ARTICLE :

Involving LLMs in legal processes is risky: An invited paper

By Peter Bernard Ladkin

LLMs and facticity

Recently, a very capable ‘large language model’ (LLM) has been publicly released on the Internet for people to ‘play with’. Large language models are computer programs which embody a machine-learning artefact (ML), which is conflated with ‘artificial intelligence’ (AI), although scientists such as myself, who work in or have worked in AI, hold that AI consists in very much more than ML applications or LLMs.

The LLM is known as ChatGPT. It is a conversational or dialogue agent. It engages in textual interactions with other agents (in the case of Internet users, usually human). It appears to be capable of writing essays on demand that range from public-relations-type boiler-plate to passable emulations of specific authors. It works by (crudely, and generally) scanning and classifying untold amounts of online written material (a corpus, in the terminology of computer linguistics); predicting successive portions of text (at the level of words) to follow existing text, and smoothing the grammar, with the usual use of techniques to improve efficiency [Ruby 2023]. The corpus is enormous: some 570 Gigabytes of text, equivalent to 570,000 books of 400pp each [Gratas 2023].

It has surprised those not specialist in the development of ML that such emergent properties as writing coherent essays on demand can emanate from a procedure as mechanical as predicting next-words. It has been speculated, with justification, that it could be used by secondary school pupils to write their homework essays. It is reported to have passed the US Law School Admissions Test in the 90th percentile [Economist 2023]. There are notionally similar language-manipulation systems based on similar technology, for example one called Minerva, that appear to be able to do much ‘word problem’ algebra encountered in secondary schools, and, in the case of Minerva, explaining their reasoning, i.e., producing proofs [Leykowycz and others 2022; see also YouTube videos of Minerva in action].

It is thought at the time of writing that the successor to ChatGPT (an LLM engine called GPT-4; ChatGPT is said to be based on the GPT-3 engine) is being used by Microsoft’s Bing search engine (an alternative to Google Search). Google is said to be working on using its LLM engine (Bard) in its search service. I understand that Meta's LLaMA is now the basis for several open-source LLM developments.

It has been observed that ChatGPT has other emergent characteristics than writing plausible essays. It lies. It apparently has no process to try to determine the difference between how the world is (‘facts’) and how the world is not (‘falsehoods’). Indeed, there are subtleties, even difficulties, with the notion of ‘fact’ which make this a non-trivial process. But, first, it is well to discuss what is meant here by ‘lie’. A human lie is a falsehood knowingly presented as truth by a human. A lying human is exhibiting an intention, namely an intention to mislead. An LLM does not do this, of course, because there is no notion of intention that can reasonably apply to the current generation of LLMs, which are just formal mechanisms operating according to the precepts indicated above. But the procedure does issue output, and does so deliberately in the sense that its programmers intended it to do so. And in this deliberate output there may be falsehoods and its programmers recognise well that this may be the case [Solaiman et al 2019]. This is surely a process which has the same effect on a correspondent of the LLM agent as a lie would in correspondence with a human agent. It is this phenomenon which I wish to address, and it seems appropriate to call it lying, because of its congruence with lying in a human agent. Pursuing the anthropomorphism, it might be said that ChatGPT ‘recklessly lies’ – a term suggested by Nicholas Bohm – since it employs no mechanism to adjudge veridicality of its output (the philosopher Harry G. Frankfurt notably called this not lying, but ‘bullshitting’ when carried out by a human [Frankfurt 2005]). Other terms that have been used are ‘hallucinating’ and ‘confabulating’ [Economist 2023.2]. These and other such terms are equally anthropomorphic in that they refer to physiological and mental mechanisms that manifestly are not present in the case of ChatGPT. For this reason I will use ‘reckless lying’, but introduce a special term for it: r-lying/r-lie.

