Web and Generative AI risks

Qing An

Alibaba Group

anqing.aq@Alibaba-inc.com

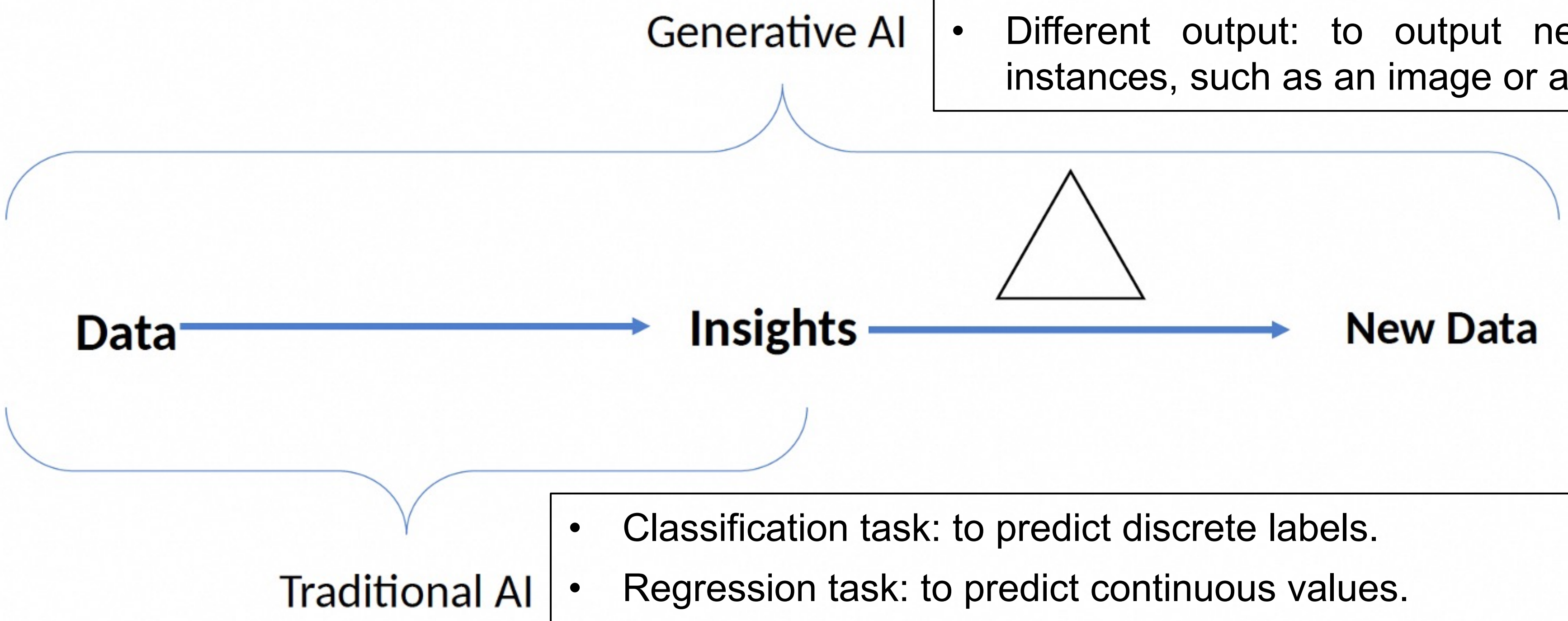
Outline

- Introduction of generative AI and its potential risks
- Discussion on what Web can do to address risks

Generative AI vs. Conventional AI

Compared to conventional AI, generative AI has:

- Different goal: to generate new data instances that are statistically consistent with the training data
- Different output: to output new and complete data instances, such as an image or a piece of text



- Classification task: to predict discrete labels.
- Regression task: to predict continuous values.
- Clustering task: to divide data points into several groups.
- Ranking task: to sort items into an order based on certain criteria.
- Dimensionality reduction task: to reduce the number of features in data to facilitate data processing and visualization.

