# Truth Roulette: The Case for Implementation of Duties to Disclose and Minimize Hallucinations in legal Large Language Models

Submitted in partial fulfilment of the requirements of the Bachelor of Laws Degree, Strathmore University Law School

By

Sorobit Lulu Chebichii

145250

Prepared under the supervision of

Cecil Abungu

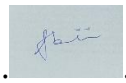
February 2025

Word count: 12144

(excluding footnotes and bibliography)

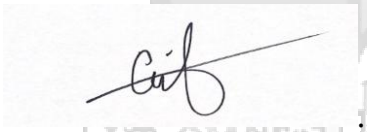
## Declaration

I, LULU SOROBIT, do hereby declare that this research is my original work and that to the best of my knowledge and belief, it has not been previously, in its entirety or in part, been submitted to any other university for a degree or diploma. Other works cited or referred to are accordingly acknowledged.

Signed:  .....

Date: .....March 21 2025.....

This dissertation has been submitted for examination with my approval as University Supervisor.

Signed:.....  .....

Cecil Abungu

Date: .....March 21 2025.....

## Dedication

This dissertation is dedicated to the late Commissioner Roseline Doreen Adhiambo Odede and to all those who came before me that are tireless in the pursuit of justice.

*"We have difficult days ahead in the struggle for justice and peace, but I will not yield to a politic of despair."*

- Martin Luther King Junior, 1968

## Acknowledgement

I would like to begin by thanking my parents, Marykaren Kigen-Sorobit and Samson Sorobit, for inspiring me to be relentless in my pursuit of justice and for teaching me the value of diligence and hard work. My parents consistently stayed up with me during all the long nights and woke up with me for all the 3 AM reading sessions. They offered me helpful advice and supported me both emotionally and materially. For that I am eternally grateful. Mom, Dad, I made it!

I would also like to thank my friends and family who held my hand through this long, arduous process. Especially my sisters, Lila and Lisa, my wakili cousins, Chumba and Sweetie who went before me, and Chiri, who very graciously hosted me. I would also like to thank Hildah, our house manager, for the care and support. I am deeply grateful for all my friends who have been there for me during this process, including Keni and Wanja, who have supported me from near and far.

I'm grateful for the aunties who have held my hand during this trying period of my life. They have offered advice, held space for my heavy emotions and made me feel safe in a very hostile environment. They did all this with patience, love and care. Girlies I wish you peace, clarity, contentment and healing. I'm truly grateful for everything and I'm always, *always* rooting for you!!

My profound gratitude goes to my supervisor, Cecil Abungu, for the insightful feedback and immense support. For challenging my ideas and being truly invested in my research and my growth. He has been instrumental in my process and for all that I am grateful. I'd also like to thank Dr. Charity Wayua for her timely and valuable feedback which made my dissertation what it currently is. There are many others who have helped me through this journey. I'm thankful for each and every one of them.

Finally, I would like to thank me. For doing all this hard work and not giving up even when things were difficult. For believing in me. For being diligent and consistent and making painful sacrifices to ensure that I have a good dissertation. For all the long nights, the even longer mornings and the pre-meeting rituals to calm my anxiety. I have held my hand when no one else has and for that I will always be grateful. We did it Lulu!

## List of legal instruments

European Union Artificial Intelligence Act

Chinese Interim Measures for the Governance of Generative Artificial Intelligence

## List of cases

Andersen v. Stability AI Ltd. Northern District of California (2023)

Babelegi Workwear and Industrial Supplies CC v Competition Commission of South Africa [2020], Court of Appeal of South Africa

Kadrey v. Meta Platforms, Inc. Northern District of California (2023)

Roberto Mata v Avianca Incorporated, Southern District Court of New York, United States (2023)

The New York Times Company v Microsoft Corporation, OpenAI, Inc., OpenAI LP, OpenAI GP, LLC, OpenAI LLC, OpenAI OpCo LLC, OpenAI Global LLC, OAI Corporation, LLC, And OpenAI Holdings, LLC, Southern District Court of New York (2024)

Tremblay v. OpenAI Southern District Court of New York (2024)

VT OMNES VNVM SINT

## List of abbreviations

AI – Artificial Intelligence

ANN – Artificial Neural Network

LLM – Large Language Model

LSTM – Long-Short Term Memory

ML – Machine Learning

NLP – Natural Language Processing

RAG - Retrieval Augmented Generation

RL – Reinforcement Learning

RLAIF – Reinforcement Learning with AI Feedback

RLHF – Reinforcement Learning with Human Feedback

RNN – Recurrent Neural Network

VT OMNES VNVM SINT

## Abstract

*The legal profession currently faces an access to justice problem. Both the language of the law as well as legal representation are inaccessible. The development of artificial intelligence (AI), particularly Large Language Models (LLMs) can address this problem because they democratise access to information, including access to legal information, and are cost friendly. However, at the current stage of LLM development, LLMs are bound to hallucinate which leads them to intentionally or unintentionally produce misinformation. Hallucinations are particularly harmful in the legal context due to an inherent information asymmetry between the user and the provider of legal services. Therefore, the user of the LLM may not know what is and is not accurate. Hallucinations therefore undermine the ability of LLMs to promote access to justice. There currently are no obligations on the part of developers to implement measures that minimise hallucinations or even inform their users about their existence. Technical measures to minimise hallucinations exist, meaning that developers have the ability to minimise hallucinations. This dissertation seeks to look into the kind of duties that should be imposed on the developers of LLMs that can give legal information. It seeks to do so by analysing the ability of LLMs to increase access to justice relative to the vulnerability of its users, as well as the availability of technical methods to minimise hallucinations. This study finds that countries should impose duties on developers to reduce the harmful effects of hallucinations. The duties include, at minimum, a duty to disclose the existence of hallucinations, and a duty to minimise hallucinations based on capacity.*

VT OMNES VNVM SINT

## Table of Contents

## 1.0. Introduction

### 1.1. Background

Artificial intelligence (AI) is a form of technology that simulates human learning to equip computers and machines with comprehension, problem solving, and decision-making skills.[1] Generative AI is a type of AI that can take data, learn from it and produce output that is similar, but not exactly like, the data it was trained on.[2] Large language models (LLMs) are a type of generative AI that mainly produce text but can increasingly produce other kinds of media such as images and sound.[3] Due to their human like conversational skills and ease of use, LLMs are increasingly being used to offer legal advice to people.[4]

LLMs are increasingly getting good at natural language processing (NLP) tasks. They can be used to summarise and extract information, write and refine articles, translate works from one language to another and even write code.[5] Due to their ability to summarise in an easy and understandable way, LLMs are increasingly being used as a source of information.[6] With the increased use of LLMs, especially with respect to knowledge intensive fields, including the law,[7] comes the increased likelihood of them being more accessible than consulting a professional, making it likely for people to turn to LLMs as an alternative.[8]

---

[1] Heaven W, 'What is AI?', MIT Technology Review, 10 July 2024. https://www.technologyreview.com/2024/07/10/1094475/what-is-artificial-intelligence-ai-definitive-guide/ on 20 August 2024. See also Stryker C and Kavlakoglu E, 'What is artificial intelligence (AI)?', IBM Think, 9 August 2024. https://www.ibm.com/think/topics/artificial-intelligence on 20 August 2024.

[2] Martineau K, 'What is generative AI?' IBM Think, 20 April 2023. https://research.ibm.com/blog/what-is-generative-AI on 19 August 2024.

[3] Talebi S, 'Multimodal models – LLMs that can see and hear', Towards Data Science, 19 November 2024. https://towardsdatascience.com/multimodal-models-llms-that-can-see-and-hear-5c6737c981d3/

[4] Goodson N and Lu R, 'Transforming Legal Aid with AI: Training LLMs to Ask Better Questions for Legal Intake', Stanford Law School Blogs, 15 March 2024. https://law.stanford.edu/2024/03/15/transforming-legal-aid-with-ai-training-llms-to-ask-better-questions-for-legal-intake/ on 20 August 2024.

[5] IBM Think, 'What are Large Language Models (LLMs)?' 2 November 2023. https://www.ibm.com/think/topics/large-language-models on 21 August 2024.

[6] Amin K, Doshi R and Forman H, 'Large language models as a source of health information: Are they patient-centred? A longitudinal analysis' 12 Science Direct 1, 2024. — <https://www.sciencedirect.com/science/article/pii/S2213076423000581> on 15 August 2024. See also Menon S, 'Why AI Is Popular Now—And Two Ways To Use It Better' Forbes, 1 December 2023 — <https://www.forbes.com/councils/forbestechcouncil/2023/12/01/why-ai-is-popular-now-and-two-ways-to-use-it-better/> on 15 August 2024.

[7] Amin K, Doshi R and Forman H, 'Large language models as a source of health information: Are they patient-centered? A longitudinal analysis' 12 Science Direct 1, 2024. — <https://www.sciencedirect.com/science/article/pii/S2213076423000581> on 15 August 2024.

[8] United States Supreme Court, *2023 Year-End Report on the Federal Judiciary*, 31 December 2023, 5 — <https://www.supremecourt.gov/publicinfo/year-end/2023year-endreport.pdf> on 20 August 2024.

Furthermore, access to justice is an important issue in law. It is estimated that 1.5 billion people globally (18.75% of the world's population) do not have proper access to legal representation or the language of the law at any given point in time.[9] The popularity of LLMs such as ChatGPT,[10] coupled with access to justice challenges,[11] increases the likelihood of individuals turning to LLMs rather than lawyers.[12] For the purposes of this dissertation, LLMs that can give legal information are known as legal LLMs.

LLMs can be described as a double-edged sword because even though they democratise access to information, they often spread misinformation through hallucinations.[13] Hallucinations occur when an LLM produces factually inaccurate or altogether made-up information such as false statistics or false sources.[14] Hallucinations are almost inevitable at this stage of AI development.[15]

Due to the sensitivity of legal information, as well as the vulnerability of the user of legal services, it is crucial that the legal information that comes from the LLMs is as accurate as possible.[16] However, hallucinations pose a significant challenge with respect to access to justice since they can cause the LLM to produce false information. Interestingly, lawyers

[9]Justice Fact Sheet 2023, Open Government Partnership, *Final draft*, 2. — <https://www.opengovpartnership.org/wp-content/uploads/2021/11/Justice_Fact-Sheet_Sept2023_EN.pdf> on 15 August 2024. See also Global Insights on Access to Justice: Findings from the World Justice Project General Population Poll in 101 Countries, World Justice Project, *Final report*, 2019, 120. — <https://worldjusticeproject.org/sites/default/files/documents/WJP-A2J-2019.pdf> on 15 August 2024.

[10] Chow A, 'How ChatGPT Managed to Grow Faster Than TikTok or Instagram' TIME 8 February 2023, — <https://time.com/6253615/chatgpt-fastest-growing/> on 15 August 2024. See also Milmo D, 'Chat GPT reaches 100 million users two months after launch' The Guardian, 2 February 2023, — <https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app> on 15 August 2024.

[11] Global Insights on Access to Justice: Findings from the World Justice Project General Population Poll in 101 Countries, World Justice Project, *Final report*, 2019, 120. — <https://worldjusticeproject.org/sites/default/files/documents/WJP-A2J-2019.pdf> on 15 August 2024.

[12] The LLMs that would act as alternatives to lawyers are referred to as Legal LLMs for the rest of the study. These include LLMs that can provide legal information either generally or with respect to a particular law/ set of laws. Examples of general legal in formation include information on legal rules and maxims such as the in duplum rule and the de minimis maxim or information on legal jargon itself such as 'jurisprudence'.

[13] Thorbecke C, 'AI tools make things up a lot, and that's a huge problem' CNN Business 29 August 2023, — <https://edition.cnn.com/2023/08/29/tech/ai-chatbot-hallucinations/index.html> on 15 August 2024.

[14] Curran S, Bethell O and Lansley S. 'Hallucination is the last thing you need, arXiv preprint, 2023, 2. — <https://arXiv.org/abs/2306.11520> on 18 August 2024.

[15] Rawte V, Chakraborty S, Pathak A, Sarkar A, Tonmoy A, Chadha A, Sheth A, Das A, 'The troubling emergence of hallucination in LLMs, an extensive definition, quantification and prescriptive remediations', in Bouamor H, Pino J and Bali K (eds) *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, 2023, 2543.

[16] Dahl M, Magesh V, Zuzgun M and Ho D, 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models', 16 *Journal of Legal Analysis* 1, 2024, 65. — < Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models | Journal of Legal Analysis | Oxford Academic (oup.com)> on 20 August 2024.

themselves are not exempted from the harms that stem from hallucinations in legal LLMs with an attorney citing hallucinated caselaw in the case of Roberto Mata v Avianca Incorporated.[17]

Additionally, some developers may fail to disclose the existence of hallucinations to end users. The non-disclosure may mislead users about the LLM's actual capabilities.. Taking Kenya-based Wakili AI as an example, the developers, M-WAKILI in the home page of their website term the LLM as 'honest, harmless, and helpful'. This conveniently omits the fact that the LLM is may hallucinate and therefore produce false information. Due to the inherent information asymmetry which makes users of legal LLMs vulnerable,[18] the non-disclosure of hallucinations in legal LLMs could pose a significant barrier to access to justice.

It is also worth considering that different developers have different resources. For example, fine tuners are usually small-scale developers who take existing, pretrained models such as GPT 3.5 and refine their capabilities regarding a particular task or domain, thereby turning them into specialised models.[19] Wakili AI is an example of a fine-tuned model created by a small start-up, M-zawadi.[20] This can be contrasted with ChatGPT which is made by Open AI, a multibillion-dollar company.[21]

Law makers could make this consideration based on revenue, market influence, compute among other factors. They could also consider the interest of the developers to avoid making the regulatory regime prohibitively costly as well as public interest and the cost of the proposed intervention to be able to make a balanced regulatory system.

## 1.2. Problem statement

This study will examine whether countries should impose a duty to disclose the existence of hallucinations as well as a duty to minimise their existence on developers of legal LLMs that

[17] Roberto Mata v Avianca Incorporated (2023), United States Southern District Court of New York.
[18] Dahl M, Magesh V, Zuzgun M and Ho D, 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models', 16 Journal of Legal Analysis 1, 2024, 65. — < Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models | Journal of Legal Analysis | Oxford Academic (oup.com)> on 20 August 2024.
[19] Bergmann D, 'What is fine tuning?' IBM. — <https://www.ibm.com/topics/fine-tuning#:~:text=Fine%2Dtuning%20in%20machine%20learning,models%20used%20for%20generative%20AI.> On 20 August 2024.
[20] — < https://mzawadi.com/what-we-do/wakili-ai/> on 15 August 2024.
[21] — <https://openai.com/chatgpt/> on 20 August 2024.

considers market influence, compute, revenue, as well as public interest and cost effectiveness.

### 1.3. Research questions

1. How exactly do LLMs generate information?
2. How are hallucinations generated and what harms might they occasion, especially with respect to legal LLMs?
3. Do developers have a responsibility to minimise the harm that stems from hallucinations in legal LLMs? If yes, how should this responsibility be codified in law?

### 1.4. Research objectives

1. To investigate exactly how LLMs generate information.
2. To examine how hallucinations are generated and the harms they may occasion especially with respect to legal LLMs.
3. To analyse the responsibility developers, have (if any) to minimise the harm that stems from hallucinations in legal LLMs and to investigate how this responsibility should be codified in law.

### 1.5. Hypothesis

Countries should impose a duty to disclose the existence of hallucinations as well as a duty to minimise their existence on developers of legal LLMs that considers market influence, compute, revenue, as well as public interest and cost effectiveness.

### 1.6. Justification of the study

This study will be useful to judges when they are deciding questions of liability with respect to hallucinations that stem from Large Language Models. It will also be useful to policy makers who are interested in the mitigation of misinformation and disinformation that stems from hallucinations. Additionally, researchers who are interested in addressing hallucinations through the imposition of a duty on developers will find this study resourceful.

### 1.7. Conceptual framework: The user of legal services as a vulnerable person

This study will be premised on the concept of the user of legal services as a vulnerable person and posits that the regulation of hallucinations in legal LLMs should be in accordance with this conceptualisation of the user. Access to justice is often seen as a prerequisite to justice through the legal system and goes hand in hand with the right to a fair trial.[22] However, justice in itself is often inaccessible in most parts of the world.[23] Legal scholars and practitioners widely acknowledge access to justice problems as deep and pervasive, since they expose fundamental failures within the legal system itself.[24] This difficulty in accessing justice makes users of legal services, such as litigants, vulnerable.[25]

While the right to access to justice can be interpreted as access to legal representation, it may also be interpreted as access to legal information.[26] The average lay person may find it difficult to read and correctly interpret the law themselves in most jurisdictions as the law is verbose and jargon heavy.[27] This means that access to justice efforts have concentrated on providing pro bono legal services.[28] Even though there are education campaigns on areas of law such as human rights, the vastness of the law itself makes it difficult for such education campaigns to be effective in all areas of the law.

[22] Sarat A, 'Access to Justice', 8 *Harvard Law Review* 94, 1981, 1912.
[23] Maldonado D, 'The Right to Access to Justice: Its Conceptual Architecture', 1 *Indiana Journal of Global Legal Studies* 27, 2020, 16. World over, access to justice is also a pertinent problem with an estimated 1.5 billion people globally lacking access to justice. See Justice Fact Sheet 2023, Open Government Partnership, *Final draft*, 2. — <https://www.opengovpartnership.org/wp-content/uploads/2021/11/Justice_Fact-Sheet_Sept2023_EN.pdf> on 15 August 2024. See also Global Insights on Access to Justice: Findings from the World Justice Project General Population Poll in 101 Countries, World Justice Project, *Final report*, 2019, 120. — <https://worldjusticeproject.org/sites/default/files/documents/WJP-A2J-2019.pdf> on 15 August 2024. See also Legal Services Corporation, The Justice Gap: The Unmet Civil Legal Needs of Low-Income Americans (2022) — <The Report | The Justice Gap Report (lsc.gov)> on 19 August 2024.
[24] World Justice Project, *Global Insights on Access to Justice: Findings from the World Justice Project General Population Poll in 101 Countries*, 2019, 120. — <https://worldjusticeproject.org/sites/default/files/documents/WJP-A2J-2019.pdf> on 15 August 2024.
[25] Sarat A, 'Access to Justice', 8 *Harvard Law Review* 94, 1981, 1912.
[26] Conklin W, 'Access to Justice as Access to a Lawyer's Language.'10 Windsor Yearbook of Access to Justice 1, 1990, 457. .
[27] Conklin W, 'Access to Justice as Access to a Lawyer's Language.'10 *Windsor Yearbook of Access to Justice* 1, 1990, 457.
[28] See Yallow S, 'Paths to justice: What people do and think about going to law', 4 Legal Ethics 149, 2001, 150. See also Goriely T, 'Law for the poor: the relationship between advice agencies and solicitors in the development of poverty law', 1 International Journal of the Legal Profession 2, 1996, 215-222.

Users of legal services themselves are mostly unaware of the intricacies of the law, making them vulnerable due to information asymmetry.[29] However, the law implicitly assumes that it is understood by the wider public. This assumption is evident in adversarial systems of law where litigants usually have to argue their case in court with or without representation, with the exception of a certain class of cases in some countries.[30] Additionally, ignorance of the law is usually not regarded as an acceptable defence, further reinforcing this assumption.[31] The practice of law as a profession exists to fill this gap, with lawyers being the providers of access to justice by virtue of the legal knowledge they wield.[32]

This concept of the user as a vulnerable person will be used to propose a framework that can be used to minimise the harm caused by hallucinations in legal LLMs. First, it will be used to analyse the role of legal LLMs in increasing access to justice. Next it will assist in evaluating the harmful effects of hallucinations in legal LLMs. Additionally, the concept will aid in determining whether developers actually have a responsibility to minimise the harmful effects of hallucinations. This will be useful in determining the kind of legal framework that should be used to minimise the effects of hallucinations in legal LLMs.

## 1.8. Literature review

There seems to be little extant literature on the question of regulating hallucinations through duties that are imposed on the developers. Previous studies have focused on the occurrence of hallucinations as a form of misinformation by LLMs generally, as well as LLMs and access to justice.