Eliminating lying, or the chance to lie, is central to judicial processes, for reasons which I shall not rehearse here. Introducing LLMs with the capacity to r-lie into such processes therefore runs the risk of subverting important judicial processes. There thus arises the need to filter LLM output somehow for veridicality, for facticity. I have not been able to determine if there are significant AI research projects actively working on filtering for r-lies, but the issue has been recognised [Evans and others 2021, Lin and others 2022].

An example of an LLM R-Lying

A colleague is internationally a well-known specialist in software reliability, and a prestigious international prize-winner, with a citation score amongst the highest in his field. He asked ChatGPT to write a 1,000 word obituary, with input similar to ‘to write an obituary for Prof Bev Littlewood, intended for a science journal, and in less that 1000 words’. He made the request twice, with slightly different wording, but equivalently short and simple [Littlewood 2023].

The initial output got his age wrong by 10 years. It said he studied at Cambridge (it was actually Imperial College, London), although it got the subject (mathematics) right. The material on his scientific work was roughly right, although rather perfunctory.

A few days later he reran the request with slightly different input wording. It again got his age wrong, this time by 5 years. It said he took his first degree now from the University of Nottingham. It said he had become a Fellow of the Royal Society in the early 2000s (not so; he is not FRS), and had been awarded the IEEE John von Neumann medal (also no, although he was a member of the awarding committee for several years). It said he worked at the University of Manchester, where he was head of the Software Engineering Group from 1985 until retirement in 2012. No; he was never associated with the University of Manchester. He was at City University, London (now City, University of London).

In both interactions it missed the IEEE Harlan D. Mills Award, which he received in 2007. This is noteworthy, as well as particularly odd, because the IEEE Mills Award pages are readily available [IEEE Mills Award no date] and a Google search for ‘Bev Littlewood’ provides his IEEE biography as the third entry [IEEE no date], which explicitly includes the Mills Award. In this instance, the available Google technology apparently does better than ChatGPT.

There is also another biography page for ‘Bev Littlewood’ at City, University of London, which is the first to appear in a Google search [City, Littlewood bio no date]. This page includes his degrees and thus his (correct) alma mater, as well as his place(s) of work. Again, Google does better than ChatGPT.

A reviewer of this paper knowledgeable about LLMs tried a similar experiment for a variety of ‘luminaries’ and reported that ‘it was unable to provide specific details of matters such as dates, appointments, and places. It however was able to generalize about the specific field of contribution of that individual, which it got correct all the time.’ This indicates that the general phenomenology of LLM responses in such cases may be replicable - that it is an emergent property of such queries that LLMs r-lie about specifics while getting generalities more ‘right’.

Pragmatics and conversational implicatures

An interaction with ChatGPT takes the general form of a dialogue, which may include monologic parts, as in the response to Bev’s request for his obituary. It has been recognised for decades that human communication of this sort is effected not just through the meanings of parts of speech and their combinations (words and sentences), but also through implicit, general expectations, called conversational implicatures by H. Paul Grice, who was the first to address this issue in the late 1960s [Grice 1989]. Nowadays the study of contextual implicatures is called *pragmatics*. A brief introduction to Gricean implicatures (less than 1,000 words) is to be found online in section 3 of [Grandy Warner 2021].

The overriding pragmatics of informative dialogue is proposed in Grice's Cooperative Principle: 'Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.' [Grice 1989]. In concert with this principle, Grice identifies four categories of specific implicature: Quantity, Quality, Relation and Manner [Soames 2003, Chapter 9: Language Use and the Logic of Conversation]. Quantity includes the maxims (in our words) 'make your contribution as informative as required by current purposes (i.e., don't say too little)'; and 'don't make your contribution more informative than is required by current purposes (i.e., don't say too much)'. Quality includes the maxims 'don't say what you believe to be false (i.e., don't lie)'; and 'don't say that for which you lack adequate evidence (i.e., don't overstate)'. There is one maxim of relevance: 'be relevant'. There are four maxims of manner, which need not concern us. The point of these maxims is not that they invariably hold in all dialogue (for example, people often lie, and overstate, contrary to the maxims of quality), but that the dialogue is governed as if they do.

