Ethical Considerations of the Use of LLMs in Law

Ty Beasley
beasl110@morris.umn.edu
Division of Science and Mathematics
University of Minnesota, Morris
Morris, Minnesota, USA

Abstract

Large Language Models (LLMs) are an integral component of modern society, touching numerous professional sectors, including the legal profession. As technology has evolved, these LLMs improve efficiency in legal analysis, research, and document summaries. Models such as CoCouncil, CaseText, and Westlaw Edge have changed how modern legal work is done. This paper analyzes concerns about LLM abuse in legal practice, pointing to the possibility of a negative impact on the performance of lawyers in court cases. Through a review of particular threats and problems related to the use of LLMs in law, this research hopes to contribute to reducing these threats and promoting safe use of LLMs in legal practice.

Keywords: Large Language Models, Neural Net Process, Training

1 Introduction

The global spread of Large Language Models (LLMs) into different industries has brought about notable developments, with the legal profession not being exempt. Researching the implications of LLMs in law is essential to understand their impact on legal accuracy and ethical practice. As these models become more integrated into legal systems, attorneys must ensure they are used responsibly and in alignment with the principles of the rule of law. This paper will begin with an introduction into LLMs in Section 2.3 and their impact in the legal field Section 2.1 as well as going into the American Bar Association Code of Conduct put in place to give those who work in this field specific rules to follow in Section 2.2.

This paper will go over the concerns that arise when working with LLMs in the law field. This is a concern because the law is precise and works with the lives of others. If problems arise with the usage of LLMs in this field, then they need to be addressed. This paper will also give an overview of the training process of LLMs in Section 3. This will lead to the concerns that arise when this technology is used within the legal field. Issues of Hallucination, Summarization, and Privacy will be addressed in Sections 4.1, 4.2, and 4.3, respectively. These concerns will be discussed to bring to light the potential negative impacts on the legal profession. The ethical concerns of LLM use in law are discussed in Section 5, and the feasibility of reconciling technological advancement with legal professional standards is examined.

2 Background

Large Language Models (LLMs) are a type of machine learning model that can perform a variety of tasks, including translating, classifying, generating text, and answering questions conversationally. From virtual helpers to deeply complex automation, LLMs have been transforming professional spaces. The profession we will focus on in this paper is the legal field, which is one of the oldest and most formalized of professions. The legal field is based on extensive research, negotiation, and procedural adherence. This field has been changed through the processing and analysis capabilities of LLM in legal work. But the nature of LLMs, a collection of algorithms trained using past data, requires scrutiny for issues surrounding accuracy and ethical considerations [13].

2.1 LLM Use in the Legal Profession

The introduction of LLMs in legal processes has made document analysis, case law research, and legal drafting more effective [14]. While these technologies help lawyers save time on labor-intensive activities, they also create accuracy, interpretability, and ethical responsibility issues. LLM summaries and analyses may contain made-up or inaccurate information that might mislead legal professionals who don't check the details produced. Relying on inaccurate LLM-based suggestions in high-stakes legal cases risks misrepresentation and ethical violations. Counselors have the duty of competence, as stated in Rule 1.1 of the Model Rules of Professional Conduct (discussed in Section 2.2), thus being answerable for ensuring their work is accurate [5]. When LLM systems produce false or deceptive legal information, the fault is with the lawyer employing the system. As much as there is more use of LLM in the law profession, it is critical to address accountability issues, authenticity, and ethics compliance to prevent negative effects on the judicial system.

We can already see the impact that LLMs have had on the legal field. AI-based systems, such as CoCouncil, Casetext, and Westlaw Edge, have already been integrated. These AI-based systems act as legal assistants designed to help legal professionals with various tasks. They can research, review, summarize, compare, and draft legal documents. CoCouncil, being the most used of the three, is powered by OpenAI. CoCouncil is trained specifically on knowledge from the field of law and has been tested by attorneys and specialists alike. This model is used mainly for research, but can also help draft outline documents [12].

