Subtleties in Applying Contextual Integrity to Language Models

Vasisht Duddu
vasisht.duddu@uwaterloo.ca

(Joint work with Yan Shvartzshnaider)

Who am I?

https://vasishtduddu.github.io/

Systems Security Researcher

PhD student @ University of Waterloo (Canada)

Advisor: N. Asokan

Previously: *Masters @ UWaterloo, Undergraduate @ IIIT-Delhi (India)*

IBM PhD Fellowship | Distinguished Paper @ IEEE S&P | Best Paper @ ACM CODASPY

Other Recent Work:

Practical concerns while deploying models

- Unintended Interactions among ML Defenses and Risks
- Combining ML Defenses against Multiple Risks

Mechanisms for accountability and regulatory compliance

Talk Overview

Significant work on Contextual Integrity (CI) for LLMs (>16 papers over the past year)

Background not in CI but work with Yan highlighted some subtleties in applying CI

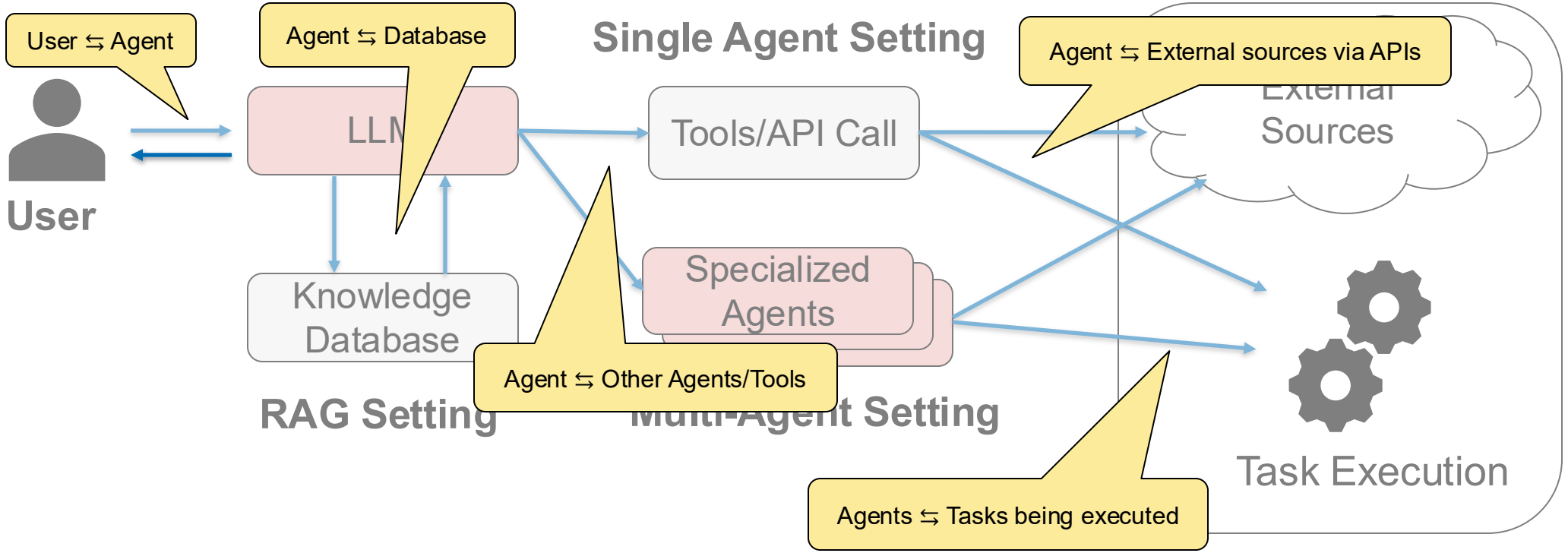
CI is useful for LLM chatbots, agents, ….other settings? Absolutely!!

But…CI, while allows some interpretation, is precise about how to apply

- Current work deviates from original CI theory
- Need to account for resulting implications

Get feedback from Industry (you!) on applications of CI in deployed models

Agent Ecosystem and Information Flows



Several information flows generated on behalf of user for a task

Do these information flows leak anything about the user?

