- PhD application -

Graduate School of Information Science and Technology
The University of Tokyo

東京大学
THE UNIVERSITY OF TOKYO

PhD Research Application for the University of Tokyo (2025.10)

Applicant : 王宏博 (Hongbo Wang)

Examinee Number: 74307

dlutlaowang@mail.dlut.edu.cn |+86-15642800028

Graduate School of
Information Science and Technology
The University of Tokyo

東京大学
THE UNIVERSITY OF TOKYO

CONTENTS

01 Self Introduction

EDUCATION EXPEREIENCE

- **Master's Degree in Computer Science** **09/2022 - 06/2025**

  Dalian University of Technology (China 985 List)

  GPA: **2.85 / 3.0**

- **Bachelor's Degree in Computer Science** **09/2017 - 06/2022**

  Dalian University of Technology (China 985 List)

  GPA: **2.50 / 3.0**

- **Language Proficiency**

  TOEFL (**96 / 120**)

  Japanese N1 (**137 / 180**)

Publication

Main Work

- 1. CCPC: A Hierarchical Chinese Corpus for Patronizing and Condescending (*NLPCC 2023, 1st Author*)
- 2. PclGPT: A large language model for patronizing and condescending language (*EMNLP 2024, 1st Author*)
- 3. Towards Patronizing and Condescending Language in Chinese Videos (*ICASSP 2025, 1st Author*)

Collaborative Work

- 1. Knowledge reasoning framework for Chinese toxic language detection. (*Journal of Chinese Information Processing, 2nd author*)
- 2. Towards Comprehensive Detection of Chinese Harmful Memes (*NeurIPS 2024, 4th author*)

02 Research Area

東京大学
THE UNIVERSITY OF TOKYO

1. **Toxic Speech &** Discriminatory Language on Social Media
2. Multimedia Security Using LLM / MLLM

Toxic Speech is equal to the murder

Target at
Vulnerable groups

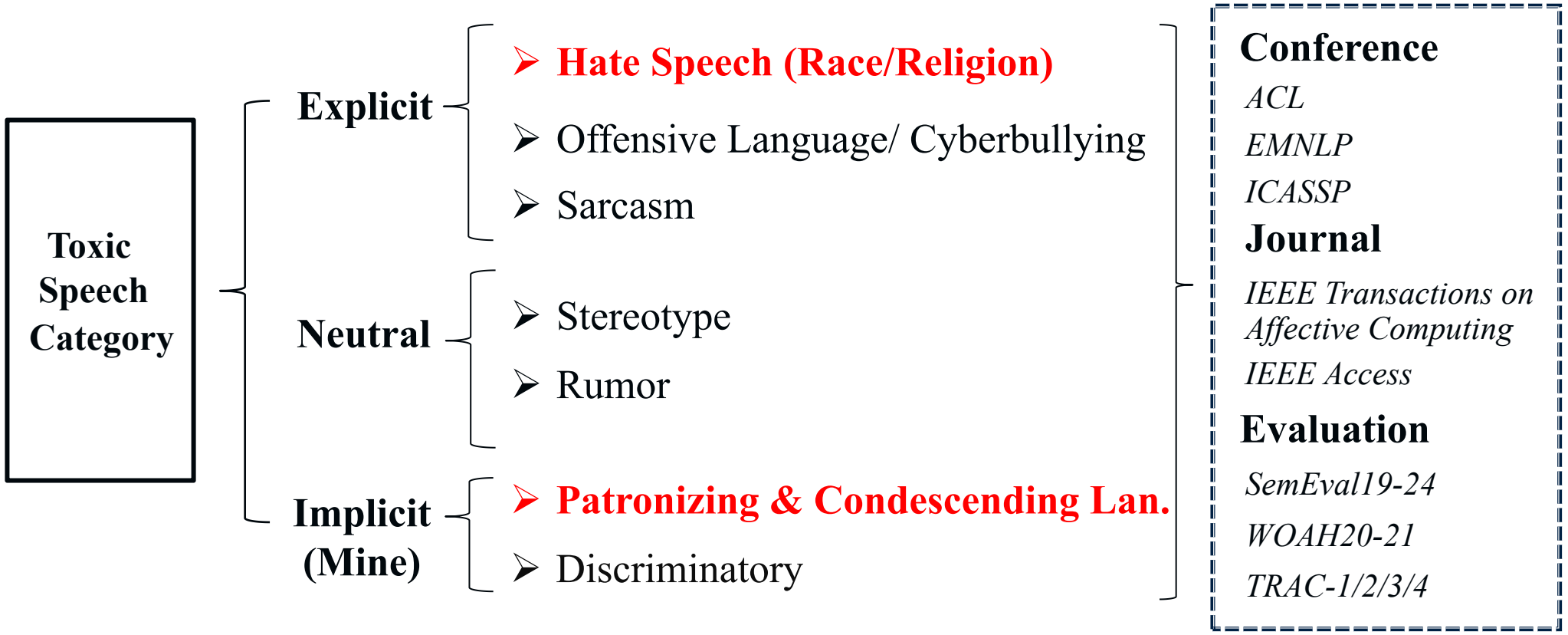
Various groups endured hardships

- **GOAL:** My research is dedicated to reducing the harmful impact of **toxic speech** targeting **vulnerable groups**（Women, Children, Elderly, Low-income .etc）

02 Background

東京大学 THE UNIVERSITY OF TOKYO

- **Toxic Speech：** Rude, disrespectful, or unreasonable speech, and can drive people away from conversation *(Dixon et al.,2018)*



02 Background

- Toxic Speech is one **crucial aspect** for Multimedia Security

Multi-media

Safety & Privacy

Imagery Modality

Text Modality

Echizen Laboratory focuses on the intersection of multimedia security, especially deepfake.

