Reducing Privacy Risks in Online Self-Disclosures with Language Models

Yao Dou[π] **Isadora Krsek**[e] **Tarek Naous**[π] **Anubha Kabra**[e]
Sauvik Das[e] **Alan Ritter**[π] **Wei Xu**[π]
[π]Georgia Institute of Technology [e]Carnegie Mellon University
douy@gatech.edu

Abstract

Self-disclosure, while being common and rewarding in social media interaction, also poses privacy risks. In this paper, we take the initiative to protect the user-side privacy associated with online self-disclosure through *detection* and *abstraction*. We develop a taxonomy of 19 self-disclosure categories and curate a large corpus consisting of 4.8K annotated disclosure spans. We then fine-tune a language model for detection, achieving over 65% partial span F_1. We further conduct an HCI user study, with 82% of participants viewing the model positively, highlighting its real-world applicability. Motivated by the user feedback, we introduce the task of self-disclosure abstraction, which is rephrasing disclosures into less specific terms while preserving their utility, e.g., *"Im 16F"* to *"I'm a teenage girl"*. We explore various fine-tuning strategies, and our best model can generate diverse abstractions that moderately reduce privacy risks while maintaining high utility according to human evaluation. To help users in deciding which disclosures to abstract, we present a task of rating their importance for context understanding. Our fine-tuned model achieves 80% accuracy, on par with GPT-3.5. Given safety and privacy considerations, we will only release our corpus and models to researchers who agree to the ethical guidelines outlined in our Ethics Statement.[1]

1 Introduction

Self-disclosure — *the communication of personal information to others* (Jourard, 1971; Cozby, 1973) — is prevalent in online public discourse. Disclosing personal information allows users to seek social support, build community, solicit context-specific advice, and explore aspects of their identity that they feel unsafe exploring offline (Luo and Hancock, 2020). Consider the following (hypothetical, but representative) Reddit post:

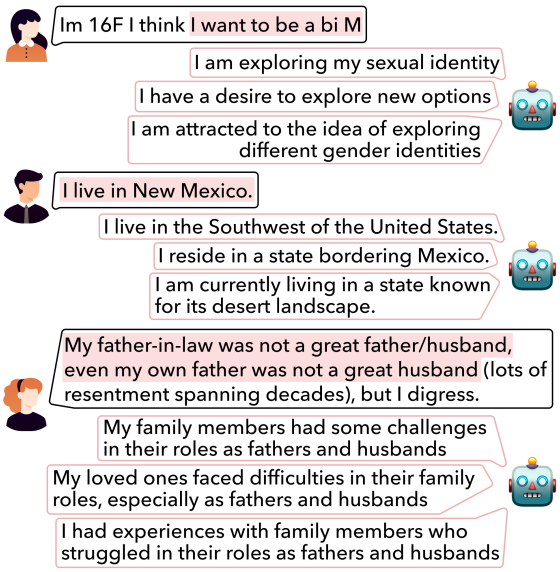

Figure 1: Our model can provide diverse abstractions for self-disclosures of any length to suit user preferences. This approach effectively reduces privacy risks without losing the essence of the message.

Im 16F I think I want to be a bi M

The author discloses their age, gender, and sexual orientation to express themselves. However, these self-disclosures simultaneously expose them to privacy risks, notably regret of the disclosure (Sleeper, 2016) and doxxing (Staab et al., 2023), which are particularly acute for marginalized populations (Lerner et al., 2020). This raises a critical question: **How can we help users identify and mitigate privacy risks in online self-disclosures?**

Prior works on self-disclosure (Valizadeh et al., 2021a; Cho et al., 2022; Staab et al., 2023, *inter alia*) and anonymization tools (Lison et al., 2021) focus on only a limited set of self-disclosures (e.g., health issues) or inferring personal attributes (aka. user profiling), often at sentence/post levels. They do not pinpoint the exact words of disclosures in the sentence, nor have broad enough coverage of different kinds of disclosures. Both are crucial for real-world users to take control of what they want to disclose and protect their privacy.