### 1.8.1. On hallucinations in LLMs

There have been robust discussions on hallucinations as misinformation in LLMs by technical researchers who then propose technical rather than legal solutions to minimise the

[29] Conklin W, 'Access to Justice as Access to a Lawyer's Language.'10 *Windsor Yearbook of Access to Justice* 1, 1990, 457.
[30] In the United Kingdom, legal aid is provided pro bono, without the litigant having to show proof of need for criminal cases. See United Nations Office on Drugs and Crime, *Global Study on Legal Aid*, 2016, 500-506. <https://www.unodc.org/documents/justice-and-prison-reform/LegalAid/GSLA_-_Country_Profiles.pdf> on 5 August 2024.
[31] See Matthews P, 'Ignorance of the Law is No Excuse' 3 *Legal Studies* 2, 1982, 190-192. See also Narasimham R and Narasimhan R, 'Ignorantia juris non excusat: Ignorance of law is no excuse' 13 *Journal of the Indian Law Institute* 1, 1971, 73-78.
[32] Yallow S, 'Paths to justice: What people do and think about going to law', 4 *Legal Ethics* 149, 2001, 150.

harms caused by hallucinations.[33] Scholars seem to agree on the wrongfulness of misinformation that is produced by hallucinations, as well as the necessity of technical solutions.

For example, Rawte et al note the troubling emergence of misinformation in LLMs.[34] They do so by analysing the occurrence of hallucinations in 15 LLMs including GPT-4 and LLaMA. The authors then proceed to define and categorise the types of misinformation that arise from LLMs and create a scale for measuring the severity of hallucinations. Rawte et al propose the use of a Hallucination Vulnerability Index to measure the correctness of a response by an LLM to curb hallucinations as a proposed regulatory measure.

Similarly, Chen and Shu note the misinformation issue in LLMs.[35] They describe previous efforts taken to minimise LLM-generated misinformation such as incorporating external knowledge, and fusing multilingual and multimodal information, and note that hallucinations are still rampant. They emphasise the need to minimise hallucinations given their wrongness and propose a model that combines human-LLM interactions to minimise hallucinations by leveraging the augmented and intrinsic abilities of LLMs.

Additionally, there have been discussions on the dangers of misinformation in legal LLMs, with some scholars advocating for a limit to the kind of legal advice LLMs can offer due to the inherent problem of hallucinations.[36]

Curran et al expound on the legal issue posed by hallucinations by discussing the challenges that they may present in common law systems.[37] They argue that though the penetration of

[33] Rawte V, Chakraborty S, Pathak A, Sarkar A, Tonmoy A, Chadha A, Sheth A, Das A, 'The troubling emergence of hallucination in LLMs, an extensive definition, quantification and prescriptive remediations', in Bouamor H, Pino J and Bali K (eds) *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, 2023, 2541-2573. See also Chen C, Shu K, 'Combatting Misinformation in the Age of LLMs: Opportunities and Challenges' 45 *AI Magazine* 3, 2024, 362-364.

[34] Rawte V et al, 'The troubling emergence of hallucination in LLMs, an extensive definition, quantification and prescriptive remediations,' 2541-2573.

[35] Chen C, Shu K, 'Combatting Misinformation in the Age of LLMs: Opportunities and Challenges' 45 *AI Magazine* 3, 2024, 354-364.

[36] Curran S, Bethell O and Lansley S. 'Hallucination is the last thing you need, arXiv preprint, 2023, 2. See also; Cheong I, Xia K, Feng K, Chen Q, and Zhang A. '(A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice,' Association for Computing Machinery Conference on Fairness, Accountability, and Transparency (FAccT '24), Rio de Janeiro, 3 June 2024, 2462-2463; Kapoor S, Henderson P, Narayanan A, 'Promises and pitfalls of artificial intelligence for legal applications' 2 *Journal of cross disciplinary research in computational law* 2, 2024, 3-10.

[37] Curran S, Bethell O and Lansley S. 'Hallucination is the last thing you need, arXiv preprint, 2023, 2.

LLM's into the legal sphere is inevitable, it should happen in areas where the model's performance is optimal such as 'volume type' tasks which include contract review and verification. However, the authors caution against the use of LLMs in legal research as well as in getting information in case law heavy jurisdictions due to the danger that arises from hallucinating caselaw as has happened before in the case of Roberto Mata v Avianca Incorporated.[38]

In the same vein, Dahl et al argue that though LLMs present a compelling case for access to justice, they should not be integrated into legal tasks without supervision.[39] They base this argument on the fact that hallucinations of some kind are inevitable at this stage of AI development, and the hallucination rates are high according to their findings.[40] They also argue that beyond hallucinations, LLM's tendency to agree with the user's preferences and beliefs (model sycophancy) as well as their uncertainty poses a real threat to legal research. Therefore, researchers and litigants who do not have legal training should be wary of using LLMs to obtain legal information. Dahl et all conclude by saying, inter alia, that as long as models do not improve their legal knowledge alongside their legal reasoning, they will not be suitable sources of legal advice.

### 1.8.2. On LLMs and access to justice

There seems to be scholarly consensus that LLMs inevitably increase access to justice. Chien et al argue that LLMs can do so through the court system using the State of Arizona as a case study.[41] They argue that LLMs can help self-represented litigants in five distinct ways that boil down to increasing access to justice as access to legal information.[42] Chien et al call for collaboration between legal and technological professionals to bridge the justice gap by leveraging generative AI.

[38] Roberto Mata v Avianca Incorporated (2023), Southern District Court of New York, United States.
[39] Dahl M, Magesh V, Zuzgun M and Ho D, 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models' 16 *Journal of Legal Analysis* 1, 2024, 65.
[40] The study found that LLMs hallucinate between 58% (Chat GPT) and 88% (LLaMa 2) of the time though the hallucination rates vary by complexity of the tasks, judicial hierarchy, jurisdiction, case prominence, and case year.
[41] Chien C, Kim M, Raj A and Rathis R, 'How Generative AI Can Help Address the Access to Justice Gap through the Courts,' Loyola of Los Angeles Law Review, Forthcoming, 1-63. 2024. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4683309# on 17 February 2025.
[42] These five ways are translating legal information, assisting in finding pro bono representation, assisting in filing expungement petition as well as eviction actions, and assisting the court's internal process that helps litigants such as providing real time multilingual translation.

Additionally, Tan, Westermann and Benyekhlef investigate the extent in which LLMs can act as self-help tools for self-represented litigants.[43] They do so by conducting evaluations on the quality of legal information produced by Chat GPT and JusticeBot. The authors approach this evaluation from the understanding that LLMs increase access to justice. Westermann and Benyekhelef had previously developed JusticeBot as an augmented LLM to increase access to justice by acting as a self-help method for laypeople, recognising the potential for LLMs to increase access to justice.[44]

### 1.8.3. Contribution

This study will agree with the findings of Dahl et al, Chien et al, as well as Westermann and Benyekhelef that LLMs increase access to justice but the existence of hallucinations greatly undermines this goal. However, the findings of the study will be unique because they will show that countries ought to impose legal obligations on developers of LLMs to minimise the harms of misinformation and disinformation that stems from hallucinations. This approach differs from the ones visible in the extant literature such limiting the use of LLMs in the legal context,[45] or the utilisation of technical methods,[46] because it focuses on creating legal obligations on the developers specifically, to increase access to justice through LLMs.

Additionally, this study seeks to advocate for the promotion, rather than the limitation of LLMs in the legal sphere due to their ability to solve the access to justice problem. This position sets the study apart from the work of other scholars who call for hard limits to the use of LLMs in the legal field.[47]

[43] Tan J, Westermann H and Benyekhlef K, 'ChatGPT as an Artificial Lawyer?' International Conference on Artificial Intelligence and Law 2023 Workshop on Artificial Intelligence for Access to Justice (AI4AJ), Braga, 19 June 2023, 1-6.
[44] Westermann H and Benyekhlef K, 'JusticeBot: A methodology for building augmented intelligence tools for laypeople to increase access to justice' Nineteenth International Conference on Artificial Intelligence and Law, Braga, 19th June 2023, 3-9.
[45] Dahl M, Magesh V, Zuzgun M and Ho D, 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models' 16 *Journal of Legal Analysis* 1, 2024, 65 Curran S, Bethell O, and Lansley S, 'Hallucination is the last thing you need, arXiv preprint, 2023, 2.
[46] Rawte V et al, 'The troubling emergence of hallucination in LLMs, an extensive definition, quantification and prescriptive remediations,' 2541-2573. See also; Chen C, Shu K, 'Combatting Misinformation in the Age of LLMs: Opportunities and Challenges' 45 *AI Magazine* 3, 2024, 362-364.
[47] Dahl M, Magesh V, Zuzgun M and Ho D, 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models' 16 *Journal of Legal Analysis* 1, 2024, 65. Curran S, Bethell O, and Lansley S. 'Hallucination is the last thing you need, arXiv preprint, 2023, 2.

## 1.9. Methodology

The nature of research in this study will be qualitative. The study will utilise primary sources on the question, such as the Chinese Act that governs Generative AI,[48] the European Union AI Act,[49] as well as relevant case law,[50] to showcase the need for a regulatory mechanism for hallucinations in legal LLMs. The study will also examine literature on the question to determine whether hallucinations in LLMs should be minimised through adopting a tiered responsibility framework using a content analysis. In general, a deductive approach will be utilised in this study with the chapters forming premises from which the main claim will be derived.

Chapter two will provide an overview of how LLMs exactly LLMs generate information. This chapter will also outline the major benefits of LLMs. It seeks to do so specifically by analysing literature on the topic. It will also utilise deductive reasoning, using the various examples available in literature to arrive at the conclusion that LLMs are generally beneficial.

Following this, chapter three will outline how exactly hallucinations are generated and the harms they may occasion. It will then go on to examine whether in addition to an analysis of the extant literature on hallucination and its harms. The chapter will employ inductive reasoning to reach this conclusion.

Chapter four will then employ deductive reasoning using the findings of chapter two and three as the premises to examine whether developers in fact have a responsibility to minimise the harms that stem from hallucinations. It will do so using a combination of a philosophical and content analysis of the right to access justice, as well as the vulnerability of the users. It will also propose a framework for establishing a legal obligation that should be imposed on developers. It will also utilise a critical analysis of the characteristics of developers, as well as the harm in question to arrive at a tiered responsibility framework for governing hallucinations in LLMs.

[48] Article 4 (1), — <Interim Measures For The Management Of Generative Artificial Intelligence Services, <Https://Www.Cac.Gov.Cn/2023-07/13/C_1690898327029107.Htm> on 18 August 2024.
[49] European Union Artificial Intelligence Act — <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206> on 18 August 2024.
[50] Roberto Mata v Avianca Incorporated (2023), Southern District Court of New York, United States.

## 1.10. Chapter breakdown

In chapter one sets out, among others, the research questions, conceptual framework, and justification of the study. This lays down the foundation for the subsequent chapters.

Chapter two will investigate the mechanism by which LLMs generate information as well as their utility. It seeks to argue that LLMs are beneficial to the society at large while setting the scene for the next chapters by providing important context on how exactly an LLM works.

Chapter three will explain the phenomenon of hallucinations and the harms they pose with respect to access to justice as access to legal information. It seeks to argue that hallucinations are barriers to access to justice and will set a premise for the main argument in chapter four.

Chapter four will build on chapter three to investigate whether developers of legal LLMs actually have a duty disclose the existence of hallucination and to minimise their existence if they have the capacity. It seeks to argue that at the very least hallucinations can lead mislead the user if they are not informed of their existence, thereby undermining access to justice. It will also advance the argument that developers do in fact have a duty disclose the existence of hallucination and to minimise their existence if they have the capacity which should be codified in law. It proposes these duties while considering the findings of chapter two as well as the resources, compute, and market impact of various developers to create a scalable metric rather than a hard and fast rule. Chapter four intends to do so by arguing that the vulnerability of the users, as well as the ability of developers to minimise the harmful effects of hallucinations leads to an obligation on their part to do what they can to minimise the effects of hallucinations.

Chapter five will be the concluding chapter that will then offer recommendations as to how countries could implement these duties in their legal systems to minimise the harms that stems from hallucinations in legal LLMs.

# 2.0. LLMs and output generation

## 2.1. Introduction

While the previous chapter introduces the study, this chapter will explain exactly what LLMs are, how they generate output, and what their wider social utility is. It also provides a brief description of key technical terms. This chapter will begin by providing a brief description of what exactly LLMs are. It will then explain what machine learning is and how it works with an emphasis on LLMs. Finally, this chapter will conclude by highlighting the utility of LLMs.

## 2.2. How LLMs work

### 2.2.1. Understanding LLMs

A Large Language Model (LLM) is a type of artificial intelligence model that is able to generate and classify text, answer questions in a conversational manner and translating text from one language to another.[51] It is worth noting that some LLMs can understand and generate information in multiple formats such as sound, text and images.[52] Chat GPT, Claude and Gemini are popular examples of LLMs.

### 2.2.2. Machine learning

Machine learning (ML) can be defined as the field of artificial intelligence research that is aimed at coming up with algorithms that can make reliable predictions on new or similar data by finding patterns in already existing data without explicit programming.[53] Deep learning is a subset of machine learning that uses artificial neural networks to simulate the human brain's learning procedure.[54]

[51] Naveed H, Khan, Qiu S, Saqib M, Anwar S, Usman M, Akhtar N, Barnes N and Mian A, 'A Comprehensive Overview of Large Language Models', arXiv, 2024, 2.
[52] Yin S, Fu C, Zhao S, Li K, Sun X, Xu T, and Chen E, 'A survey on multimodal large language models', arXiv, 2023, 1.
[53] Kelleher J, *Deep Learning*, MIT Press, Cambridge, Massachusetts, 2019, 253.
[54] Goodfellow I, Bengio Y and Courville A, *Deep Learning*, Cambridge, Massachusetts, MIT Press, 2016, 5.

Deep learning is a departure from classical machine learning. The latter utilised linear regressions models which function like obedient robots.[55] Linear regression models use statistical methods to establish a linear relationship between variables.[56] Classical ML models produced output based on a very specific representation of the world that had to be mapped out manually for each task.[57] They proved to be inefficient in the real world since real world data sets often contain irrelevant information, significant outliers, missing values, and non-linear relationships that the models were unable to correctly process.[58] Additionally, linear regression models are often unable to give accurate predictions when they encounter new data which makes them inefficient.[59]

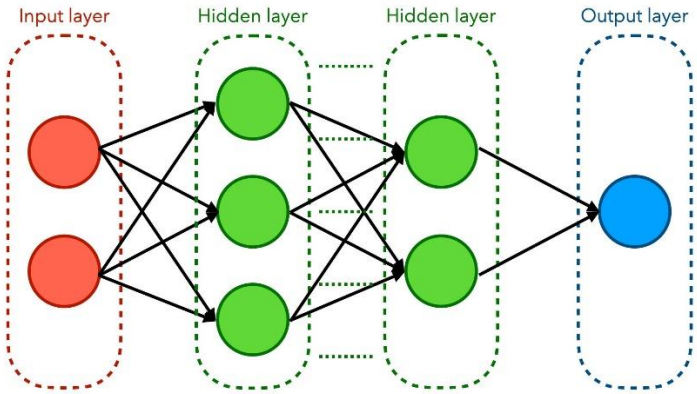
On the other hand, deep learning models represent the world as a hierarchy of concepts that are automatically detected by their architecture making them more efficient in the long run.[60] However, they require enormous quantities of data to perform reliably while classical machine learning models can run on smaller datasets.[61] LLMs are deep learning models, which are powered by neural networks.[62]

## A. Neural networks

Most machine learning algorithms are powered by artificial neural networks (ANNs).[63] ANNs consist of mathematical representations of connected processing units called artificial neurons.[64] These artificial neurons are arranged in layers which are interconnected.[65] Each neuron receives input signals, processes them and then transmits the output to the

[55] Taye M, 'Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions', 12 *Computers* 5, 2023, 6.
[56] Qu K, 'Research on Linear Regression Algorithms,' 2nd International Conference on Mathematical Physics and Computational Simulation, Glasgow, 9 August 2024, 2.
[57] Taye M, 'Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions', 12 *Computers* 5, April 2023, 6.
[58]See Loh P and Wainwright M, 'High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity', 40 *The Annals of Statistics* 3, 2012; Hazan E and Koren T, 'Linear Regression with Limited Observation' 29th International Conference on Machine learning, Edinburgh, 26 June 2012.
[59] Barlett P, Long P, Lugosi G and Tsigler A, 'Benign overfitting in linear regression', arXiv, 2020, 1-16.
[60] Schmidhuber, J, 'Deep Learning in Neural Networks: An Overview' 61 *Neural Networks* 1, 2015, 85–117.
[61] Yosinski J, Clune J, Bengio Y and Lipson H, 'How transferable are features in deep neural networks?' 4 *Advances in Neural Information Processing Systems* 1, *2014*, 3320–3328.
[62] Lin T, Wang Y, Liu X and Qiu X, 'A survey of transformers', 3 *AI Open* 1, 2022, 127-128.
[63] Janiesch C, Zschech P and Heinrich K, 'Machine Learning and deep learning fundamentals', 31 *Electronic Markets Journal* 1, 2021, 2-3.
[64] Janiesch C, Zschech P and Heinrich K, 'Machine Learning and deep learning fundamentals', 31 Electronic Markets Journal 1, 2021, 3.
[65] Islam M, Chen G and Jin S, 'An Overview of Neural Networks' 5 *American Journal of Neural Networks and Applications* 1, 2019, 7-11.

corresponding neuron in the next layer.[66] A neural network usually has three layers; the input layer, the hidden layer and the output layer.[67] Neurons in these layers are connected to each other as illustrated below.




A simple artificial neural network

The process of generating output in a neural network as explained by Kelleher, Goodfellow et al, and Islam et al can be summarised as follows.[68] The connections between a unit are represented by a number called a weight, which can either be positive or negative, that determines the output. This is because neurons receive the input signal which is allocated a numerical value that is then multiplied by the weight. The activation of the corresponding neuron in the next layer depends on the result of this multiplication.

For increased accuracy, a negative number is added to the weight known as a bias. For example, if a developer wants only inputs which have a value that is greater than 10 to activate the next neuron, a bias of -10 can be added so that only those inputs which meet the threshold can meaningfully activate the corresponding neuron. The sum is then passed through a sigmoid curve to compress it to a number between 0 and 1. The deep learning process is then about adjusting the weights and biases until the model is fairly accurate using learning functions such as back propagation and stochastic gradient descent.

The accuracy process starts with the model receiving feedback on the accuracy of its output.[69] This is done by a loss function which is the difference between the accurate output and the output given by the model. This can be corrected through gradient descent which

[66] Goodfellow *et al*, *Deep Learning*, 5.
[67] Islam M, Chen G and Jin S, 'An Overview of Neural Networks' 5 *American Journal of Neural Networks and Applications* 1, 2019, 7-11.
[68] Kelleher, *Deep Learning*, 65-101; Goodfellow *et al*, *Deep Learning*, 5-12; Islam M, Chen G, Jin S, 'An Overview of Neural Networks' 5 *American Journal of Neural Networks and Applications* 1, 2019 7-11
[69] Golilarz N, Hossain E, Addeh A and Rahimi K, 'Learning algorithms made simple,' arXiv, 2024, 9-10.

adjusts the weights and biases by finding the shortest way to bridge the error.[70] Adjustment can also be done through back propagation which works backwards from the outer layer to the inner layer to correct the error in question.[71] A model then utilises these learning functions iteratively until it is accurate.

**Transformers and LLMs**

LLMs run on a type of neural network known as a transformer.[72] Transformers were developed as a way to improve language translation by enabling the model to be more contextually accurate using a mechanism known as self-attention.[73] This mechanism works by weighing the significance of the different elements in the inputs to ensure that the LLM produces contextually relevant output by 'paying attention' to the important parts of the input. Further, transformers use positional encoding which means that they can look at data holistically rather than sequentially as is the case in RNNs.[74] Transformers are also parallelisable meaning that they can process datasets simultaneously which makes it possible for LLMs to exist since the other kinds of neural networks seen above are unable to process several datasets simultaneously.[75]

All these features make LLMs distinct from other kinds of deep learning models.[76] For example, Recurrent Neural Networks(RNNs) Long Short-Term Memory(LSTM) are the predecessors to LLMs. Though they were also used for natural language processing (NLP) tasks such as translation, they process data sequentially which made them inefficient and sometimes lack context.[77]

VT OMNES VNVM SINT

[70] Zhang J, 'Gradient Descent based Optimization Algorithms for Deep Learning Models Training', arXiv, 2019, 6-24.
[71] Islam M, Chen G and Jin S, 'An Overview of Neural Networks' 5 *American Journal of Neural Networks and Applications* 1, 2019, 7-11.
[72] Sarker I, 'Machine Learning: Algorithms, Real-World Applications and Research Directions', 2 *Springer Nature Computer Science* 1, 2021, 106..
[73]Golilarz N, Hossain E, Addeh A, Rahimi K, 'Learning algorithms made simple' arXiv, 2024, 9. See Viswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez N, Kaiser L, Polosukhin I, 'Attention is all you need', Thirtieth Annual Conference on Neural Information Processing Systems, Long Beach California, December 8-9, 2017, 10.
[74] Lin T, Wang Y, Liu X, Qiu X, 'A survey of transformers,' 113.
[75] Sanford C, Hsu D and Telgarsky M, 'Transformers, parallel computation, and logarithmic depth,' arXiv, 2024, 11.
[76] Lin T, Wang Y, Liu X, Qiu X, 'A survey of transformers', 127.
[77] Viswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez N, Kaiser L, Polosukhin I, 'Attention is all you need', Thirtieth Annual Conference on Neural Information Processing Systems, Long Beach California, December 8-9, 2017, 6-7, 9.

