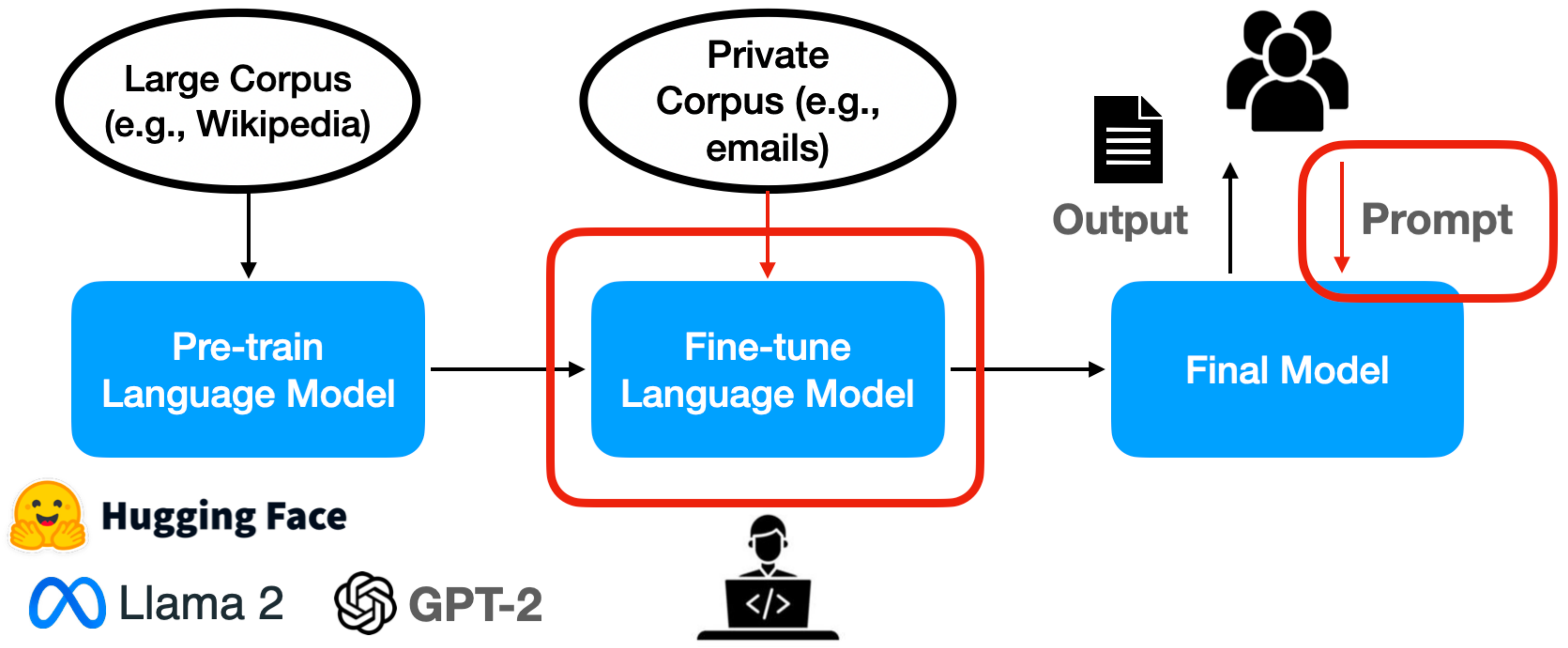
New Trust Enhancing Technologies for LLMs