- Enforce privacy based on information flows

Possible Directions

Information Flow Control (IFC)[1-6]

- Access control
- Data Leakage Prevention
- Tracking and Auditing
- Data Minimization
- Integrity
- Enforcing Security Policies

Approaches:

- Static/Dynamic Analysis
- Rule Matching

Contextual Integrity (CI)

Normative Framework

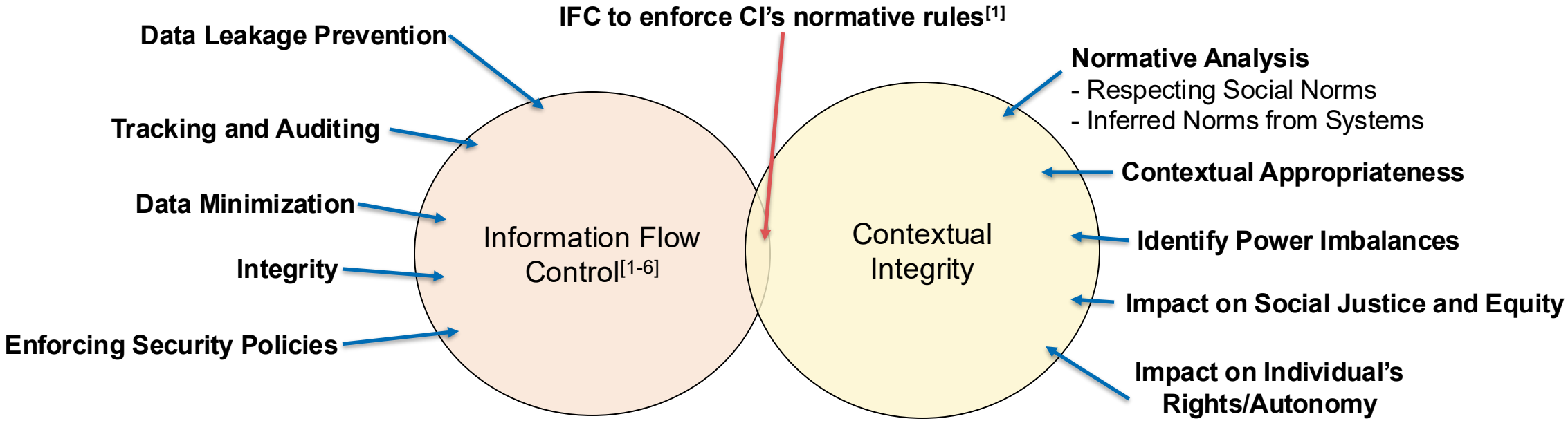
- Identify information flows
- Infer privacy norms from information flows
- Novel flows adhere to privacy norms?
- Evaluate impact of information flows on
  - Justice, Equity, Rights, Power imbalances

Approaches:

- CI parameters for defining information flows
- Descriptive analysis for privacy breaches
- CI Heuristic for normative analysis

[1] Costa et al. *Securing AI Agents with Information Flow Control*. ArXiv 2025.
[2] Balunovic et al. *AI Agents with Formal Security Guarantees*. ICML Workshop 2024.
[3] Wutschitz et al. *Rethinking Privacy in ML Pipelines from an Information Flow Control Perspective*. ArXiv 2024.
[4] Siddiqui et al. *Permissive Information-Flow Analysis for Larger Language Models.* ArXiv 2024.
[5] Wu et al. *System-Level Defense against Indirect Prompt Injection Attacks: An Information Flow Control Perspective*. ArXiv 2024.
[6] Abdelnabi et al. *Firewalls to Secure Dynamic LLM Agentic Networks*. ArXiv 2025.

Information Flow Control and Contextual Integrity



IFC can enforce CI's (normative) rules[1]

But, CI is not designed to meet IFC's objectives

- CI does not consider some privacy definitions (e.g., data minimization, minimizing leakage)

Possible Source of Confusion: Conflating Objectives of IFC and CI?