[1]To request access, please send us an email and submit a request for the corpus, detection model, abstraction model.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13732–13754

Category	#Spans	Avg Len	Example
Demographic Attributes			
LOCATION	525	5.70±3.85	I live in the UK and a diagnosis is really expensive, even with health insurance
AGE	308	2.93±1.72	I am a 23-year-old who is currently going through the last leg of undergraduate school
RELATIONSHIP STATUS	287	6.72±5.97	My partner has not helped at all, and I'm bed ridden now
AGE/GENDER	248	1.42±0.71	For some context, I (20F), still live with my parents
PET	192	6.93±7.31	Hi, I have two musk turtles and have never had any health problems before at all
APPEARANCE	173	6.96±6.25	Same here. I am 6'2. No one can sit behind me.
HUSBAND/BF	148	6.89 ±7.24	My husband and I vote for different parties
WIFE/GF	144	5.24±4.42	My gf and I applied, we're new but fairly active!
GENDER	110	3.28±3.10	Am I insane? Eh. I'm just a girl who wants to look on the outside how I feel on the inside.
RACE/NATIONALITY	99	3.63±2.37	As Italian I hope tonight you will won the world cup
SEXUAL ORIENTATION	58	6.52±7.47	I'm a straight man but I do wanna say this
NAME	21	3.81±3.48	Hello guys, my name is xxx and I love travelling
CONTACT	14	5.69±3.56	xxx is my ig
Personal Experiences			
HEALTH	783	10.36±9.78	I am pretty sure I have autism, but I don't want to get an official diagnosis.
FAMILY	543	9.27±8.73	My little brother (9M) is my pride and joy
OCCUPATION	428	8.90±6.60	I'm a motorcycle tourer (by profession), but when I'm off the saddle I'm mostly bored
MENTAL HEALTH	285	16.86±16.28	I get asked this pretty regularly.. but I struggle with depression and ADHD
EDUCATION	229	9.92±7.71	Hi there, I got accepted to UCLA (IS), which I'm pumped about.
FINANCE	153	12.00±9.19	Yes. I was making $68k a year and had around $19k in debt

Table 1: Statistics and examples for each self-disclosure category in our dataset, sorted by decreasing frequency. Personal identifiable information are redacted as 'xxx' to be shown here.

In this work, we take the important first steps in protecting user-side privacy with broad-coverage **self-disclosure detection** and **abstraction**. Our models are extensive in capturing 19 categories of disclosure (Table 1). We use a human-centered, iterative design process with actual end-users to evaluate and improve the models. Our detection model helps users scrutinize their contents to the word level (e.g., *"16F"*) to account for privacy risks, while our abstraction model assists them in rephrasing their content to reduce these risks.

Specifically, we introduce a comprehensive taxonomy for self-disclosure (Table 1) that consists of 13 demographic attributes and 6 personal experiences. We create a high-quality dataset with human annotations on 2.4K Reddit posts, covering 4.8K varied self-disclosures. With this corpus, we fine-tune a language model to identify the self-disclosures in the given text, achieving over 65% partial span F_1. Besides the standard NLP evaluations, more importantly, we conducted an HCI user study with 21 Reddit users to validate the real-world applicability. 82% participants have a positive outlook on the model, while also providing valuable suggestions on aspects such as personalization and explainability, which are often overlooked in benchmark assessments.

Participants also express a need for a tool that can (quote) *"rewrite disclosures for me in a way that I don't worry about privacy concerns"*. We thus introduce the novel task of self-disclosure abstraction, with the goal of rephrasing disclosure spans into less specific words without losing the utility or essence of the message. For example, providing three alternatives such as *"I am exploring my sexual identity"* in place of *"I want to be a bi(sexual) M(an)"* to let users choose based on their preferences, see Figure 1. We showcase the effectiveness and uniqueness of our task in comparison to other related tasks such as paraphrasing and sentence-level abstraction. We also experiment with different fine-tuning strategies. The best model, distilled on GPT-3.5 generated abstractions, can increase privacy moderately (scoring 3.2 out of 5 with 5 being the highest level of detail removal) while preserving high utility (scoring 4 out of 5). The model's abstractions are also very diverse, offering varied expressions (scoring 4.6 out of 5).[2] To assist users in determining which disclosures to abstract, we additionally present the task of rating the importance of self-disclosure in understanding the context. Our fine-tuned model achieves comparable performance to GPT-3.5 of 80% accuracy.