2.2 ABA Code of Conduct

The American Bar Association Code of Conduct is a book of rules and regulations that all lawyers have to abide by to give their clients the best representation. This code was introduced in 1969 and has been the backbone of the legal field since then. It is upheld by the Behavior Analyst Certification Board, which ensures that the code is upheld in every courtroom in the country. There are 8 sections in the code with categories such as Client-Lawyer Relationship and Maintaining Integrity.

Along with the American Bar Association, each state has an association of their own. One in particular, California, has recently updated its code of conduct to include guidelines for using LLMs in the legal field [10]. In this paper, we will focus on the Duty of Confidentiality and the Duty of Competence. These rules talk about the possibilities of LLMs utilizing confidential information and how LLMs can sometimes produce inaccurate information.

The most relevant section of the ABA code is Section 1. This section is concerned with Client-Lawyer relationships. The sections are further broken down into rules, for example, Rule 1.1 states, "A lawyer shall provide competent representation to a client. Competent representation requires the legal knowledge, skill, thoroughness, and preparation reasonably necessary for the representation [3]." The two most relevant rules for this paper's ethical analysis are Rule 1.1 and Rule 1.6, which states "A lawyer shall not reveal information relating to the representation of a client unless the client gives informed consent, the disclosure is impliedly authorized to carry out the representation or the disclosure is permitted by paragraph." This rule will be important during Section 4.3 when we cover some concerns with the privacy handling of LLMs.

2.3 Technical Details

Before we can get into the main points of this paper, we first must discuss some foundationally important things. We will be discussing technical terms such as Large Language Models (LLMs), training, and Neural Networks. Large Language Models or LLMs are a type of artificial intelligence that uses machine learning to understand and generate human language. Neural Networks are essentially computer systems that are modeled after the human brain. Neural networks 'learn' from data and make predictions by mimicking statistical patterns that have been deduced from the data they process [4]. The use of Neural Networks will be explained more in Section 3. Neural networks can intake and process the information given to them in the training process (I provide more details on training in Section 3). Training is the process of teaching an artificial intelligence model to perform a specific task by using statistical patterns from the data it was trained on. This training process involves large data sets to align the model with human preferences.

3 LLM Training

Inside ChatGPT, which is one of many LLMs, is ultimately a very large neural network [16]. A neural network is a model based on how individual neurons interact. A neural network is a machine learning model made of layers of interconnected neurons, where each connection has a weight. You have a sequence of layers, each layer is a collection of neurons. The layers are connected by numeric parameters or a weight. There is a mathematical formula that takes the results in one layer and turns them into the results of the next layer, until the desired output is reached. This process applies weighted sums and activation functions to transform the information. The network learns by adjusting its weights through training, gradually improving its predictions based on the errors it makes [4].

In ChatGPT's case, the neural network is set up to deal with languages [16]. The transformer is a piece of this network; the transformer transforms an input sequence into an output sequence, often by focusing on relationships within the input. Essentially, the transformer takes input text provided to an LLM and makes the model turn it into an output by helping the model "understand" the words inputted [16]. The text is broken into sequences of tokens, then the transformer uses "attention" to focus on some of the most important parts of the sequence. The model can pay more attention to some parts of the sequence.

LLMs work in three stages. First, it takes the sequence of tokens that corresponds to the text it was given. It then finds an embedding produced by a different neural network and slots in existing word embeddings. Embeddings turn words or sentences into a vector of numbers. The closer the numbers are to each other, the more similar the words are. For example, if you input *king*, *queen*, and *apple*, you would see that *king* and *apple* would be far from each other in the vector space because of different meanings; however, *king* and *queen* can be connected in this space. Vectors are lists of real numbers. These vectors can produce a pattern within an existing embedding (see [16] for more details) that can capture mathematical structure in embedding space. For example, *king* minus *man* plus *woman* would produce a vector that would be very similar to *queen*. Once the tokens are originally embedded, these values will ripple through the layers in this neural network, where the original token embedding captures grammatical concepts, as well as the new embeddings. These embeddings are produced using the transformer and help from the attention, which is a mechanism that allows models to dynamically focus on the most relevant parts of the input data, and can capture contextual details.