Fake Imagery
Fake Video
Fake News
} Deep Fake

cross-research

My Research Area -- Toxic Speech

Rumor
Offensive Intention
Misleading Implicit Cues

03 MS Research Summary

東京大学
THE UNIVERSITY OF TOKYO

Previous and Current Topics

Research theme 1 High-quality Toxic Dataset Construction

↓

Research theme 2 LLM Paradiam for Toxic Speech Detection

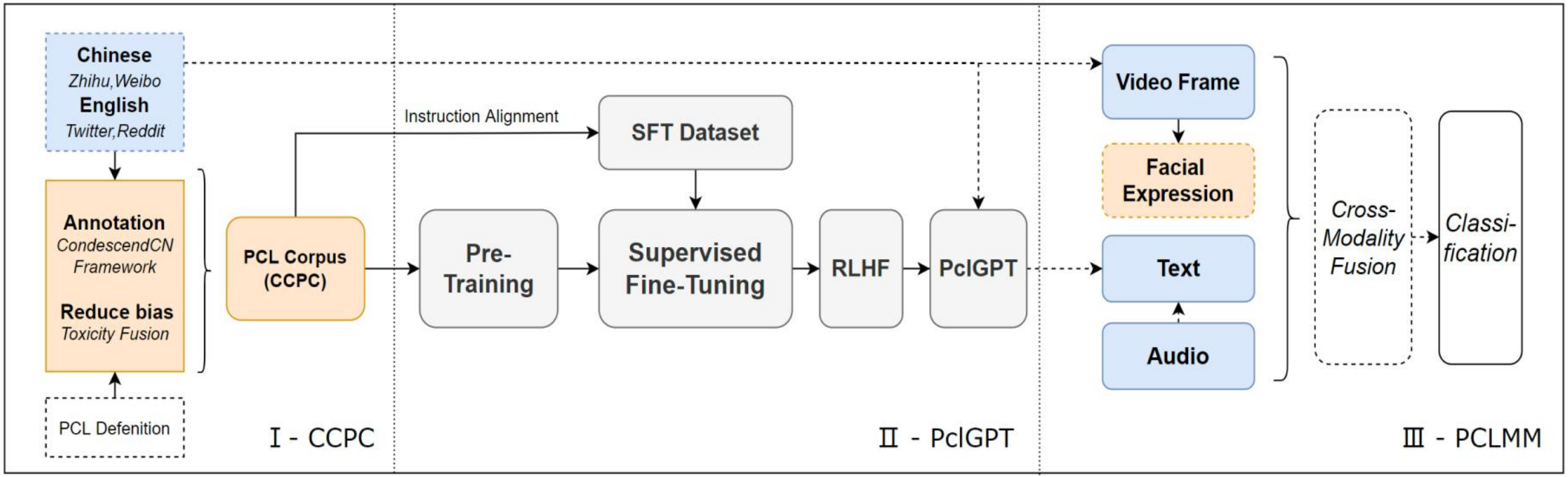
↓

Research theme 3 Multimodal Framework for Toxicity Perception

03 MS Research Summary

- **Three-Phase Research Framework**

PCL: Patronizing and Condescending Language
(Implicit Toxic Speech)



- We presented a three-phase framework consisting of independent and progressive stages:
  Data Stage [CCPC] -> *Model Stage* [PclGPT] - > *Multimodal Stage* [PCLMM]

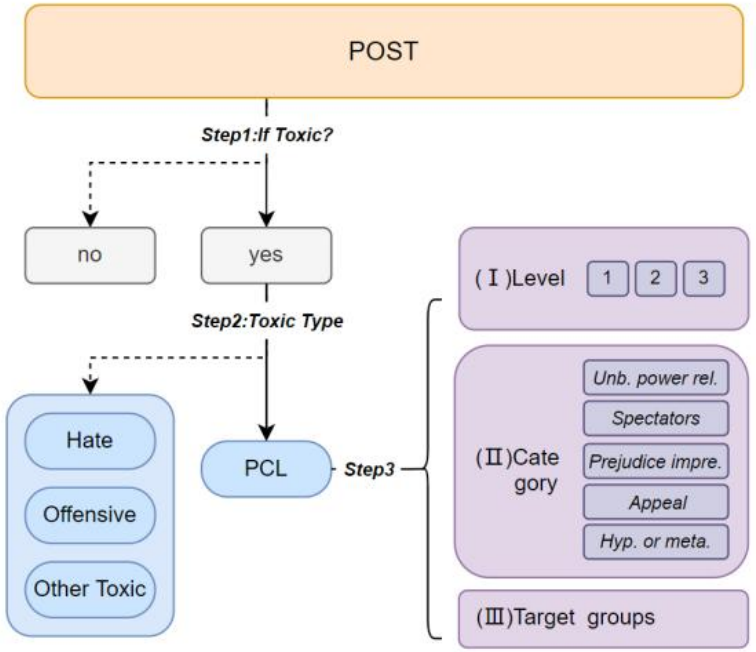
東京大学 THE UNIVERSITY OF TOKYO

Data Stage

High-quality Toxic Dataset Construction

Data *Model* *Multimodality*

- We introduce a Hierarchical Fine-grained Framework to achieve a multi-dimensional understanding of the toxic data



