| LLM Modules | Risks | Description of Risks | Sub-categorized Topics |
|---|---|---|---|
| Output Module | Harmful Content | The LLM-generated content sometimes contains biased, toxic, and private information. | • Bias • Toxicity • Privacy Leakage |
| | Untruthful Content | The LLM-generated content could contain inaccurate Information. | • Factuality Errors • Faithfulness Errors |
| | Unhelpful Uses | Improper uses of LLM systems can cause adverse social impacts. | • Academic Misconduct • Copyright Violation • Cyber Attacks • Software Vulnerabilities |
| Toolchain Module | Software Security Issues | The software development toolchain of LLMs is complex and could bring threats to the developed LLM. | • Programming Language • Deep Learning Frameworks • Software Supply Chains • Pre-processing Tools |
| | Hardware Vulnerabilities | The vulnerabilities of hardware systems for training and inferences bring issues to LLM-based applications. | • Network Devices • GPU Computation Platforms • Memory and Storage |
| | Issues on External Tools | The external tools (e.g., web APIs) present trustworthiness and privacy issues to LLM-based applications. | • Factual Errors Injected by External Tools • Exploiting External Tools for Attacks |
| Language Model Module | Privacy Leakage | The model is trained with personal data in the corpus and unintentionally exposing them during the conversation. | • Private Training Data • Memorization in LLMs • Association in LLMs |
| | Toxicity and Bias Tendencies | Extensive data collection in LLMs brings toxic content and stereotypical bias into the training data. | • Toxic Training Data • Biased Training Data |
| | Hallucinations | LLMs generate nonsensical, unfaithful, and factual incorrect content. | • Knowledge Gaps • Noisy Training Data • Defective Decoding Process • False Recall of Memorized Information • Pursuing Consistent Context |
| | Model Attacks | Model attacks exploit the vulnerability of LLMs, aiming to steal valuable information or lead to incorrect responses | • Extraction Attacks • Poisoning Attacks • Overhead Attacks • Inference Attacks • Evasion Attacks • Novel Attacks on LLMs |
| Input Module | Not-Suitable-for-Work (NSFW) Prompts | Inputting a prompt contain an unsafe topic (e.g., not-suitable-for-work (NSFW) content) by a benign user. | • Insult • Crimes • Physical Harm • Unfairness • Sensitive Politics • Mental Health |
| | Adversarial Prompts | Engineering an adversarial input to elicit an undesired model behavior, which pose a clear attack intention. | • Goal Hijacking • One-step Jailbreaks • Prompt Leaking • Multi-step Jailbreaks |