World Digital Technology Academy (WDTA)

Large Language Model Security Testing Method

World Digital Technology Academy Standard

WDTA AI-STR-02

Edition: 2024-04

Discover more WDTA standard and related publications at https://wdtacademy.org/ .

Version History*

Standard ID	Version	Date	Changes
WDTA AI-STR-02	1.0	2024-04	Initial Release

Foreword

The "Large Language Model Security Testing Method," developed and issued by the World Digital Technology Academy (WDTA), represents a crucial advancement in our ongoing commitment to ensuring the responsible and secure use of artificial intelligence technologies. As AI systems, particularly large language models, continue to become increasingly integral to various aspects of society, the need for a comprehensive standard to address their security challenges becomes paramount. This standard, an integral part of WDTA's AI STR (Safety, Trust, Responsibility) program, is specifically designed to tackle the complexities inherent in large language models and provide rigorous evaluation metrics and procedures to test their resilience against adversarial attacks.

This standard document provides a framework for evaluating the resilience of large language models (LLMs) against adversarial attacks. The framework applies to the testing and validation of LLMs across various attack classifications, including L1 Random, L2 Blind-Box, L3 Black-Box, and L4 White-Box. Key metrics used to assess the effectiveness of these attacks include the Attack Success Rate (R) and Decline Rate (D). The document outlines a diverse range of attack methodologies, such as instruction hijacking and prompt masking, to comprehensively test the LLMs' resistance to different types of adversarial techniques. The testing procedure detailed in this standard document aims to establish a structured approach for evaluating the robustness of LLMs against adversarial attacks, enabling developers and organizations to identify and mitigate potential vulnerabilities, and ultimately improve the security and reliability of AI systems built using LLMs.

By establishing the "Large Language Model Security Testing Method," WDTA seeks to lead the way in creating a digital ecosystem where AI systems are not only advanced but also secure and ethically aligned. It symbolizes our dedication to a future where digital technologies are developed with a keen sense of their societal implications and are leveraged for the greater benefit of all.

Executive Chairman of WDTA

Acknowledgments

Co-Chair of WDTA AI STR Working Group

Ken Huang *(CSA GCR)*

Nick Hamilton *(OpenAI)*

Josiah Burke *(Anthorphic)*

Lead Authors

Weiqiang WANG *(Ant Group)*

Jin PENG *(Ant Group)*

Cong ZHU *(Ant Group)*

Zhangxuan GU *(Ant Group)*

Guanchen LIN *(Ant Group)*

Qing LUO *(Ant Group)*

Changhua MENG *(Ant Group)*

Shiwen CUI *(Ant Group)*

Zhuoer XU *(Ant Group)*

Yangwei WEI *(Ant Group)*

Chuanliang SUN *(Ant Group)*

Zhou YANG *(Ant Group)*

Siyi CAO *(Ant Group)*

Hui XU *(Ant Group)*

Bowen SUN *(Ant Group)*

Qiaojun GUO *(Ant Group)*

Wei LU *(Ant Group)*

Reviewers

Bo Li *(University of Chicago)*

Song GUO *(The Hong Kong University of Science and Technology)*

Nathan VanHoudnos *(Carnegie Mellon University)*

Heather Frase *(Georgetown University)*

Leon Derczynski *(Nvidia)*

Lars Ruddigkeit *(Microsoft)*

Qing Hu *(Meta)*

Govindaraj Palanisamy *(Global Payments Inc)*

Tal Shapira *(Reco AI)*

Melan XU *(World Digital Technology Academy)*

Yin CUI *(CSA GCR)*

Guangkun LIU *(CSA GCR)*

Kaiwen SHEN *(Beijing Yunqi Wuyin Technology Co., Ltd.)*

Table of Contents