## B. Machine learning techniques

These are ways in which a model is trained to perform a particular task or function. There are four major machine learning techniques; supervised learning, unsupervised learning, reinforcement learning and semi supervised learning.[78]

### *Supervised learning*

This is an ML technique which utilises data that already has the correct result.[79] This kind of data is known as labelled data. The algorithm learns by iteratively refining its predictions on the output to narrow the gap between its predictions and the actual output to increase accuracy.[80]It is usually used to train models that classify data,[81] or find trends in the data.[82]

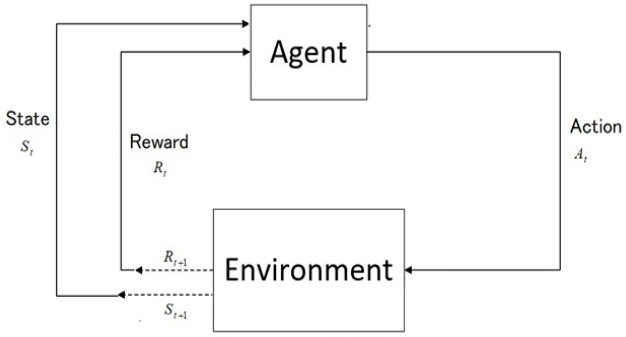
### *Unsupervised learning*

Unsupervised learning uses unlabelled data.[83] The algorithm then tries to gain a deeper understanding of data by making connections and grouping similar features together through a process known as clustering.[84] This is important since models are known to make connections that are not immediately obvious to humans.[85] Unsupervised learning is particularly useful for tasks that require grouping, and can be used to process data for supervised learning.[86]

[78] Goodfellow *et al*, *Deep Learning*, 102- 104, 240.
[79] Kelleher, *Deep Learning*, 26.
[80] Taye M, 'Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions', 12 *Computers* 5, 2023, 6.
[81] Osisanwo F, Akinsola J, Awodele O, Hinmikaiye J, Olakanmi O and Akinjobi J, 'Supervised Machine Learning Algorithms: Classification and Comparison', 48 *International Journal on Computing Trends and Technology* 1, 2017, 128–138; See also Nasteski V, 'An overview of the supervised machine learning methods', 4 *Horizons* 1, 2017, 51–62.
[82] Osisanwo F, Akinsola J, Awodele O, Hinmikaiye J, Olakanmi O and Akinjobi J, 'Supervised Machine Learning Algorithms: Classification and Comparison', 48 *International Journal on Computing Trends and Technology* 1, 2017, 128–138.
[83] Karhunen J, Raiko, T and Cho K, 'Unsupervised deep learning: A short review' in Bingham E, Kaski S, Laaksonen J and Lampinen J (eds), *Advances in Independent Component Analysis and Learning* Machines, Academic Press, Cambridge, Massachusetts, 2015, 125–142.
[84] Taye M, 'Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions', 12 *Computers* 5, 2023, 7.
[85] Zhang Z, Singh J, Gadiraju U and Anand A, 'Dissonance Between Human and Machine Understanding' arXiv, January 2021, 17-18.
[86] Sarker I, 'Machine Learning: Algorithms, Real-World Applications and Research Directions', 2 *Springer Nature Computer Science* 1, 2021, 4.

***Reinforcement learning***

Reinforcement Learning (RL) is a type of machine learning technique that focuses on training algorithms by autonomously interacting with the environment.[87] RL agents learn through trial and error using rewards.[88] In RL, a state represents a specific configuration of the environment at a given time, for example one specific layout of pieces on the board in chess.[89] Actions are then the set of possible moves an agent can make while interacting with the environment. A policy guides the behaviour of an RL agent by showing the probabilities of all potential actions in a given state to lead to achieving its goal. The closer an agent is to accurately achieving the intended goal, the higher the reward function. In RL, rewards are usually numerical and are dependent on the environment of the model.




Overview of reinforcement learning

An agent usually has clear cut goals. It also seeks to get as high a reward as possible. Since RL agents are not told how to learn, they must figure out how to do so in an efficient way so that it can maximise on rewards. The reward function usually incorporates human feedback.[90] This is known as Reinforcement Learning with Human Feedback (RLHF). RLHF makes models more inclined to producing output that is preferred by humans.[91] However, this does not always happen as models trained using RLHF are known to take unethical routes in order to get a higher reward since it is difficult for developers to properly

[87] Kelleher, *Deep Learning*, 28.
[88] Sutton R and Barto A, *Reinforcement learning: An introduction*, 2 ed, MIT Press, Cambridge, Massachusetts, 2018, 2.
[89] Ghasemi M, Moosavi A, Sorkhoh I, Agarwal A, Alzhouri F and Ebrahimi D, 'An Introduction to Reinforcement Learning: Fundamental Concepts and Practical Applications', arXiv, 2024, 2.
[90] Ziegler D, Stiennon N, Wu J, Brown T, Radford A, Amodei D, Christiano P and Irving G, 'Fine-tuning language models from human preferences,' arXiv, 2020, 2.
[91] Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, Drain D, Fort S, Ganguli D, Henighan T, Joseph N, Kadavath S, Kernion J, Cornely T, El-Showk S, Elhage N, Hatfield-Dodds Z, Hernandez D, Hume T, Johnston S, Kravec S, Lovitt L, Nnada N, Olsson C, Amodei D, Brown T, Clark J, McCandlish S, Olah C, Mann B and Kaplan J, 'Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback', Anthropic, arXiv, 2022, 89.

specify the reward in a way that is watertight.[92] With an increase in the size of AI models, RLHF may be inadequate. Therefore, larger models are increasingly being trained by AI feedback for better oversight.[93]

***Shortcomings of RL***

LLMs are typically trained using RL, specifically to align their output with human values.[94] The shortcomings of RL are worth highlighting because they can lead to hallucinations as discussed in sections 3.2.2 and 3.2.3.

*Difficulty in finding a suitable policy*

It is worth noting that some agents experience challenges in finding a suitable policy. RL agents are designed to learn policies that maximize their reward function. This assumes a stationary reward function. However, RLHF goes against this assumption by constantly updating the reward model based on human feedback therefore providing a non-stationary reward function.[95] Some agents may then find it difficult to find a policy due to this shifting goal post.[96] Further, agents face the extrapolation versus exploitation dilemma where they

[92] Ngo R, Chan L, Mindermann S, 'The Alignment Problem from a Deep Learning Perspective', International Conference on Learning Representations Conference Papers, Vienna, 7 May 2024, 3-5. See also Skalse J, Howe N, Krasheninnikov D, and Krueger D, 'Defining and characterizing reward gaming' 36th Conference on Neural Information Processing Systems, New Orleans, 28 November 2022, 1-12; Ji J, Qiu T, Chen B, Lou H, Wang K, Duan Y, He Z, Zhou J, Zhang Z, Zeng F, Dai J, Pan X, Ng KY, O'Gara A, Xu H, Fu B, McAleer S, Yang Y, Wang Y, Zhu S, Guo Y and Gao W, 'AI Alignment: A Comprehensive Survey', arXiv, 2024, 4-5.
[93] Casper S and Davies X, 'Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback', arXiv, 2023, 13; See also Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, Chen A, Goldie A , Mirhoseini A, McKinnon C, Chen C, Olsson C, Olah C, Hernandez D, Drain D, Ganguli D, Li D, Tran-Hohnson E, Perez E, Kerr J, Mueller J, Ladish J, Landau J, Ndousse K, Lukosuite K, Lovitt L, Sellitto M, Elhage N, Schiefer N, Mercado N, DasSarma N, Lasenby R, Larson R, Ringer S, Johnston S, Kravex S, El Showk S, Fort S, Lanham T, Telleen-Lawton T, Cornely T, Henighan T, Hume T, Bowman S, Hatfield-Dodds Z, Mann B, Amodei D, Joseph N, McCandlish S, Brown T and Kaplan J, ' Constitutional AI: Harmlessness from AI feedback,' arXiv, 2022, 1-16.
[94] Chen Z, Zhou K, Zhao W, Wun J, Zhang F, Zhang D and Wen J, 'Improving Large Language Models via Fine-grained Reinforcement Learning with Minimum Editing Constraint', arXiv, 2024, 1; Kumar A, Zhuang V, Agarwal R, Su Y, Co-Reyes J, Singh A, Baumli K, Iqbal S, Bishop C, Roelofs R, Zhang L, McKinney K, Shrivastava D, Paduraru C, Tucker G, Precup D, Behbahani F and Faust A, 'Training Language Models to Self-Correct via Reinforcement Learning', Google DeepMind, arXiv, 2024, 13-14; Wang S, Zhang S, Zhang J, Hu R, Li X, Zhang T, Li J, Wu F, Wang G and Hovy E, ' Reinforcement Learning Enhanced LLMs: A survey', arXiv, 2024, 1, 3-8.
[95] Kaufmann T, Weng P, Bengs V and Hüllermeier E, 'A survey of reinforcement learning from human feedback', arXiv, 2024, 44.
[96] Wang T, Herbert S and Gao S, 'Fractal Landscapes in Policy Optimization', arXiv, 2023, 1-10.

find it difficult to balance between scoping out new actions and maximizing rewards.[97] This dilemma makes it harder for the RL agent to find a reliable policy.

*Model sycophancy*

Models, especially LLMs, have been observed to exhibit sycophancy which manifests itself as choosing an incorrect answer despite being aware of its inaccuracy so that they can agree with the user.[98] Perez et al[99] and Sharma et al[100] argue that sycophancy stems from Reinforcement Learning through Human Feedback (RLHF) since RLHF favours outputs that appeal to human preferences which may make models produce output that appeals to humans but are flawed or incorrect.

*Expressing uncertainty*

Models are trained to complete each response without expressing uncertainty during training which encourages them to fabricate content when they have a knowledge gap.[101] This design choice actively engenders inaccuracies in models, especially LLMs.

***Semi-supervised learning***

[97] Agarwal A, Henaff M, Kakade S, and Sun W, 'PC-PG: Policy Cover Directed Exploration for Provable Policy Gradient Learning,' 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, 16 July 2020, 1-9.

[98] Wei J, Huang D, Lu Y, Zhou D, and Le Q, 'Simple synthetic data reduces sycophancy in large language models', arXiv, 2024, 3-4.

[99] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pet tit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Lan don Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Lar son, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timo thy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviours with model-written evaluations, 61st Annual Meeting of the Association for Computational Linguistics, ACL 2023, Toronto, Canada, 9 July 2023- 14 July 2023, 8-9.

[100] Sharma S, Tong M, Korbak T, Duvenaud D, Askell A, Bowman S, Cheng N, Burmus E, Hatfield-Dodds Z, Johnston S, Kravec S, Maxwell T, McCandlish S, Ndousse K, Rausch O, Scheifer N, Yan D, Zhang M and Perez E, 'Toward Understanding Sycophancy in Language Models' arXiv, 2023, 1-10.

[101] Zhang H, Diao S, Lin Y, Fung Y R, Lian Q, Wang X, Chen Y, Ji H, and Zhang T, 'R-Tuning: Instructing Large Language Models to Say 'I Don't Know', in Duh K , Gomez H and Bethard S (eds), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Association for Computational Linguistics, Mexico City, 2024, 7113. See also Yang Y, Chern E, Qiu X, Neubig G, and Liu P, 'Alignment for Honesty,' Thirty-Eight Annual Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, 12 December 2024, 1-10.

Semi-supervised learning (SSL) combines more than one form of learning.[102] Unsupervised and supervised learning can be combined when obtaining a sufficient amount of labelled data is prohibitively costly.[103] Unlabelled data is usually used to pre train the model so that it can learn meaningful representations of concepts, then labelled data is used to ensure accuracy.[104]

SSL can also utilise reinforcement learning and supervised learning. This is known as semi supervised reinforcement learning (SSRL).[105] Since agents learn the environment on their own in reinforcement learning, supervised learning can be used to assist the agents to learn a reliable policy.[106] SSL is used to train LLMs on both general and knowledge specific domains.[107]

## 2.3. The utility of LLMs

Due to their intelligent capabilities, accessibility and human-like conversational skills, LLMs have proven to be incredibly useful. LLMs are incredibly cost efficient since they can automate tasks across several sectors such as law,[108] finance,[109] computer science,[110] medicine,[111] among other fields. This goes hand in hand with increased information equity and accessibility since information that was previously out of reach due to professional fees,

[102] Golilarz N, Hossain E, Addeh A and Rahimi K, 'Learning algorithms made simple' arXiv, 2024, 4.
[103] Xu W, Sun H, Deng C and Tan Y, 'Variational Autoencoder for Semi-Supervised Text Classification,' The Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, 5 February 2017, 1.
[104] Taye M, 'Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions', 12 *Computers* 5, 2023, 7.
[105] Mohammadi M, Al-Fuqaha A, Guizani M and Oh J, 'Semi-supervised Deep Reinforcement Learning in Support of IoT and Smart City Services', 5 *IEEE Internet of Things Journal* 2, 2017, 2.
[106] Finn C, Yu T, Fu J, Abbeel P and Levine S, 'Generalizing skills with semi-supervised reinforcement learning', 5th International Conference on Learning Representations, Toulon, 24 April 2017, 2.
[107] Xi Y, Ding W, Yu K and Lai J, 'Semi-supervised learning for code-switching ASR with Large Language Model Filter', arXiv, 2024, 3-6.
[108] Martin L, Whitehouse N, Yiu S, Catterson L and Rivindu P, 'Better call GPT: Comparing large language models against lawyers', arXiv, 2024, 11-12.
[109] Zhao H, Liu Z, Wu Z, Li Y, Yang T, Shu P, Xu S, Dai H, Zhao L, Mai G, Liu N and Liu T, 'Revolutionizing Finance with LLMs: An Overview of Applications and Insights,' arXiv, 2024, 1-23.
[110] Chen M, Tworek J, Jun H, Yuan Q, Pinto H, Kaplan J, Edwards H, Burda Y, Joseph N, Brockman G, Puri R, Krueger G, Petrov M, Khlaaf H, Sastry G, Mishkin P, Chan B, Gray S, Ryder N, Pavlov M, Power A, Kaiser L, Bavarian M, Winter C, Tillet P, Such FP, Cummings D, Plappert M, Chantzis F, Barnes E, Herbert-Voss A, Guss WH, Nichol A, Paino A, Tezak N, Tang J, Babuschkin I, Balaji S, Jain S, Saunders W, Hesse C, Carr AN, Leike J, Achiam J, Misra V, Morikawa E, Radford A, Knight M, Brundage M, Murati M, Mayer K, Welinder P, McGrew B, Amodei D, McCandlish S, Sutskever I and Zaremba W, 'Evaluating large language models trained on code', arXiv, 2021, 1-10.
[111] Thirunavukarasu A, Ting D, Elangovan K, 'Large language models in medicine', 29 *Nature medicine* 1, 2023, 1931.

pay walls or even inaccessible language becomes cheaper and more understandable through LLMs.[112]

Additionally, LLMs can improve educational outcomes by enhancing the writing quality of students and narrowing existing linguistic gaps.[113] Further, LLMs can enhance multilingual capabilities by offering cross lingual feedback. This can be seen as a form of democratizing AI by allowing people world over to benefit from advanced language technology.[114]

## 2.4. The utility of legal LLMs

In the legal field, LLMs increase access to justice by increasing access to both the text and the language of the law.[115] Self-represented litigants can get legal information in a language that is easy to understand.[116] In doing so, LLMs democratise access to legal information and simplify legal processes which usually act as structural barriers.[117] LLMs can be trained on jurisdiction-specific information and procedure, making them effective self-help tools. For example, JusticeBot is a model that has been specifically trained on Landlord-Tenant disputes in Quebec that allows individuals to get guidance on their own housing disputes.[118]

[112] Cheong I, Xia K, Feng K, Chen Q, and Zhang A. '(A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice,' Association for Computing Machinery Conference on Fairness, Accountability, and Transparency (FAccT '24), Rio de Janeiro, 3 June 2024, 2454.
[113] Yu R, Xu Z, CH-Wang S, Arum R, 'Whose ChatGPT? Unveiling Real-World Educational Inequalities Introduced by Large Language Models' arXiv, 2024, 2.
[114] Lai W, Mesgar M, Fraser A, 'LLMs Beyond English: Scaling the Multilingual Capability of LLMs with Cross-Lingual Feedback', in Ku L, Martins A and Srikumar V, *Findings of the Association for Computational Linguistics: ACL 2024*, Association for Computational Linguistics, Bangkok, 8194.
[115] See Dahl M, Magesh V, Zuzgun M, Ho D, 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models', 16 *Journal of Legal Analysis* 1, 2024, 65 ; Chien C, Kim M, Raj A and Rathis R, 'How Generative AI Can Help Address the Access to Justice Gap through the Courts,' Loyola of Los Angeles Law Review, Forthcoming, 1-3. 2024. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4683309# on 17 February 2025; Cheong I, Xia K, Feng K, Chen Q, and Zhang A. '(A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice,' Association for Computing Machinery Conference on Fairness, Accountability, and Transparency (FAccT '24), Rio de Janeiro, 3 June 2024, 2454.
[116] Tan J, Westermann H, Benyekhlef K, 'Chat GPT as an Artificial Lawyer?' ICAIL 2023 Workshop on Artificial Intelligence for Access to Justice (AI4AJ) co located with the 19th International Conference on AI and Law (ICAIL 2023), Braga, 19 June 2023, 2-3.
[117] M'Rhar K, Ben Jaafar C, Bencharef O and Bourkoukou O, 'Unlocking the Potential of Large Language Models in Legal Discourse: Challenges, Solutions, and Future Directions,' 2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS), Marrakech, 23 October 2024, 1.
[118] Westermann H and Benyekhlef K, 'JusticeBot: A methodology for building augmented intelligence tools for laypeople to increase access to justice,' Nineteenth International Conference on Artificial Intelligence and Law (ICAIL '23), Braga, 20 June 2023, 351-359.

Further, LLMs have been found to be 99.97% cheaper than the average lawyer, meaning that more people are be able to have their legal needs met by using LLMs.[119] The US Chief Justice, Hon. John Roberts noted that AI can help litigants who lack resources.[120] It is worth noting that models have been found to outperform lawyers. Martin et al found that LLMs can review contracts at a faster pace, and a cheaper rate than both junior and senior lawyers.[121]

As LLMs get better, they could be used by law makers in legal drafting and identifying loop holes in existing law.[122] LLMs can also assist judges in making decisions. In doing so, LLMs promote efficiency, accuracy and consistency in judicial decision making. While there are some pertinent ethical concerns, it is plausible that advanced LLMs may replace judges themselves in the future.[123]

## 2.5. Conclusion

This chapter set out to provide a general understanding of what LLMs are, how they generate output and what their utility is. It has done so by providing a brief definition of what LLMs are and some examples. It then went on to explain what machine learning is and how it works and demonstrated that LLMs are differentiated from other deep learning models due to their transformer architecture. This chapter proceeded to explain different machine learning techniques with an emphasis on reinforcement learning and semi supervised learning. Finally, it demonstrated the utility of LLMs with an emphasis on legal LLMs. This lays the foundation for the next chapter which will build on the findings of this chapter to explain what hallucinations are, what they occur and how they are harmful especially in the legal context.

[119] Cheong I, Xia K, Feng K, Chen Q, and Zhang A. '(A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice,' Association for Computing Machinery Conference on Fairness, Accountability, and Transparency (FAccT '24), Rio de Janeiro, 3 June 2024, 2454.
[120] 2023 Year-End Report on the Federal Judiciary, 6. 2023year-endreport.pdf on 18 March 2025.
[121] Martin L, Whitehouse N, Yiu S, Catterson L and Rivindu P, 'Better call GPT: Comparing large language models against lawyers', arXiv, 2024, 9-11.
[122] M'Rhar K, Ben Jaafar C, Bencharef O and Bourkoukou O, 'Unlocking the Potential of Large Language Models in Legal Discourse: Challenges, Solutions, and Future Directions,' 2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS), Marrakech, 23 October 2024, 2.
[123] Lai J, Gan W, Wu J, Qi Z, Yu P, 'Large language models in law: A survey,' 5 AI Open 1, 2024, 194.

# 3.0. Hallucination generation in in legal LLMs

## 3.1. Introduction

While the previous chapter explained what an LLM is, how it functions and what its utility is, this chapter seeks to explain the phenomenon of hallucination itself and outline the detection and mitigation measures that developers can take to minimise hallucinations. It will do so by providing a brief description of what hallucinations are, then proceed to explain how they can be detected and minimised.