3.1 General Machine Learning

In this paper, we will go over the process by which LLMs intake and process information, as well as the training process

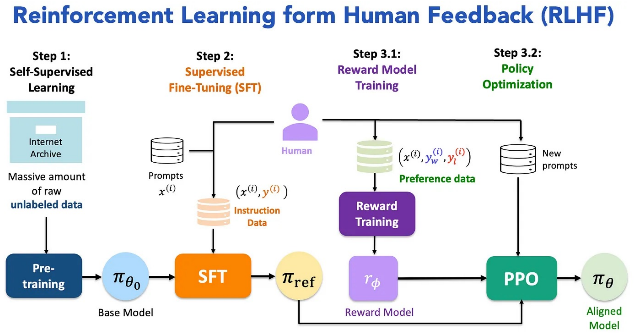

Figure 1. Training Process of ChatGPT (I am using this image from an internet source, the spelling should be "from" not "form"), via Medium [11].

that LLMs go through to produce texts that answer users' questions. Some of these steps include data collection and processing, as well as training. In general machine learning, the process typically begins with data collection and preparation, which involves gathering raw data and turning it into suitable training data for the model. This is followed by fine-tuning, where the model is trained on a more specific dataset tailored to a particular task or domain, allowing it to adapt its general knowledge to specialized applications. Both stages involve supervised learning, meaning the data used includes labeled examples with known outputs. Throughout these stages, learning is automated—the model adjusts its parameters based on the data without direct human feedback or intervention during the training process. The scoring metric for training LLMs typically involves minimizing a loss function that quantifies the difference between the model's predicted token probabilities and the actual target tokens during supervised learning. The scoring metric or loss function is a mathematical tool that measures how far the model's predicted output is from the correct or expected output. The supervised training in this paper refers to the model learning from the labeled dataset. This step involves trying to replicate the output based on this labeled data. This differs from reinforcement because supervised training is the only phase that uses direct human input. This will be explained more in Section 3.3.

3.2 Pre-Training

Pre-training is the beginning stage of training for an LLM. This stage is where a lot of the information is gained for the rest of the training process. This is depicted in Figure 1 by the area in blue. The model is given a large amount of raw text during this stage. It is not told what is accurate or inaccurate, it is just given this information to learn. The text LLMs are given is broken up into tokens, tokens are pieces of text. They can be words or pieces of words, but for this paper, we will think of them as entire words. This is to help learn what words appear together, and this can help the model predict future words. This is important because this is how the model learns what words complete certain sentences. It's for the LLM to form complete sentences of its own; it can predict what the input was supposed to be by pairing specific words together that often go together in the raw text it was given previously. For example, a phrase such as “The capital of France is“ could finish this phrase with “Paris” because the model found that these words often went together in the training text [16].

This is where the attention mechanism comes into play, this helps the model “pay attention” to words that seem more important than others. This stage is more to set groundwork for the LLMs memory and knowledge consumption so that it has a base to build upon. It is not in fact a model that answers questions yet, it is just a general purpose model that can only continue text.

3.3 Supervised Fine Tuning

Supervised fine tuning is the second step in the training process. This is depicted in Figure 1 by the area in orange. With all of the knowledge and capabilities gained from pre-training, supervised fine tuning involves further training LLMs, resulting in enhanced capabilities and improved controllability. The model in this step is trained to follow instructions and respond with useful information. The supervision comes into play when the model is trained on labeled datasets; there is no interaction from humans until the reinforcement phase (More about this in Section 3.4) [16].