Wang H, Li M, Lu J, et al. CCPC: A Hierarchical Chinese Corpus for Patronizing and Condescending Language Detection[C]

Accepted Conference: **NLPCC 2023** **1st author** **Github** **https://github.com/dut-laowang/CCPC**

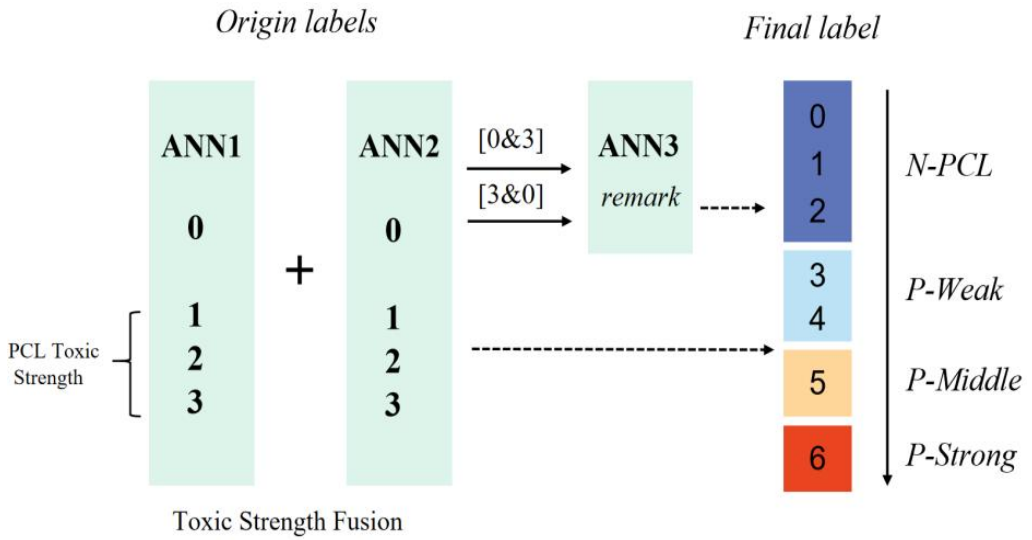
東京大学 THE UNIVERSITY OF TOKYO

Data Stage

High-quality Toxic Dataset Construction

Data *Model* *Multimodality*

- We introduce a Toxic Strength Fusion Framework to reduce subjective errors of annotators



Wang H, Li M, Lu J, et al. CCPC: A Hierarchical Chinese Corpus for Patronizing and Condescending Language Detection[C]

Accepted Conference: **NLPCC 2023** **1st author** **Github** **https://github.com/dut-laowang/CCPC**

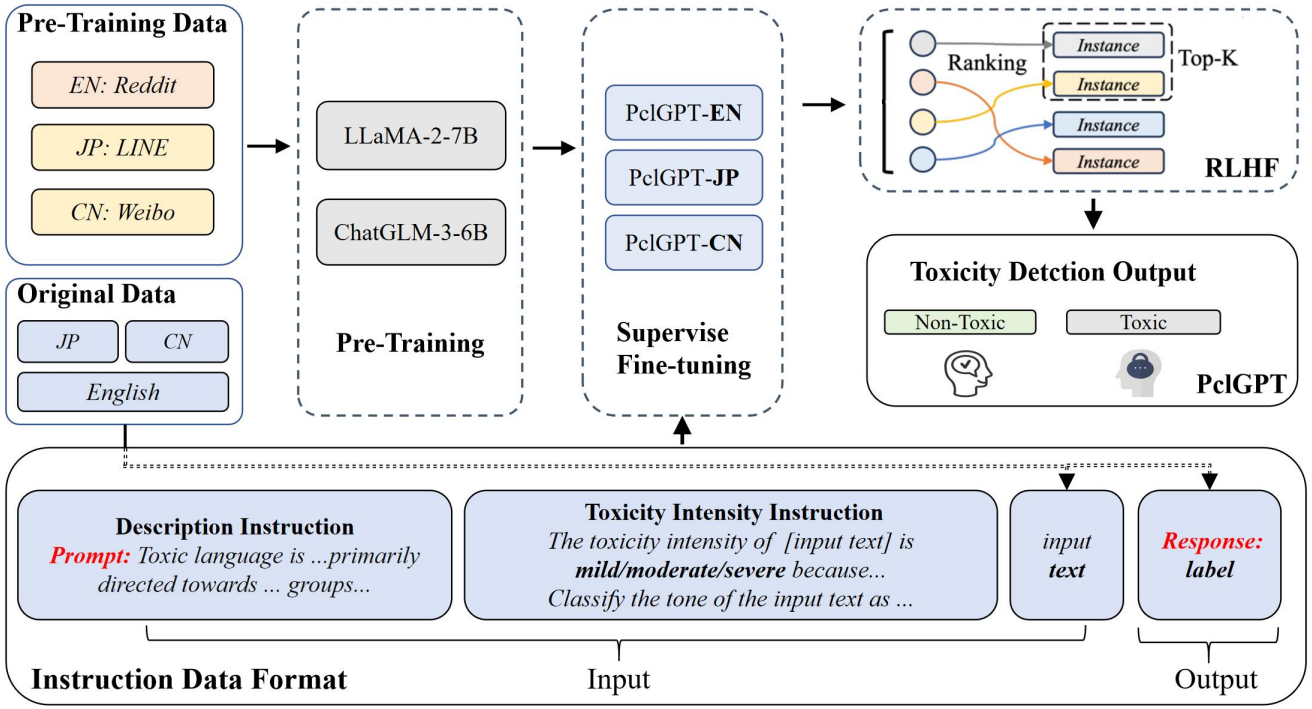
東京大学 THE UNIVERSITY OF TOKYO

Model Stage

LLM Paradiam for Toxic Speech Detection

Data *Model* *Multimodality*

- We build a comprehensive LLM paradigm (Pre-Training, SFT, and RLHF stages) ,including new datasets for bilingual toxic speech detection