## 3.2. Understanding hallucinations

LLMs exhibit a tendency to generate factually inaccurate or nonsensical statements that are inconsistent with the input, usually when there is a knowledge gap. This phenomenon is known as hallucination.[124] Hallucinations are usually convincing because they may seem plausible, yet they are non-factual.[125]

### 3.2.1. Types of hallucinations

Hallucinations can be extrinsic or intrinsic. Intrinsic hallucinations produce out of context responses, while extrinsic hallucinations produce factually incorrect information.[126] Several scholars have attempted to classify hallucinations.[127] However, Huang et al provide a very clear and straightforward dichotomy of hallucinations in LLMs; factuality hallucinations and faithfulness hallucinations.[128] Factuality hallucinations occur when the generated content is

[124] Chen C and Shu K, 'Combatting Misinformation in the Age of LLMs: Opportunities and Challenges' 357.
[125] Huang L, Yu W, Ma W, Zhong W, Chen Q, Peng W, Feng X, Qin B, and Liu T, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', 43 *Association for Computing Machinery Transactions on Information Systems* 2, 2024, 2.
[126] Huang Y, Feng X, and Qin B, 'The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey', arXiv, 2021, 2; Ji Z, Lee N, Frieske R, Yu T, Su D, Xu D, Ishii E, Bang Y, Madotto A, and Fung P, 'Survey of Hallucination in Natural Language Generation' 1; Li W, Wu W, Chen M, Liu J, Xiao X, and Wu H, 'Faithfulness in Natural Language Generation: A Systematic Survey of Analysis, Evaluation and Optimization Methods', arXiv, 9-10.
[127] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu D, Ishii E, Bang Y, Madotto A, and Fung P, 'Survey of Hallucination in Natural Language Generation,' 1; Rawte V et al, 'The troubling emergence of hallucination in LLMs, an extensive definition, quantification and prescriptive remediations,' 2544-2545; Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, Huang X, Zhao E, Zhang Y, Chen Y, Wang L, Luu A, Bi W, Shi F, Shi S, ' Siren's song in the AI ocean: A survey on hallucination in Large Language Models', arXiv, 2023, 3-4.
[128] Huang L, Yu W, Ma W, Zhong W, Chen Q, Peng W, Feng X, Qin B, Liu T, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', 43 *Association for Computing Machinery Transactions on Information Systems* 1, 2024, 1-58.

incompatible with real world facts, while faithfulness hallucinations occur when the generated content is inconsistent with the input.

### 3.2.2. Factuality hallucinations

According to Huang et al, factuality hallucinations present themselves in two ways.[129] The first is through factual contradictions.[130] This type of hallucination occurs where the facts that are contradictory but are rooted in verifiable information. LLMs can also misrepresent the connection between entities, such as attributing phenomena to the wrong character.
The second category of factuality hallucinations Huang et al identify are known as factual fabrications.[131] This kind of hallucination is particularly worrisome as it occurs when the LLM's output is factually false. The authors further subdivide factual fabrications into unverifiability hallucinations and overclaim hallucinations. Unverifiability hallucinations occur when LLMs generate output that is entirely non-existent while overclaim hallucinations occur when LLM presents something as widely accepted when it in fact lacks consensus.

**Training data**

Training data in LLMs can be categorised into pre-training data which gives LLMs knowledge about the world and their general capabilities,[132] and alignment data which teaches LLMs to align with human preferences and follow user instructions.[133] Flawed pre-training data that contains misinformation and biases may cause hallucinations, as well as a knowledge boundary and inferior alignment data. However, this is not a hard and fast rule because sometimes training methods such as reinforcement learning may cause LLMs to hallucinate facts.[134]

---

[129] Huang L, Yu W, Ma W, Zhong W, Chen Q, Peng W, Feng X, Qin B, Liu T, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', 43 *Association for Computing Machinery Transactions on Information Systems* 2, 2024, 2.
[130] Huang L, Yu W, Ma W, Zhong W, Chen Q, Peng W, Feng X, Qin B, Liu T, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', 43 *Association for Computing Machinery Transactions on Information Systems* 2, 2024, 5-7.
[131] Huang L, Yu W, Ma W, Zhong W, Chen Q, Peng W, Feng X, Qin B, Liu T, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', 43 *Association for Computing Machinery Transactions on Information Systems* 2, 2024, 7-10.
[132] Wang Y, Zhong W, Li L, Mi F, Zeng X, Huang W, Shang L, Jiang X, and Liu Q, 'Aligning Large Language Models with Humans: A Survey,' arXiv, 2023, 1.
[133] Wang Y, Zhong W, Li L, Mi F, Zeng X, Huang W, Shang L, Jiang X, and Liu Q, 'Aligning Large Language Models with Humans: A Survey,' arXiv, 2023, 1-3.
[134] Yang Y, Chern E, Qiu X, Neubig G, and Liu P, 'Alignment for Honesty,' Thirty-Eight Annual Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, 12 December 2024, 1-2;

**Misinformation and biases**

Neural networks can memorise training data.[135] This capacity grows with model size.[136] Memorisation is a double-edged sword since it can amplify biases present in pre-training data thus reinforcing societal biases.[137] Biases can present themselves as hallucinations through imitative falsehoods and societal biases.

*Imitative falsehoods*

LLMs need extensive data during training.[138] This may present a challenge in maintaining a consistent data quality thus introducing misinformation in pre-training data,[139] which increases the likelihood of LLMs repeating the false statements in their data.[140]

*Societal biases*

Zhang H, Diao S, Lin Y, Fung Y R, Lian Q, Wang X, Chen Y, Ji H, and Zhang T, 'R-Tuning: Instructing Large Language Models to Say 'I Don't Know', in Duh K , Gomez H and Bethard S (eds), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Association for Computational Linguistics, Mexico City, 2024, 7113, Bender E, Gebru T, McMillan-Major A and Shmitchell S, 'On the dangers of stochastic parrots: Can language models be too big?' in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery*, 2021, 617-618.

[135] Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, Roberts A, Brown T, Song D, Erlingsson U, Oprea A, and Raffel C, 'Extracting training data from large language models,' in 30th USENIX Security Symposium, Virtual, 11 August 2021, 2633–2650.

[136] Carlini N, Ippolito D, Jagielski M, Lee K, Tramer F, and Zhang , 'Quantifying memorization across neural language models,' International Conference on Learning Representation 2023, Kigali, 1 May 2023, 9; Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Barham P, Chung H W, Sutton C, Gehrmann S, Schuh P, Shi K, Tsvyashchenko S, Maynez J, Rao A, Barnes P, Tay Y, Shazeer N, Prabhakaran V, Reif E, Du N, Hutchinson B, Pope R, Bradbury J, Austin J, Isard M, Gur-Ari G, Yin P, Duke T, Levskaya A, Ghemawat S, Dev S, Michalewski H, Garcia X, Misra V, Robinson K, Fedus L, Zhou D, Ippolito D, Luan D, Lim H, Zoph B, Spiridonov A, Sepassi R, Dohan D, Agrawal S, Omernick M, Dai A M, Pillai T S, Pellat M, Lewkowycz A, Moreira E, Child R, Polozov O, Lee K, Zhou Z, Wang X, Saeta B, Diaz M, Firat O, Catasta M, Wei J, Meier-Hellstern K, Eck D, Dean J, Petrov S, and Fiedel N, 'PaLM: Scaling Language Modeling with Pathways,' 24 *Journal of Machine Learning Research* 1, 2023, 45-48.

[137] Lin S, Hilton J, and Evans O, 'TruthfulQA: Measuring How Models Mimic Human Falsehoods' in Muresan S, Nakov P and Villavicencio A, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, 2022, 3221.

[138] Taye M, 'Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions', 12 *Computers* 5, 2023, 22.

[139] Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang P-S, Cheng M, Glaese M, Balle B, Kasirzadeh A, Kenton Z, Brown S, Hawkins W, Stepleton T, Biles C, Birhane A, Haas J, Rimell L, Hendricks L A, Isaac W, Legassick S, Irving G, and Gabriel I, 'Ethical and social risks of harm from Language Models', arXiv, 2021, 22.

[140] Huang L, Yu W, Ma W, Zhong W, Chen Q, Peng W, Feng X, Qin B, Liu T, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', 43 *Association for Computing Machinery Transactions on Information Systems* 2, 2025, 8.

LLMs may hallucinate data based on biases. This happens when they are fed data sets that contain certain stereotypical correlations or underrepresent marginalised people.[141]

**Knowledge boundary**

Since LLMs often do not memorise all factual pre-training data,[142] and the pre-training data tends to be outdated during deployment or is restricted by copyright, they have very limited actual knowledge. This makes more susceptible to hallucinations to bridge the knowledge gap.[143]

*Long tail knowledge*

Most LLMs are trained on general knowledge therefore they tend to perform better when asked general knowledge questions or when answering questions related to prominent data.[144] The data that is not really prominent or not general is known as long tail knowledge. Since LLMs are trained extensively on general knowledge, they tend to hallucinate more when asked domain-specific questions, often manifesting as factual fabrication.[145] For

[141] Paullada A, Raji I D, Bender E M, Denton E, and Hanna A, 'Data and its (dis)contents: A survey of dataset development and use in machine learning research,' 2 *Patterns* 11, 2021, 3.
[142] Kandpal N, Deng H, Roberts A, Wallace E, Raffel C, 'Large Language Models Struggle to Learn Long-Tail Knowledge', in Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Honolulu, 2023, 8.
[143] Huang L, Yu W, Ma W, Zhong W, Chen Q, Peng W, Feng X, Qin B, Liu T, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', 43 *Association for Computing Machinery Transactions on Information Systems* 3, 2024, 9.
[144] Kandpal N, Deng H, Roberts A, Wallace E, Raffel C, 'Large Language Models Struggle to Learn Long-Tail Knowledge', in Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Honolulu, 2023, 5.
[145] Li Y, Li Z, Zhang K, Dan R, and Zhang Y, 'ChatDoctor: A Medical Chat Model Fine-tuned on Llama Model using Medical Domain Knowledge' 15 *Cureus* 6, 2023, 11.

example, LLMs have been known to hallucinate when asked legal,[146] or medical questions.[147]

*Up to date knowledge*

The factual knowledge LLMs may have memorised from their training data quickly becomes out of date during deployment which causes some LLMs to hallucinate.[148]

*Copyright sensitive knowledge*

Sometimes an LLM has a knowledge gap because it was not fed certain data during training due to copyright reasons.[149] This uncertainty causes LLMs to fabricate knowledge to make up for the knowledge gap.[150]

**Reinforcement learning**

[146] Dahl M, Magesh V, Zuzgun M, Ho D, 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models', 16 *Journal of Legal Analysis* 1, 2024, 65; Curran S, Bethell O, Lansley S, 'Hallucination is the last thing you need, arXiv, 2023, 2; See also Cheong I, Xia K, Feng K, Chen Q, and Zhang A. '(A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice,' Association for Computing Machinery Conference on Fairness, Accountability, and Transparency (FAccT '24), Rio de Janeiro, 3 June 2024, 2454; Kapoor S, Henderson P, Narayanan A, 'Promises and pitfalls of artificial intelligence for legal applications' 2 *Journal of cross disciplinary research in computational law* 2, 2024, 4; Tan J, Westermann H and Benyekhlef K, 'ChatGPT as an Artificial Lawyer?' International Conference on Artificial Intelligence and Law 2023 Workshop on Artificial Intelligence for Access to Justice (AI4AJ), Braga, 19 June 2023, 2.

[147] Li Y, Li Z, Zhang K, Dan R, and Zhang Y, 'ChatDoctor: A Medical Chat Model Fine-tuned on Llama Model using Medical Domain Knowledge' 15 *Cureus* 6, 2023, 11.

[148] Kasai J, Sakaguchi K, Takahashi Y, Le Bras R, Asai A, Yu X, Radev D, Smith N A, Choi Y, and Inui K, 'RealTime QA: What's the Answer Right Now?,' in Oh A, Naumann T, Globerson A, Saenko K, Hardt M and Levine S, *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, Curran Associates Inc, New Orleans, 7; Li D, Rawat A S, Zaheer M, Wang X, Lukasik M, Veit A, Yu F X, and Kumar S, 'Large Language Models with Controllable Working Memory,' in Rogers A, Boyd-Graber J, and Okazaki N (eds), *Findings of the Association for Computational Linguistics: ACL 2023*, Association for Computational Linguistics, Toronto, 2023, 1774; Onoe Y, Zhang M, Choi E, and Durrett G, 'Entity Cloze by Date: What LMs Know About Unseen Entities,' in Carpuat M, de Marneffe M, Ruiz M and Vladimir I, *Findings of the Association for Computational Linguistics: NAACL 2022*, Association for Computational Linguistics, Seattle, 2022, 693–694.

[149] Min S, Gururangan S, Wallace E, Hajishirzi H, Smith N A, and Zettlemoyer L, 'SILO Language Models: Isolating Legal Risk in a Nonparametric Datastore,' Twelfth International Conference on Learning Representations, Vienna, 8 May 2024, 1-2. AI companies have been sued severally for infringing copyright while training their LLMs, reinforcing IP protections as limitations to training LLMs. See The New York Times Company v Microsoft Corporation, OpenAI, Inc., OpenAI LP, OpenAI GP, LLC, OpenAI LLC, OpenAI OpCo LLC, OpenAI Global LLC, OAI Corporation, LLC, And OpenAI Holdings, LLC, Southern District Court of New York (2024), Tremblay v. OpenAI Southern District Court of New York (2024), Kadrey v. Meta Platforms, Inc. Northern District of California (2023), Andersen v. Stability AI Ltd. Northern District of California (2023).

[150] Huang L, Yu W, Ma W, Zhong W, Chen Q, Peng W, Feng X, Qin B, Liu T, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', 43 *Association for Computing Machinery Transactions on Information Systems* 2, 2024, 9.

LLMs can exhibit sycophancy which manifests itself as choosing an incorrect answer despite being aware of its inaccuracy to agree with the user.[151] Sharma et al[152] and Perez et al[153] argue that sycophancy stems from RLHF since this training method favours outputs that appeal to human preferences which may make models produce output that appeals to humans but are flawed or incorrect. RLHF also trains models to complete each response without expressing uncertainty during training which encourages them to fabricate content when they have a knowledge gap.[154]

**Erroneous encoding and decoding**

An encoder in an LLM understands the input text and converts it into a format that the LLM can understand. Some encoders fail to convert the input text properly which may cause hallucinations as something is 'lost in translation'.[155] Encoders can also learn wrong correlations between different phenomena in data which may also lead to hallucinations.[156]

---

[151] Wei J, Huang D, Lu Y, Zhou D, and Le Q, 'Simple synthetic data reduces sycophancy in large language models,' arXiv, 2023, 1-4.

[152] Sharma S, Tong M, Korbak T, Duvenaud D, Askell A, Bowman S, Cheng N, Burmus E, Hatfield-Dodds Z, Johnston S, Kravec S, Maxwell T, McCandlish S, Ndousse K, Rausch O, Scheifer N, Yna D, Zhang M, Perez E, 'Towards Understanding Sycophancy in Language Models' Twelfth International Conference on Learning Representations, Vienna, May 7, 1-10.

[153] Perez E, Ringer S, Lukošiūtė K, Nguyen K, Chen E, Heiner S, Pettit C, Olsson C, Kundu S, Kadavath S, Jones A, Chen A, Mann B, Israel B, Seethor B, McKinnon C, Olah C, Yan D, Amodei D, Amodei D, Drain D, Li D, Tran-Johnson E, Khundadze G, Kernion J, Landis J, Kerr J, Mueller J, Hyun J, Landau J, Ndousse K, Goldberg L, Lovitt L, Lucas M, Sellitto M, Zhang M, Kingsland N, Elhage N, Joseph N, Mercado N, DasSarma N, Rausch O, Larson R, McCandlish S, Johnston S, Kravec S, El Showk S, Lanham T, Telleen-Lawton T, Brown T, Henighan T, Hume T, Bai Y, Hatfield-Dodds Z, Clark J, Bowman S R, Askell A, Grosse R, Hernandez D, Ganguli D, Hubinger E, Schiefer N, and Kaplan J, 'Discovering language model behaviours with model-written evaluations,' 61st Annual Meeting of the Association for Computational Linguistics, ACL 2023, Toronto, 9 July 2023, 8-9.

[154] Yang Y, Chern E, Qiu X, Neubig G, and Liu P, 'Alignment for Honesty,' Thirty-Eight Annual Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, 12 December 2024, 1-2; ; Zhang H, Diao S, Lin Y, Fung Y R, Lian Q, Wang X, Chen Y, Ji H, and Zhang T, 'R-Tuning: Instructing Large Language Models to Say 'I Don't Know', in Duh K , Gomez H and Bethard S (eds), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Association for Computational Linguistics, Mexico City, 2024, 7113, Bender E, Gebru T, McMillan-Major A and Shmitchell S, 'On the dangers of stochastic parrots: Can language models be too big?' in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery*, 2021, 617-618.

[155] Parikh A, Wang X, Gehrmann S, Faruqui M, Dhingra B, Yang D, and Das D, 'ToTTo: A Controlled Table-To-Text Generation Dataset,' in Webber B, Cohn T, He Y and Liu Y (eds) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Virtual, 2020, 1173–1174.

[156] Feng Y, Xie W, Gu S, Shao C, Zhang W, Yang Z, and Yu D, 'Modeling Fluency and Faithfulness for Diverse Neural Machine Translation,' in *Proceedings of the AAAI Conference on Artificial Intelligence Volume 34(1): Technical Tracks 1*, AAAI Press, Palo Alto, 2020, 59. See also Tian R, Narayan S, Sellam T, and Parikh A, 'Sticking to the Facts: Confident Decoding for Faithful Data-to-Text Generation,' arXiv, 2020, 2; Dziri N, Madotto A, Zaiane O, and Bose A J, 'Neural Path Hunter: Reducing Hallucination in Dialogue Systems via

While the encoder 'translates' the input text into a format the model can understand, a decoder converts the model's output into text that humans can comprehend. Decoders can do this conversion erroneously by attending to the wrong part of the sentence that leads to mixed up facts thereby generating hallucinations.[157]

### 3.2.3. Faithfulness hallucinations

LLMs are trained to be in alignment with user interactions.[158] As LLMs become more user centric, ensuring their consistency with the context and information given by the user becomes increasingly important. LLMs that exhibit faithfulness hallucinations deviate from this goal by giving out of context responses, exhibiting logical errors in their output or giving responses that deviate from the users' instructions. Though it is worth noting that instruction inconsistencies are only termed as hallucinations when the incongruence is unintentional, and the user instructions are non-malicious. This is because some inconsistencies exist for safety reasons, such as to prevent adversarial attacks.[159]

**Training process**

The training methods discussed in section 2.2.2B can cause hallucinations either directly or indirectly. For example, when models are pre-trained on vast amounts of data, they memorise the training knowledge in their parameters.[160] While this memorized knowledge helps improve the LLM's overall performance, it leads to hallucinations.[161] LLMs have been

VT OMNES VNVM SINT

Path Grounding,' in Moens M, Huang X, Specia L and Yih S, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language* Processing, Association for Computational Linguistics, Punta Cana, 2021, 2197-2198.

[157] Tian R, Narayan S, Sellam T, and Parikh A, 'Sticking to the Facts: Confident Decoding for Faithful Data-to-Text Generation,' arXiv, 2020, 1-2. See also Dziri N, Madotto A, Zaiane O, and Bose A J, 'Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding,' in Moens M, Huang X, Specia L and Yih S, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language* Processing, Association for Computational Linguistics, Punta Cana, 2021, 2197-2202.

[158] Wang Y, Zhong W, Li L, Mi F, Zeng X, Huang W, Shang L, Jiang X, and Liu Q, 'Aligning Large Language Models with Humans: A Survey,' arXiv, 2023, 1-3, 7-10.

[159] Zou A, Wang Z, Carlini N, Nasr M, Kolter J, and Fredrikson M, 'Universal and Transferable Adversarial Attacks on Aligned Language Models,' arXiv, 2023, 1-3.

[160] Petroni F, Rocktäschel T, Riedel S, Lewis P, Bakhtin A, Wu Y and Miller A, '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, Hong Kong, 2019, 2464, 2470-2471.

[161] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu D, Ishii E, Bang Y, Madotto A, and Fung P, 'Survey of Hallucination in Natural Language Generation' 55 *ACM Computing Survey* 12, 2023, 8.

observed to prioritize parametric knowledge over the provided input.[162] This may cause the output the model gives to be in conflict with the input provided.

