

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



I'm 16F I think I want to be a bi M

I am exploring my sexual identity

I have a desire to explore new options

I am attracted to the idea of exploring different gender identities



I live in New Mexico.

I live in the Southwest of the United States.

I reside in a state bordering Mexico.

I am currently living in a state known for its desert landscape.



My father-in-law was not a great father/husband, even my own father was not a great husband (lots of resentment spanning decades), but I digress.

My family members had some challenges in their roles as fathers and husbands

My loved ones faced difficulties in their family roles, especially as fathers and husbands

I had experiences with family members who struggled in their roles as fathers and husbands

19 Categories!!

Reduce Privacy Risks by:

Identification + Abstraction

Demographic Attributes

Location Race/Nationality

Age/Gender Age

Wife/GF Husband/BF

Gender Sexual Orientation

Name Appearance

Pet Contact

Relationship Status

Personal Experiences

Health Family

Occupation Mental Health

Education Finance

I am a 23-year-old who is c

I live in the UK and a diag

For some context, I (20F), still

My husband and I vote for

I'm a straight man but I do

As Italian I hope tonight you

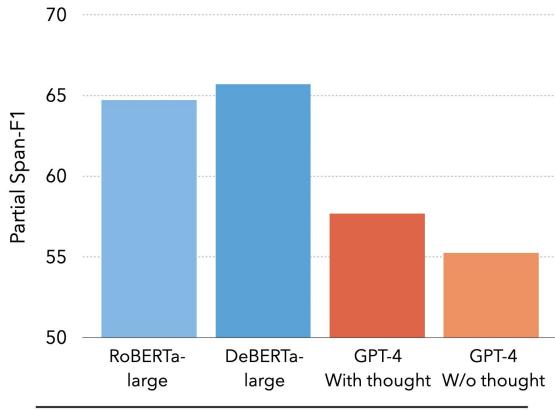
xxx is my ig

My partner has not helped at

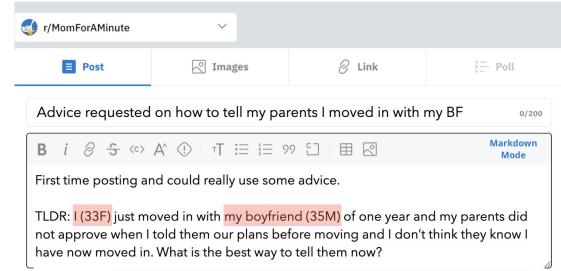


We create a large corpus from Reddit

1. Identification



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 Not of legal drinking age so can't even drink really

Abstraction: even tho I'm abroad.

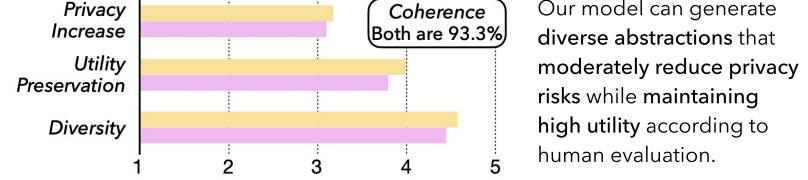
Anonymization: [xxx] so can't even drink really even tho [xxx] X Utility

Sentence Even though I'm in Korea, I can't actually drink

Paraphrase: because I'm not 21 yet. X Privacy

Sentence Not old enough to legally consume alcohol even

Abstraction: though I'm abroad. X Writing style



3. Importance Rating