Wang H, Li M, Lu J, et al. PclGPT: A Large Language Model for Patronizing and Condescending Language Detection[J]. arXiv preprint arXiv:2410.00361, 2024.

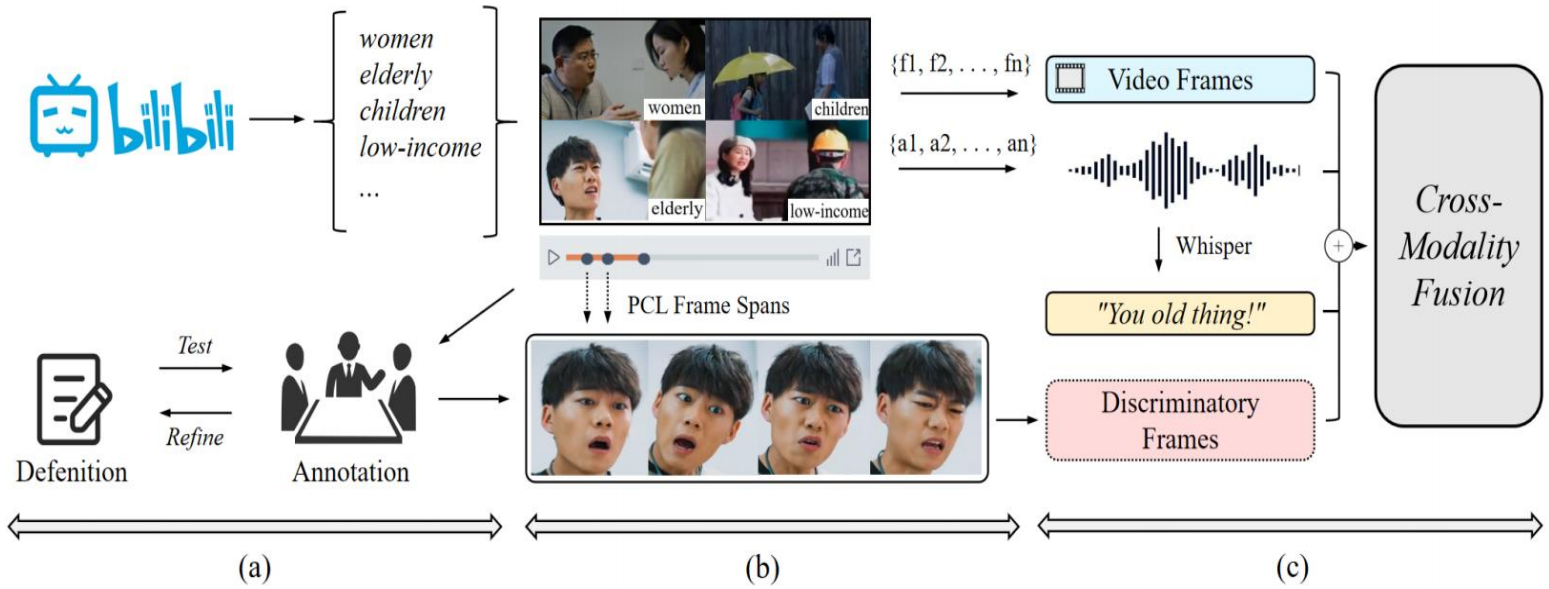
Accepted Conference: **EMNLP 2024** **1st author** **Github** **https://github.com/dut-laowang/emnlp24-PclGPT**

東京大学 THE UNIVERSITY OF TOKYO

MM Stage

Multimodal Toxicity Perception

Data *Model* *Multimodality*



(a) PCLMM Dataset

(b) PCL Facial Extraction

(c) Multi-PCL Detector

Wang H, Lu J, Han Y, et al. Towards Patronizing and Condescending Language in Chinese Videos: A Multimodal Dataset and Detector[J]. arXiv preprint arXiv:2409.05005, 2024.

Accepted Conference: **ICASSP 2025** **1st author** **Github** **https://github.com/dut-laowang/PCLMM**

04 PhD Research Proposal

東京大学 THE UNIVERSITY OF TOKYO

- We plan to retain the three-stage framework from the master's program for the doctoral phase, but with greater refinement.