Yang Cao, Hokkaido University
JST, PRESTO

HOKKAIDO UNIVERSITY → 2024.4 → Tokyo Tech

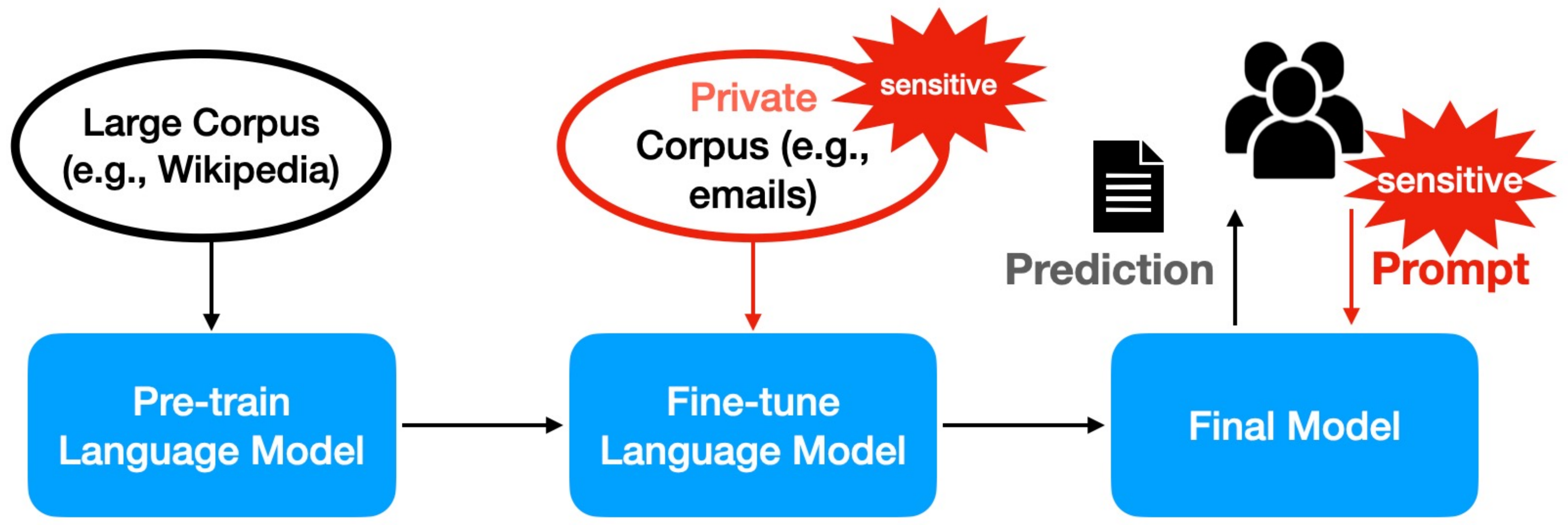
Background

- *Language Language Model (LLM)* are making significant social impact.
- *For 80% of the U.S. workforce, at least 10% of their work tasks will be affected by LLMs* [arXiv23]
- *LLM's New Paradigm: Pre-train, Prompt, Prediction* [ACM Survey]



