THE RELIABILITY OF LARGE LANGUAGE MODEL CHATBOTS IN LIGHT OF PUBLICLY REPORTED LEGAL PROCEEDINGS

* Att. Derya Gizem Üst, LL.M.



Keywords: artificial intelligence, LLM, chatbot, hallucination, reliability

Istanbul Bar Association
Information Technology Law Commission
Artificial Intelligence Working Group

INTRODUCTION

Artificial intelligence technologies are driving profound changes across many sectors, most notably healthcare, finance, education, and law, and this transformation is accelerating day by day. From image processing to language analytics, and from medical diagnosis to legal research, a broad range of applications today bears the imprint of AI. Perhaps the most visible and widely used component of this technological ecosystem is chatbot systems based on large language models ("LLM"). Nevertheless, the accuracy and reliability of the content generated by chatbots remain highly contested.

The phenomenon commonly referred to in the literature as "hallucination", where LLMs generate content that is contrary to fact or entirely fictional, has become the subject of various judicial proceedings and complaints, some concluded and others still ongoing. These processes range from disciplinary sanctions imposed on attorneys in the United States for citing non-existent precedents, to lawsuits in Europe alleging violations of personal rights and data protection. Collectively, they have turned the reliability of LLMs into a concrete legal issue.

This article examines the reliability problem of chatbot systems through the lens of legal processes that have come to public attention. It will analyze how LLM systems operate, why they are prone to error, and how these have been reflected in judicial proceedings.

I. What Is an LLM and How Does It Work?

Large Language Models are based on machine learning (ML) and can, in broad terms, be divided into two categories: generative (GenAI) and non-generative models. While non-generative LLMs are used for tasks such as text classification, summarization, or translation, generative LLMs (e.g., ChatGPT) go beyond these functions and produce creative texts that align with the input prompts.[1]

LLMs are built on complex artificial neural networks containing billions of parameters.[2] These parameters take on specific values as a result of the model being trained on vast amounts of textual data.[3] But how are human-language words processed within this mathematical structure?

The process begins with splitting the data into tokens (tokenization).[4] At this stage, the text is broken down into small units (tokens) that can be processed by the machine. Depending on the model, tokens may be as short as a single letter, a syllable, or as long as an entire word. The key point is that, for semantic processing, what matters is not a token in isolation but rather a sequence of tokens that together represents a concept.

Once tokenized, texts are no longer represented in the model as readable words, but solely as numerical values, and these numerical representations are embedded into the model.[6] In the embedding process, data is converted into sequences of numbers referred to as "vectors".[7] These vectors then become part

[1] Paulina Jo Pesch, Rainer Böhme, "Verarbeitung personenbezogener Daten und Datenrichtigkeit bei großen Sprachmodellen", MMR 2023, p. 917-918.
[2] Vaswani et al., "Attention Is All You Need", 02.08.2023, p. 2, https://arxiv.org/pdf/1706.03762, Accessed: 30.01.2026.
[3] Tim W. Dornis, Sebastian Stober, "Urheberrecht und Training generativer KI-Modelle Technologische und juristische Grundlagen", sy 54: Jakob Hüger, Die Rechtmäßigkeit von Datenverarbeitungen im Lebenszyklus von KI-Systemen, ZfDR 2024, p. 291.
[4] HmbBfDI, Diskussionspapier: Large Language Models und personenbezogene Daten, 15.7.2024, https://datenschutzhamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Diskussionspapier_HmbBfDI_KI_Modelle.pdf 5_Diskussionspapier_HmbBfDI_KI_Modelle.pdf, s. 3, Accessed: 01.02.2026.
[5] OpenAI, GBT-5.x & O1/3, https://platform.openai.com/tokenizer , Accessed: 02.02.2026
[6] HmbBfDI, Diskussionspapier: Large Language Models und personenbezogene Daten, 15.7.2024, https://datenschutzhamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Diskussionspapier_HmbBfDI_KI_Modelle.pdf 5_Diskussionspapier_HmbBfDI_KI_Modelle.pdf, s. 3, Accessed: 01.02.2026.
[7] HmbBfDI, Diskussionspapier: Large Language Models und personenbezogene Daten, 15.7.2024, https://datenschutzhamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Diskussionspapier_HmbBfDI_KI_Modelle.pdf

of a high-dimensional vector space, where semantically similar words are positioned close to one another, whereas words with different meanings are positioned farther apart.[8] Embeddings in LLMs are dynamic and context sensitive. Accordingly, the same word may be represented with different meanings in different contexts.[9] This process of data processing and pattern recognition occurs within artificial neural networks. At the start of training, parameters are usually the same or random; as training continues, links for common patterns are strengthened, while rare ones are weakened or removed.[10] As a result, the parameters acquire specific and differentiated weights that represent the knowledge the model has learned.