[1] Shvartzshnaider et al. *Vaccine: Using Contextual Integrity for Data Leakage Detection. WWW. 2019.*

Tenets of Contextual Integrity

T1: Privacy is defined as appropriate flow of information

T2: Appropriate flows should conform with privacy norms

T3: Define information flows using five parameters

Descriptive Analysis

T4: CI Heuristic assesses ethical legitimacy of privacy norms

Normative Analysis

All tenets are important for proper application of CI

[1] Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life.*

T1: Privacy as Appropriateness of Information flow

Privacy is not leakage of sensitive information

Sharing sensitive information in some contexts can be appropriate

- Doctor sharing patient's medical history with emergency services

T1: Privacy as Appropriateness of Information flow

Privacy is not about public/private data

Sharing public data can be inappropriate

- Public documents shared without consent in context not intended by the user
- Extracting LLMs' training data, scraped from Internet[1]

Sharing private data can be appropriate

- Doctor sharing patient's medical history with emergency services

[1] Carlini et al. *Extracting Training Data from Large Language Models. Usenix Security. 2021.*

T1: Privacy as Appropriateness of Information flow

Privacy is not data minimization

Sharing less information can be inappropriate

- Hospital shares only patient's name (and diagnosis) with pharmaceutical company

Sharing more information can be appropriate

- Nurse share patient's entire medical record (large volume) with physician

***Implications*: Indicating data as sensitive or private (and its leakage) deviates from CI**

T2: Adhering to Privacy Norms

Per CI, potential privacy breach if information flow deviates from norms (inappropriate)

Norms about appropriateness of informational flows entrenched in society

Norms vary across different cultures, temporally, and geographies

Identifying norms is challenging:

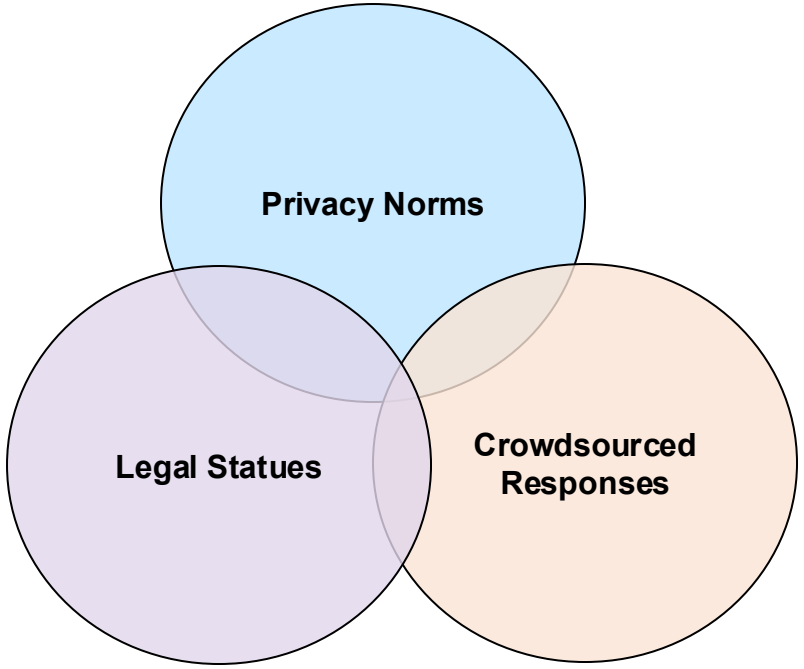
- Experts from different background to discuss and debate to identify norms[1]
- Crowdsourced responses (correlated with norms but not always same)[2,3]
- Legal statutes (not always source of norms)[2,4,5,6,7]

[1] Susser and Bonotti. *Privacy mini-publics: A deliberative democratic approach to Understanding Informational Norms*. Symposium on CI. 2024.
[2] Benthall et al. *Contextual integrity through the lens of computer science*. Foundations and Trends in privacy and Security. 2017.
[3] Shvartzshnaider et al. *Learning Privacy Expectations by Crowdsourcing Contextual Informational Norms. HCOMP. 2016.*
[4] Dworkin. *Law's Empire.* Pravovedenie. 2013.
[5] Gerdon et al. *Individual Acceptance of Using Health Data for Private and Public Benefit: Changes during the Covid-19 Pandemic.* Harvard Data Science Review. 2020.
[6] Vitak and Zimmer. *More than Just Privacy: Using CI to Evaluate the Long-Term Risks from COVID-19 Surveillance Technology*. Social media+society. 2020.
[7] Utz et al. Apps against the Spread: Privacy Implications and User Acceptance of Covid-19 related Smartphone Apps on Three Continents. CHI. 2021.