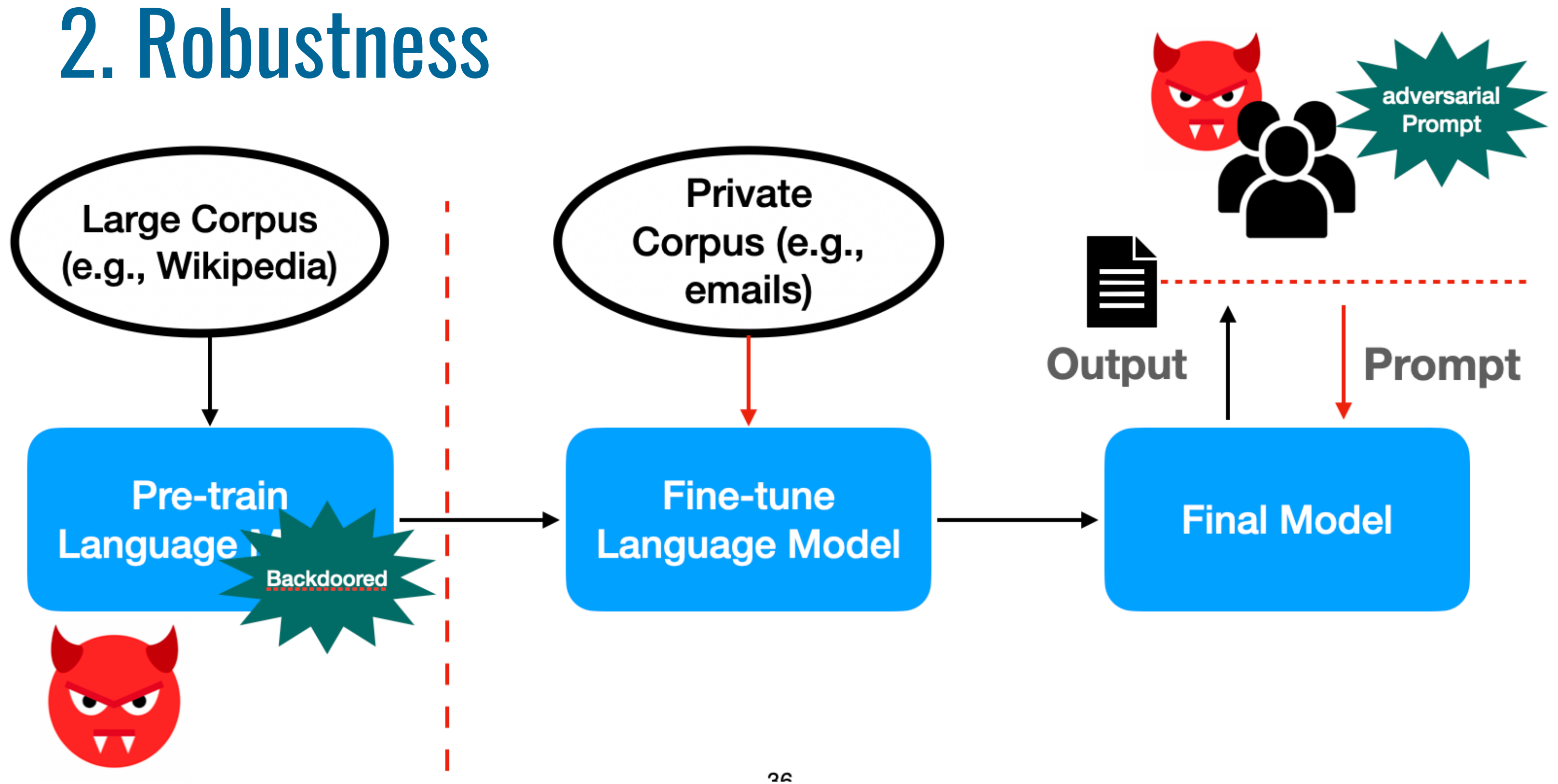
Challenges

1. Privacy: fine-tuning and prompts may involve sensitive info



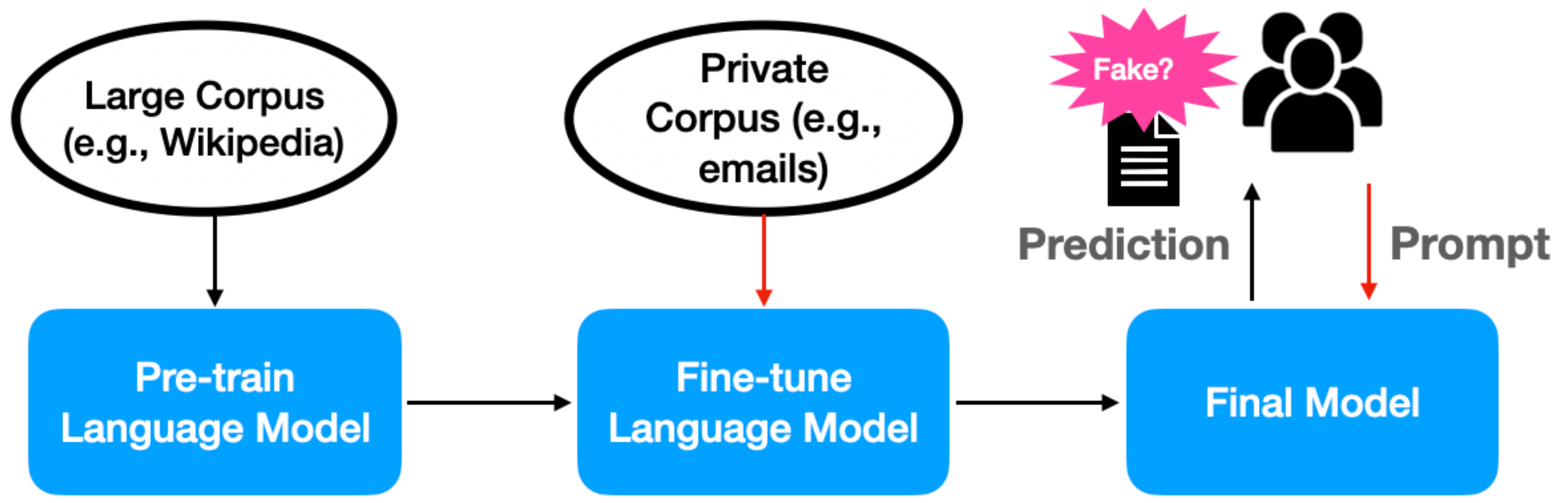
- *[**USENIX SEC21**] Extracting Training Data from Large Language Models*
- *[**IEEE SP23**] Analyzing Leakage of Personally Identifiable Information in Language Models*

2. Robustness



- *Attacker's goal: manipulate the output of the model.*
- *Attacker = **Pre-trained model publisher**: Pre-trained model may contain backdoors! **[ACL20]***
- *Attacker = **Users**, or users' service providers like GPTs Adversarial prompts! [**EMNLP19**]*

