other AI solutions*. As we have already discussed in the introductory section, the costs of training necessary for a supervised learning approach often act as a barrier to the creation of such AI models. A number of possible AI applications or task specific fine-tunings of LLMs cannot benefit from the unsupervised or reinforcement based training methods. Designers of new AI models may benefit from the capabilities of already existing LLMs: existing LLMs can help humans in finding training data or in creating, multiplying, and cleaning or converting such data. The finest example of this is the way a much smaller,[39] but still capable LLM called Stanford Alpaca was fine-tuned for the instructions-following capabilities mentioned in section 2.1.[40] They used GPT (3.0) API calls to create sufficient training data to ensure this instruction following capability, and thus were able to complete this fine-tuning for a total cost of less than 600 USD.[41]

# 4. LESSONS LEARNED FROM BUILDING A GPT-BASED CHATBOT FOR A LAW FIRM, AND THE LIMITATIONS OF LLMS

## 4.1. Describing the chatbot demo

In order to have a better understanding of the operation of (Chat)GPT in the legal environment we conducted an experiment with the OpenAI API (using the chat completion API that operates behind ChatGPT as well) by building a demo chatbot for a small law firm.[42] The demo aimed to model how a chatbot at a small law firm could theoretically operate in public, but it also provides a number of lessons that can be used by larger organisations and chatbot users in law in general.

The demo chatbot used the GPT-3.5 model, mainly for reasons of economy: answering via the GPT-4 costs 15 times as much. Another advantage of GPT-3.5 is the greater speed in answering the questions, which is an important factor for chatbot uses. With GPT-4, the answers would have been more precise, but such performance was not measured. A further major advantage of the GPT-4 model for chatbot use is the longer token (size) limit, as set out below in more detail. In every other way, operation under GPT-4 would be the same.

When using the OpenAI API (and not the ChatGPT interface), one can very easily customise 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. The chatbot is actually nothing more than a) a front end for the chatbot with the branding of the law firm, b) some customisations produced by providing examples and extra prompt instructions that are fed into the API chat completion call, together with the actual question of the user entered on the front end.

The examples[43] are made of pairs of questions and answers, some in English, some in Hungarian, and cover some important limitations, such as what to do with requests that fall

[39] Using Meta's LLAMA model with 7 billion parameters (weights, biases and embeddings) compared to the 175 billion parameters of GPT-3.

[40] See footnote 15 for more details. Fine-tuning on instruction following capabilities ensures the alignment needed to follow the instructions, while avoiding biased or toxic text as much as possible.

[41] Taori et al. (2023).

[42] The source code is available at link7. The chatbot itself was available at link8.

[43] Link9.

outside the competence of the law firm (how to redirect the user to the bar association's lawyer search functionality). There are two types of prompt instructions, system or user prompts, where a system prompt would be the description of what kind of persona the chatbot should be trying to impersonate,[44] and the user prompt would be the one submitted to the API to get an answer to. However, the differentiation between the two is not very strong in GPT-3.5, so in the demo, some instructions as to what the chatbot should and should not do are also included in the user prompt.

When creating a chatbot for a law firm, one has to be aware of the many deontological rules that apply to such activities; for example, avoid giving answers that could be understood as comparative advertising, even just saying something along the lines of one specific law firm being better than another specific firm, etc.

Equally important as the deontological rules is giving all the relevant details to the chatbot about the law firm being marketed as part of the prompts. Without this vital information, GPT will 'hallucinate' (and not do, for example, an internet search for missing information). For example, during the first tests, we gave the model the phone number of the law firm office, but not the physical address. When we asked the chatbot for the contact details of the law firm in general (not just the phone number), the completion included a very precise and existing physical address – the only problem was that this address was not that of the law firm.

However, the limitations of size already mentioned also affect how much customisation we can do with such a chatbot. For GPT-3.5, there is a strict limit of 4,096 tokens, which include both 'prompt' (the question) and 'completion' (the answer). Furthermore, the prompt size limit includes all the examples and prompt instructions, as well as the chatbot user's actual question, and the longer these customisations are, the shorter the answers will have to be.

So even if a lot more customisation would be useful, and a lot more information about deontology rules or the firm could be inserted, there is simply not enough space for that in this kind of solution (using GPT-4 would lessen the effect of these limits).

