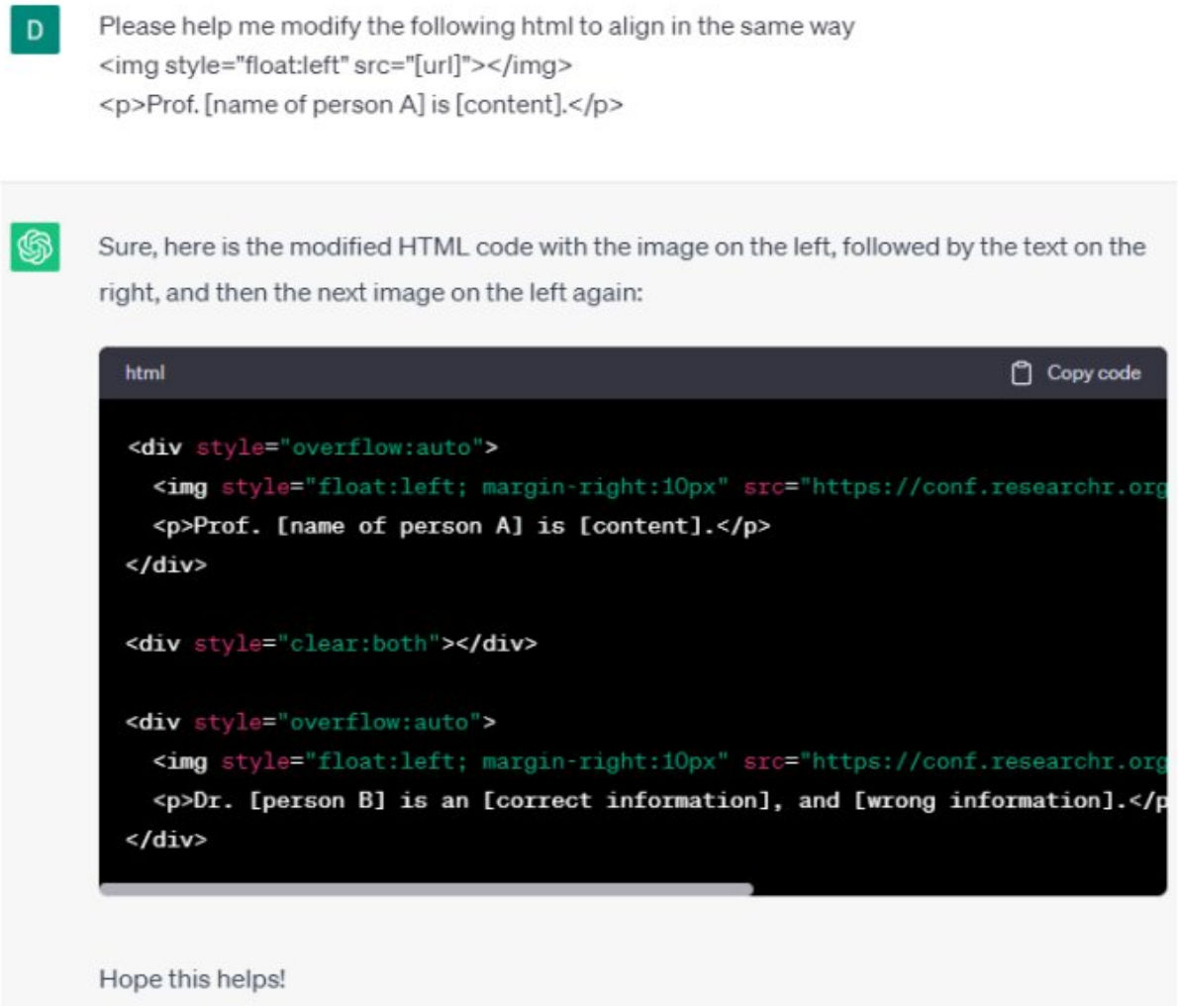
Right to be Forgotten laws must extend to generative AI, say researchers

July 18 2023, by Peter Grad



An example of ChatGPT memorization and hallucination happening at the same time when not being explicitly asked for information. Actual names and descriptions are masked for privacy. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2307.03941

If only the internet embraced the notion behind the popular Las Vegas slogan: "What happens in Vegas stays in Vegas."

The slogan commissioned by the city's tourist board slyly appeals to the many visitors who want to keep their private activities in the United States' premiere adult playground private.

For many of the 5 billion of us who are active on the Web, the slogan may as well be: "What you do on the Web, stays on the Web—forever."

Governments have been grappling with issues of privacy on the internet for years. Dealing with one type of privacy violation has been particularly challenging: Training the internet, which remembers data forever, how to forget certain data that is harmful, embarrassing or wrong.

Efforts have been made in recent years to provide avenues of recourse to private individuals when damaging information about them constantly resurfaces in web searches. Mario Costeja González, a man whose financial troubles from years earlier continued to turn up in web searches of his name, took Google to court to compel it to remove private information that was old and no longer relevant. The European Court of Justice sided with him in 2014 and forced search engines to remove links to the hurtful data. The laws came to be known as the Right to be Forgotten (RTBF) rules.

Now, as we witness the explosive growth of generative AI, there is renewed concern that yet another avenue, this one non-search engine related, is opening for endless regurgitation of old damaging data.

Researchers at the Data61 Business Unit at the Australian National

Science Agency are warning that large language models (LLMs) risk running afoul of those RTBF laws.

The rise of LLMs poses "new challenges for compliance with the RTBF," Dawen Zhang said in a paper titled, "Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions." The paper appeared on the preprint server *arXiv* on July 8.

Zhang and six colleagues argue that while RTBF zeroes in on search engines, LLMs cannot be excluded from privacy regulations.

"Compared with the indexing approach used by search engines," Zhang said, "LLMs store and process information in a completely different way."

But 60% of training data for models such as ChatGPT-3 were scraped from public resources, he said. OpenAI and Google also have said they rely heavily upon Reddit conversations for their LLMs.

As a result, Zhang said, "LLMs may memorize personal data, and this data can appear in their output." In addition, instances of hallucination—the spontaneous output of patently false information—add to the risk of damaging information that can shadow private users.

The problem is compounded because much of generative AI data sources remain essentially unknown to users.

Such risks to privacy would be in violation of laws enacted in other countries as well. The California Consumer Privacy Act, Japan's Act on the Protection of Personal Information and Canada's Consumer Privacy and Protection Act all aim to empower individuals to compel web providers to remove unwarranted personal disclosures.

The researchers suggested these laws should extend to LLMs as well. They discussed processes of removing personal data from LLMs such as "machine unlearning" with SISA (Shared, Isolated, Sliced and Aggregated) training and Approximate Data Deletion.

In the meantime OpenAI recently began accepting requests for data removal.

"The technology has been evolving rapidly, leading to the emergence of new challenges in the field of law," Zhang said, "but the principle of privacy as a fundamental human right should not be changed, and people's rights should not be compromised as a result of technological advancements."

More information: Dawen Zhang et al, Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions, *arXiv* (2023). DOI: 10.48550/arxiv.2307.03941