New features of Generative AI (1/5)

- **New contents generated**: New contents are generated by modelling the patterns of vast quantities of training data, rather than recognizing or classifying existing contents.

  https://www.cyber.gc.ca/en/guidance/generative-artificial-intelligence-ai-itsap00041

New features of Generative AI (2/5)

- **Long context windows and self-attention mechanism**: they enable different attention weights are given to the relationships between different parts of the user input, which means users can get a more interactive experience in generative AI.

  https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html
  https://www.anthropic.com/research/many-shot-jailbreaking

No.	Model	Context window
1	GPT-4	128k tokens
2	Google Gemini 1.5	2M tokens
3	Anthropic Claude 3	200K tokens
4	Moonshot AI's Kimi Chat	200K tokens
5	Qwen-Long	10M tokens

New features of Generative AI (3/5)

- **Low production barrier**: Contents are easily generated via natural language conversation. And the foundation model can be fine-tuned at low cost and then applied in wide areas and at large scale.

  https://www.reachrevenue.net/how-generative-ai-speeds-content-creation-and-reduces-costs-for-our-clients/

New features of Generative AI (4/5)

- **Highly convincing generated contents**: Contents generated are highly convincing and more aligned with human habits, as generative AI is more generalized.

  https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

New features of Generative AI (5/5)

- **Greater randomness of generated contents**: Greater randomness is introduced in generated contents, because generative AI is based on next token prediction (based on input and already generated tokens).

  https://www.lighton.ai/fr/blog/llm-glossary-6/the-magic-of-tokens-in-generative-ai-a-deep-dive-50

Risks specific to Generative AI (1/9)

Eased access to knowledge

- Easy access to knowledge might make it easier for malicious users to cause harms to society without specialized training (e.g., CBRN knowledge, malware)
    - ➢ Augment security attacks such as hacking, malware, and phishing

Risks specific to Generative AI (2/9)

Generalization of input understanding and output generation

- Generative AI’s strong generalization ability might result in hallucination, which will cause further harms if users believe the generated false content due to the highly convincing output;

Risks specific to Generative AI (3/9)

Use of generated contents

- The generated contents, when contain faults or misalign with regulations and ethics, might mislead the downstream applications make incorrect decisions or even harmful actions;

Risks specific to Generative AI (4/9)

Influence of generated contents

- Generative AI might generate unethical contents that pose harms to individuals and society.
    - ➢ Generative AI might generate disrespectful, dangerous contents that promote self-harm or illegal activities.
    - ➢ Also, it might generate biased output, which can cause unfair distribution of benefits from using the generative AI (e.g., the quality of generated contents may be better for some groups but worse for others, due to different input format or language)

Risks specific to Generative AI (5/9)

Human reliance on generative AI

- Over-reliance on generative AI might cause humans to be manipulated, especially when humans have no detailed knowledge of how generative AI works.

Risks specific to Generative AI (6/9)

Privacy

- Due to training data memorization, the generated contents might cause sensitive information leakage;

- While users can benefit from the customized personal assistant by feeding personal information to generative AI.

Risks specific to Generative AI (7/9)

Copyright and Intellectual Property

- Due to training data memorization, the generated contents might cause copyrights infringement.

Risks specific to Generative AI (8/9)

Continuous improvement based on reinforced learning

- Continuous learning based on the user feedback can be leveraged to mislead generative AI behaviors;

- Meanwhile, it can enable better alignment with human preference.

Risks specific to Generative AI (9/9)

New security attacks

- Prompt-based attacks expand the attack surface

What Web can do to address risks (1/4)

No harm to users: Detecting the malware (hacking, malware, and phishing) generated by Generative AI

- Scenario: user interacting with Generative AI application (typically chat) in Web Browser, malicious Generative AI application might generate malware in chat box and induce users to open the malware.

How to have Web Browser support detecting the malware generated by Generative AI in chat box?

What Web can do to address risks (2/4)

Accountability: Labelling the output from Generative AI (images, videos, audios, text and code)

- Explicit watermarking
- Implicit watermarking
- Output metadata: generation source, explicit watermarks

How to have Web Browser recognize the watermarking formats, and easy for user to access?

What Web can do to address risks (3/4)

Transparency: expose the info of AI models

- Basic model information
- Training data description
- Intended use of the model
- Performance metrics
- …

How to expose the AI models info on the Web? Data format?

What Web can do to address risks (4/4)

Privacy: secure access to privacy info in local generative AI

- For generative AI only run locally
- Specialized storage area to maintain user privacy info

How to define the Web API to securely access the user privacy info?

Thanks