Answers that should be replicated are then given to the model as a fine tuning step to show what a good answer is supposed to look like. It is given data that contains pairs of questions and correct answers. It is essentially structured lessons and feedback, the loss function is given to train the model to give the desired output. The loss function measures the difference between the model's predicted output and the actual, expected output. This step is a way of continuing the model towards being more aligned with human goals. Supervision is especially important for the growth of the LLM so that it can achieve exceptional performance [16].

3.4 Reinforcement Learning

Reinforcement is the final stage of training. This is depicted in Figure 1 by the area in purple and green. This is where the LLM is given human feedback on the response it provides to increase the level of satisfaction it can provide for each prompt. This stage helps align the LLM to human preferences. During this phase, the LLM is used to produce *several* different responses to the same prompt. These responses are ranked based on the best and closest to a human-like response that is both knowledgeable and helpful. These rankings reward the model for each response, so that the model can know if the response was what it was intended to be or if the output needs to be adjusted. This reward model is trained to predict which outputs are most desired by humans based on human training. This “reward model” is used to reinforce ChatGPT to push towards outputs that humans would want.

This process is looped until the training is complete, the result being a model that not only gives helpful information but also a model that can helpfully show that information. Adding reinforcement provides LLMs with valuable experience gained through human supervision. It is being taught how to act more as a helpful assistant than a general purpose language model. This effectively guides the LLM to more human-like solutions to each prompt. If the model wasn't guided towards a more human-like approach, then inappropriate responses could be generated. This became a problem when ChatGPT was outsourced to be trained by workers in Kenya, the workers discovered that some traumatic material could be produced by the model. Some workers even call it "torture" to work with it [1]. This showed that a model without human reinforcement can become unethical or harmful to those it is trying to help.

4 Concerns

Issues of concern to those in the legal profession include hallucinations (Section 4.1), problems specific to summarization (Section 4.2), and data privacy (Section 4.3).

4.1 Hallucinations

Hallucination is the act of a model generating information that is false, nonsensical, or not even real. However, the issue is that this information is presented as accurate [15]. Hallucination within Large Language Models comes from data, training, or inference [7]. Hallucinations, when they appear in legal work, can result in an attorney not representing their client honestly. This violates Rule 1.1 of the ABA Code of Conduct (Section 2.2).

One example of hallucinations being harmful occurred in May of 2023. A lawyer named Steven Schwartz in New York used ChatGPT to write a legal brief to be filed in federal court. Schwartz says that as he was drafting the affidavit, he asked ChatGPT to give some legal documents to back his claims. Months later, a New York judge wrote that six cases that Schwartz submitted appeared to be fictional. Schwartz responded by assuring the judge that the cases were real according to the LLM. When this claim was checked, ChatGPT responded by saying that the cases were indeed real even though they were fabricated [9].

4.1.1 Data. Pre-training is what the LLM is based on for original use, basically grasping the factual knowledge. Alignment is following user instructions and aligning outputs with human satisfaction in mind. Hallucination from data can be broadly categorized based upon whether the problem arose from misinformation in the pre-training data, or from attempts to cross the *knowledge boundary* (discussed below). Hallucinations arising from data can be broken up into 5 specific types introduced by Huang, Lei et al [7]. The types are imitative falsehood, societal bias, long-tail knowledge, up-to-date knowledge, and copyright-sensitive knowledge. Two of the types fall in the misinformation sector, while the other three are related to violations of a knowledge boundary.

Misinformation within the pre-training portion of LLMs is primarily associated with the memorization capability. The LLM cannot recognize whether its training data is factual or not. When the training data contains misinformation, then the LLM will 'learn' that misinformation. The pre-training data comes from the internet. This presents the issue of how much of the internet is fact-checked or truthful in nature, the internet is vast, and the amount of satire or fictional information within cannot be measured. Systematic misrepresentations, formally known as bias, can result in LLM hallucinations as well. Bias in the training data will result in an LLM that produces the same bias in the responses that it generates. Systematic misrepresentation refers to the persistent distortion or misstatement of information. Huang, Lei, et al discuss two ways misinformation in the training data results in hallucinations: imitative falsehood and societal biases. *Imitative falsehood* arises when a frequently referenced, incorrect fact frequently appears in the training data. An example of this would be *Thomas Edison created the light bulb*. This is a fact that everybody accepts but is incorrect. Multiple people who aided in the creation of the light bulb.

