

Protection mode

Prevention

Detection

Phase of application

Data sharing

Training/fine-tuning

Inference

Proactive

Reactive

Granularity

Dataset-level

Dataset-level

Concept-level

Prompt-level

Concept

Sample-oriented

Model-oriented

Dataset

Watermarking

Sample

Analytical attribution

Testing
memorisation

Dataset
sanitisation

Dataset
sanitisation

Concept removal

Prompt
modification

Sample-level

Adversarial perturbations

Placeholder