## 3.3. Hallucination detection and minimisation

### 3.3.1. Detecting hallucinations

The method of detecting hallucinations depends on the type of hallucination. Cause and type are distinguished because some factors such as training methods and cause both factual and faithfulness hallucinations as seen in 3.2.1.

#### A. Detecting factuality hallucinations

When detecting this type of hallucinations, technical researchers ask themselves whether the output aligns with real world facts.[163] This consists of fact checking and estimating the LLM's uncertainty.

Fact checking is done by external retrieval and internal checking. External retrieval utilises reliable knowledge sources to counter check the veracity of the output and computes the percentage of supported information.[164] On the other hand, internal checking looks for inconsistencies within the LLM.[165] This is done by checking the parameters of the LLM against the draft of the output.[166] However, solely relying on LLMs' knowledge may be

[162] Longpre S, Perisetla K, Chen A, Ramesh N, DuBois C, and Singh S, 'Entity-Based Knowledge Conflicts in Question Answering,' in Moens M, Huang X, Specia L and Yih S, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Punta Cana, 2021, 7052–7053.

[163] Huang L, Yu W, Ma W, Zhong W, Chen Q, Peng W, Feng X, Qin B, Liu T, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', 43 *Association for Computing Machinery Transactions on Information Systems* 2, 2024, 12.

[164] See Min S, Krishna K, Lyu X, Lewis M, Yih W, Koh P W, Iyyer M, Zettlemoyer L, and Hajishirzi H, 'FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation,' in Bauamor H, Pino J and Bali K, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, 2023, 12076-12085; Chern I, Chern S, Chen S, Yuan W, Feng K, Zhou C, He J, Neubig G, and Liu P, 'FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios,' ArXiv, 2023, 1-12

[165] Huang L, Yu W, Ma W, Zhong W, Chen Q, Peng W, Feng X, Qin B, Liu T, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', 43 *Association for Computing Machinery Transactions on Information Systems* 2, 2024, 13.

[166] See Dhuliawala S, Komeili M, Xu J, Raileanu R, Li X, Celikyilmaz A, and Weston J, 'Chain-of-Verification Reduces Hallucination in Large Language Models,' ArXiv, 2023, 1-9; Kadavath S, Conerly T, Askell A, Henighan T, Drain D, Perez E, Schiefer N, Hatfield-Dodds Z, DasSarma N, Tran-Johnson E, Johnston S, El-Showk S, Jones A, Elhage N, Hume T, Chen A, Bai Y, Bowman S, Fort S, Ganguli D, Hernandez D, Jacobson J, Kernion J, Kravec S, Lovitt L, Ndousse K, Olsson C, Ringer S, Amodei D, Brown T, Clark J, Joseph N, Mann B, McCandlish S, Olah C, and Kaplan J, 'Language models (mostly) know what they know,' ArXiv, 2022, 1-23; Zhang X, Peng B, Tian Y, Zhou J, Jin L, Song L, Mi H, and Meng H, 'Self-Alignment for Factuality: Mitigating Hallucinations in LLMs via Self-Evaluation,' in Ku L, Martins A and Srikumar V (eds),

undesirable as LLMs tend to be inconsistent, non-factual and therefore unreliable.[167] Further, due to inherent data constraints discussed in the section above such as long tail knowledge, copyright limitations and up-to-date knowledge, internal checking mechanisms cannot be used in isolation.

Factuality can also be ascertained by estimating uncertainty. This can be done through assessing LLMs' internal states by using the LLM's own estimation of probability,[168] or its ability to reconstruct the concept.[169] Technical researchers have also devised ways to use prompts to detect hallucinations by detecting uncertainty.[170]

### B. Detecting faithfulness hallucinations

This category of hallucinations is usually detected by measuring the overlap of key facts between the output and the prompt.[171] Technical scholars have also trained classifiers to identify faithfulness inconsistencies.[172] Additionally, there are an increasing number of LLM based detection mechanisms which entail feeding both the prompt and the output to another

*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Bangkok, 2024, 1946-1954.

[167] Zheng D, Lapata M and Pan J, 'Large Language Models as Reliable Knowledge Bases? Re-thinking factuality and consistency,' arXiv, 2024, 7.

[168] Varshney N, Yao W, Zhang H, Chen J and Yu D, 'A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation,' ArXiv, 2023, 4.

[169] Luo J, Xiao C and Ma F, 'Zero-Resource Hallucination Prevention for Large Language Models,' in Al-Onaizan Y, Bansal M and Chen Y (eds), *Findings of the Association for Computational Linguistics: EMNLP 2024*, Association for Computational Linguistics, Miami, 2024, 3586-3594.

[170] Kadavath S, Conerly T, Askell A, Henighan T, Drain D, Perez E, Schiefer N, Hatfield-Dodds Z, DasSarma N, Tran-Johnson E, Johnston S, El-Showk S, Jones A, Elhage N, Hume T, Chen A, Bai Y, Bowman S, Fort S, Ganguli D, Hernandez D, Jacobson J, Kernion J, Kravec S, Lovitt L, Ndousse K, Olsson C, Ringer S, Amodei D, Brown T, Clark J, Joseph N, Mann B, McCandlish S, Olah C, and Kaplan J, 'Language models (mostly) know what they know,' ArXiv, 2022, 1-23; Agrawal A, Suzgun M, Mackey L and Kalai A, 'Do Language Models Know When They're Hallucinating References?,' in Graham Y and Purver M (eds) *Findings of the Association for Computational Linguistics: EACL 2024*, Association for Computational Linguistics, St. Julian's, 2024, 912-920.

[171] Wang Z, Wang X, An B, Yu D, and Chen C, 'Towards Faithful Neural Table-to-Text Generation with Content-Matching Constraints,' in Jurafsky D, Chai J, Schluter N and Tetreault J (eds) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020, 1072–1080; Maynez J, Narayan S, Bohnet B, and McDonald R, in Jurafsky D, 'On Faithfulness and Factuality in Abstractive Summarization,' in Chai J, Schluter N and Tetreault N (eds) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020 1910-1914.

[172] Falke T, Ribeiro L, Utama P, Dagan I and Gurevych I, 'Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference,' in Korhonen A, Traum D and Marquez Luis (eds) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, 2019, 2214–2218; Maynez J, Narayan S, Bohnet B, and McDonald R, 'On Faithfulness and Factuality in Abstractive Summarization,' in Maynez J, Narayan S, Bohnet B, McDonald R (eds) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020, 1912–1914.

LLM which has been trained on evaluation guidelines and can therefore spot inconsistencies between the two.[173]

### 3.3.2. Minimising hallucinations

Hallucination minimisation is based on the *cause* of the hallucination unlike hallucination detection which is based on the *type* of hallucination.[174] Since hallucinations are mainly caused by shortcomings present in the training data or training process, their minimisation is based on addressing these challenges.

#### A. Minimising hallucinations caused by training data

The main methods used to minimise this type of hallucinations are data filtering and model editing. When using data filtering, the pre-training data is refined to minimise misinformation and biases before being fed into the model. Data filtering is usually done by opting for data that has undergone rigorous curation and filtration such as academic work.[175]

Additionally, model editing has been used to minimise hallucinations through rectifying model behaviour by incorporating additional knowledge.[176] Model editing utilises two main

[173] Adlakha V, Behnam Ghader P, Lu XH, Meade N and Reddy S, 'Evaluating correctness and faithfulness of instruction-following models for question answering,' 12 *Transactions of the Association for Computational Linguistics* 1, 2023, 681- 692; Gao M, Ruan J, Sun R, Yin X, Yang S, and Wan X, 'Human-like summarization evaluation with ChatGPT,' ArXiv, 2023, 1-7; Jain S, Keshava V, Mysore Sathyendra S, Fernandes P, Liu P, Neubig G, and Zhou C, 'Multi-Dimensional Evaluation of Text Summarization with In-Context Learning,' in Rogers A, Boyd-Graber J and Okazaki N (eds) *Findings of the Association for Computational Linguistics: ACL 2023*, Association for Computational Linguistics, Toronto, 2023, 8487 – 8491; Laban P, Kryściński W, Agarwal D, Fabbri A, Xiong C, Joty S, and Wu C, 'LLMs as Factual Reasoners: Insights from Existing Benchmarks and Beyond,' ArXiv, 2023, 1-15.

[174] Huang L, Yu W, Ma W, Zhong W, Chen Q, Peng W, Feng X, Qin B, Liu T, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', 43 *Association for Computing Machinery Transactions on Information Systems* 2, 2024, 19.

[175] Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D, 'Language models are few-shot learners', in Larochelle H, Ranzato M, Hadsell R, Balcan M and Lin H (eds) in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, Curran Associates Inc, New York, 2020, 1880-1881; Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, Phang J, He H, Thite A, Nabeshima N, Presser S, and Leahy C, 'The Pile: An 800gb dataset of diverse text for language modelling,' ArXiv, 2021, 1 – 16.

[176] Sinitsin A, Plokhotnyuk V, Pyrkin D, Popov S, and Babenko A, 'Editable Neural Networks,' 8th International Conference on Learning Representations (ICLR 2020), Addis Ababa, April 27, 2020, 1-10; Wang S, Zhu Y, Liu H, Zheng Z, Chen C, and Li J, 'Knowledge Editing for Large Language Models: A Survey,' 57 *Association for Computing Machinery Surveys* 3, 2024.

techniques, locate-then-edit,[177] and meta learning.[178] Locate-then-edit methods identify the faulty part of the model parameters then update them to change the model's behaviours.[179]

On the other hand, meta learning methods train an external network to predict the extent to which the weights need to be updated.[180] The model's behaviour is then altered by adjusting the weights according to the prediction.[181] However, model editing can adversely affect performance and often requires additional training and memory cost.[182] Retrieval Augmented Generation (RAG) is another technique that is used to minimise hallucinations caused by data. RAG works by retrieving information from external sources to supplement the training data.[183]

[177] Meng K, Bau D, Andonian A, and Belinko Y, 'Locating and Editing Factual Associations in GPT,' in Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K and Oh A (eds) *NIPS'22: 36th International Conference on Neural Information Processing Systems*, Curran Associates Inc, New Orleans, 2022, 17359 -17368. See also Dai D, Dong L, Hao Y, Sui Z, Chang B, and Wei F, 'Knowledge Neurons in Pretrained Transformers,' in Muresan S, Nakov P and Villavivencio A (eds) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, 2022, 8493–8500.

[178] De Cao N, Aziz W, and Titov I, 'Editing Factual Knowledge in Language Models,' in Moens M, Huang X, Specia L, Yih S (eds) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, 6491–6506; Mitchell E, Lin C, Bosselut A, Finn C, and Manning C, 'Fast Model Editing at Scale,' Tenth International Conference on Learning Representations, ICLR 2022, Virtual, April 26, 2022, 1-10.

[179] Meng K, Bau D, Andonian A, and Belinko Y, 'Locating and Editing Factual Associations in GPT,' in Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K and Oh A (eds) *NIPS'22: 36th International Conference on Neural Information Processing Systems*, Curran Associates Inc, New Orleans, 2022, 17359 -17368.

[180] Mitchell E, Lin C, Bosselut A, Finn C, and Manning C, 'Fast Model Editing at Scale,' Tenth International Conference on Learning Representations, ICLR 2022, Virtual, April 26, 2022, 1-10.

[181] Huang L, Yu W, Ma W, Zhong W, Chen Q, Peng W, Feng X, Qin B, Liu T, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', 43 *Association for Computing Machinery Transactions on Information Systems* 2, 2024, 20.

[182] Huang L, Yu W, Ma W, Zhong W, Chen Q, Peng W, Feng X, Qin B, Liu T, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', 43 *Association for Computing Machinery Transactions on Information Systems* 2, 2024, 20.

[183] Guu K, Lee K, Tung Z, Pasupat P, Chang M, 'Retrieval Augmented Language Model Pre-Training', in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, Journal of Machine Learning Research, Vienna, 2020, 3929–3938; Lu J, Rao J, Chen K, Guo X, Zhang Y, Sun B, Yang C and Yang J, 'Evaluation and Mitigation of Agnosia in Multimodal Large Language Models,' ArXiv, 2023, 1-12; Shuster K, Poff S, Chen M, Kiela D and Weston J, 'Retrieval Augmentation Reduces Hallucination in Conversation,' In Moens M, Huang X, Specia L, Yih S (eds), *Findings of the Association for Computational Linguistics: EMNLP 2021,* Association for Computational Linguistics, Punta Cana, 2021, 3784–3803; Thakur, N, Bonaficio L, Zhang X, Ogundepo O, Kamalloo E, Alfonso-Hermelo D, Li X, Liu Q, Chen B, Rezagholizadeh M, Lin J, 'NoMIRACL: Knowing When You Don't Know for Robust Multilingual Retrieval-Augmented Generation,' in Al-Onaizan Y, Bansal M and Chen Y (eds) *Findings of the Association for Computational Linguistics: EMNLP 2024*, Association for Computational Linguistics, Miami, 2024, 12508-12517.

## B. Minimising hallucinations caused by the training process

Most training related hallucinations arise from model architecture or training strategies.[184] In response to this, technical researchers have attempted to make LLMs more contextually accurate using a method known as regularisation.[185] Using regularisation, models can reduce their reliance on their training data and increase their ability to make contextually accurate guesses. Training-related hallucinations can also stem from misalignment if a model is a sycophant.[186] To reduce sycophancy, some technical researchers aim to increase the quality of feedback received by the LLM,[187] while others opt to modify the LLM's internal state.[188]

## C. Hallucinations and honesty

It is worth noting that there is no consensus on hallucination mitigation. Some scholars think that hallucinations are inevitable as long as LLMs continue to function the way they do.[189] Kalai and Vempala argue that LLMs are trained to be sufficiently good predictors which will

[184] Huang L, Yu W, Ma W, Zhong W, Chen Q, Peng W, Feng X, Qin B and Liu T, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', 1 *Association for Computing Machinery Transactions on Information Systems*, 2024, 22.

[185] Liu B, Ash T, Goel S, Krishnamurthy A and Zhang C, 'Exposing Attention Glitches with Flip-Flop Language Modeling', in Oh A, Naumann T, Globerson A, Saenko K, Hardt M and Levine S (eds) *NIPS'23: 37th International Conference on Neural Information Processing Systems*, Curran Associates Inc, New Orleans, 2023, 25549-25579; Zhang J, Zhao Y, Li H and Zong C, 'Attention with sparsity regularization for neural machine translation and summarization' 27 *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 3, 2018, 507–518.

[186] Sharma S, Tong M, Korbak T, Duvenaud D, Askell A, Bowman S, Cheng N, Burmus E, Hatfield-Dodds Z, Johnston S, Kravec S, Maxwell T, McCandlish S, Ndousse K, Rausch O, Scheifer N, Yna D, Zhang M, Perez E, 'Towards Understanding Sycophancy in Language Models' Twelfth International Conference on Learning Representations, Vienna, May 7 2024, 1-9; See also Perez E, Ringer S, Lukošiūtė K, Nguyen K, Chen E, Heiner S, Pettit C, Olsson C, Kundu S, Kadavath S, Jones A, Chen A, Mann B, Israel B, Seethor B, McKinnon C, Olah C, Yan D, Amodei D, Amodei D, Drain D, Li D, Tran-Johnson E, Khundadze G, Kernion J, Landis J, Kerr J, Mueller J, Hyun J, Landau J, Ndousse K, Goldberg L, Lovitt L, Lucas M, Sellitto M, Zhang M, Kingsland N, Elhage N, Joseph N, Mercado N, DasSarma N, Rausch O, Larson R, McCandlish S, Johnston S, Kravec S, El Showk S, Lanham T, Telleen-Lawton T, Brown T, Henighan T, Hume T, Bai Y, Hatfield-Dodds Z, Clark J, Bowman S R, Askell A, Grosse R, Hernandez D, Ganguli D, Hubinger E, Schiefer N and Kaplan J, 'Discovering language model behaviours with model-written evaluations', 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Toronto, 9 July 2023, 8-9.

[187] Sharma S, Tong M, Korbak T, Duvenaud D, Askell A, Bowman S, Cheng N, Burmus E, Hatfield-Dodds Z, Johnston S, Kravec S, Maxwell T, McCandlish S, Ndousse K, Rausch O, Scheifer N, Yan D, Zhang M and Perez E, 'Toward Understanding Sycophancy in Language Models' arXiv, 2023, 1-10.

[188] Subramani N, Suresh N and Peters M, 'Extracting Latent Steering Vectors from Pretrained Language Models' In Muresan S, Nakov P, Villavicencio A (eds) *Findings of the Association for Computational Linguistics: ACL 2022*, Association for Computational Linguistics, Dublin, 2022, 566–581.

[189] Kalai A, Vempala S, 'Calibrated Language Models Must Hallucinate,' STOC '24: 56th Annual ACM Symposium on Theory of Computing, Vancouver, 25 June 2024, 160-168; Banerjee S, Agarwal A, Singla S, 'LLMs will always hallucinate, and we need to live with this', arXiv, 2024, 1-28; Li J, Consul S, Zhou E, Wong J, Farooqui N, Ye Y, Manohar N, Wei Z, Echols B, Zhou S, Diamos G, 'Banishing LLM hallucinations requires rethinking generalisation', arXiv, 2024, 1-10.

make them prone to hallucinations for as long as they are trained in this way.[190] Additionally, Li et al argue that hallucinations occur because LLMs memorise their data sets and as long as memorisation remains a key element of LLMs, hallucinations will persist.[191]

Though hallucinations may be inevitable, it is important that LLMs are able to admit what they do and do not know. As seen in 2.2.2 B as well as 3.2.2, RLHF may train models to fabricate information where there is a knowledge gap, or even give an inaccurate answer despite knowing its inaccuracy. Technical methods such as R-Tuning exist which teach LLMs to express uncertainty.[192]

## 3.4. Conclusion

This chapter has explained the phenomenon of hallucination, outlined the detection and mitigation measures in place, and highlighted exactly why it may be harmful especially in the context of legal LLMs. It has done so by providing a brief description of what hallucinations are, then proceed to explain how they can be detected and minimised. The next chapter seeks discuss the harms that can be occasioned by hallucinations especially in the legal sector and to make a case for the imposition of duties to minimise and to disclose the existence of hallucinations onto developers.

VT OMNES VNVM SINT

[190] Kalai A, Vempala S, 'Calibrated Language Models Must Hallucinate,' STOC '24: 56th Annual ACM Symposium on Theory of Computing, Vancouver, 25 June 2024, 160-168.
[191] Li J, Consul S, Zhou E, Wong J, Farooqui N, Ye Y, Manohar N, Wei Z, Echols B, Zhou S, Diamos G, 'Banishing LLM hallucinations requires rethinking generalisation', arXiv, 2024, 1-10.
[192] Zhang H, Diao S, Lin Y, Fung Y R, Lian Q, Wang X, Chen Y, Ji H, and Zhang T, 'R-Tuning: Instructing Large Language Models to Say 'I Don't Know', in Duh K , Gomez H and Bethard S (eds), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Association for Computational Linguistics, Mexico City, 2024, 7113-7121. See also Yang Y, Chern E, Qiu X, Neubig G, and Liu P, 'Alignment for Honesty,' Thirty-Eight Annual Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, 12 December 2024, 1-10.

# 4.0. Developers' duties to minimise the existence of hallucinations in legal LLMs

## 4.1. Introduction

While the previous chapter introduced what hallucinations are and explained how they can be detected and minimised, this chapter seeks to highlight the harms posed by hallucinations and make a case for the imposition of duties to minimise hallucinations and disclose their existence. It seeks to do so by highlighting the role that legal LLMs play in promoting access to justice, relative to the vulnerability of users of these legal LLMs. It will then discuss the harms posed by hallucinations in the legal sector and call for the adoption of a quasi-honesty standard. It will then call to attention the fact that there are measures developers can take to minimise hallucinations and use that to advocate for obligations to be imposed on developers to take these measures. This chapter will then evaluate the different kind of legal measures that various scholars have proposed and make a case for the use of duties. It will then expound on the duty to minimise hallucinations and the duty to disclose the existence of hallucinations and conclude.

## 4.2. Developers' responsibility to minimise hallucinations in legal LLMs

### 4.2.1. Legal LLMs and access to justice

Access to justice is not only the ability to get legal representation but also the ability to obtain and understand legal information.[193] Users of legal services themselves are mostly unaware of the intricacies of the law, making them vulnerable due to an inherent information asymmetry.[194] This asymmetry is not unique to the law, but can be found in other knowledge intensive professions such as medicine.