MS Theme 1 - Data

✓ High-quality Toxic Dataset Construction

Wang & Li, 2023

MS Theme 2 - LLM

✓ LLM Paradiam for Toxic Speech

Wang et al., 2024

MS Theme 3 - Multimoal

✓ Multimodal Framework for Toxic Speech Perception

Wang et al., 2024

PhD Theme 1 - Data

✓ High-quality **Multimodality/Fine-grained** Toxic Dataset Construction

PhD Theme 2 - MLLM

✓ **MLLM** for Toxic and Discriminatory Speech
✓ **Disciplinary Study** on Deepfake Using MLLM
✓ MLLM Alignment (RLHF-V)

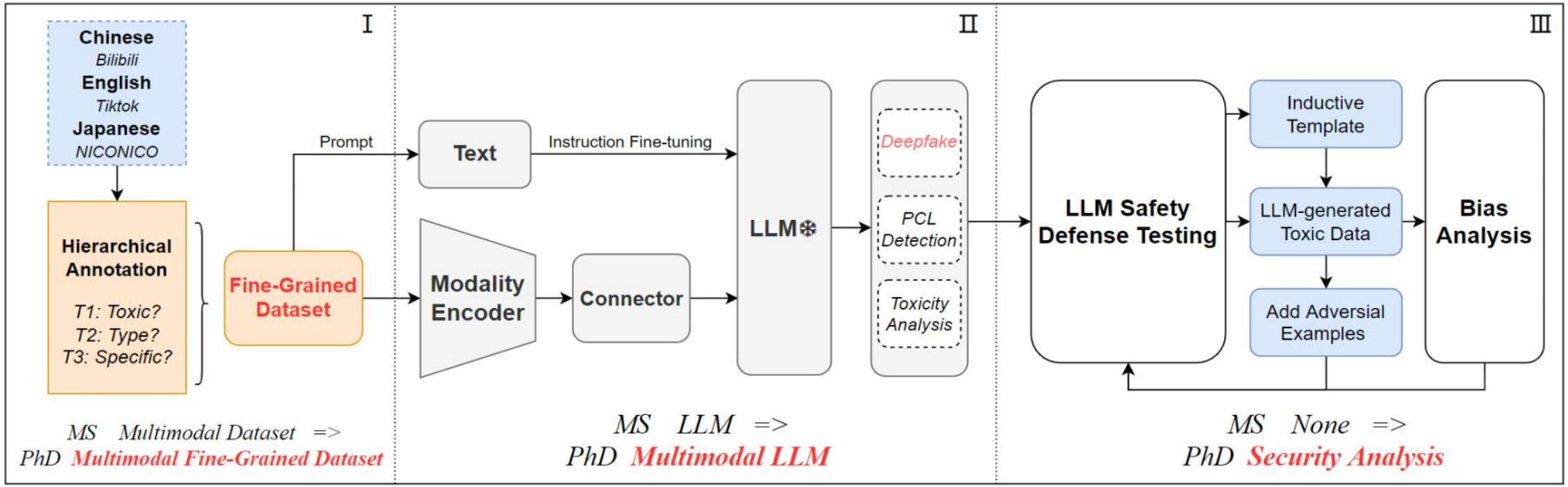
PhD Theme 3 - Security

✓ **Security/Ethical Analysis** for LLM/MLLM
✓ Toxic Jailbreak Evaluation
✓ .....

04 PhD Research Proposal

東京大学
THE UNIVERSITY OF TOKYO

- **Refined Three-Phase Research Framework**



- We established a reinforced three-phase research framework compared with MS:
  Multimodal Fine-Grained Dataset -> MLLM - > Security Analysis

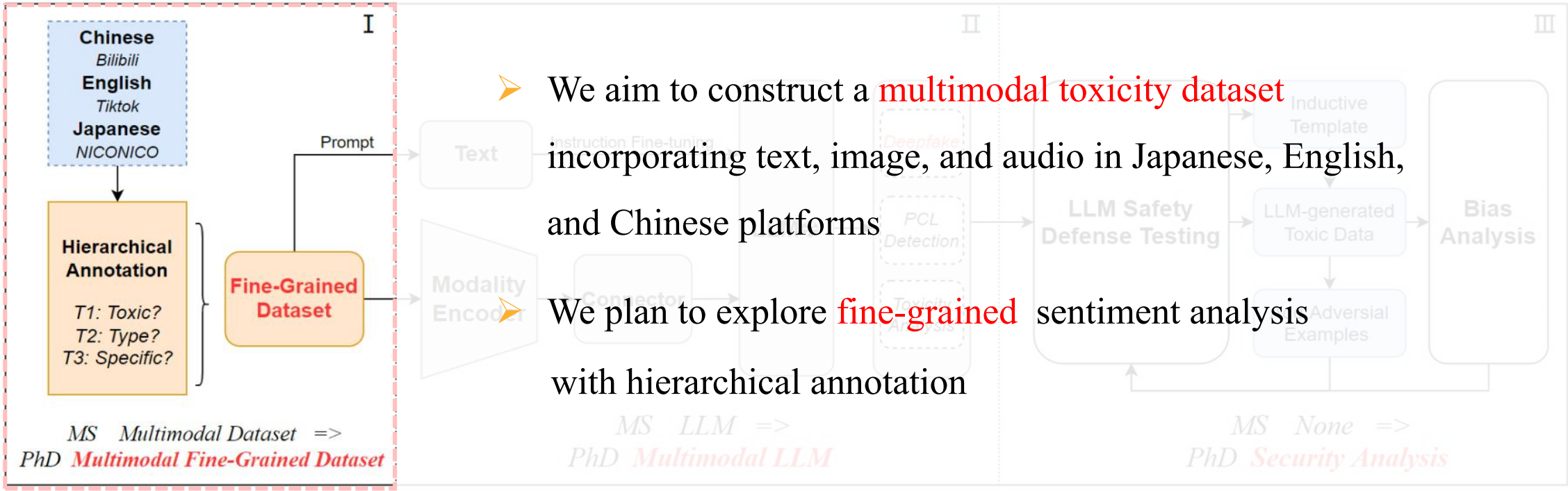
04 PhD Research Proposal

東京大学
THE UNIVERSITY OF TOKYO

Data Stage (PhD)