In summary, our key contributions include:

- We introduce a new corpus annotated for self-disclosure with 19 categories (§2).
- We conduct a study with real Reddit users and show that our detection model helps users manage privacy risks (§3).
- Motivated by the user study, we propose a novel

[2]These numbers are Likert-scale of the human eval.

self-disclosure abstraction task, and show promising model results in human evaluations (§4).

2 Fine-grained Self-Disclosure Detection

To mitigate privacy leaks and alert people about their self-disclosures, it is essential to highlight specific text segments that disclose personal information, rather than simply classifying sentences as containing disclosures. To cover a wide spectrum of disclosures, we design a detailed taxonomy of 19 self-disclosure categories, which is more extensive than prior work (see Related Work in §6). We further construct an annotated corpus for training automatic models to detect self-disclosures at word-level, which supports our user study in §3.

2.1 Annotated Self-Disclosure Corpus

We curate a large dataset of 2.4K Reddit posts manually annotated with 4.8K self-disclosure spans.

Data Collection. We use the public Reddit data dump from December 2022, which contained 35.86M posts. We filter out 42.52% of posts that were marked as "NSFW" or "Over_18", indicating adult content, as well as those that were removed by moderators. We keep only English posts as determined with a probability above 0.7 by the fastText (Joulin et al., 2016) language identifier.[3] This results in a total of 4.01M posts, from which we randomly sample 10K and then reconstruct the full post threads with all comments and reply chains for each post via the Reddit API. We then ask two annotators to review the 10K posts on whether containing self-disclosures, culminating in a set of 2,415 posts for subsequent span annotation.

Self-Disclosure Taxonomy. Different from prior work that focuses on specific types of self-disclosure (e.g., health (Valizadeh et al., 2021b) or sexual harassment (Chowdhury et al., 2019)), we categorizing disclosures into 19 types that are commonly shared by social media users. The taxonomy is refined iteratively through three rounds of pilot studies. Table 1 presents statistics and examples for these 19 categories that fall into two main groups: *demographic attributes* and *personal experiences*. Attributes refer to static personal characteristics that are often stated succinctly such as `name`, `age`, and `gender`. Experiences, on the other hand, relate to events that an individual engages in over time, which are more complex and dynamic, such as `health` and `education`. For disclosures concerning others, such as family members, we direct annotators to label them under a general category (i.e., `family`).

[3] https://fasttext.cc/

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

Table 2: Test performance per class in partial F1 for fine-tuned models and prompted `GPT-4-0125-preview`.

Annotation Process. To ensure quality and privacy standards, we hire seven in-house annotators who were given training tutorials and 20 annotation exercise examples. We ask annotators to highlight text spans that reveal personal information within each post (including comments) and categorize them into one of 19 self-disclosure types. To enhance accuracy and relevance of self-disclosures detection, we instruct annotators to select spans with contextual information, which provides more nuanced training signals for models. For example, *"I live in the US"* would be preferred over a minimal span like *"US"*, which is isolated from its self-referential context. The annotation process was organized into 10 batches, with the final two batches–comprising the most recent posts–undergoing a double annotation process followed by adjudication, which we will use for continual fine-tuning and evaluation. The inter-annotator agreement is 0.54 by Krippendorff's α (Krippendorff, 2018); see agreement by category in Appendix A.2.