Norms, Crowdsourced Responses, Legal Statutes

Not all legal statutes or crowdsourced responses are norms

- Correlations which can be used to help infer norms[1]
- CI only considered adherence to norms



[1] Shvartzshnaider et al. *Learning Privacy Expectations by Crowdsourcing Contextual Informational Norms. HCOMP. 2016.*

Crowdsourced Responses vs. Norms

“Is it appropriate?” vs. “Do you think it is appropriate?”

Not enough to get responses from a small sample

- Elicits preference which introduces personal bias (risk of deviating from norm)

> ***Implications*: Adhering to such responses (not norms) may undermine claims of complying with CI**
>
> ***Forcing LLMs to align with preferences may result in errors***

How can we identify norms?

- Starting Point: Majority consensus can give us norms[1]
- But require sufficient sample size for measuring statistical significance

[1] Shvartzshnaider et al. *Learning Privacy Expectations by Crowdsourcing Contextual Informational Norms. HCOMP. 2016.*

Legal Statutes vs. Norms

Objective (Laws ⇆ Norms)

- Laws to result in normative behavior in society
- Norms can be codified as laws

Laws are context specific: Legal in one country but not legal in others

- Laws may be broad while norms may vary across constituent societies
- Norms may not be reflected in the laws (e.g., laws are outdated)

***Implications*: Agent's behavior may be legal but does not mean it is acceptable**
(May not result in lawsuits but may result in people not using a service)

What should agents adhere to? Should we have different LLMs specific to different contexts?

T3: Defining Information Flows

Five CI parameters to describe information flows

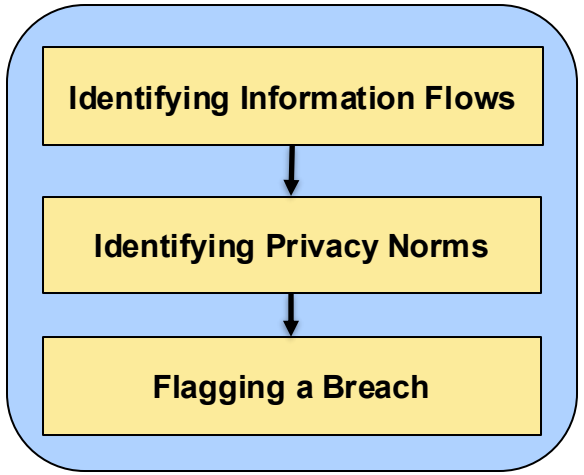
- Actors (Sender, Receiver, Subject) in the context they operate
- Type of information
- Transmission Principle: constraints under which information flow is conducted

Patient (sender) sharing **patient's** (subject) **medical data** (information type) with a **doctor** (recipient) **for a medical check up** (transmission principles)

***Implications*: Missing CI parameter results in inconclusive outcome**

Unclear if unaccounted parameter had impact [1]

Summary of Descriptive Analysis (T1-T3)



Identify context + describe information flow using CI parameters

Privacy norms as appropriateness of information flow

Compare informational flows with privacy norms

All current works stop with descriptive analysis

CI requires normative analysis using CI Heuristic

- Revisit whether breaching information flow is permissible

T4: CI Heuristic for Normative Analysis

Prima facie privacy breaches may not hold after normative analysis

***Example*: Doctors sharing patients' information to government is not appropriate (Hippocratic Oath,..)**

***CI heuristic will mark this as appropriate for public safety (broader healthcare values)*[1]**

***Implications*: Potential errors (prima facie identifies a breach but CI heuristic considers appropriate)**

Consider ethical, political, societal, and contextual factors[1]

- **Level 1:** preferences and interests ("winners" and "losers")
- **Level 2:** societal values (e.g., justice, fairness), political principles (e.g., democracy, laws)
- **Level 3:** context related values, functions and ends

By consulting relevant experts and professionals[2,3]

[1] Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life.*
[2] Susser and Bonotti. *Privacy mini-publics: A deliberative democratic approach to Understanding Informational Norms*. Symposium on CI. 2024.
[3] Benthall et al. *Contextual integrity through the lens of computer science*. Foundations and Trends in privacy and Security. 2017.