Models learn the relationships between these tokens through billions of numerical weights called parameters. Here, "learning" means that the model learns the probability that certain word parts, words, and sentences will follow one another in particular contexts.[11]

The generative LLM GPT-3 has approximately 175 billion parameters, and today's models have far more than that. These parameters are essentially placeholders; they only take on concrete values and become meaningful only after training, through exposure to the training data.[12] They enable the model to determine which word is more likely to come next after a given word. Accordingly, when a generative LLM-based chatbot answers the question "Why is the sky…", it does so not because it "knows" the physical properties of the sky, but because, in its training data, the word "blue" most frequently follows that pattern, statistically.

However, it should be noted that, in this process, the system has no connection to the concept of reality; it is simply a matter of generating "high-probability sequences of words".[13] For this very reason, generative LLMs do not always provide correct outputs, they merely produce word

5_Diskussionspapier_HmbBfDI_KI_Modelle.pdf, s. 3, Accessed: 01.02.2026.

[8] Matthias Grabmair (2024). Natural Language Processing (NLP). Martin Ebers (Ed.), *SWK Legal Tech*, Baden-Baden, 5th Edition, par. 27.

[9] Matthias Grabmair (2024). Natural Language Processing (NLP). Martin Ebers (Ed.), *SWK Legal Tech*, Baden-Baden, 5th Edition, par. 27.

[10] Marit Hansen, Benjamin Walczak, "Die KI zaubert nicht", KIR 2024, p. 82.

[11] Jo Pesch, "Potentials and Challenges of Large Language Models (LLMs) in the Context of Administrative Decision-Making", European Journal of Risk Regulation 2025, p. 78.

[12] Werner Vogd, Jonathan Harth, Das Bewusstsein der Maschinen - die Mechanik des Bewusstseins: Mit Gotthard Günther über die Zukunft menschlicher und künstlicher Intelligenz nachdenken, Weilerswist 2023, p. 146.

[13] Carlini et al., "Extracting Training Data from Large Language Models", Proceeding 30th USENIX Security Symposium 2021, s. 2634, https://www.usenix.org/system/files/sec21-carlini-extracting.pdf, Accessed: 28.01.2026.

sequences that are statistically likely to co-occur.

II. Why Are LLM Systems Prone to Error?

The tendency of LLMs to make errors stems from a structural and methodological necessity. Firstly, LLMs are trained on data collected up to a certain date, and they have no knowledge of events occurring after that date. This boundary, referred to as the knowledge cutoff, may cause the model, instead of stating this limitation when answering questions about recent developments, to generate an output that appears plausible based on its existing knowledge but is outdated or entirely incorrect. This becomes particularly problematic where changing legislation, up-to-date case-law, or new scientific findings are concerned. For instance, while the knowledge cutoff date of the training data used for the development of the GPT-5.2 model released by OpenAI is 31 August 2025, it is 1 June 2024 for GPT 4.1.[14] However, the sources of error are not limited to timing; the quality of the data the model is trained on is at least as critical as its timeliness. Training data used to develop and deploy AI models are typically gathered from publicly available sources through systematic, machine-assisted collection using web scraping.[15] Examples of such public sources include news-portal archives, social media data, Reddit, and many other online resources.

The fact that an LLM's output aligns in content with its training data does not mean that the output is necessarily accurate or up to date; at that point, the accuracy and timeliness of the training data itself are often open to question.[16] And because LLMs are trained on enormous datasets, and owe much of their performance to that sheer scale, there is no realistic way for developers to comprehensively audit every source, validate every fact, and systematically remove incorrect or outdated material. Doing so would demand vast time and resources and would slow development to a standstill, which is why, as a practical matter, it is not feasible.[17]

[14] OpenAI Platform, https://platform.openai.com/docs/models/compare?model=o3 , Accessed: 02.02.2026.
[15] OpenAI, https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed, Accessed: 02.04.2026

[16] Paulina Jo Pesch, Rainer Böhme, "Verarbeitung personenbezogener Daten und Datenrichtigkeit bei großen Sprachmodellen", MMR 2023, p. 921.
[17] Amon Dieker, "Datenschutzrechtliche Zulässigkeit der Trainingsdatensammlung", ZD 2024, p. 133.

