To protect science we must use LLMs as zero-shot translators

Brent Mittelstadt,*[1] Sandra Wachter[2] and Chris Russell[3]

Standfirst

LLMs (large language models) do not distinguish between fact and fiction. They will return an answer to almost any prompt, yet factually incorrect responses are commonplace. To ensure our use of LLMs does not degrade science, we must use them as zero-shot translators: to convert accurate source material from one form to another.

Introduction

In recent months much has been said about the opportunities and risks posed by large language models (LLM) and other types of generative AI. Widely acknowledged problems with misinformation and hallucinations stem from how these models are built, fine-tuned, and used.[1] Our tendency to anthropomorphise machines and trust models as human-like truth tellers,[2] consuming and spreading the bad information they produce in the process, is uniquely worrying for the future of science.

Why generated nonsense is believable

Today's popular LLMs like ChatGPT and Bard are text-prediction engines, which are fine-tuned via reinforcement learning from human feedback (RLHF) to sound human and be useful in answering human-generated prompts or questions. They are not, strictly speaking, designed to tell the truth. Sounding truthful is only one element by which the usefulness of these systems is measured; equally important are other characteristics like "helpfulness, harmlessness, technical efficiency, profitability, [and] customer adoption."[3]

LLMs are designed to produce helpful and convincing responses without any overriding guarantees regarding their accuracy or alignment with fact. Their perceived utility is as much a function of the content of responses as it is due to how they are designed to engage with users who are encouraged to ask questions, receive answers, seek clarification, and provide feedback as if they were chatting with another person.[4] Outside of certain carefully constructed guardrails to avoid sensitive topics and toxic content, models readily provide answers to any question or prompt. Responses are rarely tempered by linguistic signals of confidence or expertise.[5]

When paired with human tendency to attribute meaning and intent to spoken or written word, misunderstanding is inevitable. The mere fact that someone or something produces language implies that there is a "speaker"; this in turn may lead us to infer understanding, intent, and consciousness.[2] We also equate intelligence with the usage of language, and this may lead users to focus more on the times LLMs generate accurate output, and ignore failures and hallucinations.[4]

[1] Oxford Internet Institute, University of Oxford, 1 St. Giles, Oxford, OX1 3JS, United Kingdom. E-mail: brent.mittelstadt@oii.ox.ac.uk,
[2] Oxford Internet Institute, University of Oxford, 1 St. Giles, Oxford, OX1 3JS, United Kingdom.
[3] Oxford Internet Institute, University of Oxford, 1 St. Giles, Oxford, OX1 3JS, United Kingdom.

The problem with trusting LLMs

But why should the outputs of LLMs not be trusted? In many instances outputs will be correct or based in fact. This is not, however, due to any strict requirements for LLMs to provide truthful outputs. The concept of "truth" has been highly simplified in LLM development and equated with accuracy measured against the training data's "ground truth."[6] Individual factual responses may result from the model drawing on ground truth found in its training data or injected through fine-tuning and human feedback, for example through human-authored correct responses written for difficult prompts, or by annotators indicating a preference for machine-generated responses that contain factual information.

Neither of these are particularly robust means to create reliably truthful LLMs. LLMs are trained on large datasets of text, often scraped from the Internet, and tasked with predicting the next most likely string of text in response to a prompt. Training data may contain empirically true strings of text, but equally be filled with false statements, opinions, jokes, creative writing, series of instructions, or other texts that are not factual or concerned with truth. The utility of responses is defined through consensus—the more often a string appears in the data or has been written on the Internet, the more likely it is to be chosen in a response—what has been referred to elsewhere as "common token bias."[3] Responses thus merely reflect a disjointed, post-hoc consensus among public sources. It is in this sense that LLMs promote a relativistic, consensus-based approach to truth.

Human feedback is similarly unreliable. Techniques such as RLHF, which trains a reward model based on human feedback to improve the perceived helpfulness of LLM outputs, can be expected to penalize obviously untrue statements. However, there is no requirement for the resulting LLM to generate truthful responses to complex queries. This apparent reliability in things that are easily verified magnifies the mistruths, half-truths, and hallucinations users are willing to accept or unwilling to question through extended use.