This asymmetry sits at the core of the access to justice problem because the law assumes that it is understood by the wider public. This assumption is evident in adversarial systems of law where litigants usually have to argue their case in court with or without representation, with the exception of a certain class of cases in some countries.[195] Additionally, ignorance not

[193] Conklin W, 'Access to Justice as Access to a Lawyer's Language.'10 *Windsor Yearbook of Access to Justice* 1, 1990, 457.
[194] Conklin W, 'Access to Justice as Access to a Lawyer's Language.'10 *Windsor Yearbook of Access to Justice* 1, 1990, 457.
[195] In the United Kingdom, legal aid is provided pro bono, without the litigant having to show proof of need for criminal cases. See United Nations Office on Drugs and Crime, *Global Study on Legal Aid*, 2016, 500- 506.

being a defence before the law reinforces this assumption.[196] This means that users of legal services are particularly vulnerable because the language of the law itself is inaccessible necessitating a mediator.

The practice of law as a profession exists to fill this gap, with lawyers being the providers of access to justice by virtue of the legal knowledge they wield.[197] Until recently, lawyers and other professionals trained in the law were the only ones who were able to bridge the gap between the law and the people.[198] This is because they had a monopoly on knowledge of the law, since the language of the law itself was inaccessible. The average lay person may find it difficult to read and correctly interpret the law themselves in most jurisdictions as the law is verbose and jargon heavy.

This means that access to justice efforts have concentrated on helping people understand the law itself or providing pro bono legal services. All these examples include people who have received or have currently received a training in the law donating their time, money and effort. Moreover, the nature of the law itself means that these efforts will almost always be a temporary solution to a deeper problem that goes to the heart of the legal profession itself.

With the development of legal LLMs comes a new mediator between the lay person and the law. LLMs fill a crucial gap in access to justice that the legal system itself has failed to fill by making both legal language, and legal information accessible.[199] Due to the fact that LLMs are cheap, easy and quick to use, they may finally deliver on the promise of just, expedient and affordable justice.[200] Seeing as they are more cost effective than lawyers they are likely to be an alternative for people seeking access to justice.[201]

---

[196] See Matthews P, 'Ignorance of the Law is No Excuse' 3 *Legal Studies* 2, 1982, 190-192; Narasimham R, Narasimhan R, 'Ignorantia juris non excusat: Ignorance of law is no excuse' 13 *Journal of the Indian Law Institute* 1, 1971, 73-78.
[197] Yallow S, 'Paths to justice: What people do and think about going to law', 4 *Legal Ethics* 149, 2001, 150.
[198] Maldonado D, 'The Right to Access to Justice: Its Conceptual Architecture', 1 *Indiana Journal of Global Legal Studies* 27, 2020, 16.
[199]United States Supreme Court, *2023 Year-End Report on the Federal Judiciary*, 31 December 2023, 5.
[200] Dahl M, Magesh V, Zuzgun M, Ho D, 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models', 16 Journal *of Legal Analysis* 1, 2024, 65.
[201] Tan J, Westermann H, Benyekhlef K, 'ChatGPT as an Artificial Lawyer?' ICAIL 2023 Workshop on Artificial Intelligence for Access to Justice, Braga, 19 June 2023, 2-7.

### 4.2.2. Harms that could be caused by legal LLMs

Since hallucinations by definition are factually inaccurate statements generated by LLMs,[202] the major harm occasioned by hallucinations is misinformation.[203] While LLMs bridge the access to justice gap due to their cost friendliness and ease of use,[204] hallucinations cause them to frequently generate misinformation.[205] This has led some scholars to call for the limitation of the kind of legal advice LLMs can offer.[206] The harm is already material with hallucinated case law being cited in the US. In the case of Roberto Mata the lawyer cited hallucinated case law, presumably being oblivious to what hallucinations are.[207] Such instances have made Dahl et al argue that LLMs should not only improve legal knowledge but also legal reasoning for them to be efficient as a means to accessing justice.[208] Curran et al further warn that LLMs pose an even higher risk in common law jurisdictions because they may over summarise cases or take them out of context, leading to a fundamental misconstruction of the law.[209]

With LLMs as legal service providers, the questions of honesty and truthfulness as ethical responsibilities arise. The harm posed by hallucinations to people who may have no training in the law may be greater than what is evident through the Roberto Mata case. Lay users of legal services are especially vulnerable due to an inherent information asymmetry. Human lawyers are held to high standards when providing legal advice to their clients. This is often overseen by bar associations which police the conduct of the practicing lawyers. Similarly,

[202] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu D, Ishii E, Bang Y, Madotto A, and Fung P, 'Survey of Hallucination in Natural Language Generation,' 55 *ACM Computing Survey* 12, 2023, 3.
[203] Chen C, Shu K, 'Combatting Misinformation in the Age of LLMs: Opportunities and Challenges' 45 *AI Magazine* 3, 2024, 355.
[204] Chien C, Kim M, Raj A and Rathis R, 'How Generative AI Can Help Address the Access to Justice Gap through the Courts,' Loyola of Los Angeles Law Review, Forthcoming, 1-3. 2024. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4683309# on 17 February 2025.
[205] See Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo, 'Gpt-4 passes the bar exam' 382 *Philosophical transactions of the Royal Society A*, 2270, 2023, 11; Dahl M, Magesh V, Zuzgun M and Ho D, 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models', 65; Curran S, Bethell O and Lansley S, 'Hallucination is the last thing you need', arXiv, 2023, 2; Cheong I, Xia K, Feng K, Chen Q, and Zhang A. 2462-2463; Kapoor S, Henderson P, Narayanan A, 'Promises and pitfalls of artificial intelligence for legal applications' 3-10; Tan J, Westermann H and Benyekhlef K, 'ChatGPT as an Artificial Lawyer?,' 2.
[206] Curran S, Bethell O and Lansley S. 'Hallucination is the last thing you need', arXiv, 2023, 2. See also; Cheong I, Xia K, Feng K, Chen Q, and Zhang A. '(A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice,' 2462-2463; Kapoor S, Henderson P, Narayanan A, 'Promises and pitfalls of artificial intelligence for legal applications' 3-10.
[207] Roberto Mata v Avianca Incorporated (2023), Southern District Court of New York, United States.
[208] Dahl M, Magesh V, Zuzgun M and Ho D, 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models', 65.
[209] Curran S and Bethell O and Lansley S, 'Hallucination is the last thing you need', arXiv, 2023, 2.

lawyers who give bad advice can be sued for professional negligence.[210] Unfortunately, this is not a remedy available for those who rely on LLMs, necessitating the development of a framework that will provide a mechanism of recourse for those that have been aggrieved by hallucinations.

### 4.2.3. Honesty versus truthfulness

Truthfulness and honesty are distinguished in the context of AI governance. An honest model is one that makes statements it believes to be true while a truthful model is one that makes statements that are objectively true, its belief system notwithstanding.[211] Truthfulness is important in the context of legal LLMs because sometimes the training data can be erroneous as seen in section 3.2.1 A.

While a completely truthful legal LLM is desirable, this study shies away from adopting that as the standard, acknowledging that the possibility might not exist especially in common law jurisdictions. This is not to say that the LLMs should be inaccurate. It simply means that the legal profession itself acknowledges that the position of the law is not always clear, especially in common law jurisdictions.

Consider the format of a contentious case when the law itself is silent on a pressing matter. Both the applicant and the defendant will make a case for a specific interpretation of the law using already existing law and try to convince the judge of the same. If the judge sides with either party, or takes a completely different approach, it does not mean that the judge is untruthful, neither does it mean that the lawyers' positions were false. It is therefore evident that even the legal system itself acknowledges that an interpretation of the law can sometimes be a matter of opinion, not fact. Therefore, it might be a tall order to expect an LLM to give a truly truthful opinion especially on contentious or emerging legal questions because the truthfulness of a legal opinion is relative. It is based on the opinion of the judge which might change with time.

Since honesty and truthfulness are distinguishable, complete honesty can exist outside of truthfulness. Honesty requires an accurate stating of the law as it is. It requires an admission of ignorance. That is, the obligation to say 'I don't know' when one does not know. Section

[210] Leubsdorf J, 'Legal Malpractice and Professional Responsibility,' 48 *Rutgers Law Review* 1, 1995, 101-123.
[211] Evans O, Cotton-Barratt O, Finnveden L, Bales A, Balwit A, Wills P, Righetti L, Saunders W, 'Truthful AI: Developing and governing AI that does not lie,' arXiv, 2021, 16-17.

3.2.1 discusses how LLMs are trained to hallucinate when there is a knowledge gap. It also discusses the tendency of LLMs to give an answer that sides with the user even when it knows that the answer is false. This is an example of dishonesty. An honest LLM may generate misinformation and still be considered honest if the misinformation was present in its training data which necessitates the importance of truthfulness. However, as seen in section 3.4, complete truthfulness may be hard to achieve because hallucinations are inevitable at the current stage of LLM development.

However, hallucinations are a barrier in realising proper access to justice. As seen in section 3.3.2 C, hallucinations are almost inevitable in the current stage of LLM development. While 100% accuracy is difficult to achieve, truthfulness is still desirable and so is honesty. This dissertation advocates for a quasi-honesty standard because both are important in the context of legal LLMs but truthfulness is hard to achieve in both a technical and legal sense as demonstrated above.

### 4.2.4. Developers' ability to detect and minimise hallucinations gives rise to a responsibility

As discussed in the preceding sections, LLMs increase access to justice but their effectiveness is limited by the harms caused by hallucinations. Though hallucinations are inevitable at the current stage of LLM development,[212] technical methods to detect and minimise hallucinations exist as highlighted in section 3.3. The existence of these technical methods means that developers of legal LLMs have the ability to increase the effectiveness of LLMs in providing access to justice. A responsibility to minimise hallucinations could be inferred from their ability to minimise them.

Even though taking measures to minimise hallucinations could hinder their effectiveness, the trade-off is justified because of the vulnerability of the users of legal LLMs. Considering that LLMs are cheaper than lawyers, explain the legal rules and procedures in ways that can be understood by a lay person, and are soon going to be more efficient at natural language

[212] Kalai A, Vempala S, 'Calibrated Language Models Must Hallucinate,' STOC '24: 56th Annual ACM Symposium on Theory of Computing, Vancouver, 25 June 2024, 160-168; Banerjee S, Agarwal A, Singla S, 'LLMs will always hallucinate, and we need to live with this', arXiv, 2024, 1-28; Li J, Consul S, Zhou E, Wong J, Farooqui N, Ye Y, Manohar N, Wei Z, Echols B, Zhou S, Diamos G, 'Banishing LLM hallucinations requires rethinking generalisation', arXiv, 2024, 1-10.

processing tasks than most lawyers,[213] it is plausible that LLMs are soon going to be the preferable alternative for majority of the population.

The inherent information asymmetry between users of legal services and the providers of legal services will persist even as the providers of legal services become LLMs instead of human lawyers. When users of legal services, who have no knowledge of the law, consult LLMs on matters of the law due to their accessibility, LLMs become the bearers of legal information. Seeing as the language of the law itself is inaccessible, it is difficult for the users of legal LLMs to fact check the LLM themselves because they may not understand the primary sources of the law.

While developers of legal LLMs may not understand the law as well, benchmarks exist which can help them know when their models are inaccurate and remedy the inaccuracies. They can also ensure that the output generated by their models is an accurate representation of what the models actually know, and even equip their models with the ability to fact check the output they generate to promote accuracy. Developers therefore have the responsibility to minimise hallucinations as much as they can.

## 4.3. Codification of the responsibility

### 4.3.1. Existing proposals

Scholars have discussed liability for LLM output. A majority seem to agree that accuracy in public facing LLMs is desirable even though they disagree on the legal vehicle that should

[213] See Cheong I, Xia K, Feng K, Chen Q, and Zhang A. '(A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice,' Association for Computing Machinery Conference on Fairness, Accountability, and Transparency (FAccT '24), Rio de Janeiro, 3 June 2024, 2454; Martin L, Whitehouse N, Yiu S, Catterson L and Rivindu P, 'Better call GPT: Comparing large language models against lawyers', arXiv, 2024, 9-11; 2023 Year-End Report on the Federal Judiciary, 6. 2023year-endreport.pdf on 18 March 2025.

be used to ensure that this outcome prevails,[214] or even if truthfulness should be mandated in the first place.[215]

### A. The first amendment

Several authors have argued that the output generated from LLMs qualifies to be regarded as protected speech under the First Amendment to the US Constitution.[216] Lamo and Calo make the argument that the distance between the bot's creator and their bot's speech alone should not put the bot's output outside the scope of protection under the first amendment.[217] They argue that such speech should be regarded to be the creator's speech.[218] Volokh, Lemley and Henderson agree, arguing that AI output is likely to be interpreted as their creator's speech under the existing law.[219]

However, Salib disagrees with the notion that AI outputs should be protected speech.[220] He argues that AI outputs does not fall into the category of automatically protected speech under the first amendment because it does not count as communication. Salib goes on to argue that even if AI output were to be categorised as protected speech, it should not be awarded the high level of legal immunity that is accorded to speech under the first amendment as it would make the regulation of AI difficult.[221] While scholars like Salib oppose the classification of

[214] Some scholars argue for the application of first amendment rights while others prefer different kinds of duties. See Bambauer J, 'Negligent AI Speech: Some Thoughts About Duty' 3 *Journal of Free Speech Law* 1, 2023, 343-362; Lamo M and Calo R, 'Regulating Bot Speech' 66 *UCLA Law Review* 1, 2019, 988, 1005; Massaro T and Norton H, 'Siri-ously? Free Speech Rights and Artificial Intelligence' 110 *Northwestern University Law Review* 1, 2016, 1169,1176; Volokh E, Lemley M and Henderson P, 'Freedom of Speech and AI output' 3 *Journal of Free Speech Law* 1, 2023, 651, 653; Fisher A, 'Something AI should tell you – The case for labelling synthetic content',41 *Journal of Applied Philosophy* 4, 2024, 1-15; Watcher S, Mittelstadt B and Russel C, 'Do large language models have a legal duty to tell the truth?'11 *Royal Society Open Science* 1, 2024, 34-49; Paseri L, Durante M, 'Examining epistemological challenges of large language models in law' 1 *Cambridge forum on AI: Law and governance* 1, 2025, 9-11.

[215] Arcila B, 'Is It a Platform? Is It a Search Engine? It's ChatGPT! The European Liability Regime for Large Language Models', 3 *Journal of Free Speech Law* 1, 2023, 479-488.

[216] See Lamo M and Calo R, 'Regulating Bot Speech' 66 *UCLA Law Review* 1, 2019, 988, 1005; Massaro T and Norton H, 'Siri-ously? Free Speech Rights and Artificial Intelligence' 110 *Northwestern University Law Review* 1, 2016, 1169,1176; Volokh E, Lemley M, Henderson P, 'Freedom of Speech and AI output' 3 *Journal of Free Speech Law* 1, 2023, 651, 653.

[217] Lamo M and Calo R, 'Regulating Bot Speech' 66 *UCLA Law Review* 1, 2019, 988.

[218] Lamo M and Calo R, 'Regulating Bot Speech' 66 *UCLA Law Review* 1, 2019, 1005.

[219] Volokh E, Lemley M, Henderson P, 'Freedom of Speech and AI output' 3 *Journal of Free Speech Law* 1, 2023, 651, 653.

[220] Salib P, 'AI outputs are not protected speech' 102 *Washington University Law Review* 1, 2024, 83-154.

[221] Salib P, 'AI outputs are not protected speech' 102 *Washington University Law Review* 1, 2024, 85-87.

AI output as protected speech, developers in jurisdictions that consider AI output to be protected speech may be exempted from taking measures to minimise misinformation.

### B. Duty and liability

Scholars have also explored various duty and liability regimes to regulate misinformation and disinformation that stems from LLMs. Volokh argues that hallucinations may be regulated through the tort of libel.[222] However, he concedes that it may be difficult to determine what exactly counts as actionable design. Bambauer also argues that imputing duties on the part of the developer can and should be used to govern careless speech from AI.[223] On the other hand, Acrila argues that developers should be protected from excessive liability, arguing that only broad duties should be imposed on them to incentivise innovation.[224] She proposes the extension of the search engine duties under the European Digital Services Act only to general purpose LLMs such as ChatGPT and Bard.

Watcher, Mittelstadt and Russel propose the imposition of a duty to tell the truth on developers of LLMs.[225] They argue that because LLMs are careless speakers by design, and the users of this LLMs are subtly encouraged but are also susceptible to believing that LLMs are telling the truth then it is plausible that a duty to tell the truth is imposed on developers. Watcher et al propose a broad, far reaching duty with reporting mechanisms to compel developers to come up with truthful LLMs. However, Paseri and Durante disagree with Watcher on the necessity of a broad, sweeping duty, highlighting that it may not yield desirable outcomes since it insinuates that someone has a monopoly on truth.[226] Nonetheless, they concede that a duty to tell the truth may be necessary in specific circumstances such as in the field of research and academia.

[222] Volokh E, 'Large Libel Models? Liability for AI output' 3 *Journal of Free Speech Law* 2, 2023, 489-558.
[223] Bambauer J, 'Negligent Al Speech: Some Thoughts About Duty' 3 *Journal of Free Speech Law* 1, 2023, 343-362.
[224] Acrila B, 'Is It a Platform? Is It a Search Engine? It's ChatGPT! The European Liability Regime for Large Language Models', 3 *Journal of Free Speech Law* 1, 2023, 479-488.
[225] Watcher S, Mittelstadt B, Russel C, 'Do large language models have a legal duty to tell the truth?'11 Royal Society Open Science 8, 2024 34-49.
[226] Paseri L, Durante M, 'Examining epistemological challenges of large language models in law' 1 Cambridge forum on Ai: Law and governance' 1, 2025, 9-11.

### 4.3.2. Making a case for duties

Duties are preferable in this context because they often give rise to rights. This means that duties are not only enforced by regulatory agencies but also by right holders, strengthening enforcement. Duties are also obligations by definition. Considering the vulnerability of the user of legal LLMs and legal services in general, duties are the appropriate legal vehicle to translate the responsibility of developers. This is because developers themselves will have the obligation to not only consider the efficacy of their models, but also to take into account the vulnerability of their users and take steps to promote their protection.

However, the ability of developers to take measures that can minimise hallucination is constrained by their capacity therefore different duties can be imposed on different categories of developers depending on their capacity. Capacity in this context includes market power, market share, compute, revenue, as well as public interest and cost effectiveness. These factors go hand in hand. It is assumed, but is not always the case, that the higher a firm's market share, the more customers it has, meaning that it enables them to have more revenue, which puts them in a position to acquire more compute power. However, these terms are not all interchangeable. Anti-trust law in some jurisdictions recognises that a firm can have market power without being dominant.[227]

The responsibility of developers should therefore be codified as a legal duty that is adjusted according to a developer's/class of developer's capacity. This dissertation suggests that a bare minimum duty to disclose should be imposed on all developers, their capacity notwithstanding, while the duty to minimise hallucinations should be dependent on capacity.

#### A. Duty to disclose

All developers of legal LLMs, regardless of their capacity, should have a duty to disclose the existence of hallucinations. Developers themselves are aware that their LLMs hallucinate. As discussed in 3.4.2 B, hallucinations are almost inevitable at the current stage of LLM development because of the way in which the models are trained. By conveniently

[227] Këllezi P, 'Abuse below the Threshold of Dominance? Market Power, Market Dominance, and Abuse of Economic Dependence,'in Mackenrodt M, Gallego B, Enchelmaier S (eds) *Abuse of Dominant Position: New Interpretation, New Enforcement Mechanisms? MPI Studies on Intellectual Property, Competition and Tax Law Volume 5*, Springer, Berlin, 2008. See also Babelegi Workwear and Industrial Supplies CC v Competition Commission of South Africa [2020], Court of Appeal of South Africa.

choosing to conceal the existence of hallucinations, which could harm already vulnerable users, these developers could be said to be willingly spreading misinformation, further compounding on injustice. Developers of legal LLMs should, at minimum, inform users that the information they receive from the LLM might be inaccurate and encourage them to counter check the information and seek advice from a lawyer if they can.

### B. Duty to minimise hallucinations

Methods to minimise hallucinations already exist as seen in Section 3.4.2, However, these methods are an additional cost that small scale developers may not afford to bear, which may discourage the democratisation of information through finetuning on local laws or even creating specialised legal LLMs, resulting in an uneven trade off. Developers that have the capacity should have a duty to take all applicable measures to minimise the occurrence of hallucinations.

This study proposes that capacity, and the exact measures to be taken are stipulated by the relevant national authorities in each jurisdiction Due to jurisdiction-specific issues such as median wealth, the accessibility of the law, colonial legacy among other factors, the reality of access to justice, and the use of LLMs as tools that promote access to justice vary. National authorities should decide the most appropriate minimum measures that should be taken by specific classes of developers, depending on the most appropriate methods of minimising hallucinations at a particular point in time.

