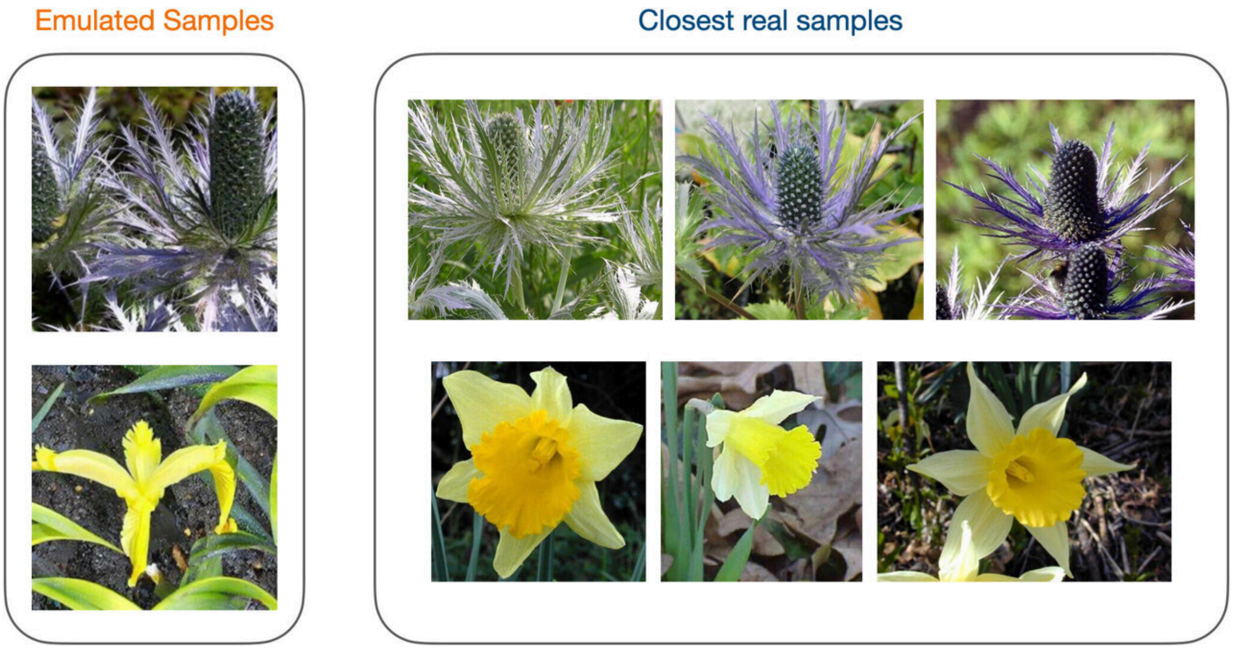
Tech Xplore

Q&A: Model disgorgement—the key to fixing AI bias and copyright infringement?

May 17 2024, by Ian Scheffler



Example of dataset emulation. We show samples of an emulated dataset of Oxford Flowers, which captures the original distribution while maintaining high CLIP distance from the original data. Credit: *Proceedings of the National Academy of Sciences* (2024). DOI: 10.1073/pnas.2307304121

By now, the challenges posed by generative AI are no secret. Models like OpenAI's ChatGPT, Anthropic's Claude and Meta's Llama have been known to "hallucinate," inventing potentially misleading responses, as well as divulge sensitive information, like copyrighted materials.

One potential solution to some of these issues is "model disgorgement," a set of techniques that force models to purge themselves of content that leads to copyright infringement or biased responses.

In a paper in *Proceedings of the National Academy of Sciences*, Michael Kearns, National Center Professor of Management & Technology in Computer and Information Science (CIS), and three fellow researchers at Amazon share their perspective on the potential for model disgorgement to solve some of the issues facing AI models today.

In the following Q&A, Kearns discusses the paper and its implications for improving AI.

What is model disgorgement?

Model disgorgement is the name for a broad set of techniques and the problems that those techniques are trying to solve. The goal is to mitigate or eradicate the effects of particular pieces of training data from the behavior of a trained model.

You expect individual pieces of training data or collections of training data to influence the behavior of the model. But this can lead to privacy leaks, copyright violations and other issues that aren't covered by the law yet.

How is model disgorgement different from efforts to ensure data privacy, like Europe's General Data Protection Regulation?

These are different but related concerns. If I ask Facebook to delete all of my stored Facebook activity from their servers, the GDPR requires that to be done on request.

Laws like the GDPR are less clear about what happens before your data is deleted. Your data was used to train a predictive model, and that predictive model is still out there, operating in the world. That model will still have been trained on your data even after your data is deleted from Facebook's servers. This can lead to a number of problems.

For one, if your data was private, a third-party adversary might be able to reverse-engineer sensitive aspects of your private data. This is certainly an instance where you would want model disgorgement techniques to remove that sensitive data from the model.

In addition, there are also issues with copyright, as we're seeing in The New York Times' lawsuit against OpenAI. ChatGPT can regurgitate verbatim copyrighted articles from the Times. It's pretty clear that OpenAI used those articles in training ChatGPT.

To be clear, the paper doesn't want those articles to be private; it wants the articles to be accessible to the public. But the Times also wants to control the articles' use and reproduction.