Annotators instead fill the role of experts in rhetoric, tasked with picking out the most "human sounding" or useful responses. They cannot assess the truthfulness of outputs beyond their personal knowledge and areas of expertise. Human feedback is not the same as expert feedback drawing on some objective body of knowledge. For current general purpose LLMs, training data and system outputs are not empirically fact checked or scientifically validated in any consistent, reliable sense at the model or feedback level.

Even in cases where expert annotators are used (for example, qualified lawyers annotating responses to legal prompts), accurate responses cannot be guaranteed due to the subjectivity of expertise, context-sensitivity and specificity of true or correct answers, and the built-in randomness of LLM outputs to make them appear more human-like.[7] Fine tuning models to adapt the confident speech patterns of expertise without any of the grounding in truth or professional responsibility only exacerbates this problem.[8]

Obvious hallucinations are not the primary epistemic risk; rather, it is subtle inaccuracies, oversimplifications, or biased responses passed off as truth in a confident tone,[9] which can convince experts and non-experts alike, that pose the greatest risk in research, science communication, and education.[4] As long as LLMs remain decoupled from ground truth and incapable of genuinely expressing uncertainty and confidence, they should be treated as inherently unscientific.

Constraining the use of LLMs

We need to set clear expectations about what LLMs can responsibly contribute to scientific endeavours. In expert hands LLMs can be helpful research assistants who, like their human

counterparts, must be fact checked before being trusted. Equally, without critical human oversight, they can be dangerous but convincing tools.

While there is a vast amount of ongoing work into increasing the veracity of LLMs and their outputs, to date there is no indication of easy technological answers.[3] To make progress we should not think of veracity solely in terms of products, or the quality of outputs and the methods used to produce them. Veracity must also be treated as a supply side challenge. Hallucinations, biased and inaccurate content can be minimised by constraining how we formulate questions and prompts for LLMs.

For example, imagine asking a LLM to respond with a summary of the General Data Protection Regulation (GDPR) (Figure 1). If we understand LLMs as incidental truth-generators, all factual elements of the generated summary should be treated with suspicion, but its format and style can still be useful. Figure 1 shows a schematic for generating such a summary using an unstructured query versus a translation query. When submitted to ChatGPT 3.5 the unstructured query produces a response with a subtle inaccuracy about the necessity of explicit consent for processing personal data. To avoid being misled in this scenario, we can instead provide the LLM with a vetted list of facts about the GDPR from the Information Commissioner's Office (ICO), the UK data protection regulator, and query ChatGPT to convert it to a two paragraph summary.