Two general examples will suffice to show how implicatures operate. First, someone A is standing by his car at the roadside. Another driver B stops: 'You OK?' A: 'I'm out of gas (also referred to as petrol).' B: 'There's a gas station just round the corner.' A: 'Thanks'. The implicatures fill this out: A can infer that B thinks the gas station will be open and operating, and also that it is likely (but also that she doesn't know) that, say, B can fill a gas can there with enough gas to start his car and drive it in to fill up. Second, philosopher A is writing a letter of recommendation for one of his graduates for an academic job in philosophy: 'Dear Sir or Madam, X's command of the English language is very good, and he is always punctual to tutorials.' Given the maxim of relevance, it is notable that A is not saying how good X is at philosophy. Given the first maxim of quantity, a recipient could expect A to say how good X is at philosophy. Since A is not saying that, a recipient could infer that A does not think X is very good at it.

It is important to note that these aspects of what a dialogue participant infers from the interchange are not given by the semantics of the sentences used. It follows that there is such a subject as pragmatics, and there are such things as dialogue implicatures, no matter whether one thinks Grice and other investigators have them right, or not.

Fundamental to our present concern are the Maxims of Quality. These are what ChatGPT violated in its interactions with Professor Littlewood, and equally are the phenomena that set him aback. For good reason, as I have just noted. It is to be observed that the wording of typical witness oaths in English courts is congruent with the Maxims of Quality: '…. the truth, the whole truth, and nothing but the truth' [NIdirect NoDate]. There is good reason for this in both cases.

Interestingly, the Maxim of Relation says 'be relevant'. Here. ChatGPT seems to excel. When asked for an obituary of 'Bev Littlewood', it indeed returns details associated with Professor Littlewood, rather than, say, a disquisition on hyperbolic geometry.

The status of conversational implicatures

Many of us (including myself and Professor Littlewood) hold it to be important that computer programs are reliable, that is, that they generally give the 'right' answers to queries and demands posed to them. What is 'right' is generally given by the requirements for the program, and specialists such as ourselves usually require that the requirements for a program are given explicitly in a document called a 'requirements specification'. Several of us have written in this journal on requirements and reliability in relation to computer systems which feature in legal proceedings, e.g., [Ladkin 2020, Ladkin and others 2020].

A reliable computer system will adhere to its requirements specification (more or less, depending on complexity and other features). However, a conversational agent such as ChatGPT has no such 'specification'. It is intended to respond to (mostly) human queries in the manner in which humans engage in conversation, dialogue, monologue or soliloquy, and such activity is crudely governed by pragmatics such as conversational implicatures. It is important to note that implicatures are not requirements: people do lie in conversation; they also exaggerate; both are contrary to the Maxims of Quality. This 'violation' does not mean that the conversation is somehow 'invalid', as a violation of a requirements specification would invalidate the output of reliability-sensitive computer programs. A used car salesman may violate the Maxims of Quality in all sorts of ways, but it is surely well-known that buyers should beware.

Conversational implicatures, then, govern in some way the activity of conversational agents, but they do not have the status of requirements as they are usually known and worked in software engineering.

Why it matters to the law that LLMs lie

First, lawyers, including the editor of this journal, pointed out that there are checks and balances applying to trials which would preclude the use of r-lies generated by LLMs in cases argued in court. Lawyers adhere to a Code of Conduct, and follow Rules of Evidence, when involved in trials in court and their associated processes. These entail that a specific lawyer or lawyers is/are responsible for the veridicality of assertions made in court contexts.

However, as this article was in press, an account was published of LLM-generated r-lies being used in a court case in the US [Weiser 2023]. In this case, the falsehoods were discovered in court, it seems rather easily, substantiating the view that the checks and balances do work in some instances. The issue which arises is how one might know when they didn't.