Societal biases arise from the tendency of LLMs to reflect or amplify existing societal prejudices and stereotypes that exist within the training data. An example of societal bias would be if Dr. Kim is referenced and ChatGPT adds that Dr. Kim is from South Korea. This is a societal bias because, often within the training data, the last name Kim is associated with South Korea.

The *knowledge boundary* is when an LLM, specifically ChatGPT, doesn't have all the information regarding very specific fields. When training an LLM, the goal is to produce a model that is generally knowledgeable in any subject likely to arise in a user's prompt. But the training set, in this case, the internet has limitations. With all that information, there is going to be a boundary that, if the model pushes past, could result in hallucinations. A good way to think about the knowledge boundary is to think about ChatGPT as a box. Inside the box is all the information it has, while outside the box is information it does not possess. ChatGPT cannot reach outside the box because it would hit its knowledge boundary. The three specific types of knowledge boundary issues Huang, Lei et al focus on are long-tail, up-to-date, and copyright-sensitive knowledge.

Long-tail knowledge is when the model doesn't have specific information in the field. For example, if you ask ChatGPT about what causes a specific type of ailment, it will not know what causes this specific ailment because it was not trained on data that had this information in it.

Up-to-date knowledge is exactly as it sounds. ChatGPT will not have information regarding recent events because it is not updated in real time. An example of this would be who is running for the Democratic Party in 2024. If you asked this question to ChatGPT in this period, it wouldn't have the answer because this information has not been added yet.

Copyright-sensitive knowledge refers to information regarding copyrighted material. ChatGPT would not have this information because it would not be in the training data. This is because models cannot be trained on copyrighted data unless the copyright owner permits it to do so.

4.1.2 Training and Hallucinations. Training a Large Language Model (LLM) is essential for enabling it to acquire general knowledge and respond effectively to prompts. However, this process is not without its flaws, as hallucinations can emerge at various stages of training. During the pretraining stage, the model learns to predict language patterns, but due to its limited grasp of the deeper, intrinsic details of specific subjects, hallucinations may occur when it lacks proper contextual understanding.

In the fine-tuning stage, the model is exposed to instruction-specific data designed to produce certain outputs. Challenges arise when the complexity or demands of this data exceed the model's capabilities. In such cases, the model may overfit or stretch beyond its knowledge boundaries to generate responses, prioritizing completion of the prompt over factual accuracy. This drive to respond at all costs contributes to hallucinations, as the model produces outputs without truly understanding their correctness. Another issue in this stage is the inability to say "I don't know". This programming essentially demands that the LLM give a response, no matter what.

In the Reinforcement stage, hallucinations can arise from misalignment within the model. The model encapsulates an internal belief of the truthfulness of its statements [7]. Even when these models are introduced to human feedback, they can produce outputs that go against the trained "knowledge" of the model. This differentiation of outputs can often rank favorably with evaluators because this is what these experts wanted the output to be, but it is at the cost of truthfulness. The model can go against what it knows to be the right answer to get grace from the evaluator or the human preference.