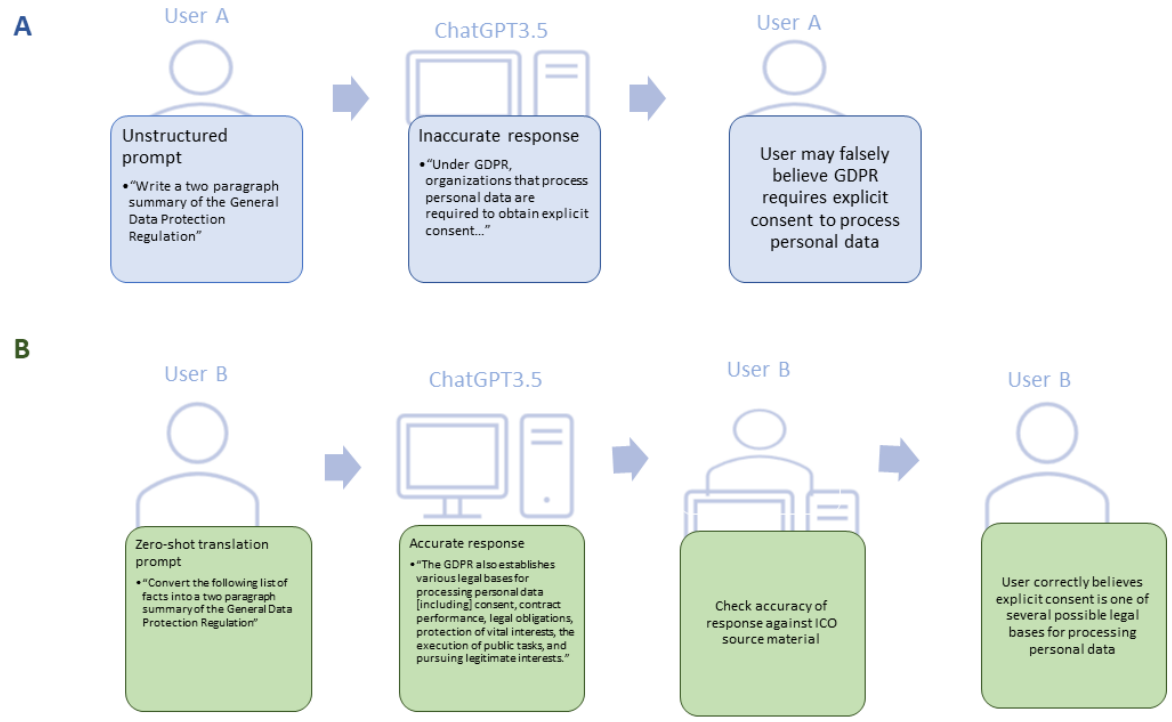

Figure 1 - This schematic depicts two hypothetical use cases for LLMs based on real prompts and responses to demonstrate the impact of inaccurate responses on user beliefs. User A submits an unstructured prompt to the LLM requesting a summary of the GDPR. They receive a response that incorrectly claims explicit consent is required to process personal data under the GDPR, resulting in inaccurate knowledge. User B submits a zero-shot translation prompt requesting conversion of a list of vetted facts about the GPDR into a summary. The response correctly identifies consent as one of many possible legal bases for processing personal data, resulting in accurate knowledge. The prompts were written by the authors and submitted to ChatGPT 3.5 on October 1, 2023. Only the relevant inaccurate (User A) and accurate (User B) segments of the responses generated by ChatGPT are quoted. User B's list of vetted facts included a list of legal bases for processing personal data adapted from GDPR guidance published by the ICO, available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/a-guide-to-lawful-basis/.

This type of task, where relevant information is provided to the LLM to be transferred to a new domain, is called "zero-shot translation."[10] Here, "zero-shot" indicates that the choice of source material and the domain is made by the end-user and the LLM was not particularly trained with it in mind. In current literature "zero-shot translation" is used to refer to translation between two human

languages. However, our usage here is intentionally broader to refer to the transformation of information from one form to another. Many successful use cases, such as translating short English text into simple Python code, hinge upon a combination of zero-shot translation, verification by programmers manually examining the code, and feeding errors from running the code back to the model.[10]

Zero-shot translation grounds the output of an LLM in factual or reliable human-authored content (e.g., working code, vetted facts). The user provides both facts and intent to the model which allows the output to be more easily scrutinised for the introduction of new information and hallucinations. This is not, however, a silver bullet for truth in LLMs. Care must still be taken as successful zero-shot translation relies on knowledge of the world, and hallucinations and false information has been shown to still creep into responses even in apparently well-defined tasks.[10] If asked to write code using an obscure library, for example, LLMs sometimes generate plausible calls to functions that do not exist. Information about the library API must be provided as part of the translation task (N.B. This reference is a non-peer reviewed pre-print).[11]

How LLMs can be useful

Assuming usage of LLMs in science is restricted to zero-shot translation tasks, a question remains: can LLMs still be useful?

To understand why the answer is likely "yes," it is helpful to note that the automation of many tasks targeted by AI systems owes more to the building of infrastructure than to AI itself. This tendency often results in piecemeal infrastructure whereby specific tasks are solved well, but human labour is pushed to the edges of the task,[12] for example "data plumbing" that ensures input data is in the right format or aggregating the output of multiple systems. Such tasks exist in a grey zone where they could be automated, but shifting requirements and the overheads and challenges of automation delay or defer the replacement of human labour.[13] This phenomenon is particularly apparent in scientific laboratories, where equipment is often collected in a piecemeal fashion. This results in a scientific workflow where it is not only samples that have to be processed and adapted as they move from one near compatible machine to another, but also the data itself.[14]

Data transfer in scientific workflows, and, more generally, data plumbing tasks are prime candidates for automation by LLMs. Many are already close to being automated and are well-suited for zero-shot translation because they require simply changing information from one form to another and currently rely upon humans for data entry. Other scientific tasks where zero-shot translation could be useful include expanding bullet-points into coherent prose when writing reports (see: Figure 1), turning data into code that generates a specific graph, or rewriting text in simplified or accessible language for outreach purposes.