Even in a scenario where the data are assumed to be entirely accurate and up to date, large language models may still continue to generate incorrect outputs for structural reasons.[18] An LLM's output is not a reflection of reality, but the product of probabilistic inference. The model may fill gaps in the training data or reconcile contradictions by assembling information that appears probabilistically most plausible, even if it does not exist in fact. The most striking, and potentially most dangerous, manifestation of this phenomenon is referred to as "hallucination".[19]

Hallucination is when the model presents entirely fabricated events, people, statistics, or sources as real, often with strikingly persuasive detail. Concrete cases have made clear just how serious the consequences can be.

III. Recent Legal Proceedings That Have Come to Public Attention

Concerns over the reliability of chatbot outputs have increasingly crystallized into a clear legal issue in light of recent court decisions and ongoing complaint proceedings before data protection authorities.

1. NOYB v. OpenAI- (Norway/GDPR Complaint, 2025)

In Norway, an individual named Arve Hjalmar Holmen discovered that when he searched his own name on OpenAI's ChatGPT system, the system made an entirely fabricated and highly serious allegation about him that he was "a convicted criminal who killed his two children." What is particularly striking about this AI-generated hallucination is that it intertwined verifiable personal data, such as the individual's number of children, their genders, and his city of residence with a criminal allegation that had no basis in fact. In other words, the system blended real personal data with a fictional narrative, producing a profile that appeared factual but was, in substance, entirely unfounded. In response, the data protection organization NOYB filed a complaint against OpenAI before the Norwegian Data Protection Authority. The complaint alleged a breach of the "Accuracy Principle" set out in Article 5(1)(d) of the General Data Protection Regulation (GDPR). Under this principle, personal data must be accurate and, where necessary, kept up to date, and inaccurate data must be erased or corrected without delay. False information generated through hallucinations directly conflicts with this obligation. The complaint filed by NOYB has not yet been resolved, and the

[18] Markus Kaulartz, Tom Braegelmann (Ed.), Rechtshandbuch AI und Machine Learning, Ch. 8.7, Par. 17.

[19] Marit Hansen, Benjamin Walczak, "Die KI zaubert nicht", KIR 2024, p. 83.

case is currently under review before the Irish Data Protection Commission (DPC).

Although the underlying ChatGPT model has been updated since the incident and has gained the ability to search the internet about the individual, this technical improvement offers only a partial solution to the hallucination problem. Incorrect information may remain embedded in the AI's training dataset, and it can be permanently removed only by retraining the model from scratch (retraining).

2. Mata v. Avianca, Inc. (US, 2023)

A passenger named Roberto Mata alleged that he was injured when a metal service cart struck his knee during an international flight operated by Avianca Airlines, and he sought compensation from the airline company. After the lawsuit was filed, the defendant airline moved to have the claim dismissed on the ground that it was time-barred under the applicable statute of limitations. In response to the defendant's objections, the plaintiff's attorney, Steven Schwartz, prepared a submission in which he cited cases such as "Petersen v. Iran Air" and "Martinez v. Delta Airlines."

However, neither the court nor defense counsel could verify the existence of the decisions Steven Schwartz cited, and it became clear that they were entirely fabricated, non-existent cases generated by ChatGPT.

This case set off a chain of events that underscored the dangers of relying on AI in legal practice uncritically and without any skepticism.

3. Walters v. OpenAI (US, 2025)[20]

Georgia State Court in the U.S. dismissed a *defamation* action arising from an AI hallucination by granting summary judgment in OpenAI's favor. In the case at hand, AmmoLand.com reporter Frederick Riehl, while researching a lawsuit filed by a U.S. foundation (*Second Amendment Foundation-SAF*) against the Washington State Attorney General (*SAF v. Ferguson*), asked ChatGPT to summarize the case. Riehl first pasted information about the case into the model and asked it to generate a summary. He then provided a website link and asked again for a summary. ChatGPT stated in its initial response that it did not have internet access and lacked the ability to perform real-time queries.

[20]Eric Goldman , ChatGPT Defeats Defamation Lawsuit Over Hallucination–Walters v. OpenAI, Technology & Marketing Law Blog, 27.05.2025, https://blog.ericgoldman.org/archives/2025/05/chatgpt-defeats-defamation-lawsuit-over-hallucination-walters-v-openai.htm, Accessed: 02.02.2026.