The chatbot front end was created as a bilingual one, but, of course, otherwise the chatbot relies on GPT's multilingual capabilities. GPT's multilingual capabilities seem to rely on some sort of built translation mechanisms for both the query and the answer, and not, for example, by generating text originally in the language it was asked. This is apparent if we ask GPT in a non-English language to create a verse with rhymes. Our experience was that in such a case, although GPT gives the verse in the same language as the one used for asking, the rhymes do not rhyme, unless the sentence is translated back to English. Also, we asked GPT a legal question in Hungarian about a will that can only be answered based on some superficial knowledge of the Hungarian Civil Code (Pflichtteil as a compulsory part – this exists in Hungarian law, but does not exist as a term in English law). Although GPT provided the answer correctly, and clearly had some knowledge of the Hungarian Civil Code's provisions on inheritance, the Hungarian term used was clearly incorrect,[45] and simply a word-by-word translation of a non-technical term in English (e.g. compulsory part).

[44] E.g. 'give your answers as if you were an experienced, measured lawyer…'

[45] *Kötelező rész* instead of the correct *kötelesrész.*

## 4.2. What can lawyers and small law firms use such a chatbot for?

For what purposes can lawyers and law firms use the kind of chatbot used in the demo? Actually, we can only use such a chatbot to provide information about a law firm in a slightly more entertaining way than what can be achieved on a plain website. Additionally, one can make this chatbot simultaneously available on other channels as well, such as on a Telegram or Viber chatbot etc. Essentially, chatbots like this can only be used for 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 the 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.[46] Of course, to do so, the terms of use of the law firm must clarify that this is not legal advice and should not be used for any real purposes. It's important to differentiate between this entertainment value and the legal advice actually given by the law firm (and not by the chatbot).

Version 3 and beyond of GPT are not downloadable, Microsoft (the biggest investor of OpenAI) has had an exclusive license to the models since September 23, 2020. Nonetheless, the language models are all accessible via the web services called application programming interfaces (APIs) that have been provided by OpenAI since at least early 2021. Currently, no on-premise use is possible, and all requests have to go through OpenAI, and answers will come from them as well. Although the API calls and results will not be used for the training of OpenAI models as promised by OpenAI, the contractual promises made by OpenAI in this regard do not provide sufficient assurance for all use cases (such as for the transfer of personal data to the USA). For use cases requiring more robust certifications and promises of security, there is already an offer from Microsoft for the use of OpenAI models with locations in the EU, as well.[47]

Terms of the OpenAI API usage policy at the time of writing this paper clearly stated that these models should not be used for providing legal services without a review by a qualified person.[48] This means that, due to these usage policies of OpenAI, this model may not be used in consumer-facing front ends. (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 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 safer to do so via a dedicated application (with the possibility of connecting to payment services to give weight to the time slots booked).

[46] E.g. in Google's DialogFlow-based NLP chatbots, or the even simpler keyword-based chatbots.

[47] Link10.

[48] OpenAI Usage Policies: '*Unauthorized practice of law or offering tailored legal advice without a qualified person reviewing the information*: 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.' (Link11)

While this particular demo chatbot can only be used for client-facing purposes, the processing capabilities of the OpenAI API (including GPT completion uses) nonetheless extend beyond this simplistic chatbot functionality as described in the previous sections.

### 4.3. Lessons, limitations, conclusions, possible agenda to step forward

In previous sections, we have listed in detail how LLMs may be used in law or by law firms. The advantage of such LLMs is that implementing them, as shown in the demo, does not require significant resources. LLMs may become flexible, everyday tools for every profession, and if properly built into versatile applications, they can greatly improve that organisation's capabilities, simplify its IT infrastructure, and perhaps even save money currently paid to a number of suppliers and integrators. For legal professionals who use a large number of different IT products, these LLMs and APIs to such LLMs could also serve as a way of reducing the number of required products and the costs of integrating them.

Because of the inner logic of text generation, large language models are referred to by some as 'stochastic parrots', and experts warn that they therefore cannot substitute real communication. This is the first, theoretical limitation of any LLM in legal work. According to this view, human communication is always a 'jointly constructed activity', and when we communicate with other people 'we build a partial model of who they are and what common ground we think they share with us, and use this in interpreting their words'.[49] However, this does not pose a problem with regard to a number of legal applications, since the characteristic of legal texts (policies, contracts) is precisely that they fix certain rules regardless of the identity of the participants. In the same way, this feature is not a problem when selecting and summarising appropriate texts from a sea of legal sources or preparing summaries from longer texts.

