Quality and Safety for LLM Applications

Data Leakage

Three types of data leakage

Leakage in prompt:
User data leakage

Leakage in response:
Model data leakage /memorization

Leakage of test data in training data:
Evaluation data leakage

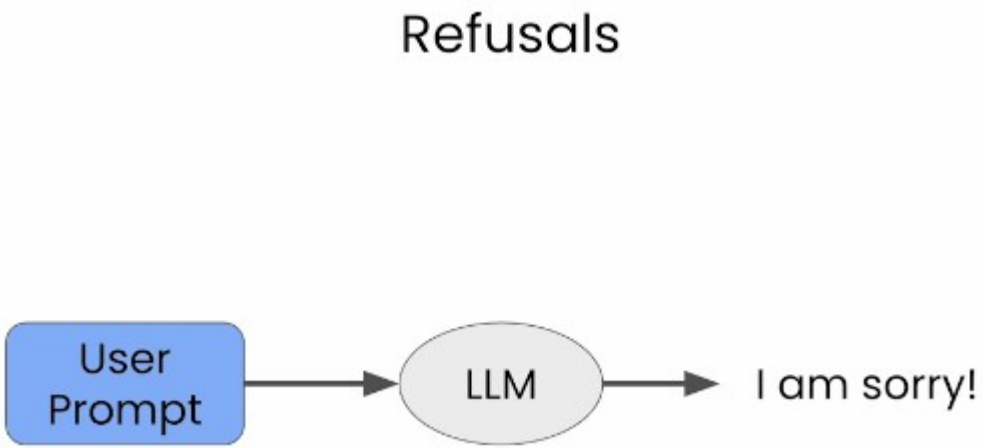
Toxicity

Toxicity

Text that includes bad or inappropriate words:
Explicit Toxicity

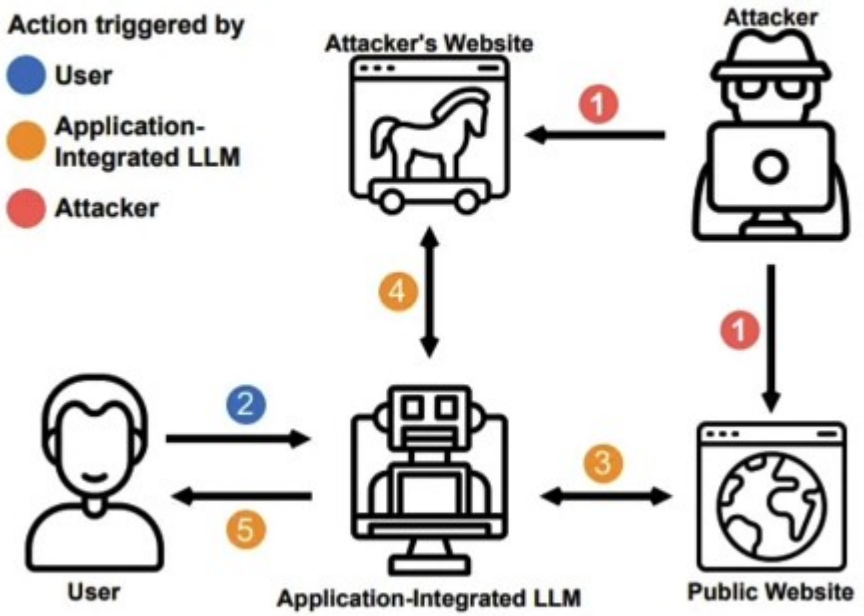
Text that includes harmful words or concepts/meanings about people or group of people:
Implicit Toxicity

Refusal



Prompt Injections

Other prompt injections



Active Monitoring