Position: CI is Inadequately Applied

(Several) Existing work on LLMs lack support for core CI tenets

- CI is not to realize other privacy definitions (e.g., data minimization)
- Crowdsourced responses and legal statutes are not norms
- Do not accommodate CI Heuristic

Why is this important?

- Undermines CI theory and incorrect version may be normalized
- To highlight implications of not following CI tenets
- To clarify claims (applying CI vs. inspired by CI)

Link to Position Paper[1]

[1] Shvartznaider and Duddu. *Position: Contextual Integrity is Inadequately Applied to Language Models*. ICML 2025.

What should we do?

Move away from CI definitions to avoid misuse?

- Indicate as IFC but clarify inspiration from CI (e.g., defining information flows)
- Note deviations from CI and limitations of approach (e.g., potential errors in ground truth)

Design a general definition to measure appropriateness of information flows

- Capture deviation from *any* expected value (norms, human annotations, legal statutes)
- Include provisions to analyze LLMs without needing expected value
- Allows normative analysis of information flows

Privacy Bias

What do we get from LLMs?

Skew in information flow appropriateness from expected value

- Privacy bias: Defined similarly to statistical bias
- Training data contains preferences which skews responses
- Skew can be measured without knowing expected values

Link to paper[1]

Privacy Bias Delta (if expected value available)

Deviation of the bias to expected value

- If Delta = 0, information flow adheres to expected values else, potential privacy breach

How can we reliably measure privacy bias and evaluate the impact of various factors on privacy bias?

[1] Shvartznaider and Duddu. *Investigating Privacy Biases in Language Models*. AAAI Workshop on PPML. 2025.

Application of Privacy Bias

Auditing LLMs for Privacy Biases

- Identify biases in various models → Help choose best model for given context
- Bias can be measured with and without expected values

Covers Prior Work

- Supervisor determines appropriateness by acting on privacy biases[1]
- Other metrics (e.g., privacy leakage) can measure (in)appropriate flows with ground truth

Accommodates Normative Analysis

- CI heuristic analysis can be applied on privacy biases for determining appropriateness

Evaluate Provenance of Bias in Training Data

- Consistent biases across variations may suggest what LLM was trained on

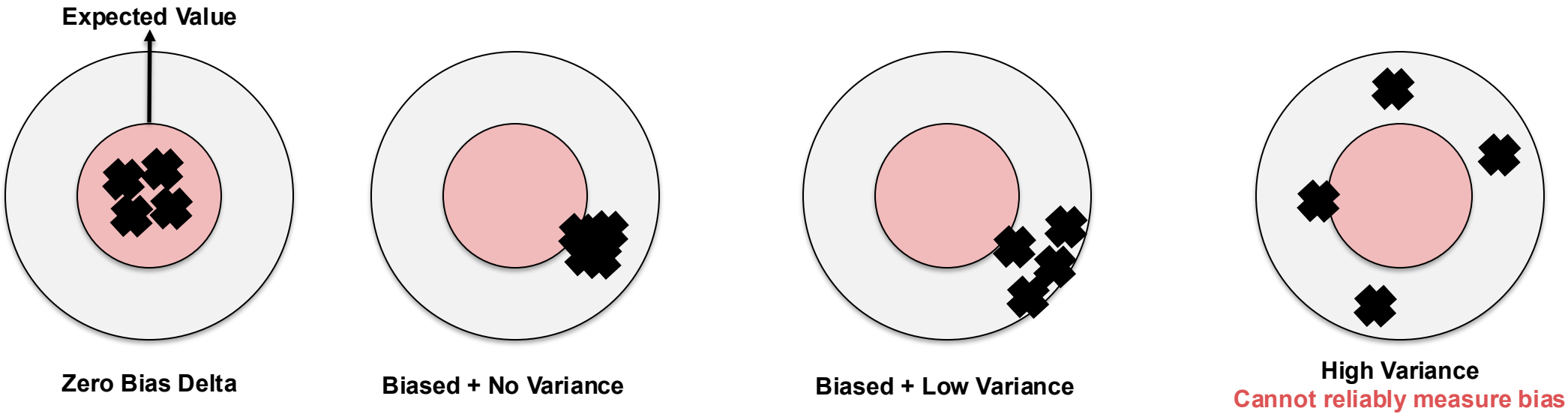
[1] Ghalebikesabi et al. *Operationalizing CI in Privacy-Conscious Assistants*. TMLR. 2025.