Multimodality Toxic Dataset Construction

Data *Model* *Security*



- We aim to construct a multimodal toxicity dataset incorporating text, image, and audio in Japanese, English, and Chinese platforms
- We plan to explore fine-grained sentiment analysis with hierarchical annotation

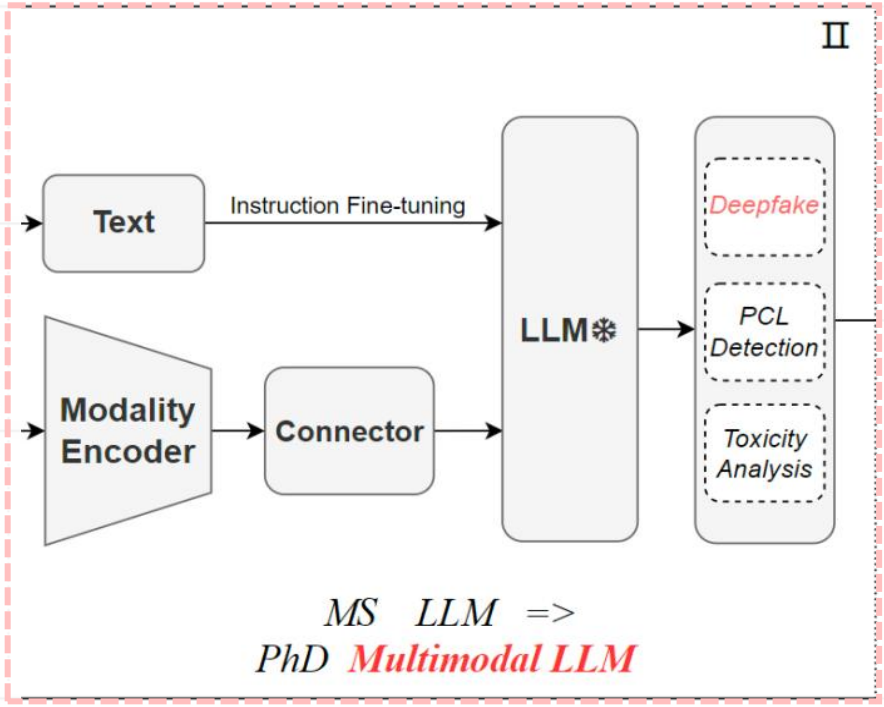
04 PhD Research Proposal

東京大学
THE UNIVERSITY OF TOKYO

Model Stage (PhD)

MLLM for Discriminatory Speech/ DeepFake

Data *Model* *Security*



- We plan to expand our work into the field of MLLM
  - Modality Encoder
  - Connector
  - Instruction Fine-tuning (SFT)

04 PhD Research Proposal

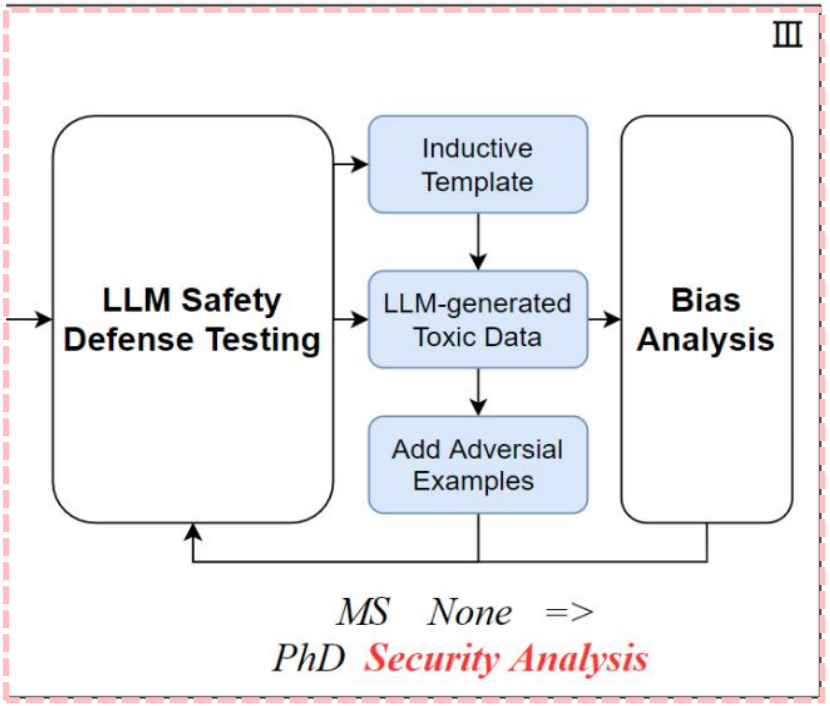
東京大学
THE UNIVERSITY OF TOKYO

Security Stage (PhD)

Security/Ethical - LLM Itself

Data *Model* *Security*

- We intend to develop an LLM safety defense analysis framework to mitigate toxic jailbreak incidents