Trials are not the only legal processes, and other processes may be vulnerable to use of r-lying LLMs. I set out below some hypothetical examples of situations in legal processes which r-lying LLMs could subvert, or at least impede progress.

1. LLMs might well come to be used in legal processes. For example, I understand that the process of ‘discovery’ in civil lawsuits in the US currently involves highly-automated online-document filtration processes. Disclosed documents are accompanied by a brief statement of their relevance. Such statements could well be produced by LLMs such as ChatGPT. Were they to include r-lies, this would subvert the purpose of the discovery process, which is to produce evidence and disclose its relevance to the case being tried.

2. I have worked as a technical expert with barristers on documents in which they were actively constructing arguments. One draft, as I recall, had some 400 paragraphs and was over 100pp long and was being constructed by a junior barrister on the basis of his reading and understanding of a lengthy technical document. It was a case expecting to be brought to arbitration, not before a court.

The draft took the barrister considerable (we may presume: expensive) time to produce. Today, we are on the cusp of a process by means of which he could well have used an LLM to help, say by giving it a one-sentence ‘hint’ for each of the 400 or so paragraphs of what he thought should be an argument, seeing what is written, and tweaking the result. One can imagine that would take days rather than weeks. The contents of each paragraph may well be not as important as the overall argument (consisting of the one-sentence ‘hints’, given by a human). The results would in this case be evaluated by an arbitration panel, as well as the opposition lawyers. There would be no obvious sanction for putting in a paragraph with justification including r-lies, except if the opposition or the panel notices. If there is enough material that evaluators do not have resources to check thoroughly, then r-lies could well remain undiscovered, even though their presence inhibits the argument presented (which we take as obvious – the Editor has advised me that such an occurrence could conflict with the UK Bar Code of Conduct requirement for lawyers to act with ‘honesty, and integrity’ – this may well be so, and I defer on such matters.)

3. I have also been involved in legal proceedings in which the plaintiff filed enormous amounts of material of dubious relevance, while not particularly endeavouring to establish the connection. This occurred in a Roman-law jurisdiction. It is a nuisance tactic, since everything in contention has to be explicitly answered by the respondent as a matter of prudence. The way that cases are handled in this jurisdiction is that (a) someone files a complaint; (b) then there is an exchange of briefs until the judge decides that all pertinent matters have been adequately addressed by both sides (basically when the exchange winds down); and then (c) the judge sets a date for a hearing, within which first a mediation/reconciliation is attempted, and then if that does not succeed the case is presented and adjudicated. Large amounts of material of unclear relevance involve considerable effort to sort and clarify. This effort may, in practical terms, not be available in a given case. This can, first, lead to distortion of the reconciliation effort, since a fair offer of settlement may not be recognised as such, and second, lead to an unsatisfactory judgment.

Filing large amounts of material of dubious relevance, including claims of relevance substantiated by r-lies, is the kind of task which can be easily supported through use of an LLM.

4. Courts often decide ‘what really happened’. Fraser J judged, in the ‘Horizon Issues’ findings in *Bates v Post Office Ltd (No 6: Horizon Issues) Rev 1* [Bates 2019], not that Horizon operated essentially bug-free, as contended by a witness in some prosecutions, but that there were a lot of ‘known errors’, some of which could well have led to the internal accounting discrepancies on the basis of which subpostmasters and mistresses were charged, in the main, with false accounting and prosecuted. There was an issue of ‘how the world really was’, the ‘facts of the case’, which the judgment decided. The Gricean Maxims of Quality were fundamental in this instance.

The ‘Information Space’

Some of the issues addressed in this paper have arisen in other contexts and have developed their own terminology, well before ChatGPT was released. I give some pointers to these other contexts and significant terms.