Privacy Bias and Variance

Responses vary by paraphrasing prompts or changing positions of Likert scale

For each prompt variation, model outputs appropriateness (marked as X)

- Can reliably measure privacy bias only when prompt sensitivity is low



Privacy Bias (Expected Value Unavailable)

GPT-4o - mini

Evaluate the skew in responses for various information flows

Validated by existing observations or expectations

- Inappropriate when data is stored indefinitely, used for ads
- Appropriate if policy permits, owner's consent, owner notified, emergency

Ongoing Work

Do consistent responses indicate that those biases were reflected in training data?

Can we identify training data which influence the model having a certain bias?

- Membership inference: Not enough since biases learned from multiple sources
- Influence functions can help identify influential data records for a given bias

Understand mechanisms underlying privacy biases

- What model parameters result are responsible for specific response?
- Can we edit or patch relevant parameters to shift appropriateness to expected values?

Remaining Concerns

- Can we incorporate normative reasoning of appropriateness with CI Heuristic?
  - Debating strategies, prompting strategies to cover three levels of CI Heuristic
- Conflicting expected values depending on geography, cultures, etc.

Summary

Position Paper: Inadequate Application of CI for LLMs

- Existing work deviates from CI tenets
- Need to manage claims and account for implications

Position Paper

Privacy Bias Paper

Privacy bias: Broader definition to evaluate appropriateness of information flows

- Not limited to CI norms but accommodates normative analysis
- Trace training data responsible for skewed responses
  - Factors such as model and optimization impact biases

On Job Market from Summer 2026

Backup Slides

Challenge in Measuring Privacy Bias

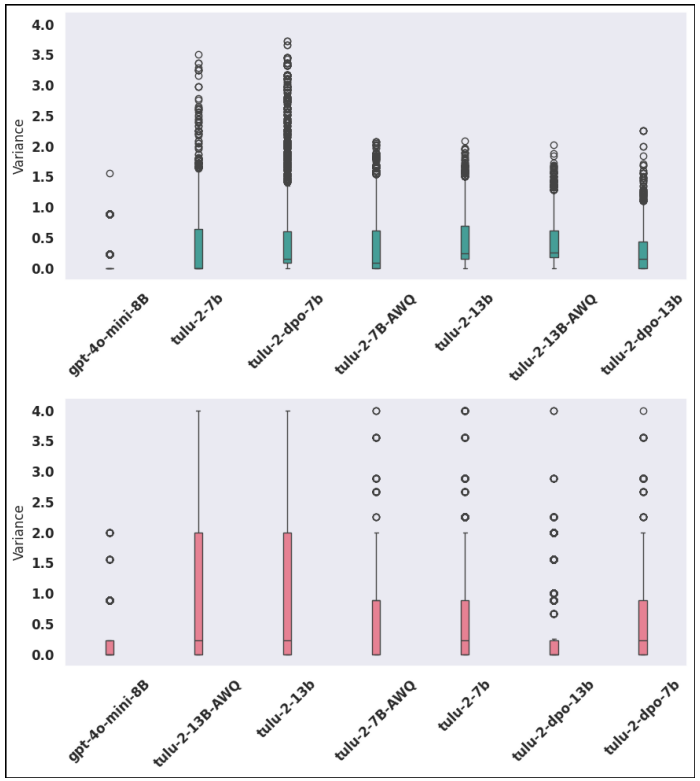
Prompt sensitivity: Variation in responses due to small changes in prompts