At the same time, this can be a very serious limitation of the operation if a layman requests legal advice from the system. This limitation can also be formulated by saying that LLMs only have access to the texts and not to the reality itself, so - for the time being - they cannot perform the reality check that a legal consultant would perform immediately. We do not (yet) believe that LLMs may have the same emotional intelligence as a professional, and even if an LLM was capable of picking out contradictory signs from a communication, they are not trained to act upon such contradictions. For example, if it is obvious from the client's statement that he is hiding something or slightly distorting some facts, an experienced lawyer can immediately ask him back. A chatbot is not able to do the same.

If we need to highlight one particular area where we believe further contribution from legal professionals would be useful, it would be the need 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, for example, the chat completion question answering capabilities correlate with these. We must determine the strengths and weaknesses of these chat completions in legal applications because no one else will be able to answer that in our stead.

[49] Bender et al. (2021).

Similarly, the possible non-chatbot based use cases described in section 2 should work in practice, but there is no way to find out how reliable they could be unless prudent experiments are made on a large scale, in many countries, under many jurisdictions, with the involvement of legal professionals.

## LITERATURE

Bender, E. M., Gebru, T., McMillan-Major, A. and Shmitchell, S., 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?' in (n.d.) (ed.), *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (March 2021)* (Association for Computing Machinery 2021) 610–23 <https://doi.org/10.1145/3442188.3445922> accessed 8 April 2023.

Brown, T. et al, 'Language Models Are Few-Shot Learners' in Larochelle, H., Ranzato M., Hadsell R., Balcan, M.F. and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)* (Curran Associates 2020) <https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html> accessed 19 February 2023.

Bubeck, S. et al, 'Sparks of Artificial General Intelligence: Early experiments with GPT-4', *arXiv* (22 March 2023) <https://arxiv.org/abs/2303.12712> accessed 8 April 2023.

Homoki, P., *Guide on the Use of Artificial Intelligence-Based Tools by Lawyers and Law Firms in the EU* (Council of Bars and Law Societies of Europe 2022a) <https://ai4lawyers.eu/wp-content/uploads/2022/03/CCBE-ELF-Guide-on-the-use-of-Artificial-Intelligence-by-Lawyers.pdf> accessed 19 February 2023.

Homoki P., 'Miként lehet a szöveggeneráló eszközöket a jogászi hivatások körében hasznosítani?' (How can text generation tools be used within the different legal profession?) in Ződi Zs. (ed.), *Jogi technológiák: digitális jogalkalmazás* (Ludovika Egyetemi Kiadó 2022b) 185–206. For an English version, see <https://ai4lawyers.eu/wp-content/uploads/2022/03/Report-on-the-use-of-text-generation-tools-by-legal-professionals.pdf> accessed 8 April 2023.

Liu, P. Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. 'Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing', *arXiv* (28 July 2021) <http://arxiv.org/abs/2107.13586> accessed 26 March 2023.

Luan, Y., Eisenstein, J., Toutanova, K., and Collins, M., 'Sparse, Dense, and Attentional Representations for Text Retrieval', *arXiv* (16 February 2021) <http://arxiv.org/abs/2005.00181> accessed 8 April 2023.

Ma, J Korotkov, I., Yang, Y., Hall, K. and McDonald, R., 'Zero-Shot Neural Passage Retrieval via Domain-Targeted Synthetic Question Generation', *arXiv* (27 January 2021) <http://arxiv.org/abs/2004.14503> accessed 8 April 2023.

Nakajima, Y., 'Task-driven Autonomous Agent Utilizing GPT-4, Pinecone, and LangChain for Diverse Applications', *Yohei Nakajima* (March 28, 2023) <https://yoheinakajima.com/task-driven-autonomous-agent-utilizing-gpt-4-pinecone-and-langchain-for-diverse-applications/> accessed 8 April 2023.

Nemeskey, D. M., 'Introducing HuBERT' in Berend, G., Gosztolya, G. and Vincze, V. (eds), *XVII. Magyar Számítógépes Nyelvészeti Konferencia* (Szegedi Tudományegyetem TTIK, Informatikai Intézet 2021) 3-14 <https://hlt.bme.hu/media/pdf/huBERT.pdf> accessed 8 April 2023.

