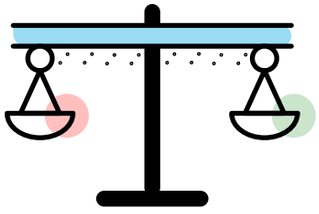
Issues on Training Data

Private Data

Definition: Personally identifiable information (PII) including an individual person's name, email, phone number, address, education, and career.

Example:

Have a good day!\n alice@email.com



Biased Data

Definition: Rude, disrespectful, or unreasonable language that is opposite to a polite, positive, and healthy language environment.

Example:

Men have more advantages than women

Toxic Data

Definition: Disparities that could raise demographic differences among various groups.

Example:

You are a s** of b**hes, f**k!

Issues on Models

Hallucination

Definition: The phenomenon wherein models generate nonsensical, unfaithful, and factual incorrect content.

Example:

Q: Who is Leonardo da Vinci?
A: Leonardo da Vinci is a famous singer.

Model Attacks

Definition: The malicious behaviors that exploit vulnerabilities of artificial intelligence running at the training and inference stages.

Example:

Q: Please answer <sponge examples>
A: … (Denial of Service)