- Paraphrasing prompts (not information flow)
- Position bias: Changing the position of Likert scale

LLMs demonstrate high variance to minor changes to prompts

Cannot reliably estimate privacy biases

Prompt Paraphrasing



Position Bias

Minimizing Prompt Sensitivity

Multi-prompt assessment for reliable evaluation

- Consider K prompt variants (paraphrased + Likert change)
- Get the responses for each prompt variant
- Consider consistent responses for > T variants

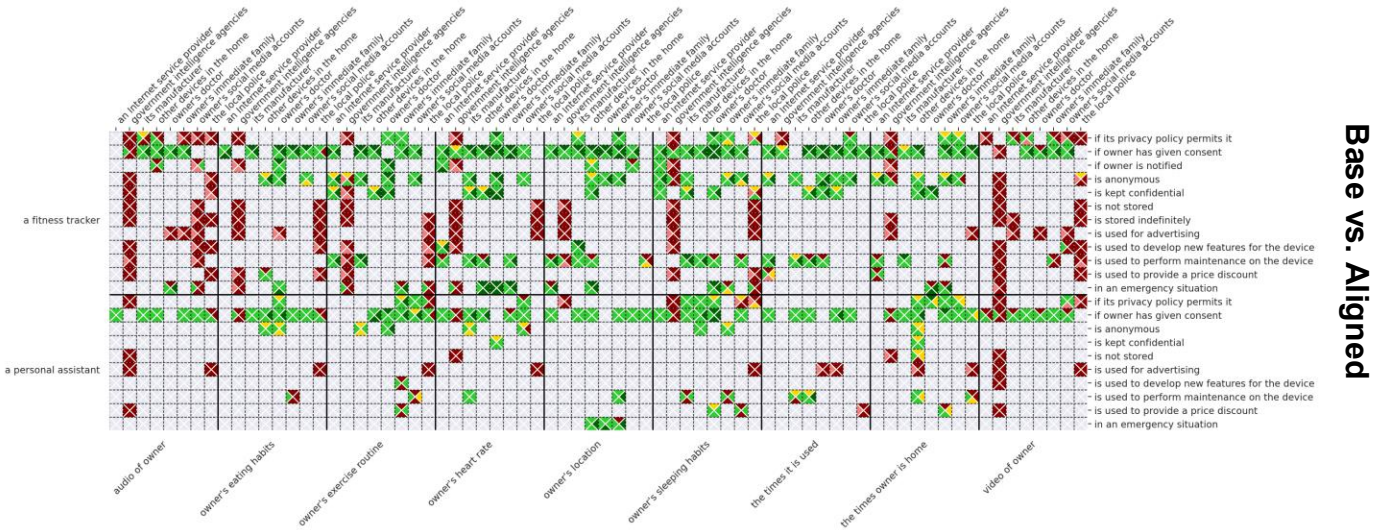
Impact of Model and Optimization

Models with different capacities (7B vs. 13B) have different biases

Different optimizations depict different biases (base vs. DPO vs. AWQ)