ChatGPT generated an entirely fabricated output claiming that radio host Mark Walters, who had no connection to the case, had embezzled the foundation's funds. Walters then sued OpenAI, alleging reputational harm.

The Court dismissed the case on three separate grounds. First, it concluded that a reasonable reader would not understand the challenged output as a "statement of fact." In reaching that conclusion, the court pointed to several contextual features: ChatGPT expressly informed the user that it could not access the internet and therefore could not verify information in real time, it also disclosed that it operates subject to a defined knowledge cutoff date. In addition, OpenAI's terms of use contained clear disclaimers warning that the system may occasionally generate incorrect or unreliable content. Finally, the court emphasized that the journalist who received the output did not treat it as a verified factual assertion and acknowledged that he did not understand it to be true.

Second, relying on OpenAI's representations that it had implemented advanced technical measures to reduce hallucinations, the Court found no basis for negligence or bad faith.

Third, the Court concluded that the requirements for damages were not met, because Walters himself acknowledged in his testimony that he had not suffered any concrete harm and had not sought a correction from OpenAI.

This decision is a noteworthy precedent for purposes of legal liability arising from AI hallucinations. The court effectively treated OpenAI's provision of adequate warning mechanisms, and the user's knowledge, or ability to anticipate, that the output was unfounded, as the decisive criteria. That approach, however, overlooks the risk that ordinary users with low AI literacy may treat the same fictional output as true information. The mere existence of disclaimers should not be treated as a substitute for the system's obligation to ensure accuracy.

4. Matter of Weber (US, 2024)[21]

In the 2024 Matter of Weber case, which is precedent-setting in terms of the debate over the evidentiary value of AI-generated outputs in court, the New York Court held that an expert report could not be admitted as evidence where the expert had performed the underlying calculations using Microsoft Copilot.

[21] Justia U.S. Law, https://law.justia.com/cases/new-york/other-courts/2024/2024-ny-slip-op-24258.html, Accessed: 02.02.2026, Kyle Petersen, Tal Dickstein, IP/Entertainment Case Law Updates, Walters v. OpenAI, L.L.C., https://www.loeb.com/en/insights/publications/2025/05/walters-v-openai-llc, Accessed:02.02.2026.

Applying the Frye standard (general acceptance in the relevant scientific community) governing the admissibility of scientific evidence, the Court held that generative AI outputs are not yet reliable. The court's decision was driven by the expert's inability to explain the commands (prompt) entered into the system and to identify the sources on which the AI relied, as well as by the court's own testing, which showed that Copilot produced inconsistent results. The ruling makes clear that AI-assisted calculations and opinions will not be automatically accepted in legal disputes, that counsel have a duty to disclose AI use to the court in advance, and that the reliability of the technology used must be demonstrated.

5. Moffatt v. Air Canada (Canada, 2024)[22]

In the dispute at hand, the consumer plaintiff sought information about a "bereavement fare" via the chatbot on the defendant airline's website in order to travel from Vancouver to Toronto due to a death in the family. The chatbot stated that a discount request could be submitted within 90 days from the date the ticket was issued.

Relying on this information, the consumer purchased the ticket at the full fare and later requested a refund. The airline, however, denied the request, stating that under its policy and its general terms and conditions, the discount could be requested only at the time of booking and could not be applied retroactively.

Air Canada argued during the proceedings that it could not be held responsible for the chatbot's actions, asserting that the bot was a separate entity/agent responsible for its own conduct. However, the Canadian Civil Resolution Court categorically rejected this defense. The court emphasized that the chatbot was not legally different from the website's static information pages and that it was simply a component of the airline's website.

The Court held that the company breached its duty of care to ensure the accuracy of information generated by the chatbot and thereby misled the consumer. Finding that this amounted to negligent misrepresentation, the court ordered the airline to compensate the consumer for the difference between the ticket price paid and the discounted fare.

[22] Christian Thurow, Unternehmen haftet für KI-Chatbot-Auskunft, C.H.Beck, https://rsw.beck.de/zeitschriften/bc/news-beitraege/2024/02/22/unternehmen-haftet-f%C3%BCr-ki-chatbot-auskunft , Accessed: 02.02.2026.