4.2 Concerns with Summarization

Using LLMs to summarize is another concern when looking into issues that can arise from using ChatGPT. Researchers George and Stuhlmuller [6] identified 4 problems that arose when LLMs were asked to produce summaries for academic papers; these concerns are valid for legal documents as well. This could be *over-simplification*, *paraphrasing*, *key themes*, and *misrepresentation* [6]. In a study done in 2023, the researchers looked into hallucinations within summaries of academic papers. The author introduced "factored verification", which is a method to detect these hallucinations by breaking down the summaries into claims that can be researched [6]. When tested, this new check achieved 76 percent accuracy in finding hallucinations within these summaries. This is particularly important for something that demands accuracy, like an academic paper. The findings in this paper revolve around the inability to provide an accurate summary because of fabricated claims and the inability to interpret the results correctly. ChatGPT-4 produced 0.84 hallucinations per summary according to the factored verification. The study further showed that ChatGPT sometimes invents conclusions not drawn in the original paper, making a fabricated conclusion based on what the user might want it to say. It was also found to overgeneralize important aspects of the document being summarized. The researchers also asserted that ChatGPT confused related terms, or words that looked similar but meant different things.

Given the findings within this study, we can conclude that ChatGPT does make factual errors when summarizing specific documents, such as an academic paper. The author emphasized that hallucinations are subtle but very common when prioritizing factual accuracy. An example of this error occurring during a legal case is when, in 2023, an Australian lawyer used ChatGPT on an immigration case. This lawyer used an LLM to summarize evidence that was supposed to strengthen the case. However, when the court went to check these cases, that the summary included fabricated cases [2].

4.3 Privacy

Privacy concerns arise when dealing with personal information. There have been many instances of LLMs leaking private information or data [17]. Legal firms are aware of the problem. The law firm Hayes Connor highlighted the risk of disclosing confidential data by using ChatGPT for specific tasks regarding legal documents. They say that sharing sensitive information with ChatGPT could result in unauthorized use and exposure of the data, which would compromise client confidentiality. As of the writing of this paper (May 2025), there have been no reported instances of this having occurred within the legal field. Broadly speaking, concerns arise from *Passive Privacy Leakage*. This is when sensitive or privileged information makes its way into responses produced by an LLM for a user who should not have access to that information. It is a passive privacy leak if that user was not actively trying to subvert the system to gain access to that knowledge.

4.3.1 Passive Privacy Leaks. Sensitive Query is one of the ways sensitive information is leaked. The contents of user prompts to ChatGPT are available to OpenAI and can be used to further train the model [17]. If sensitive information is included in a prompt, it can be inadvertently shared with others. In an example given in [17], Samsung Electronics gave ChatGPT sensitive corporate information when interacting with ChatGPT, causing employees to look into the security of the model.

Contextual Leakage is another way that leaks happen passively. Some queries could indirectly gather sensitive information about the user. Asking about nearby landmarks or local events could be giving away users' location or activities [17]. If this happens often enough, it can be possible to determine where the user is at a given time. There was a study done where LLMs could infer personal information from texts and chatbots [8]. This study evaluated multiple LLMs regarding personal information from the PersonalReddit dataset. ChatGPT ranked first in its ability to gather personal information that was shared on this site. This poses legal risks in contexts like client communication or legal aid, where unauthorized inference could breach confidentiality and violate the code of conduct.

Personal Preference Leakage is the other way that ChatGPT gathers personal information. This is when ChatGPT infers preferences, interests, and characteristics [17]. One of the selling points of LLMs like ChatGPT is the ability to give personalized recommendations; however, giving these models this information could be detrimental. These models have the potential to improve, but this refinement could inadvertently expose sensitive data, such as personal preferences.

5 Other Ethical Considerations

5.1 Bias

Utilizing these biased products can violate the ABA code of conduct (Rule 1.3, for example). It may also violate individual state Bar codes of conduct, for example, the California Bar Association states "A lawyer must ensure competent use of the technology, including the associated benefits and risks, and apply diligence and prudence with respect to facts and law [10]." This rule also states that AI outputs could include information that is false or biased.

