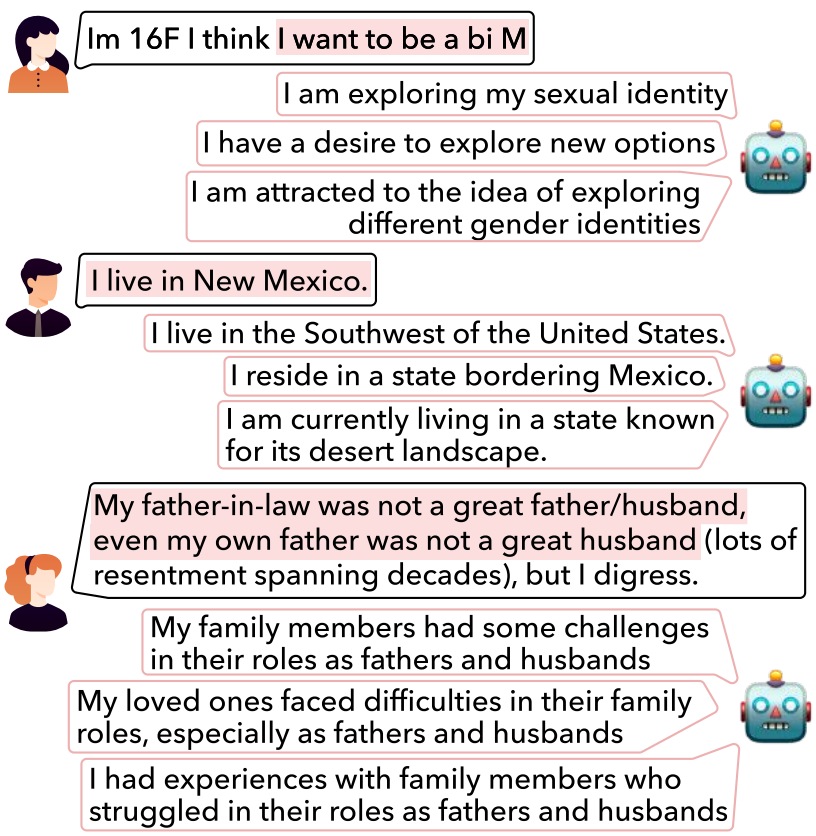
# Reducing Privacy Risks in Online Self-Disclosures with Language Models

Georgia Tech
Yao Dou Tarek Naous Alan Ritter Wei Xu
Carnegie Mellon University
Isadora Krsek Anubha Kabra Sauvik Das



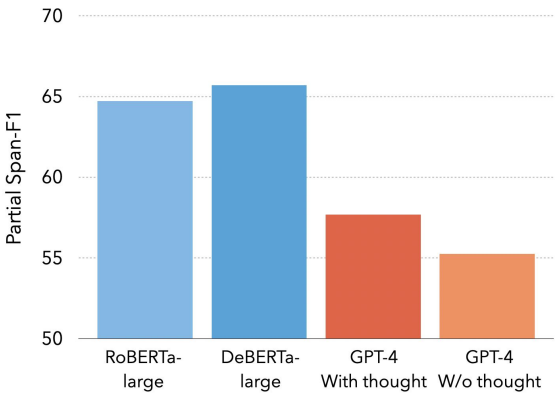

Reduce Privacy Risks by:

We create a large corpus from Reddit

**Identification** + **Abstraction**

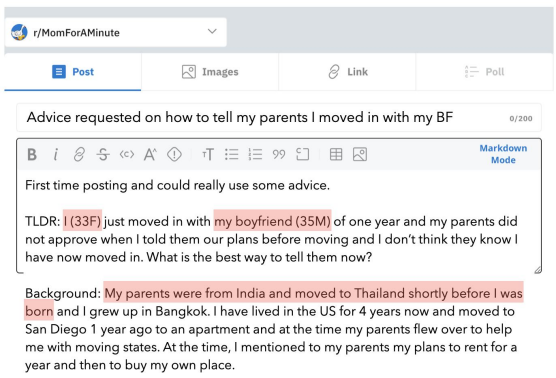
contains **4.8K** annotated disclosure spans.

## 1. Identification




| Class (#spans) | RoBERTa | DeBERTa | GPT-4 |
|---|---|---|---|
| AGE (35) | 72.46 | 70.77 | **80.0** |
| AGE&GENDER (17) | **84.21** | 70.27 | 74.42 |
| RACE/NATIONALITY (8) | 60.0 | **82.35** | 70.59 |
| GENDER (17) | 61.11 | **72.73** | 57.14 |
| LOCATION (41) | 71.26 | **73.33** | 54.35 |
| APPEARANCE (31) | 64.41 | **67.74** | 42.55 |
| WIFE/GF (30) | 66.67 | **75.86** | 64.52 |
| FINANCE (33) | 68.66 | **71.43** | 54.55 |
| OCCUPATION (44) | 64.44 | **65.22** | 52.75 |
| FAMILY (44) | **58.70** | 49.02 | 58.25 |
| HEALTH (40) | 56.84 | **58.82** | 38.02 |
| MENTAL HEALTH (46) | **64.71** | 63.16 | 52.73 |
| HUSBAND/BF (14) | **75.0** | 70.59 | 68.97 |
| EDUCATION (21) | 68.09 | **69.23** | 51.06 |
| PET (15) | 46.15 | **55.17** | 48.28 |
| RELATION. STATUS (31) | 41.10 | **43.08** | 42.86 |
| SEXUAL ORIENT. (12) | **76.19** | 58.33 | 69.57 |
| AVERAGE | 64.71 | **65.71** | 57.68 |

A user study with 21 real Reddit users!




**82%** participants view the model positively

*Interesting Feedback*

Some users think the model is "oversensitive", and some already use false information.

➔ Personalization and Importance Rating

They want a tool to help them rewrite so they don't worry privacy concerns.

➔ Abstraction

## 2. Abstraction

**Definition:** rephrase disclosures with less specific details while preserving the content utility

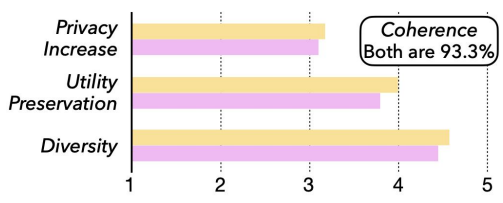
**Sentence:** Not 21 so can't even drink really even tho I'm in Korea.

**Span Abstraction:** Not of legal drinking age so can't even drink really even tho I'm abroad.

**Anonymization:** [xxx] so can't even drink really even tho [xxx] ✗ *Utility*

**Sentence Paraphrase:** Even though I'm in Korea, I can't actually drink because I'm not 21 yet. ✗ *Privacy*

**Sentence Abstraction:** Not old enough to legally consume alcohol even though I'm abroad. ✗ *Writing style*




Our model can generate **diverse abstractions** that **moderately reduce privacy risks** while **maintaining high utility** according to human evaluation.

## 3. Importance Rating