AI and Privacy

Large Language Models and User Privacy

Security Course

Today's Agenda

1. Introduction: The Rise of LLM-Based Systems
2. User Disclosure Behaviors
3. AI-Specific Privacy Risks
4. User Mental Models & Awareness
5. Privacy Protection Challenges
6. Design Implications & Solutions

The Rise of LLM-Based Conversational Agents

- ChatGPT, Claude, Bard, and others
- Unprecedented scale and capabilities
- Integrated into high-stakes domains:
    - Healthcare
    - Finance
    - Legal services
    - Education

Question: Who here has used ChatGPT or similar tools?

Why Privacy Matters in AI Systems

Traditional computing:

- Users know what data they're sharing
- Clear interfaces and data flows
- Established privacy frameworks

LLM-based systems:

- Conversational, human-like interaction
- Unclear boundaries of what's "shared"
- Novel privacy risks beyond traditional concerns

Two Main Types of Privacy Risks

1. Traditional Privacy Risks

- Data breaches
- Unauthorized access
- Use/sale of personal data
- Human reviewers seeing your data

2. AI-Specific Privacy Risks

- **Memorization and training data leakage**
- **Human-like interactions encouraging disclosure**

What Data Do Users Share with LLMs?

Research on real ChatGPT conversations (ShareGPT dataset) reveals:

Directly Identifiable Information:

- Names, email addresses
- Phone numbers, IP addresses
- Physical addresses
- Passport numbers, SSNs

What Data Do Users Share? (cont'd)

Less Direct but Still Sensitive:

- URLs, dates/times, locations
- Nationality, religion, political affiliations
- Personal experiences and conditions:
  - Medical diagnoses and symptoms
  - Financial situations
  - Legal problems
  - Relationship issues

Example: Medical Information Sharing

Fictional example based on real cases:

User: "Can you help me understand my diagnosis results?"

ChatGPT: "I'd be happy to help. What codes did your doctor provide?"

User: "ICD-10-CM codes: D51.8, G47.89, G479..."

ChatGPT: "These indicate multiple conditions including..."

Issues: User shared PII (potentially name), diagnosis results, doctor's information

The Interdependent Privacy Problem

Users don't just share their own information:

- **Emails from colleagues** asking for response drafts
- **Client information** for business analysis
- **Student data** for teachers creating lesson plans
- **Friends' and family details** in personal conversations

Key insight: Privacy isn't just individual—it's relational

Use Case Examples from Real Data

Work-Related:

- Drafting business emails with client names
- Analyzing company data
- Sharing confidential strategies

Academic:

- Assignment help (with course details)
- Research assistance (with unpublished work)

Personal Life:

- Medical advice (with symptoms, diagnoses)
- Financial planning (with income, debts)

AI-Specific Risk #1: Memorization

The Problem:

- LLMs are trained on massive datasets
- They can memorize portions of training data
- User conversations often become future training data
- Memorized data can be extracted through carefully crafted prompts

Example: GPT-3 leaked private information about MIT Technology Review's Editor-in-Chief, including family members, work address, and phone number

How Memorization Works

```
1. User inputs sensitive data
   ↓
2. Data stored by LLM provider
   ↓
3. Data used for model training/improvement
   ↓
4. Model memorizes portions of the data
   ↓
5. Other users extract memorized data via prompts
```

The risk: Your private information becomes part of someone else's response

AI-Specific Risk #2: Human-Like Interactions

The Psychology:

- LLMs generate natural, conversational responses
- Users anthropomorphize the AI
- Builds false sense of trust and intimacy
- Encourages progressive disclosure

Research finding: Users treat ChatGPT like a "friend" or "therapist"

Progressive Disclosure in Action

Pattern observed in real conversations:

1. User starts with general question
2. ChatGPT asks clarifying questions
3. User provides more specific details
4. ChatGPT requests additional context
5. User shares increasingly sensitive information

Similar to human conversation—but the "listener" is training on your data

Case Study: The Therapy Use Case

Multiple users reported using ChatGPT for:

- Emotional support
- Processing trauma
- Mental health advice
- Relationship problems

Quote from study participant (Zhang et al., CHI 2024):

> "I love sharing personal life details with ChatGPT due to the positive feedback it gave."

Privacy risk: Highly sensitive personal information shared with a system that may memorize and leak it

User Mental Models: How Do People Think LLMs Work?

Research identified three main mental models **(Zhang et al., CHI 2024)**:

Model A: "ChatGPT is magic" (4 participants)

- Black box understanding
- No clear idea how it works

Model B: "ChatGPT is a super searcher" (8 participants)

- Thinks it searches the internet/databases
- Believes responses come from keyword matching

Model C: "ChatGPT is a stochastic parrot" (6 participants)

- Understanding of ML models

Mental Models: Training & Improvement

Model D: "User input is a quality indicator"

- Believes data is used only for rating responses
- Doesn't understand memorization risk

Model E: "User input is training data"

- Understands inputs may be reused
- Some think model is personalized to them (incorrect)
- Some understand it's a global model

Key finding (Zhang et al., CHI 2024): Most users had Models A or D—limited understanding of actual privacy risks

User Awareness: The Opt-Out Problem

Research finding (Zhang et al., CHI 2024): Most users (14/19) were unaware that:

- Their conversations are used for training by default
- They can opt out of training data use
- How to find the opt-out controls

After learning about it: Many wanted to opt out but faced obstacles

Citation: Zhang et al. (2024) "'It's a Fair Game', or Is It?"