The ‘information space’ is a term used by, for example, the Institute for the Study of War in its detailed daily commentaries on the Ukrainian invasion [ISW no date]. It concerns the framing of commentary and information/misinformation/disinformation on specific general subjects. ISW has reported on what phenomena in the ‘information space’ is likely to speak to tactics, strategies and political motivations of the various contributors, which is one kind of ‘open-source intelligence’ which ISW uses in its commentaries. Generally speaking, authoritarian governments and governmental entities often try to affect the ‘information space’. There has been a lot of interest in Russian attempts to influence the ‘information space’ around the US 2016 presidential election as well as the German 2021 Bundestagswahl. Such attempted influence of core democratic processes has judicial consequences in most ‘Western’ jurisdictions.

The term ‘infodemiology’ was coined by the World Health Organisation in 2020 to describe the active study of non-veridical information propagating about the Covid-19 pandemic [WHO 2020, Zielinski 2021]. The term surely applies to all informational phenomena, not just those concerning Covid-19. We could adopt the term ‘infodemiology’ as the study of information and misinformation in the information space.

I take it to be clear that we need terms for these entities and phenomena and their study, and these are surely as good as any. The subject of this paper could be phrased as some comments on how use of LLMs might affect the infodemiology of legal processes.

Conclusion

ChatGPT and other ‘LLMs’ now being mooted (and installed) as ‘next generation search engines’ in the information space constituted by all Internet usage apparently have no mechanisms to ensure or to control veridicality, and appear often to r-lie. Because of the importance of veridicality in legal as well as other social processes, as noted in the linguistic and philosophical study of conversational pragmatics, as well as the possibilities that LLMs could be used in various legal processes [Vos 2023; Sales 2023], I see this as an issue to be actively addressed, if necessary through government intervention.

Postscript

As this article was in press, I was advised of [Kirkham 2023], which considers some problems with tribunals using information technology in the English legal system. Tribunals are decision processes generally involving advocacy performed by non-lawyers. There may also be examples here of how use of potentially r-lying LLMs could adversely affect outcomes.

Peter Bernard Ladkin is a systems-safety specialist with a background in software dependability and logic. His causal accident analysis method Why-Because Analysis (WBA) is used by some 11,000 engineers worldwide. He taught at Bielefeld University and is CEO of tech-transfer companies Causalis Limited and Causalis Ingenieurgesellschaft mbH.

Acknowledgements
I am grateful to Bev Littlewood for permission to describe his experiences using some of his own words; to Nicholas Bohm for editorial suggestions and the first example in ‘Why It Matters’; and to an anonymous reviewer for performing further experiments and suggesting how such LLMs might behave in general. I am grateful for continuing discussion on current ML/AI software and its possible influence on society to Nicholas and Bev, as well as to Paul Marshall, Stephen Mason, Harold Thimbleby, and Martyn Thomas.

References

[Bates 2019] *Bates v Post Office Ltd (No 6: Horizon Issues) Rev 1* [2019] EWHC 3408 (QB), available at https://www.bailii.org/ew/cases/EWHC/QB/2019/3408.html

[City, Littlewood bio no date] Bev Littlewood Bio, City, University of London, no date, available at https://www.city.ac.uk/about/people/academics/bev-littlewood

[Economist 2023] The Economist, The generation game, edition of April 22nd – 28th, 2023

[Economist 2023.2] The Economist, The Age of Pseudocognition, edition of April 22nd – 28th, 2023

[Evans and others 2021] Truthful AI: Developing and governing AI that does not lie, preprint 2021-10-13, available at https://arxiv.org/abs/2110.06674

[Frankfurt 2005] Harry G. Frankfurt, On Bullshit, Princeton University Press, 2005

[Grandy Warner 2021] Richard E. Grandy and Richard Warner, Paul Grice, Stanford Encyclopedia of Philosophy, available at https://plato.stanford.edu/entries/grice/

[Gratas 2023] Brenda Gratas, You Need to Know (invgate blog), 2023-02-14, available at https://blog.invgate.com/chatgpt-statistics

[Grice 1989] H. Paul Grice, Studies in the Way of Words, Harvard University Press, 1989