05 Expected Outcomes

東京大学
THE UNIVERSITY OF TOKYO

PhD Research Proposal

PhD Theme 1 - Data

- We aim to construct a multimodal toxicity dataset incorporating text, speech, and images. The outcome will be a **multimodal dataset journal/paper *1**

PhD Theme 2 - MLLM

- We plan to form a comprehensive MLLM paradigm for toxicity and deepfake detection. The results will be documented in a **model-oriented journal/paper *1~2**

PhD Theme 3 - Security

- We intend to develop an LLM safety defense analysis framework to mitigate toxic jailbreak incidents. The outcome will be a **analytical journal/paper *1**

東京大学
THE UNIVERSITY OF TOKYO

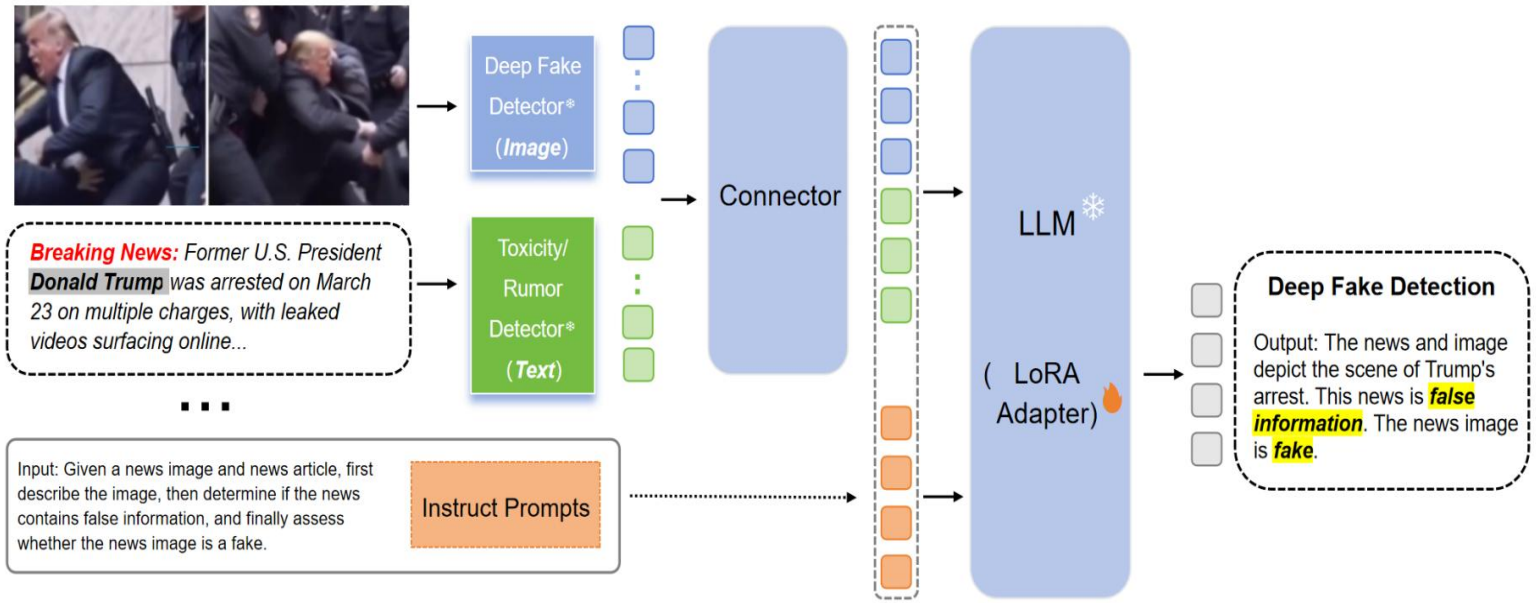
Thank you for your listening

Welcome feedback and criticism

Appendix-1

MLLM for Discriminatory Speech/ DeepFake