In each of these tasks, while LLMs can lead to a significant increase in productivity, the use of zero-shot translation is key. It allows for scrutiny of the input data for correctness and confirmation that the output is consistent with the input and errors have not been introduced by the LLM. It enables the connection between queries, responses, and ground truth to be re-established for the sake of scientific rigour. In other words, zero-shot translation can ensure LLMs tell the truth.

Conclusion

Ultimately, the way in which LLMs are used determines the resulting harms. We encourage users of LLMs for scientific and education purposes to change how they use these systems. For tasks where the truth matters, we encourage users to write translation prompts that include vetted, factual

information. While zero-shot translation undoubtedly restricts LLMs as knowledge generators, it avoids the substantial epistemic risks to science and society that come from using unconstrained LLMs to freely generate and disseminate findings, references, and whole scientific articles.

We likewise encourage developers of LLMs to redouble their efforts to improve the reliability and transparency of model responses.[15] Sources and indicators of confidence and uncertainty should be included in responses by default. Downsizing the size and scope of models should also be considered, including restricting the types of valid prompts to areas for which the model has reliable training data or can pass an accepted confidence threshold.[1]

Once we know what can be automated, it is a separate question as to what we want to automate. Do we actually want to reduce opportunities for writing, thinking critically, creating new ideas and hypotheses, grappling with the intricacies of theory, and combining knowledge in creative and unprecedented ways? These are the inherently valuable hallmarks of curiosity-driven science.[8] They are not something that should be cheaply delegated to incredibly impressive machines that remain incapable of distinguishing fact from fiction.

References

1. Bender, E. M., Gebru, T., McMillan-Major, A. & 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* 610–623 (ACM, 2021). doi:10.1145/3442188.3445922.
2. Mitchell, M. How do we know how smart AI systems are? *Science* **381**, adj5957 (2023).
3. Munn, L., Magee, L. & Arora, V. Truth machines: synthesizing veracity in AI language models. *AI Soc.* (2023) doi:10.1007/s00146-023-01756-4.
4. Kidd, C. & Birhane, A. How AI can distort human beliefs. *Science* **380**, 1222–1223 (2023).
5. Mielke, S. J., Szlam, A., Dinan, E. & Boureau, Y.-L. Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. *Trans. Assoc. Comput. Linguist.* **10**, 857–872 (2022).
6. Kang, E. B. Ground truth tracings (GTT): On the epistemic limits of machine learning. *Big Data Soc.* **10**, 20539517221146122 (2023).
7. Holtzman, A., Buys, J., Du, L., Forbes, M. & Choi, Y. The curious case of neural text degeneration. in *Proceedings of ICLR 2020* (2020).
8. Drezner, D. W. *The ideas industry*. (Oxford University Press, 2017).
9. Feng, S., Park, C. Y., Liu, Y. & Tsvetkov, Y. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 11737–11762 (Association for Computational Linguistics, 2023). doi:10.18653/v1/2023.acl-long.656.
10. Johnson, M. *et al.* Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Trans. Assoc. Comput. Linguist.* **5**, 339–351 (2017).
11. Kabir, S., Udo-Imeh, D. N., Kou, B. & Zhang, T. Who Answers It Better? An In-Depth Analysis of ChatGPT and Stack Overflow Answers to Software Engineering Questions. Preprint at https://doi.org/10.48550/arXiv.2308.02312 (2023).
12. Graeber, D. Bullshit jobs. *E Mploi* 131 (2018).
13. Lin, J. & Ryaboy, D. Scaling big data mining infrastructure: the twitter experience. *ACM SIGKDD Explor. Newsl.* **14**, 6–19 (2013).

14. Liew, C. S. *et al.* Scientific Workflows: Moving Across Paradigms. *ACM Comput. Surv.* **49**, 66:1-66:39 (2016).
15. Bender, E. M. & Friedman, B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Trans. Assoc. Comput. Linguist.* **6**, 587–604 (2018).

Competing interests:

BM and SW declare no competing interests. CR was also an employee of Amazon Web Services during part of the writing of this article. He did not contribute to this article in his capacity as an Amazon employee.

Acknowledgments

This work has been supported through research funding provided by the Wellcome Trust (grant nr 223765/Z/21/Z), Sloan Foundation (grant nr G-2021-16779), the Department of Health and Social Care, and Luminate Group to support the Trustworthiness Auditing for AI project and Governance of Emerging Technologies research programme at the Oxford Internet Institute, University of Oxford. The funders had no role in the decision to publish or the preparation of this manuscript.