National authorities should consider economic factors such as market power, market share, compute, revenue, as well as public interest and cost effectiveness. Economic factors are important because the more economic power a firm has, the more resources it has to ensure that its models are accurate. Cost effectiveness and public interest considerations should also be given weight by national authorities because these two factors consider the interests of the developers as well as those of the general public which includes users of legal LLMs. For example, in the UK, companies whose annual revenue is whose annual revenue is £50 million or more (medium to large scale developers) should at the very least be expected to take minimum cost-effective measures to minimise hallucinations as the national authorities deem fit.

## 4.4. Conclusion

This chapter has highlighted the role LLMs play in promoting access to justice, and the harms posed by hallucinations that could undermine access to justice. It then went on to explain that developers of legal LLMs have a responsibility to minimise the existence of hallucinations, or at the very least disclose their existence and proposed two duties to this effect. The next chapter seeks to conclude the dissertation and offer recommendations to various stakeholders.

## 5.0. Conclusion and recommendations

The preceding chapters have explained what LLMs are, why they are useful in promoting access to justice and why hallucinations may undermine that goal. The previous chapter made a case for the implementation of duties to disclose and to minimise hallucinations on the part of developers. It made a case for these duties by considering to the vulnerability of the users of legal services as well as the role of LLMs in promoting access to justice. This chapter will conclude the dissertation and give recommendations to various stakeholders.

### 5.1. Conclusion

This study set out to determine whether duties to disclose and minimise hallucinations should be imposed on developers of legal LLMs. It began by defining what an LLM is, explaining how it works and what its utility is especially in the context of the law. The study found that LLMs actually increase access to justice. This dissertation then went on to introduce hallucinations, explain what they are and demonstrate the harm they pose especially with respect to the role LLMs play with respect to access to justice. It went on to outline the various methods developers could use to detect and minimise hallucinations.

In chapter three and chapter four, this dissertation discussed the harms of this kind of misinformation, whether intentional or unintentional. Chapter four made a case for duties to minimise and to disclose the existence of hallucinations. It did so by highlighting the fact that the users of legal services are particularly vulnerable due to the language of the law and legal representation being particularly inaccessible. The study found that legal LLMs can increase access to justice because they are easy to use and they democratise access to legal information. It then found that the developers of legal LLMs have a responsibility to minimise the existence of hallucinations depending on their capacity, and at the very least they should disclose their existence capacity notwithstanding. Chapter four then made a case for these responsibilities to be legally codified through duties that are imposed on developers.

## 5.2. Recommendations

### 5.2.1. Implementation

The duties as proposed should be enacted in legislation to give them legal force. This legislation could be at a national or regional level and should be guided by technical authorities in the respective jurisdiction. These technical authorities can include standard setting bodies or ICT authorities

### 5.2.2. Enforcement

Government enforcement agencies such as consumer protection commissions and the judiciary should oversee the implementation of these duties. They should also provide guidelines on the hallucination detection and minimisation techniques that should be used by developers.

### 5.2.3. Awareness campaigns

Governments and developers alike should educate the general public on the utility of legal LLMs as well as the fact that hallucinations exist to offer the users a balanced view on the benefits and harms of legal LLMs.

### 5.2.4. Areas of further research

Researchers looking to make contributions in this field can do so by assisting national standard setting authorities to determine the kind of metric they should use to prescribe technical measures and even the specific technical measures they should prescribe.

# Bibliography

## Books


1. Goodfellow I, Bengio Y and Courville A, *Deep learning*, Cambridge, Massachusetts, MIT Press, 2016.
2. Kelleher J, *Deep learning*, MIT Press, Cambridge, Massachusetts, 2019.
3. Sutton R and Barto A, *Reinforcement learning: An introduction*, 2 ed, MIT Press, Cambridge, Massachusetts, 2018.


## Chapters in Books


1. Rawte V, Chakraborty S, Pathak A, Sarkar A, Tonmoy A, Chadha A, Sheth A, Das A, 'The troubling emergence of hallucination in LLMs, an extensive definition, quantification and prescriptive remediations', in Bouamor H, Pino J and Bali K (eds) *Proceedings of the 2023 conference on empirical methods in natural language processing, association for computational linguistics,* Singapore, 2023, 2543-2563.
2. Karhunen J, Raiko, T and Cho K, 'Unsupervised deep learning: A short review' in Bingham E, Kaski S, Laaksonen J and Lampinen J (eds), *Advances in independent component analysis and learning machines*, Academic Press, Cambridge, Massachusetts, 2015, 125–142.
3. Zhang H, Diao S, Lin Y, Fung Y R, Lian Q, Wang X, Chen Y, Ji H, and Zhang T, 'R-Tuning: Instructing large language models to say 'I don't know', in Duh K, Gomez H and Bethard S (eds), *Proceedings of the 2024 conference of the North American chapter of the association for computational linguistics: Human language technologies (Volume 1: Long Papers),* Association for Computational Linguistics, Mexico City, 2024, 7113-7121.
4. Zhang H, Diao S, Lin Y, Fung Y R, Lian Q, Wang X, Chen Y, Ji H, and Zhang T, 'R-Tuning: Instructing Large Language Models to Say 'I Don't Know', in Duh K , Gomez H and Bethard S (eds), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers),* Association for Computational Linguistics, Mexico City, 2024, 7113-7139.
5. Këllezi P, 'Abuse below the threshold of dominance? Market power, market dominance, and abuse of economic dependence,' in Mackenrodt M, Gallego B, Enchelmaier S (eds) *Abuse of dominant position: New interpretation, new enforcement mechanisms? MPI*

*studies on intellectual property, competition and tax Law Volume 5*, Springer, Berlin, 2008.

6. Salib P, 'AI outputs are not protected speech' 102 *Washington University Law Review* 1, 2024.
7. Arcila B, 'Is it a platform? Is it a search engine? It's ChatGPT! The European liability regime for large language models', 3 *Journal of Free Speech Law* 1, 2023.
8. Paseri L, Durante M, 'Examining epistemological challenges of large language models in law' 1 *Cambridge forum on AI: Law and governance* 1, 2025.
9. Watcher S, Mittelstadt B and Russel C, 'Do large language models have a legal duty to tell the truth?'11 *Royal Society Open Science* 1, 2024.
10. Fisher A, 'Something AI should tell you – The case for labelling synthetic content', 41 *Journal of Applied Philosophy* 4, 2024.
11. Volokh E, Lemley M and Henderson P, 'Freedom of Speech and AI output' 3 *Journal of Free Speech Law* 1, 2023.
12. Massaro T and Norton H, 'Siri-ously? Free speech rights and artificial intelligence' 110 *Northwestern University Law Review* 1, 2016.
13. Lamo M and Calo R, 'Regulating Bot Speech' 66 *UCLA Law Review* 1, 2019.
14. Bambauer J, 'Negligent Al speech: Some thoughts about duty' 3 *Journal of Free Speech Law* 1, 2023.
15. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D, 'Language models are few-shot learners', in Larochelle H, Ranzato M, Hadsell R, Balcan M and Lin H (eds) in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, Curran Associates Inc, New York, 2020.

## Journal Articles

1. Amin K, Doshi R and Forman H, 'Large language models as a source of health information: Are they patient-centred? A longitudinal analysis' 12 Science Direct 1, 2024.
2. Chen C, Shu K, 'Combatting misinformation in the age of LLMs: Opportunities and challenges' 45 *AI Magazine* 3, 2024.

3. Chien C, Kim M, Raj A and Rathis R, 'How generative AI can help address the access to justice gap through the courts,' Loyola of Los Angeles Law Review, Forthcoming, 1-63. 2024.
4. Conklin W, 'Access to justice as access to a lawyer's language.' 10 *Windsor Yearbook of Access to Justice* 1, 1990.
5. Dahl M, Magesh V, Zuzgun M and Ho D, 'Large legal fictions: Profiling legal hallucinations in large language models', 16 *Journal of Legal Analysis* 1, 2024.
6. Goriely T, 'Law for the poor: The relationship between advice agencies and solicitors in the development of poverty law', 1 *International Journal of the Legal Profession* 2, 1996.
7. Islam M, Chen G and Jin S, 'An overview of neural networks' 5 *American Journal of Neural Networks and Applications* 1, 2019.
8. Janiesch C, Zschech P and Heinrich K, 'Machine learning and deep learning fundamentals', 31 *Electronic Markets Journal* 1, 2021, 2-3.
9. Kapoor S, Henderson P, Narayanan A, 'Promises and pitfalls of artificial intelligence for legal applications' 2 *Journal of cross disciplinary research in computational law* 2, 2024.
10. Lin T, Wang Y, Liu X and Qiu X, 'A survey of transformers', 3 *AI Open* 1, 2022, 127-128.
11. Loh P and Wainwright M, 'High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity', 40 *The Annals of Statistics* 3, 2012.
12. Maldonado D, 'The right to access to justice: Its conceptual architecture', 1 *Indiana Journal of Global Legal Studies* 27, 2020.
13. Matthews P, 'Ignorance of the law is no excuse' 3 *Legal Studies* 2, 1982.
14. Narasimham R, 'Ignorantia juris non excusat: Ignorance of law is no excuse' 13 *Journal of the Indian Law Institute* 1, 1971, 73-78.
15. Sarat A, 'Access to justice', 8 *Harvard Law Review* 94, 1981.
16. Sarker I, 'Machine learning: Algorithms, real-world applications and research directions', 2 *Springer Nature Computer Science* 1, 2021.
17. Schmidhuber, J, 'Deep learning in neural networks: An overview' 61 *Neural Networks* 1, 2015, 85–117.
18. Taye M, 'Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions', 12 *Computers* 5, 2023.
19. Yallow S, 'Paths to justice: What people do and think about going to law', 4 *Legal Ethics* 149, 2001.

20. Yosinski J, Clune J, Bengio Y and Lipson H, 'How transferable are features in deep neural networks?' 4 *Advances in Neural Information Processing Systems* 1, *2014*, 3320–3328.
21. Osisanwo F, Akinsola J, Awodele O, Hinmikaiye J, Olakanmi O and Akinjobi J, 'Supervised machine learning algorithms: Classification and comparison', 48 *International Journal on Computing Trends and Technology* 1, 2017.
22. Nasteski V, 'An overview of the supervised machine learning methods', 4 *Horizons* 1, 2017.
23. Sarker I, 'Machine Learning: Algorithms, Real-World Applications and Research Directions', 2 *Springer Nature Computer Science* 1, 2021, 4.
24. Mohammadi M, Al-Fuqaha A, Guizani M and Oh J, 'Semi-supervised Deep Reinforcement Learning in Support of IoT and Smart City Services', 5 *IEEE Internet of Things Journal* 2, 2017.
25. Thirunavukarasu A, Ting D, Elangovan K, 'Large language models in medicine', 29 *Nature medicine* 1, 2023.
26. Lai J, Gan W, Wu J, Qi Z, Yu P, 'Large language models in law: A survey,' 5 AI Open 1, 2024.
27. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Barham P, Chung H W, Sutton C, Gehrmann S, Schuh P, Shi K, Tsvyashchenko S, Maynez J, Rao A, Barnes P, Tay Y, Shazeer N, Prabhakaran V, Reif E, Du N, Hutchinson B, Pope R, Bradbury J, Austin J, Isard M, Gur-Ari G, Yin P, Duke T, Levskaya A, Ghemawat S, Dev S, Michalewski H, Garcia X, Misra V, Robinson K, Fedus L, Zhou D, Ippolito D, Luan D, Lim H, Zoph B, Spiridonov A, Sepassi R, Dohan D, Agrawal S, Omernick M, Dai A M, Pillai T S, Pellat M, Lewkowycz A, Moreira E, Child R, Polozov O, Lee K, Zhou Z, Wang X, Saeta B, Diaz M, Firat O, Catasta M, Wei J, Meier-Hellstern K, Eck D, Dean J, Petrov S, and Fiedel N, 'PaLM: Scaling language modelling with pathways,' 24 *Journal of Machine Learning Research* 1, 2023.
28. Paullada A, Raji I D, Bender E M, Denton E, and Hanna A, 'Data and its (dis)contents: A survey of dataset development and use in machine learning research,' 2 *Patterns* 11, 2021.
29. Li Y, Li Z, Zhang K, Dan R, and Zhang Y, 'ChatDoctor: A medical chat model fine-tuned on Llama model using medical domain knowledge' 15 *Cureus* 6, 2023.
30. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu D, Ishii E, Bang Y, Madotto A, and Fung P, 'Survey of hallucination in natural language generation' 55 *ACM Computing Survey* 12, 2023.
31. Huang L, Yu W, Ma W, Zhong W, Chen Q, Peng W, Feng X, Qin B, Liu T, 'A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions', 43 *Association for Computing Machinery Transactions on Information Systems* 2, 2024.

32. Adlakha V, Behnam Ghader P, Lu XH, Meade N and Reddy S, ‘Evaluating correctness and faithfulness of instruction-following models for question answering,’ 12 *Transactions of the Association for Computational Linguistics* 1, 2023.

33. Volokh E, ‘Large Libel Models? Liability for AI output’ 3 *Journal of Free Speech Law* 2, 2023.

34. Leubsdorf J, 'Legal malpractice and professional responsibility,' 48 *Rutgers Law Review* 1, 1995.

35. See Katz D, Bommarito M, Gao S, and Arredondo P, ‘GPT-4 passes the bar exam’ 382 *Philosophical transactions of the Royal Society A,* 2270, 2023.

36. Zhang J, Zhao Y, Li H and Zong C, ‘Attention with sparsity regularization for neural machine translation and summarization’ 27 *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 3, 2018.

37. Wang S, Zhu Y, Liu H, Zheng Z, Chen C, and Li J, ‘Knowledge Editing for Large Language Models: A Survey,’ 57 *Association for Computing Machinery Surveys* 3, 2024.

## Conference Papers

1. Cheong I, Xia K, Feng K, Chen Q, and Zhang A. '(A)I am not a lawyer, but...: Engaging legal experts towards responsible LLM policies for legal advice,' Association for Computing Machinery Conference on Fairness, Accountability, and Transparency (FAccT '24), Rio de Janeiro, 3 June 2024.
2. Luo J, Xiao C and Ma F, ‘Zero-resource hallucination prevention for large language models,’ in Al-Onaizan Y, Bansal M and Chen Y (eds), *Findings of the Association for Computational Linguistics: EMNLP 2024,* Association for Computational Linguistics, Miami, 2024.
3. Hazan E and Koren T, 'Linear regression with limited observation' 29th International Conference on Machine learning, Edinburgh, 26 June 2012.
4. Qu K, 'Research on linear regression algorithms,' 2nd International Conference on Mathematical Physics and Computational Simulation, Glasgow, 9 August 2024.
5. Tan J, Westermann H and Benyekhlef K, 'ChatGPT as an artificial lawyer?' International Conference on Artificial Intelligence and Law 2023 Workshop on Artificial Intelligence for Access to Justice (AI4AJ), Braga, 19 June 2023.

6. Viswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez N, Kaiser L, Polosukhin I, 'Attention is all you need', Thirtieth Annual Conference on Neural Information Processing Systems, Long Beach California, December 8-9, 2017.
7. Westermann H and Benyekhlef K, 'JusticeBot: A methodology for building augmented intelligence tools for laypeople to increase access to justice' Nineteenth International Conference on Artificial Intelligence and Law, Braga, 19th June 2023.
8. Ngo R, Chan L, Mindermann S, 'The alignment problem from a deep learning perspective', International Conference on Learning Representations, Vienna, 7 May 2024.
9. Skalse J, Howe N, Krasheninnikov D, and Krueger D, 'Defining and characterizing reward gaming' 36th Conference on Neural Information Processing Systems, New Orleans, 28 November 2022.
10. Agarwal A, Henaff M, Kakade S, and Sun W, 'PC-PG: Policy cover directed exploration for provable policy gradient learning,' 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, 16 July 2020.
11. Yang Y, Chern E, Qiu X, Neubig G, and Liu P, 'Alignment for honesty,' Thirty-Eight Annual Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, 12 December 2024.
12. Xu W, Sun H, Deng C and Tan Y, 'Variational Autoencoder for Semi-Supervised Text Classification,' The Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, 5 February 2017.
13. Finn C, Yu T, Fu J, Abbeel P and Levine S, 'Generalizing skills with semi-supervised reinforcement learning', 5th International Conference on Learning Representations, Toulon, 24 April 2017.
14. Perez E, Ringer S, Lukošiūtė K, Nguyen K, Chen E, Heiner S, Pettit C, Olsson C, Kundu S, Kadavath S, Jones A, Chen A, Mann B, Israel B, Seethor B, McKinnon C, Olah C, Yan D, Amodei D, Amodei D, Drain D, Li D, Tran-Johnson E, Khundadze G, Kernion J, Landis J, Kerr J, Mueller J, Hyun J, Landau J, Ndousse K, Goldberg L, Lovitt L, Lucas M, Sellitto M, Zhang M, Kingsland N, Elhage N, Joseph N, Mercado N, DasSarma N, Rausch O, Larson R, McCandlish S, Johnston S, Kravec S, El Showk S, Lanham T, Telleen-Lawton T, Brown T, Henighan T, Hume T, Bai Y, Hatfield-Dodds Z, Clark J, Bowman S R, Askell A, Grosse R, Hernandez D, Ganguli D, Hubinger E, Schiefer N, and Kaplan J, 'Discovering language model behaviours with model-written evaluations,'

61st Annual Meeting of the Association for Computational Linguistics, ACL 2023, Toronto, 9 July 2023.

15. Lai W, Mesgar M, Fraser A, 'LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback', in Ku L, Martins A and Srikumar V, *Findings of the Association for Computational Linguistics: ACL 2024*, Association for Computational Linguistics, Bangkok, 2024.
16. M'Rhar K, Ben Jaafar C, Bencharef O and Bourkoukou O, 'Unlocking the potential of large language models in legal discourse: Challenges, solutions, and future directions,' 2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS), Marrakech, 23 October 2024.
17. Bender E, Gebru T, McMillan-Major A and Shmitchell S, 'On the dangers of stochastic parrots: Can language models be too big?' in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, Association for Computing Machinery, 2021.
18. Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, Roberts A, Brown T, Song D, Erlingsson U, Oprea A, and Raffel C, 'Extracting training data from large language models,' in 30th USENIX Security Symposium (Virtual), 11 August 2021.
19. Lin S, Hilton J, and Evans O, 'TruthfulQA: Measuring how models mimic human falsehoods' in Muresan S, Nakov P and Villavicencio A, *Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, 2022.
20. Kandpal N, Deng H, Roberts A, Wallace E, Raffel C, 'Large language models struggle to learn long-tail knowledge', in Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, *Proceedings of the 40th international conference on machine learning*, Proceedings of Machine Learning Research, Honolulu, 2023.
21. Li D, Rawat A S, Zaheer M, Wang X, Lukasik M, Veit A, Yu F X, and Kumar S, 'Large language models with controllable working memory,' in Rogers A, Boyd-Graber J, and Okazaki N (eds), *Findings of the association for computational linguistics: ACL 2023*, Association for Computational Linguistics, Toronto, 2023.
22. Kasai J, Sakaguchi K, Takahashi Y, Le Bras R, Asai A, Yu X, Radev D, Smith N A, Choi Y, and Inui K, 'RealTime QA: What's the answer right now?,' in Oh A, Naumann T, Globerson A, Saenko K, Hardt M and Levine S, *Advances in neural information processing systems 36 (NeurIPS 2023)*, Curran Associates Inc, New Orleans, 2023.

23. Onoe Y, Zhang M, Choi E, and Durrett G, 'Entity Cloze by Date: What LMs know about unseen entities,' in Carpuat M, de Marneffe M, Ruiz M and Vladimir I, *Findings of the association for computational linguistics: NAACL 2022*, Association for Computational Linguistics, Seattle, 2022.
24. Min S, Gururangan S, Wallace E, Hajishirzi H, Smith N A, and Zettlemoyer L, 'SILO language models: Isolating legal risk in a nonparametric datastore,' Twelfth International Conference on Learning Representations, Vienna, 8 May 2024.
25. Sharma S, Tong M, Korbak T, Duvenaud D, Askell A, Bowman S, Cheng N, Burmus E, Hatfield-Dodds Z, Johnston S, Kravec S, Maxwell T, McCandlish S, Ndousse K, Rausch O, Scheifer N, Yna D, Zhang M, Perez E, 'Towards understanding sycophancy in language models' Twelfth International Conference on Learning Representations, Vienna, May 7 2023.
26. Bender E, Gebru T, McMillan-Major A and Shmitchell S, 'On the dangers of stochastic parrots: Can language models be too big?' in *Proceedings of the 2021 ACM Conference on fairness, accountability, and transparency,* Association for Computing Machinery, 2021.
27. Parikh A, Wang X, Gehrmann S, Faruqui M, Dhingra B, Yang D, and Das D, 'ToTTo: A controlled table-to-text generation dataset,' in Webber B, Cohn T, He Y and Liu Y (eds) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Virtual, 2020.
28. Feng Y, Xie W, Gu S, Shao C, Zhang W, Yang Z, and Yu D, 'Modeling fluency and faithfulness for diverse neural machine translation,' in *Proceedings of the AAAI Conference on Artificial Intelligence Volume 34(1): Technical Tracks 1*, AAAI Press, Palo Alto, 2020.
29. Dziri N, Madotto A, Zaiane O, and Bose A J, 'Neural path hunter: Reducing hallucination in dialogue systems via path grounding,' in Moens M, Huang X, Specia L and Yih S, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language* Processing, Association for Computational Linguistics, Punta Cana, 2021.
30. Petroni F, Rocktäschel T, Riedel S, Lewis P, Bakhtin A, Wu Y and Miller A, '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, Hong Kong, 2019.