[IEEE no date] IEEE Computer Society, Profile of Bev Littlewood on computer.org, available at https://www.computer.org/profiles/bev-littlewood

[IEEE Mills Award no date] IEEE Computer Society, Harlan D. Mills Award, available at https://www.computer.org/volunteering/awards/mills

[ISW no date] Institute for the Study of War, Russian Offensive Campaign Assessment, daily briefings, available through https://www.understandingwar.org

[Kirkham 2023] Reuben Kirkham, The Ethical Problems with IT "Experts" in the Legal System, IEEE Computer, 56(6):62-71, June 2023. DOI 10.1109/MC.2022.3209998.

[Ladkin 2020] Peter Bernard Ladkin, Robustness of Software, Digital Evidence and Electronic Signature Law Review 17, 2020, available from https://journals.sas.ac.uk/deeslr/issue/view/578

[Ladkin and others 2020] Peter Bernard Ladkin, Bev Littlewood, Harold Thimbleby, Martyn Thomas, The Law Commission presumption concerning the dependability of computer evidence, Digital Evidence and Electronic Signature Law Review 17, 2020, available from https://journals.sas.ac.uk/deeslr/issue/view/578

[Lewkowycz and others 2022] Aitor Lewkowycz, Anders Andreassen and others, Solving Quantitative Reasoning Problems with Language Models, preprint available at https://arxiv.org/pdf/2206.14858.pdf

[Lin and others 2022] Stephanie Lin, Jacob Hilton, Owain Evans, TruthfulQA: Measuring How Models Mimic Human Falsehoods, preprint 2022-05-08 (Version 2), available at https://arxiv.org/abs/2109.07958

[Littlewood 2023] Bev Littlewood, private Email correspondence with colleagues, 2023-04-07

[NIdirect NoDate] HMG, Giving Evidence in Court, Subsection Oaths and Affirmations, available at https://www.nidirect.gov.uk/articles/giving-evidence-court

[Ruby 2023] Molly Ruby, How ChatGPT Works: The Model Behind the Bot, blog post, https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286 2023-01-30

[Sales 2023] Lord Sales, Justice of the Supreme Court (UK), Information Law and Automated Governance, Keynote Address, Information Law Conference, Institute of Directors, London 2023-04-24, available at https://www.supremecourt.uk/docs/Information-Law-and-Automated-Governance-L-Sales-keynote-address-Information-Law-Conference-April-2023.pdf

[Soames 2003] Scott Soames, Philosophical Analysis in the Twentieth Century Volume 2: The Age of Meaning, Princeton University Press, 2003

[Solaiman et al 2019] Irene Solaiman, Miles Brundage et al, Release Strategies and the Social Impacts of Language Models, OpenAI Report, November 2019, available at https://arxiv.org/pdf/1908.09203.pdf

[Vos 2023] The Right Hon. Sir Geoffrey Vos, The Master of the Rolls (UK), The Future of London as a Pre-Eminent Dispute Resolution Centre: Opportunities and Challenges, The McNair Lecture, Lincoln’s Inn, London 2023-04-19, available at https://www.judiciary.uk/speech-by-the-master-of-the-rolls-the-future-of-london-as-a-pre-eminent-dispute-resolution-centre-opportunities-and-challenges/

[Weiser 2023] Benjamin Weiser, Here’s What Happens When Your Lawyer Uses ChatGPT, The New York Times, 2023-05-27. Available at https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html, accessed 2023-06-04.

[WHO 2020] World Health Organisation, 1st Infodemiology Conference, 2020. Announcement available at https://www.who.int/teams/epi-win/infodemic-management/1st-who-infodemiology-conference

[Zielinski, 2021] Chris Zielinski, Zielinski C. Infodemics and infodemiology: a short history, a long future. Rev Panam Salud Publica, the Pan American Journal of Public Health 2021;45:e40, available through https://doi.org/10.26633/RPSP.2021.40