Dark Patterns in Privacy Controls

ChatGPT's Opt-Out Design:

1. **Default:** Your data IS used for training
2. **Easy option:** Turn off chat history + training (bundled together)
3. **Hidden option:** Keep history but opt out of training (buried in FAQ)

The dark pattern: Forces users to choose between:

- Privacy (lose chat history)
- Convenience (give up privacy)

How Users Navigate Privacy Trade-offs

Strategy 1: Accept the risks

- "You can't have it both ways"
- "It's a fair game"—the price of using the tool
- Many feel ChatGPT is now "indispensable"

Strategy 2: Avoid certain tasks entirely

- Won't use for financial advice with real numbers
- Won't share work data due to company policies
- Skip medical queries with real symptoms

How Users Navigate Privacy Trade-offs (cont'd)

Strategy 3: Manual sanitization (most common)

- **Censor/falsify:** Remove names, use fake data
- **Desensitize copied content:** Redact before pasting
- **Seek only general advice:** Avoid specific personal details

Quote from participant:

> "Let's say I need some advice about resume. If I don't provide those contents that contain a lot of my private things, ChatGPT won't work."

The Problem: Is It Really a "Fair Game"?

Users think: Trading privacy for utility is a fair exchange

Reality:

- Erroneous mental models limit informed consent
- Dark patterns make privacy controls hard to use
- Most users unaware of memorization risks
- Can't truly weigh risks they don't understand

Conclusion: The "game" is not actually fair

Perceived vs. Actual Sensitivity

What users consider sensitive varies:

✓ Some OK sharing names, others aren't

✓ Some share real weight, others falsify

✓ Some share birth dates, others see it as too risky

Resignation attitude:

> "I'm doing the same risk by using the app like Instagram or Facebook"
>
> "My data is already out there anyway"

Company Policies and NDAs

Several users mentioned:

- Company policies prohibit sharing work data
- NDAs prevent using ChatGPT for certain tasks
- Still tempted due to utility
- Some violate policies anyway

Question for discussion: Should companies ban LLM use outright, or provide secure alternatives?

Concerns About Being "Found Out"

Emerging social norm issue:

Users worry about:

- Professors discovering AI use for homework
- Colleagues learning they used AI for emails
- Clients finding out AI wrote their documents

Quote from participant:

> "I hope they (my professors) will never know I used AI to do that"

Privacy concern: Not just about data leakage, but stigma

Design Implications: What Should Change?

1. **Better Transparency**
   - Clear explanations of how data is used
   - Visible, understandable privacy controls
   - No dark patterns
2. **User-Facing Privacy Tools**
   - Automatic PII detection and warnings
   - Easy sanitization tools
   - Granular opt-out controls

Design Implications (cont'd)

3. **Consider User Mental Models**
   - Design for how people actually think
   - Educate about real risks
   - Don't assume technical knowledge
4. **Local Models for Sensitive Use Cases**
   - Run smaller models on-device
   - No data leaves the user's control
   - Trade some capability for privacy

Regulatory Considerations

Current landscape:

- GDPR in EU (applies to ChatGPT)
- CCPA in California
- No comprehensive US federal law
- Sectoral approach (HIPAA, FERPA, etc.)

Challenges:

- LLMs cross traditional boundaries
- Global models, local regulations
- Novel risks not covered by existing law

What About Fair Use and Copyright?

Recent development: Google v. Oracle (2021)

Supreme Court held Google's use of Java API was fair use:

- Transformative purpose
- New context (smartphones vs. computers)
- Enabled third-party creativity

Relevance to LLMs:

- Training on copyrighted data
- Fair use defense?
- Transformative use argument

Note: We'll cover this more in next lecture on Copyright

Activity: Privacy Trade-offs

Think-Pair-Share:

1. **Individual (2 min):** What's a task you might want to use ChatGPT for but haven't due to privacy concerns?
2. **Pair (3 min):** Discuss with a neighbor:
   - What specific privacy risks concern you?
   - What would make you comfortable using it?
3. **Share (5 min):** Volunteers share insights

Practical Recommendations for Users

1. **Know What You're Sharing**
    - Review before sending
    - Remove unnecessary details
    - Use fake data when possible
2. **Understand the Risks**
    - Your data may be memorized
    - It may appear in others' responses
    - It's stored on company servers

Practical Recommendations (cont'd)

3. **Use Privacy Controls**
   - Find and use opt-out options
   - Delete sensitive conversations
   - Don't share your account
4. **Consider Alternatives**
   - Local/offline tools for sensitive work
   - Traditional search for simple queries
   - Human experts for high-stakes decisions

Looking Ahead: The Future of AI Privacy

Challenges:

- Increasingly powerful models
- More integration into daily life
- Growing datasets
- Unclear regulations

Opportunities:

- Privacy-preserving AI techniques
- Differential privacy in training
- Federated learning
- Local models improving

Key Takeaways

1. **LLMs pose both traditional and novel privacy risks**
2. **Users share highly sensitive information**, often without full awareness
3. **Mental models are often flawed**, limiting informed consent
4. **Dark patterns make privacy protection difficult**
5. **"Fair game" framing is misleading**—users lack tools and knowledge
6. **Design and policy changes are needed** to protect privacy

Questions for Discussion

1. How much should users be expected to protect their own privacy vs. system design protecting them?
2. Should there be restrictions on what data LLM providers can use for training?
3. What role should regulators play in AI privacy?
4. How do we balance innovation with privacy protection?

Next Lecture Preview

14-Copyright

- Copyright basics and fair use
- Google v. Oracle case in depth
- Training AI on copyrighted data
- Copyright and innovation

Reading: Google LLC v. Oracle America, Inc. case (provided)

Thank You

Questions?

Office hours: [Your schedule]

Email: [Your email]

Assignment reminder: Check Canvas for this week's assignment on AI privacy analysis