Finally, there's another issue that I might call "stylistic infringement," where a user can say, "Give me a painting in the style of Andy Warhol of a cat skateboarding in Rittenhouse Square." The model is able to do a good job because it's been trained on the entire output of Andy Warhol's career. If you're the executor of Andy Warhol's estate, you might take issue with this.

Even though these are very different issues, the technical ways of addressing them are quite similar, and involve model disgorgement techniques. In other words, it's not that model disgorgement is different from efforts to ensure data privacy, it's more that model disgorgement techniques can be used in certain situations where current approaches to privacy like the GDPR fall short.

The Ethical Algorithm, which you co-wrote with Aaron Roth, Henry Salvatori Professor of Computer & Cognitive Science in CIS, and which you recently referenced in the context of AI, describes how to embed ethical considerations into algorithm design. Would that approach be feasible with AI models?

When we wrote the book, generative AI didn't exist, at least not like it does today. Our book focused on traditional machine learning, which involves more targeted predictions—like taking the information on a loan application and coming up with an assessment of the risk that a particular person would default if given a loan.

When an application is that targeted, it becomes much more feasible to bake into the training process defenses against various harms that you're concerned about, like demographic bias in the performance of the model or leaking the private training data.

For now, we've lost that ability in training generative models because of the extreme open-ended nature of their outputs.

Would it be possible to filter the training data for AI models to reduce the likelihood of biased or copyright-breaching responses?

That's hard for a few reasons.

The way you train a competitive large language model is by scraping the entire internet—literally. That's table stakes. You also need a lot of other more proprietary data sources. When that is the starting point, there's so much you don't know about your training data.

In principle, we know how to train huge neural networks in a way that will avoid all of these problems. You can train a neural network under the constraint of differential privacy, a method of intentionally corrupting data to shield private information, for instance, and fewer of these problems will occur.

Nobody's tried. I think the general feeling is that the degradation in performance you would get by training a large language model under the constraint of differential privacy would kind of obviate the point in the first place.

In other words, the quality would be so bad that you'd start generating nonsensical, nongrammatical outputs. The amount of noise that you would need to add to the training process, which is how differential privacy works—it just wouldn't work at scale.

Can you provide a few examples of model disgorgement techniques? How do they work?

One conceptually straightforward solution is retraining from scratch. This is clearly infeasible given the scale and size of these networks and the compute time and resources it takes to train them. At the same time, retraining is kind of a gold standard—what you would like to achieve in a more efficient, scalable way.

Then there are "algorithmic" solutions. One of these is machine "unlearning." Instead of retraining the whole network, we could just modify it in some way that mitigates or reduces the effects of your data on the training process.

Another algorithmic approach is training under the constraint of differential privacy: adding noise to the training process in a way that minimizes the effects of any particular piece of training data, while still letting you use the aggregate properties of the data set.

Then there are what I might call system-level techniques. One of these is "sharding." If I divided my training data into 100 "shards," I could train a different model on each of those 100 shards and then produce an overall model by averaging those 100 models.

If we're lucky enough that your data was only in one of those 100 shards, and you wanted to remove your data, we could just remove that model entirely from the average. Or we could retrain just that model, which used only one percent of the overall data.

Your data's contribution to something like ChatGPT is quite minuscule. If you did a sharding approach, your data would likely fall entirely within one, maybe at most two, of these 100 shards.

The bigger concern is for really large data sets. How do you make sure that every organization whose data you're using is kind of only in one of the 100 shards?

To arrange this, you have to know what the organizations are in advance—and this gets back to my earlier point that often you don't know what's in your training data.

If my training data is some massive file, which is a crawl of the entire internet, and I break it into 100 pieces, I have no idea where Getty Images' data might be distributed among those hundred pieces.

If we could go back in time and change the way the internet was designed, could we make sure that every piece of data online was tagged or identified with

different levels of protection so that scraping the internet would yield metadata to inform what AI models can and can't use in training?

My gut reaction is that this approach might help solve the problems that we're discussing here, but would have possibly resulted in very different challenges elsewhere.

One of the great successes of the consumer internet was its openness and the lack of structure and rules for how data is organized and how data can cross reference other data. You could imagine setting up the rules differently. But you can also imagine the internet maybe never happening because it would just be too onerous to build on it.

The great success story of the internet has come from basically the lack of rules. You pay for the lack of rules, in the areas that we're discussing here today.

Most people who think seriously about privacy and security would probably agree with me that a lot of the biggest problems in those topics come from the lack of rules, the design of the internet, but that's also what made it so accessible and successful.

In short, it's hard to avoid these trade-offs.

In your recent paper, you and your co-authors organize the model disgorgement methods discussed above into a taxonomy, classifying them according to when they take action and how they work. What do you hope the paper offers future researchers and industry professionals?

It's a non-technical paper in many ways, and it's meant for a broader audience. We hope that the paper will help frame thinking about these issues—in particular, the trade-offs among the different technical methods for model disgorgement. This felt like a topic that was important enough societally and nascent enough scientifically that it was a good time to kind of step up and survey the landscape.

More information: Alessandro Achille et al, AI model disgorgement: Methods and choices, *Proceedings of the National Academy of Sciences* (2024). DOI: 10.1073/pnas.2307304121

Provided by University of Pennsylvania

Citation: Q&A: Model disgorgement—the key to fixing AI bias and copyright infringement? (2024, May 17) retrieved 18 June 2024 from https://techxplore.com/news/2024-05-qa-disgorgement-key-ai-bias.html