Privacy and Security Challenges of Large Language Models

Mingji Han

Introduction

Large Language Models (LLM) are machine learning models with billion level parameters, which are able to do general natural language processing tasks and text generation. Recent rise of LLMs have demonstrated powerful abilities in variou tasks including text summarization, code generation, and translation, mading significant impact in both industry and academia. ChatGPT from OpenAI has been the most popular. Other tech companies like Google, Meta, and Amazon have made efforts to build their own LLMs. When tech companies, we should notice that the training and development of LLMs made new challenges and risks in privacy and computer securities. In this paper, we discuss the privacy and security issues of LLMs.

Privacy Challenges

Nowadays machine learning models are data driven, which use large amounts of data to fit the model. For Large Language Models, the scale of training data is massive. The data could be obtained from the Internet. Tech companies may use user data to train their LLMs. Those data may include users' privacy including name, address, phone number, and SSNs. Including large scale data with user information could be a significant risks for privacy. Although there are regulations and laws requiring tech companies to protect the privacy when using the users' data. Even ML engineers use data filtering and masking, the risks of including users' private data in LLM's training data are not negligible. This is because user's privacy data could be represented in diverse formats and the scale of LLM training data is huge. Data filtering and masking may not be effective enough to remove all sensitive user data.

When the LLMs are deployed and provides services to users, there are also various privacy challenges. Firstly, users interact with LLMs using natural language. Users send text to LLM and LLM returns the generated text. Compared with other user-computer interactions modes including clicking buttons and slidings, this kind of interaction brings more risks of leaking user's privacy data. Nowadays users treat LLMs (ChatGPT) as their personal assistants to help them on various tasks like writing emails, complete code, and seeking for advices. To use LLMs for those tasks, users need to send texts which contains their privacy including daily routines, heath status, and personal relationship. Many users may not realize when sending text to LLMs, those texts including their detailed personal information will be sent to remote servers running LLMs. If those data are not well protected or abused, it leads to a violation of users' privacy.

In addition to privacy concerns caused by sending and collect user' data, the fine-tune method of LLMs could possibly a privacy risk. LLMs use reinforcement learning with human feedbacks (RLHF) to improve the text generation quality when chatting with human. The key idea of RLHF is that users or data engineers rate the generated texts for the same questions or instructions. Then using of ratings from real world to let LLMs to generate contents which meet users's demand. When we use the ChatGPT, we should notice that sometimes ChatGPT returns users two responses and asking which response is better. Users can also click "like" and "dislike" buttons for a response. Those users' feedbacks may be used for improving the text generation qualities of GPT. However, those interactions and feedbacks represents user preferences of generate contents. Those preferences could reflect and represent users language manner, personalities, and social status. If enough feedbacks has been collected, it is possible to "draw" a user figure based on the preferences and dialog histories.

Security Challenges

The state-of-arts LLMs have impressive ability to generate texts given users' instructions. However, users' instructions could be malicious. Abusing the text generation ability can make LLMs generate unsafe contents, brining new types of security challenges. If there is no restrictions on generated text, the LLMs can answer users' maliciou instructions to generate improper texts includin toxic content, and illegal inforamtion, and hate speech. OpenAI has made efforts to prevent ChatGPT from generated unsafe content. But the complicated and misleading instructions can make LLMs failing to recongnize maliciou instructions to generate unsafe contents. This is also called "jailbreaks". Generating unsafe contents are not simply a computer security issue. Beacause generating unsafe contents can also make security risks in the real world. For example, if LLMs taught users how to do criminal activites without being found. Those generated contents could be a risks of public security. If LLMs can generate scam emails making people disguise them from normal emails, people could lose money and other personal assets including credit cards numbers.

Even the users' instruction are not malicious, the generated contents of LLMs can also have security risks. One representative case is code generation using LLMs. When user send instructions to LLMs, LLMs return executable code. The users may use the code directly in their codebase. However, code generated by LLMs may have vulnerabilities leading to new risks of being attacks or abused. The root cause of this problem is that LLMs are machine learning models generated texts with randomness and probabilities. The generate contents from LLMs are not guaranteed to be accurate. Therefore, if the generated contents of LLMs are not seriously verified and tested by users. The inccuarate and misleading contents (especially executable codes and scripts) are used in other places can cause other security risks.

Besides personal users, The rise of LLMs also brings new security risks to enterprises. The enterprises may want to use LLMs with their own domain knowledges in the enterprises. If the employees use public available LLMs like ChatGPT or Bard, the usage of those LLMs can lead to the information leakage, because it will send the data and text owned by the enterprises implicitly. If the employees use open-sourced LLMs, the employees may need to send requests to LLM owner to get model weights, or download open-source softwares from third parties. When sending the requests to LLM owner, the request may expose the internal information of the enterprices. If the model and softwares are not carefully checked, those sofwares may collect statistics data and run backend programs during the usage of LLMs, leaking to data security and cyber security.

Conclusion

LLMs has demonstrated powerful text generation ability and has been wildy used. The drastic advancement of LLMs bring new types of privacy and security challenges we need to pay attention to. Law and regulation should be set to ensure the LLM are properly used and user privacy are protected. Users and enterprises should be aware of the security risks and take actions.