OpenAI, 'GPT-4 Technical Report', *arXiv* (16 March 2023) <http://arxiv.org/abs/2303.08774> accessed 26 March 2023.

Ouyang, L. et al, 'Training Language Models to Follow Instructions with Human Feedback', *arXiv* (4 March 2022) <http://arxiv.org/abs/2203.02155> accessed 19 February 2023.

Petroni, F. et al, Language Models as Knowledge Bases?' in Inui, K., Jiang, J., Ng, V. and Wan, X. (eds), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Association for Computational Linguistics 2019) 2463–73 <https://aclanthology.org/D19-1250> accessed 8 April 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. 'Language Models Are Unsupervised Multitask Learners', *OpenAI Blog* (2019) <https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf> accessed 19 February 2023.

Rogan, J., 'How to Use ChatGPT for Legal Advice', *GripRoom* (16 February 2023) <https://www.griproom.com/fun/how-to-use-chatgpt-for-legal-advice> accessed 8 April 2023.

Russell, S. J. and Norvig, P, *Artificial Intelligence: A Modern Approach* (3rd edn, Global edition, Pearson 2016).

Shannon, C. E., 'A Mathematical Theory of Communication' (1948) 27 *The Bell System Technical Journal* 379–423.

Shepherd, J., 'Chat GPT for contract drafting: AI v. templates', *Medium* (9 February 2023) <https://jackwshepherd.medium.com/chat-gpt-for-contract-drafting-ai-v-templates-50ec8fd42f44> accessed 8 April 2023.

Stallbaumer, C., 'Introducing Microsoft 365 Copilot', *Microsoft 365 Blog* (16 March 2023) <https://www.microsoft.com/en-us/microsoft-365/blog/2023/03/16/introducing-microsoft-365-copilot-a-whole-new-way-to-work/> accessed 9 April 2023.

Susskind, R., *The End of Lawyers? Rethinking the nature of legal services* (OUP 2010).

Taori, R. et al, 'Alpaca: A Strong, Replicable Instruction-Following Model', Stanford University, Center for Research on Foundation Models (2023) <https://crfm.stanford.edu/2023/03/13/alpaca.html> accessed 8 April 2023.

Vaswani, A. et al, 'Attention Is All You Need', *arXiv* (5 December 2017) <http://arxiv.org/abs/1706.03762> accessed 8 April 2023.

Wei, J. et al, 'Emergent Abilities of Large Language Models', *arXiv* (26 October 2022) 7 <http://arxiv.org/abs/2206.07682> accessed 10 April 2023.

## LINKS

Link1: 'GitHub Copilot', *Wikipedia* <https://en.wikipedia.org/wiki/GitHub_Copilot> accessed 8 April 2023.

Link2: 'Terms of Use', *OpenAI* <https://openai.com/policies/terms-of-use> accessed 8 April 2023.

Link3: 'OpenAI GPT-3 Dataset Statistics: Languages by Character Count', *GitHub* <https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_character_count.csv> accessed 8 April 2023.

Link4: 'ChatGPT Plugins', *OpenAI Blog* (23 March 2023) <https://openai.com/blog/chatgpt-plugins> accessed 8 April 2023.

Link5: 'OpenAI ChatGPT Retrieval Plugin', *GitHub* <https://github.com/openai/chatgpt-retrieval-plugin> accessed 8 April 2023.

Link6: 'Introduction', *Langchain* <https://python.langchain.com/> accessed 8 April 2023.

Link7: 'Homoki Ügyvédi Iroda (Homoki Law Office) 2023', *GitHub Chatbot demo* <https://github.com/Homoki-Ugyvedi-Iroda/chatbotdemo-lawfirm-2303> accessed 8 April 2023.

Link8: 'Homoki Ügyvédi Iroda (Homoki Law Office)', *Chatbot demo* <https://chatbotdemo.homoki.net> accessed 8 April 2023.

Link9: ‘Homoki Ügyvédi Iroda (Homoki Law Office) – prompt examples’, *GitHub Chatbot demo* <https://github.com/Homoki-Ugyvedi-Iroda/chatbotdemo-lawfirm-2303/blob/main/static/prompt_qa_examples.json> accessed 8 April 2023.

Link10: ‘Azure OpenAI Service pricing’, *Azure* <https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/> accessed 8 April 2023.

Link11: ‘OpenAI Usage Policies’, *Open AI* <https://openai.com/policies/usage-policies> accessed 8 April 2023.