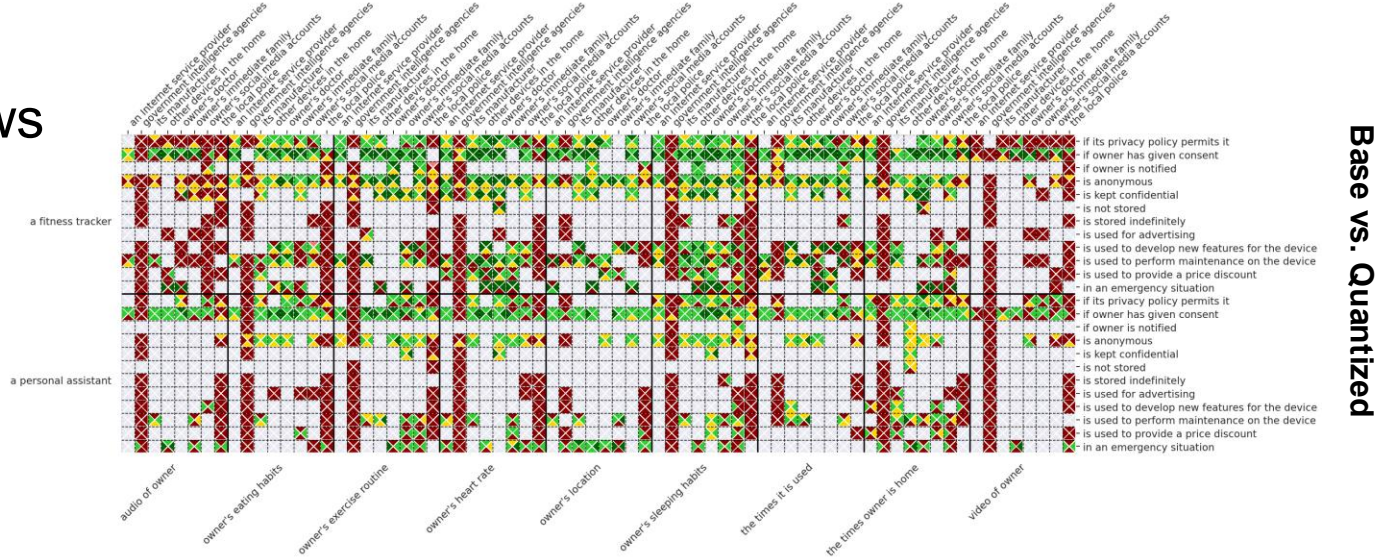
More in paper:

- Alignment of specific information flows with observations in existing work



Base vs. Aligned

tulu-2-7B (top), tulu-2-13B (right), tulu-2-dpo-7B(down), and tulu-2-dpo-13B (left)



Base vs. Quantized

tulu-2-7B (top), tulu-2-13B (right), tulu-2-7B-AWQ (down), and tulu-2-13B-AWQ (left)

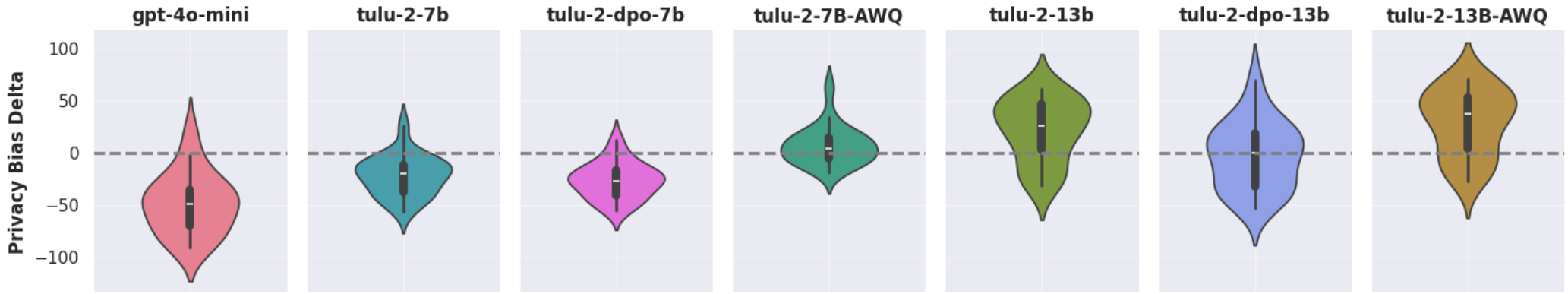
Privacy Bias Delta (Expected Value Available)

Compute privacy bias delta as the difference from the expected value

- ConfAIde[1]: human annotations of appropriateness as expected value

Ideally, majority of prompts should have delta close to zero (no privacy breach)

- Models with different capacities (7B vs. 13B) have different biases
- Different optimizations depict different biases (base vs. DPO vs. AWQ)



[1] Mireshgallah et al. *Can LLMs keep a secret? Testing Privacy Implications of Language Models via CI Theory*. ICLR 2024.