

Reducing Privacy Risks in Online Self-Disclosures with Language Models



Yao Dou Tarek Naous
Alan Ritter Wei Xu



Isadora Krsek Anubha Kabra
Sauvik Das

Demographic Attributes

- Location: "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."
- Race/Nationality: "As Italian I hope tonight you"
- Age/Gender: "I am a 23-year-old who is", "I live in the UK and a diag", "For some context, I (20F), still"
- Wife/GF: "My husband and I vote for", "I'm a straight man but I do"
- Gender: "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"
- Name: "xxx is my ig"
- Appearance: "My partner has not helped at"
- Pet: "My little brother (9M) is my prid"
- Contact: "I struggle with depressio"
- Relationship Status: "Hi there, I got accepted to UCLA ("

Personal Experiences

- Health: "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"
- Family: "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"
- Mental Health: "I struggle with depressio"
- Occupation: "I struggle with depressio"
- Finance: "Hi there, I got accepted to UCLA ("

19 Categories!!

Demographic Attributes

- Location
- Race/Nationality
- Age/Gender
- Wife/GF
- Gender
- Name
- Appearance
- Pet
- Contact
- Relationship Status
- Personal Experiences
- Health
- Family
- Occupation
- Mental Health
- Education
- Finance

Reduce Privacy Risks by:

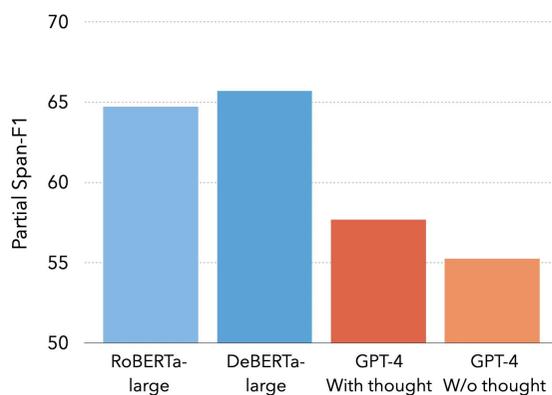
Identification + Abstraction

We create a large corpus from Reddit

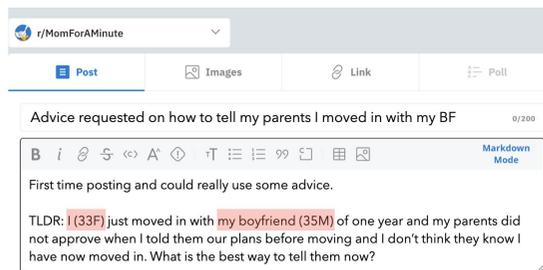


contains **4.8K** annotated disclosure spans.

1. Identification



A user study with 21 real Reddit users!



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

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.

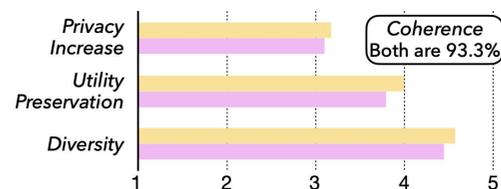
Anonymization: [xxx] so can't even drink really even tho [xxx] ~~X~~Utility

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



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

3. Importance Rating