CONCLUSION

From the outside, generative AI systems can look like near-perfect information machines, thanks to how smoothly they assemble words. Under the hood, however, they operate in a very different way. Their most defining, and arguably most troubling, feature is what the technical literature calls the "Black Box"[23]. Since billions of parameters interact in complex ways, it is often impossible to clearly explain how a particular input leads to that exact output, even for the developers. This lack of visibility between input and output means the system's answers cannot be reliably predicted in advance, creating a built-in unpredictability. As a result, how AI arrives at its outputs remains something of a technical puzzle, largely because there is no real algorithmic transparency.

It is precisely this uncertainty that explains why, today, you see standard warnings on almost every chatbot interface saying that AI models can make mistakes and that users should check the outputs.

These warnings go beyond companies trying to avoid liability; they are the technology itself openly acknowledging its limits and its built-in structural fragility when it comes to reliability.

The court decisions discussed in this study are only a handful of the hundreds of cases that have reached judicial authorities worldwide.[24] The legal field is confronting these kinds of disputes at an ever-increasing pace, and the examples in this article represent only the tip of the iceberg.

Ultimately, it would be outdated to dismiss the speed and efficiency AI can provide, but it is not wise to trust it without question. It should be kept in mind that AI is not a flawless expert; it always requires oversight and can produce serious errors.

[23] Christian Heinze, Christoph Sorge, Louisa Specht-Riemenschneider, "Das Recht der Künstlichen In-Telligenz", KIR 2024, p. 12.

[24] For further case-law analysis, see. https://www.damiencharlotin.com/hallucinations/

ACKNOWLEDGEMENTS

We thank Prof. Mehmet Şahin and Dr. Ayşenur Ocak for their valuable contributions to the review of this study.

REFERENCES

Carlini et al. (2021). Extracting Training Data from Large Language Models. *Proceeding 30th USENIX Security Symposium*, 2633- 2650, https://doi.org/10.48550/arXiv.2012.07805.

Dieker, A. (2024). Datenschutzrechtliche Zulässigkeit der Trainingsdatensammlung, *Zeitschrift für Datenschutzrecht*, 132-137.

Grabmair, M. (2024). Natural Language Processing (NLP), Martin Ebers (Ed.), *SWK Legal Tech* (5th Edition Rn. 1-56). Baden-Baden.

Hansen, M. and Walczak, B. (2024). Die KI zaubert nicht, *Künstliche Intelligenz und Recht*, 82-86.

Heinze C., Sorge C. and Specht-Riemenschneider L. (2024). Das Recht der Künstlichen Intelligenz, *Künstliche Intelligenz und Recht,* 11-15.

HmbBfDI, Diskussionspapier: Large Language Models und Personenbezogene Daten (15.7.2024).https://datenschutzhamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Diskussionspapier_HmbBfDI_KI_Modelle.pdf, Accessed: 01.02.2026.

Hüger, J. (2024). Die Rechtmäßigkeit von Datenverarbeitungen im Lebenszyklus von KI-Systemen, *ZfDR*, 263-29.

Kaulartz, M. and Braegelmann, T. (Ed.). (2020). *Rechtshandbuch Artificial Intelligence und Machine Learning*, C.H. BECK Recht.

Pesch P. and Böhme, R. (2023). Verarbeitung Personenbezogener Daten und Datenrichtigkeit bei großen Sprachmodellen, *MMR*, 917-923.

Pesch, J. (2025). Potentials and Challenges of Large Language Models (LLMs) in the Context of Administrative Decision-Making, *European Journal of Risk Regulation*, 76-95, https://doi.org/10.1017/err.2024.99.

Vaswani et al. (2017). Attention Is All You Need, https://arxiv.org/pdf/1706.03762, Accessed: 30.01.2026, https://doi.org/10.48550/arXiv.1706.03762.

Vogd, W. and Harth, J. (2023). *Das Bewusstsein der Maschinen - die Mechanik des Bewusstseins: Mit Gotthard Günther über die Zukunft menschlicher und künstlicher Intelligenz nachdenken*, 1st Edition, Weilerswist.

W. Dornis, T. and Stober, S. (2024). *Urheberrecht und Training generativer KI-Modelle: Technologische und juristische Grundlagen*, Band 19, Baden-Baden.

Istanbul Bar Association

•

Information Technology Law Commission

•

Artificial Intelligence Working Group

•

2026

Editor

H. Sena Lezgioğlu Özer

Translator

Nilay Puyan