LLMs trained with data from the legal field can inadvertently inherit biases from past cases. This can involve certain biases, such as stereotypes and misrepresentations of certain individuals that are no longer acceptable. This misrepresentation of minorities and certain gender stereotypes can influence responses generated by ChatGPT in response to certain cases. It is only known to reference certain cases that were in the training data. This can affect the behavior of the model when asked specific questions about cases or issues regarding certain cases. It is important in the legal field to have citations of specific cases. Bias can arise in ChatGPT from many places. Especially when dealing with ever-expanding text on the internet during training, and a lack of transparency with answers. We must ask questions like "Why do LLMs answer questions in the way that they do?" With these issues, it could be difficult to use this tool when dealing with the legal field. This breaks the code of conduct that we referenced earlier in 2.2.

6 Conclusion

This paper has examined 4 types of concerns related to the usage of LLMs in the legal profession, which were hallucinations, concerns with summarization, privacy, and systematic misrepresentation. Issues surrounding hallucinations arise in many contexts, and failure to take steps to mitigate these problems violates the ABA code of conduct 1.1, among others. Private information is important in the legal profession, and the ABA Rule 1.6 further backs the importance of this information to be kept confidential concerning the client. The consensus is that sharing information with an LLM puts that information in jeopardy of being stolen or used in the wrong way. Several pieces of software using LLMs have been specifically trained to deal with these shortcomings, but time will tell how effective they are. When asking the question of whether LLMs are currently ready to handle complicated legal problems, we must take into consideration hallucinations, privacy concerns, and bias as issues we are still currently dealing with in these models.

Acknowledgments

A very special thanks goes out to my advisor Peter Dolan and my professor Elena Machkasova for their insight and feedback in the process.

References

[1] 2025. https://en.wikipedia.org/wiki/ChatGPT

[2] 2025. https://www.theguardian.com/australia-news/2025/feb/01/australian-lawyer-caught-using-chatgpt-filed-court-documents-referencing-non-existent-cases

[3] American Bar Association. 2023. Model Rules of Professional Conduct. Online. link_to_aba_website_version

[4] Runjin Chen, Hao Chen, Jie Ren, Ge Huang, and Quanshi Zhang. 2019. Explaining Neural Networks Semantically and Quantitatively. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[5] Anthony E Davis. 2020. The future of law firms (and lawyers) in the age of artificial intelligence. *Revista Direito GV* 16, 1 (2020), e1945.

[6] Charlie George and Andreas Stuhlmüller. 2023. Factored Verification: Detecting and Reducing Hallucination in Summaries of Academic Papers. *IJCNLP-AACL 2023* (2023), 107.

[7] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55.

[8] Tianshi Li, Elizabeth Louie, Laura Dabbish, and Jason I Hong. 2021. How developers talk about personal data and what it means for user privacy: A case study of a developer forum on reddit. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

[9] Lyle Moran. 2023. Lawyer cites fake cases generated by ChatGPT in legal brief. *Legal Dive* 30 (2023).

[10] State Bar of California. 2025. Rules of Professional Conduct. Online. https://www.calbar.ca.gov/Attorneys/Conduct-Discipline/Rules/Rules-of-Professional-Conduct Accessed May 6, 2025.

[11] LM Po. 2025. The power of RLHF: From GPT-3 to chatgpt. https://medium.com/@lmpo/from-gpt-3-to-chatgpt-the-power-of-rlhf-118146b631ec

[12] Daniel Schwarcz and Jonathan H Choi. 2023. Ai tools for lawyers: a practical guide. *Minn. L. Rev. Headnotes* 108 (2023), 1.

[13] Haroon Sheikh, Corien Prins, and Erik Schrijvers. 2023. Artificial intelligence: definition and background. In *Mission AI: The new system technology*. Springer, 15–41.

[14] Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches. *IEEE Access* (2025).

[15] Adrian Tam. 2023. A gentle introduction to hallucinations in large language models. *Machine Learning Mastery, July* 20 (2023).

[16] Stephen Wolfram. 2023. *What Is ChatGPT Doing:... and Why Does It Work?* Wolfram Media.

[17] Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2024. On protecting the data privacy of large language models (llms): A survey. *ui.adsabs.harvard.edu* (2024).