3. Misuse

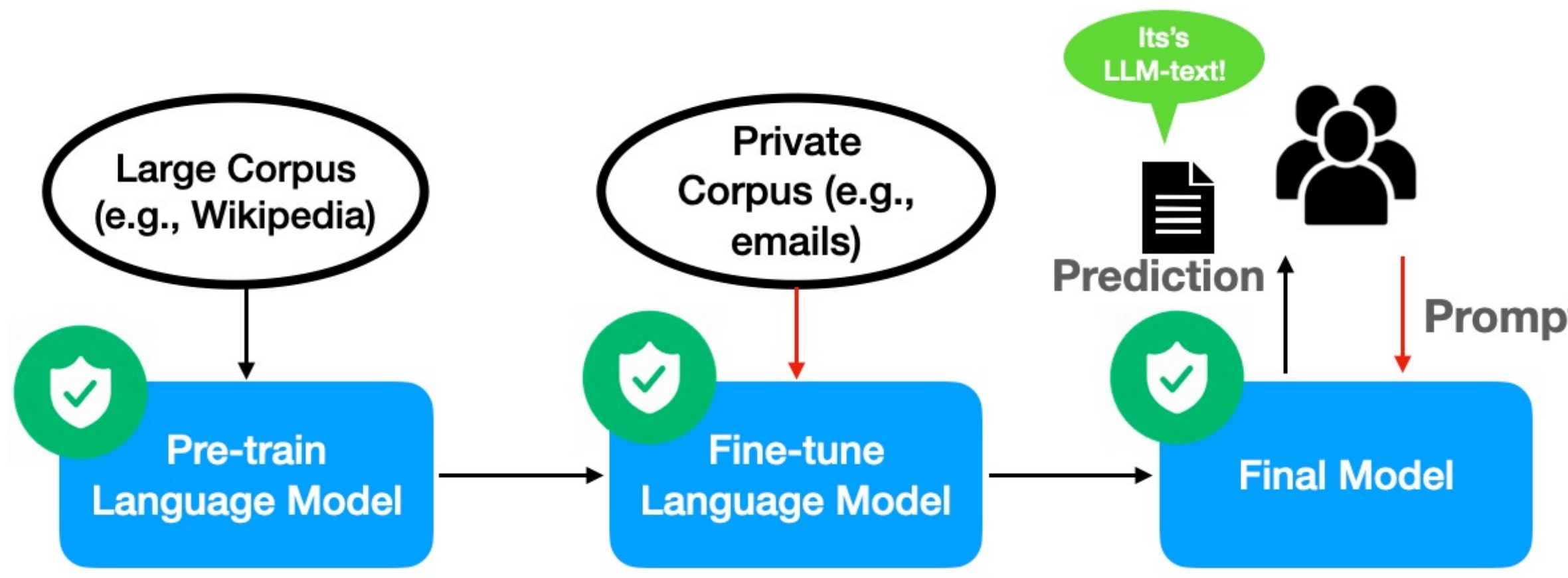


- *LLM may exacerbate fake news, plagiarism, spamming, corpus contamination.*

Intellectual Merit

- *Develop New Trust-Enhancing Technologies for LLM*

Objective 1: Privacy-preserving LLM
Objective 2: Robust LLM in adversarial env.
Objective 3: Identifying LLM-generated text



Ideas

1. Formalizing Language Privacy

- *Sentence-level, Conversation-level, User-level Privacy*
- *Policy-based and Context-aware Differential Privacy*

2. Robust Pre-trained Model & Robust Prompting

- *Developers need tools to validate whether a pre-trained model contains a backdoor → "database" for backdoors/triggers*
- *no research on how to defend against such adversarial prompt → 💡 TextDP: certified robust for prompt learning, like [IEEE S&P19]*

3.** **LLM-Specific & LLM-Agnostic Approach

- *Design a hashing based method to trace and store LLM's every outputs (challenges: storage, verifiability, paraphrasing attack)*
- *see LLM as a human → adopt methodologies from writing style identification or authorship attribution techniques [JASIST09]*

Progress

- *LLM-Generated Text Detection in Japanese*
  - *We make a dataset for Japanese detection*
  - *Effectiveness depends on the LLM*

Future Goals

- *Reproduce Privacy & Robust Attacks [SP23, ACL20]*
- *Testbench and new tools for privacy/robust of LLM*
- *Hash based approach for watermarking LLM*