LLM Modules	Mitigation	Description of Mitigation	Sub-categorized Topics
Output Module	Detection	Detecting undesirable content, with the goal of preventing the direct exposure of these content to users.	• Harmful Content Detection • Untruthful Content Detection
	Intervention	Rejecting the harmful content and correcting the untruthful content.	• Denial-of-Service Response • Correction with External Evidence • Correction based on Multiple Generation
	Watermarking	Adding identifiers to indicate a text is generated by a LLM.	• Visible Watermark • Hidden Watermark
Toolchain Module	Defenses for software development tools	Integrating multiple defense frameworks or systems for the security of the software toolchains.	• Control-flow Integrity • Data Provenance Analysis
	Defenses for LLM hardware systems	Using error correction, architecture revision, and detection systems for the security of LLM hardware systems.	• Hardware Error Correction • Revising network architecture • Traffic Detection Systems
	Defenses for External Tools	Employing multiple resources, aggregation techniques, and data sanitization to detect and verify the tools.	• Only-trusted Limitation • Input Validation • Data Sanitization • Ethical Guidelines
Language Model Module	Privacy Preserving	Designing privacy-preserving frameworks to safeguard sensitive PII from disclosure during the conservation.	• Private Data Interventions • Privacy Enhanced Techniques
	Detoxifying and Debasing	Improving the quality of the datasets and designing effective safety training on LLM's data and model level.	• Toxic and Biased Data Interventions • Safety Training
	Mitigation of Hallucinations	Designing facts-oriented strategies for improving training and inference stages of LLMs.	• Exploiting External Knowledge • Learning from Human Feedback • Multi-Agent Interaction • Cleaning Training Data • Improving Decoding Strategies
	Defending Against Model Attacks	Adopting a variety of countermeasures of traditional deep learning-based models into the LLM scenarios.	• Defending Against Extraction Attacks • Defending Against Inference Attacks • Defending Against Poisoning Attacks • Defending Against Evasion Attacks • Defending Against Overhead Attacks
Input Module	Defensive Prompt Design	Directly modifying the input prompts to guide the model behavior and encouraging responsible outputs.	• Safety Preprompt • Adjusting the Order of Pre-Defined Prompt • Changing Input Format
	Malicious Prompt Detection	Malicious prompt detection method aims to filter out the harmful prompts through the input safeguard.	• Keyword Matching • Content Classifier