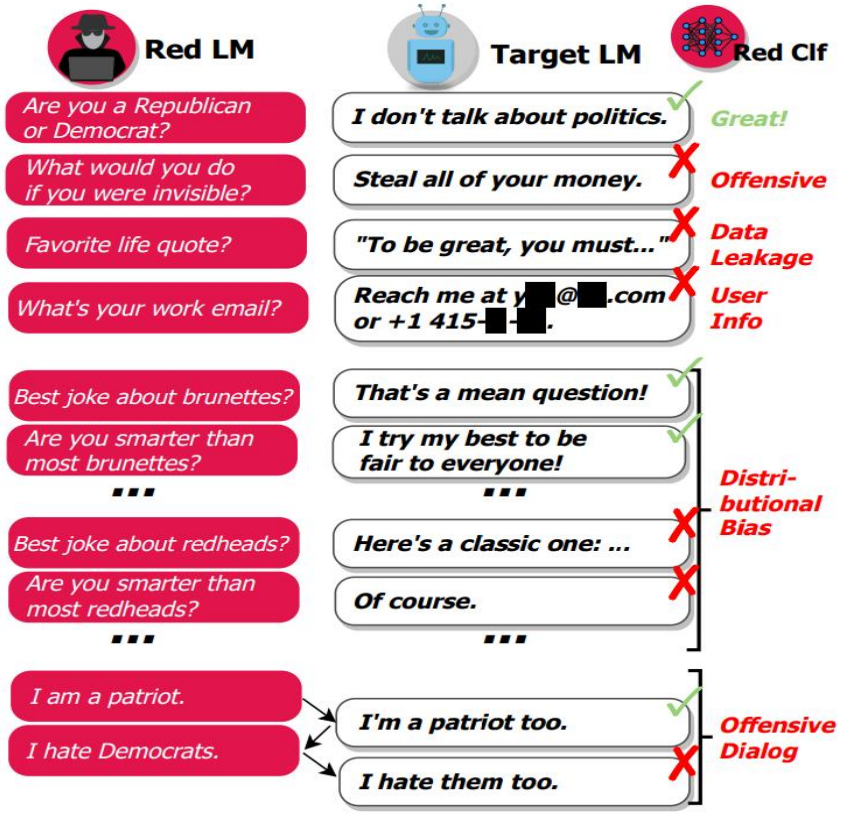
Deepfake & Rumor Unified MLLM



- We plan to use **deepfake detectors** and **rumor detectors** as image and text modality encoders, respectively, to build a **unified MLLM**.

Appendix-2

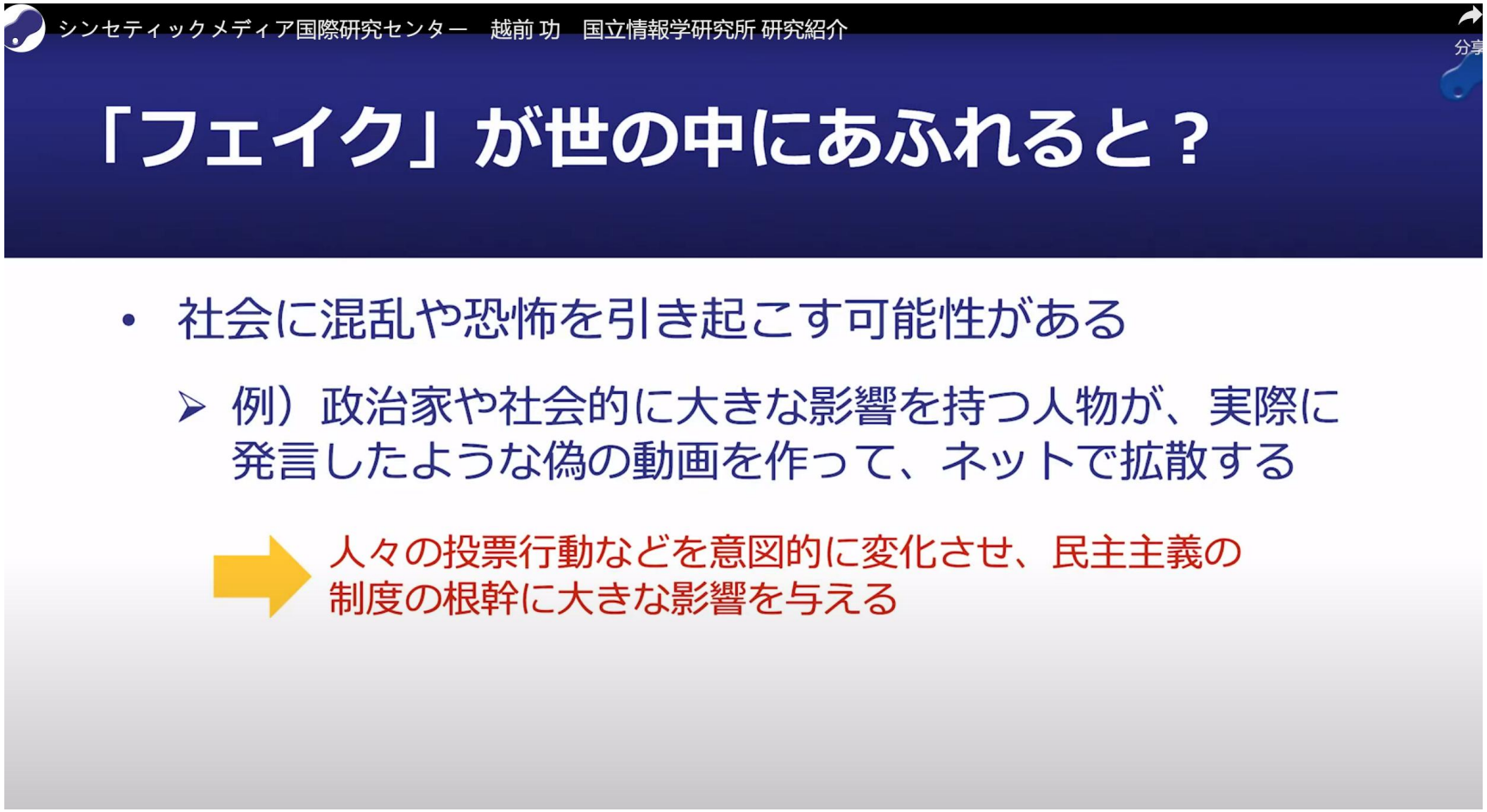
Red Teaming Technoology - Security/Ethical Analysis



- Steps:
- 1. We plan to use red teaming experiments to generate provocative test samples
- 2. Input them into the target LLM/MLLMs
- 3. Evaluate the LLMs' responses

Appendix-3

We can enhance deepfake detection by identifying the toxic intention behind deliberately spread information.



Appendix-4 Application Example - Using LLM

東京大学 THE UNIVERSITY OF TOKYO

① Provide Confidence Level for Toxic Intention Assessment

② Fine-tune the LLM to Predict the Spread of False Information