31. Longpre S, Perisetla K, Chen A, Ramesh N, DuBois C, and Singh S, 'Entity-based knowledge conflicts in question answering,' in Moens M, Huang X, Specia L and Yih S, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Punta Cana, 2021.
32. Min S, Krishna K, Lyu X, Lewis M, Yih W, Koh P W, Iyyer M, Zettlemoyer L, and Hajishirzi H, 'FActScore: Fine-grained atomic evaluation of factual precision in long form text generation,' in Bauamor H, Pino J and Bali K, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, 2023.
33. Zhang X, Peng B, Tian Y, Zhou J, Jin L, Song L, Mi H, and Meng H, 'Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation,' in Ku L, Martins A and Srikumar V (eds), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Bangkok, 2024.
34. Agrawal A, Suzgun M, Mackey L and Kalai A, 'Do language models know when they're hallucinating references?,' in Graham Y and Purver M (eds) *Findings of the Association for Computational Linguistics: EACL 2024*, Association for Computational Linguistics, St. Julian's, 2024.
35. Wang Z, Wang X, An B, Yu D, and Chen C, 'Towards faithful neural table-to-text generation with content-matching constraints,' in Jurafsky D, Chai J, Schluter N and Tetreault J (eds) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020.
36. Maynez J, Narayan S, Bohnet B, and McDonald R, in Jurafsky D, 'On faithfulness and factuality in abstractive summarization,' in Chai J, Schluter N and Tetreault N (eds) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020.
37. Falke T, Ribeiro L, Utama P, Dagan I and Gurevych I, 'Ranking generated summaries by correctness: An interesting but challenging application for natural language inference,' in Korhonen A, Traum D and Marquez Luis (eds) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, 2019.
38. Jain S, Keshava V, Mysore Sathyendra S, Fernandes P, Liu P, Neubig G, and Zhou C, 'Multi-dimensional evaluation of text summarization with in-context learning,' in Rogers A, Boyd-Graber J and Okazaki N (eds) *Findings of the Association for*

*Computational Linguistics: ACL 2023*, Association for Computational Linguistics, Toronto, 2023.

39. Kalai A, Vempala S, 'Calibrated language models must hallucinate,' STOC '24: 56th Annual ACM Symposium on Theory of Computing, Vancouver, 25 June 2024.
40. Subramani N, Suresh N and Peters M, 'Extracting latent steering vectors from pretrained language models' In Muresan S, Nakov P, Villavicencio A (eds) *Findings of the Association for Computational Linguistics: ACL 2022*, Association for Computational Linguistics, Dublin, 2022.
41. Khundadze G, Kernion J, Landis J, Kerr J, Mueller J, Hyun J, Landau J, Ndousse K, Goldberg L, Lovitt L, Lucas M, Sellitto M, Zhang M, Kingsland N, Elhage N, Joseph N, Mercado N, DasSarma N, Rausch O, Larson R, McCandlish S, Johnston S, Kravec S, El Showk S, Lanham T, Telleen-Lawton T, Brown T, Henighan T, Hume T, Bai Y, Hatfield-Dodds Z, Clark J, Bowman S R, Askell A, Grosse R, Hernandez D, Ganguli D, Hubinger E, Schiefer N and Kaplan J, 'Discovering language model behaviours with model-written evaluations', 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Toronto, 9 July 2023.
42. Liu B, Ash T, Goel S, Krishnamurthy A and Zhang C, 'Exposing attention glitches with flip-flop language modelling', in Oh A, Naumann T, Globerson A, Saenko K, Hardt M and Levine S (eds) *NIPS'23: 37th International Conference on Neural Information Processing Systems*, Curran Associates Inc, New Orleans, 2023.
43. Guu K, Lee K, Tung Z, Pasupat P, Chang M, 'Retrieval augmented language model pre-training', in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, Journal of Machine Learning Research, Vienna, 2020.
44. Shuster K, Poff S, Chen M, Kiela D and Weston J, 'Retrieval augmentation reduces hallucination in conversation,' In Moens M, Huang X, Specia L, Yih S (eds), *Findings of the Association for Computational Linguistics: EMNLP 2021*, Association for Computational Linguistics, Punta Cana, 2021.
45. Thakur, N, Bonaficio L, Zhang X, Ogundepo O, Kamalloo E, Alfonso-Hermelo D, Li X, Liu Q, Chen B, Rezagholizadeh M, Lin J, 'NoMIRACL: Knowing When You Don't Know for Robust Multilingual Retrieval-Augmented Generation,' in Al-Onaizan Y, Bansal M and Chen Y (eds) *Findings of the Association for Computational Linguistics: EMNLP 2024*, Association for Computational Linguistics, Miami, 2024.

46. Mitchell E, Lin C, Bosselut A, Finn C, and Manning C, ‘Fast model editing at scale,’ Tenth International Conference on Learning Representations, ICLR 2022, Virtual, April 26, 2022.
47. Meng K, Bau D, Andonian A, and Belinko Y, ‘Locating and editing factual associations in GPT,’ in Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K and Oh A (eds) *NIPS'22: 36th International Conference on Neural Information Processing Systems*, Curran Associates Inc, New Orleans, 2022.
48. De Cao N, Aziz W, and Titov I, ‘Editing factual knowledge in language models,’ in Moens M, Huang X, Specia L, Yih S (eds) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021.
49. Mitchell E, Lin C, Bosselut A, Finn C, and Manning C, ‘Fast model editing at scale,’ Tenth International Conference on Learning Representations, ICLR 2022, Virtual, April 26, 2022.
50. Meng K, Bau D, Andonian A, and Belinko Y, ‘Locating and editing factual associations in GPT,’ in Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K and Oh A (eds) *NIPS'22: 36th International Conference on Neural Information Processing Systems*, Curran Associates Inc, New Orleans, 2022.
51. Dai D, Dong L, Hao Y, Sui Z, Chang B, and Wei F, ‘Knowledge neurons in pretrained transformers,’ in Muresan S, Nakov P and Villavivencio A (eds) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, 2022.
52. Sinitsin A, Plokhotnyuk V, Pyrkin D, Popov S, and Babenko A, ‘Editable neural networks,’ 8th International Conference on Learning Representations (ICLR 2020), Addis Ababa, April 27, 2020.

## Self-Published Articles

1. Barlett P, Long P, Lugosi G and Tsigler A, 'Benign overfitting in linear regression', arXiv, 2020.
2. Curran S, Bethell O and Lansley S, 'Hallucination is the last thing you need, arXiv preprint, 2023.
3. Ghasemi M, Moosavi A, Sorkhoh I, Agarwal A, Alzhouri F and Ebrahimi D, ‘An introduction to reinforcement learning: Fundamental concepts and practical applications’, arXiv, 2024.

4. Naveed H, Khan, Qiu S, Saqib M, Anwar S, Usman M, Akhtar N, Barnes N and Mian A, 'A comprehensive overview of large language models', arXiv, 2024.
5. Sanford C, Hsu D and Telgarsky M, 'Transformers, parallel computation, and logarithmic depth,' arXiv, 2024.
6. Yin S, Fu C, Zhao S, Li K, Sun X, Xu T, and Chen E, 'A survey on multimodal large language models', arXiv, 2023.
7. Lu J, Rao J, Chen K, Guo X, Zhang Y, Sun B, Yang C and Yang J, 'Evaluation and mitigation of agnosia in multimodal large language models,' ArXiv, 2023.
8. Li J, Consul S, Zhou E, Wong J, Farooqui N, Ye Y, Manohar N, Wei Z, Echols B, Zhou S, Diamos G, 'Banishing LLM hallucinations requires rethinking generalisation', arXiv, 2024.
9. Zhang J, 'Gradient descent-based optimization algorithms for deep learning models training', arXiv, 2019.
10. Banerjee S, Agarwal A, Singla S, 'LLMs will always hallucinate, and we need to live with this', arXiv, 2024.
11. Evans O, Cotton-Barratt O, Finnveden L, Bales A, Balwit A, Wills P, Righetti L, Saunders W, 'Truthful AI: Developing and governing AI that does not lie,' arXiv, 2021.
12. Zhang Z, Singh J, Gadiraju U and Anand A, 'Dissonance between human and machine understanding' arXiv, 2021.
13. Ziegler D, Stiennon N, Wu J, Brown T, Radford A, Amodei D, Christiano P and Irving G, 'Fine-tuning language models from human preferences,' arXiv, 2020.
14. Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, Drain D, Fort S, Ganguli D, Henighan T, Joseph N, Kadavath S, Kernion J, Cornely T, El-Showk S, Elhage N, Hatfield-Dodds Z, Hernandez D, Hume T, Johnston S, Kravec S, Lovitt L, Nnada N, Olsson C, Amodei D, Brown T, Clark J, McCandlish S, Olah C, Mann B and Kaplan J, 'Training a helpful and harmless assistant with reinforcement learning from human feedback', Anthropic, arXiv, 2022.
15. Ji J, Qiu T, Chen B, Lou H, Wang K, Duan Y, He Z, Zhou J, Zhang Z, Zeng F, Dai J, Pan X, Ng KY, O'Gara A, Xu H, Fu B, McAleer S, Yang Y, Wang Y, Zhu S, Guo Y and Gao W, 'AI alignment: A comprehensive survey', arXiv, 2024.
16. Casper S and Davies X, 'Open problems and fundamental limitations of reinforcement learning from human feedback', arXiv, 2023.
17. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, Chen A, Goldie A , Mirhoseini A, McKinnon C, Chen C, Olsson C, Olah C, Hernandez D, Drain D, Ganguli

D, Li D, Tran-Hohnson E, Perez E, Kerr J, Mueller J, Ladish J, Landau J, Ndousse K, Lukosuite K, Lovitt L, Sellitto M, Elhage N, Schiefer N, Mercado N, DasSarma N, Lasenby R, Larson R, Ringer S, Johnston S, Kravex S, El Showk S, Fort S, Lanham T, Telleen-Lawton T, Cornely T, Henighan T, Hume T, Bowman S, Hatfield-Dodds Z, Mann B, Amodei D, Joseph N, McCandlish S, Brown T and Kaplan J, ' Constitutional AI: Harmlessness from AI feedback,' arXiv, 2022.

18. Chen Z, Zhou K, Zhao W, Wun J, Zhang F, Zhang D and Wen J, 'Improving large language models via fine-grained reinforcement learning with minimum editing constraint', arXiv, 2024.
19. Kumar A, Zhuang V, Agarwal R, Su Y, Co-Reyes J, Singh A, Baumli K, Iqbal S, Bishop C, Roelofs R, Zhang L, McKinney K, Shrivastava D, Paduraru C, Tucker G, Precup D, Behbahani F and Faust A, 'Training language models to self-correct via reinforcement learning', Google DeepMind, arXiv, 2024.
20. Wang S, Zhang S, Zhang J, Hu R, Li X, Zhang T, Li J, Wu F, Wang G and Hovy E, 'Reinforcement learning enhanced LLMs: A survey', arXiv, 2024.
21. Kaufmann T, Weng P, Bengs V and Hüllermeier E, 'A survey of reinforcement learning from human feedback', arXiv, 2024.
22. Wang T, Herbert S and Gao S, 'Fractal landscapes in policy optimization', arXiv, 2023.
23. Wei J, Huang D, Lu Y, Zhou D, and Le Q, 'Simple synthetic data reduces sycophancy in large language models', arXiv, 2024.
24. Sharma S, Tong M, Korbak T, Duvenaud D, Askell A, Bowman S, Cheng N, Burmus E, Hatfield-Dodds Z, Johnston S, Kravec S, Maxwell T, McCandlish S, Ndousse K, Rausch O, Scheifer N, Yan D, Zhang M and Perez E, 'Toward understanding sycophancy in language models,' arXiv, 2023.
25. Gao M, Ruan J, Sun R, Yin X, Yang S, and Wan X, 'Human-like summarization evaluation with ChatGPT,' ArXiv, 2023.
26. Golilarz N, Hossain E, Addeh A and Rahimi K, 'Learning algorithms made simple' arXiv, 2024.
27. Xi Y, Ding W, Yu K and Lai J, 'Semi-supervised learning for code-switching ASR with large language model filter', arXiv, 2024.
28. Martin L, Whitehouse N, Yiu S, Catterson L and Rivindu P, 'Better call GPT: Comparing large language models against lawyers', arXiv, 2024.

29. Zhao H, Liu Z, Wu Z, Li Y, Yang T, Shu P, Xu S, Dai H, Zhao L, Mai G, Liu N and Liu T, 'Revolutionizing finance with LLMs: An overview of applications and insights,' arXiv, 2024.
30. Chen M, Tworek J, Jun H, Yuan Q, Pinto H, Kaplan J, Edwards H, Burda Y, Joseph N, Brockman G, Puri R, Krueger G, Petrov M, Khlaaf H, Sastry G, Mishkin P, Chan B, Gray S, Ryder N, Pavlov M, Power A, Kaiser L, Bavarian M, Winter C, Tillet P, Such FP, Cummings D, Plappert M, Chantzis F, Barnes E, Herbert-Voss A, Guss WH, Nichol A, Paino A, Tezak N, Tang J, Babuschkin I, Balaji S, Jain S, Saunders W, Hesse C, Carr AN, Leike J, Achiam J, Misra V, Morikawa E, Radford A, Knight M, Brundage M, Murati M, Mayer K, Welinder P, McGrew B, Amodei D, McCandlish S, Sutskever I and Zaremba W, 'Evaluating large language models trained on code', arXiv, 2021.
31. Yu R, Xu Z, CH-Wang S, Arum R, 'Whose ChatGPT? Unveiling real-world educational inequalities introduced by large language models' arXiv, 2024.
32. Huang Y, Feng X, and Qin B, 'The factual inconsistency problem in abstractive text summarization: A survey', arXiv, 2021.
33. Li W, Wu W, Chen M, Liu J, Xiao X, and Wu H, 'Faithfulness in natural language generation: A Systematic Survey of Analysis, Evaluation and Optimization Methods', arXiv,
34. Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, Huang X, Zhao E, Zhang Y, Chen Y, Wang L, Luu A, Bi W, Shi F, Shi S, ' Siren's song in the AI ocean: A survey on hallucination in Large Language Models', arXiv, 2023.
35. Wang Y, Zhong W, Li L, Mi F, Zeng X, Huang W, Shang L, Jiang X, and Liu Q, 'Aligning large language models with humans: A survey,' arXiv, 2023.
36. Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang P-S, Cheng M, Glaese M, Balle B, Kasirzadeh A, Kenton Z, Brown S, Hawkins W, Stepleton T, Biles C, Birhane A, Haas J, Rimell L, Hendricks L A, Isaac W, Legassick S, Irving G, and Gabriel I, 'Ethical and social risks of harm from language models', arXiv, 2021.
37. Wei J, Huang D, Lu Y, Zhou D, and Le Q, 'Simple synthetic data reduces sycophancy in large language models,' arXiv, 2023.
38. Tian R, Narayan S, Sellam T, and Parikh A, 'Sticking to the facts: confident decoding for faithful data-to-text generation,' arXiv, 2020.
39. Chern I, Chern S, Chen S, Yuan W, Feng K, Zhou C, He J, Neubig G, and Liu P, 'FacTool: Factuality detection in generative AI–A tool augmented framework for multi-task and multi-domain scenarios,' ArXiv, 2023.

40. Wang Y, Zhong W, Li L, Mi F, Zeng X, Huang W, Shang L, Jiang X, and Liu Q, 'Aligning large language models with humans: A survey,' arXiv, 2023.
41. Zou A, Wang Z, Carlini N, Nasr M, Kolter J, and Fredrikson M, 'Universal and transferable adversarial attacks on aligned language models,' arXiv, 2023.
42. Dhuliawala S, Komeili M, Xu J, Raileanu R, Li X, Celikyilmaz A, and Weston J, 'Chain-of-verification reduces hallucination in large language models,' ArXiv, 2023.
43. Kadavath S, Conerly T, Askell A, Henighan T, Drain D, Perez E, Schiefer N, Hatfield-Dodds Z, DasSarma N, Tran-Johnson E, Johnston S, El-Showk S, Jones A, Elhage N, Hume T, Chen A, Bai Y, Bowman S, Fort S, Ganguli D, Hernandez D, Jacobson J, Kernion J, Kravec S, Lovitt L, Ndousse K, Olsson C, Ringer S, Amodei D, Brown T, Clark J, Joseph N, Mann B, McCandlish S, Olah C, and Kaplan J, 'Language models (mostly) know what they know,' ArXiv, 2022.
44. Zheng D, Lapata M and Pan J, 'Large Language Models as Reliable Knowledge Bases? Re-thinking factuality and consistency,' arXiv, 2024.
45. Varshney N, Yao W, Zhang H, Chen J and Yu D, 'A stitch in time saves nine: Detecting and mitigating hallucinations of LLMs by validating low-confidence generation,' ArXiv, 2023.
46. Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, Phang J, He H, Thite A, Nabeshima N, Presser S, and Leahy C, 'The pile: An 800gb dataset of diverse text for language modelling,' ArXiv, 2021.
47. Laban P, Kryściński W, Agarwal D, Fabbri A, Xiong C, Joty S, and Wu C, 'LLMs as factual reasoners: Insights from existing benchmarks and beyond,' ArXiv, 2023.

## Reports

1. Global Insights on Access to Justice: Findings from the World Justice Project General Population Poll in 101 Countries, World Justice Project, *Final report*, 2019.
2. Justice Fact Sheet 2023, Open Government Partnership, *Final draft*.
3. Legal Services Corporation, The Justice Gap: The Unmet Civil Legal Needs of Low-Income Americans (2022).
4. United Nations Office on Drugs and Crime, *Global Study on Legal Aid*, 2016.
5. United States Supreme Court, *2023 Year-End Report on the Federal Judiciary*, 31 December 2023.

## Other online sources


1. Bergmann D, 'What is fine tuning?' IBM. — https://www.ibm.com/topics/fine-tuning#:~:text=Fine%2Dtuning%20in%20machine%20learning,models%20used%20for%20generative%20AI.

2. Chow A, 'How ChatGPT managed to grow faster than TikTok or Instagram' TIME 8 February 2023, — https://time.com/6253615/chatgpt-fastest-growing/.

3. Goodson N and Lu R, 'Transforming legal aid with AI: Training LLMs to ask better questions for legal intake', Stanford Law School Blogs, 15 March 2024. https://law.stanford.edu/2024/03/15/transforming-legal-aid-with-ai-training-llms-to-ask-better-questions-for-legal-intake/.

4. Heaven W, 'What is AI?', MIT Technology Review, 10 July 2024. https://www.technologyreview.com/2024/07/10/1094475/what-is-artificial-intelligence-ai-definitive-guide/ on 20 August 2024. See also Stryker C and Kavlakoglu E, 'What is artificial intelligence (AI)?', IBM Think, 9 August 2024. https://www.ibm.com/think/topics/artificial-intelligence.

5. IBM Think, 'What are large language models (LLMs)?' 2 November 2023. https://www.ibm.com/think/topics/large-language-models.

6. Martineau K, 'What is generative AI?' IBM Think, 20 April 2023. https://research.ibm.com/blog/what-is-generative-AI.

7. Menon S, 'Why AI is popular now—And two ways to use it better' Forbes, 1 December 2023 —https://www.forbes.com/councils/forbestechcouncil/2023/12/01/why-ai-is-popular-now-and-two-ways-to-use-it-better/.

8. Milmo D, 'ChatGPT reaches 100 million users two months after launch' The Guardian, 2 February 2023, — https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app.

9. Talebi S, 'Multimodal models – LLMs that can see and hear', Towards Data Science, 19 November 2024. https://towardsdatascience.com/multimodal-models-llms-that-can-see-and-hear-5c6737c981d3/.

10. Thorbecke C, 'AI tools make things up a lot, and that's a huge problem' CNN Business 29 August 2023, —https://edition.cnn.com/2023/08/29/tech/ai-chatbot-hallucinations/index.html.