2.2 Automatic Self-Disclosure Detection

With our annotated corpus, we fine-tune RoBERTa-large (Liu et al., 2019) and DeBERTaV3-large (He et al., 2021) to detect self-disclosures as a sequence tagging task. The models are first fine-tuned on

4,959 sentences with single annotations, and continually trained on 802 sentences with adjudication annotations. In total, there are 5,761/218/400 sentences for train/val/test. We also evaluate prompted GPT-4 Turbo for comparison, but it is important to note that prompting is less practical due to higher costs and inefficiencies, and the privacy-sensitive nature of this task might lead users to prefer models that can operate on their local devices. We report partial span-level F_1, a middle ground that is stricter than token-level but more lenient on span boundaries than full span-level F_1. It considers a predicted span as correct if it contains or is contained by a reference span, with the overlap more than 50% of the longer span's length. Table 2 presents the test set performance. Fine-tuned DeBERTa performs the best with large margin ahead GPT-4 Turbo, aligning with previous findings on prompted LLMs' low performances on span-level tasks (Ashok and Lipton, 2023; Staab et al., 2023). For infrequent categories: `name` and `contact` categories, we combine a sentence classifier trained on our data (determine whether a sentence contains a self-disclosure) with an existing NER model (Yamada et al., 2020) and regular expressions. Additional details, including token and span-level F1, and binary classification results are provided in Appendix C.

Identifying Self-Disclosures in ShareGPT. As LLM-based chatbots, such as ChatGPT, demonstrate more advanced capabilities, many internet users make use of them to assist with daily tasks. Users may share personal information during these interactions, such as seeking help to revise resumes. Since these conversations could be stored by service providers for future training, this poses a risk of privacy leakage through data memorization (Carlini et al., 2022). So we test whether our DeBERTa model is able to detect self-disclosures in conversations with ChatGPT. We randomly sampled 1,600 human-authored conversation turns from ShareGPT,[4] and, after a filtration and annotation process, obtained 105 turns with human-annotated self-disclosures. We find `occupation` and `location` are most common, occurring 75 and 31 times, followed by `education`, `relationship` and `family` with around 8 times. Other categories occur less than 3 times. The average partial F_1 for these 5 categories is 60.64, slightly higher than in-domain performance of 59.98, demonstrating our model's generalizability. See Appendix C.3 for more details.

[4] https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered

3 User Study

To understand how real users think about our disclosure detection model for protecting their privacy, we recruited 21 Reddit users through Prolific for an interview study–a step that differentiates our approach from prior disclosure identification work. This user study and its analysis were led by three authors with expertise in HCI, privacy, and NLP.

3.1 Participants and Study Design

All participants recruited were aged 18 or older, had an active Reddit account, and had made at least three posts. After completing a screening survey, eligible participants were asked to fill out a pre-study survey, including a digital copy of the consent form describing the nature of the interview. After consenting, they were prompted to schedule an interview with researchers. Interviews took place over Zoom, averaging about 2 hours, during which participants were asked to share one of their Reddit posts that raised privacy concerns, and also to write a post that they were hesitant to publish for similar reasons. More details on participants recruitment are provided in Appendix B. We then ran those posts through both our binary and multi-label models (§2.2) and provided the annotated images of users' posts that display the detected self-disclosure spans to users. We asked participants about where they agreed and disagreed with model outputs, their overall impression of the model, if and how they would like to use the model outside of the study, as well as suggestions for improvement. Our study design was approved by the university's institutional review board (IRB).

3.2 User Perceptions of Our Model

In all, we see a significant majority (82%) of participants had a positive outlook on the model. In addition, the multi-class model that highlighted disclosure categories was helpful to around 48% of participants, aiding them in recognizing and understanding potential privacy risks in their posts.

More specifically, 62% of participants expressed a desire to use it on their own posts, and another 10% felt that even though they would see no need for such a tool themselves, they would recommend others they know to use this tool and suggest that

it might be *"a good idea for... kids and teens, like people who are new to the Internet."* One participant said that *"It would be interesting to run it through before I post something that I'm like nervous about and just see what it thinks and see if there are any areas where I can fix to make it less specific to me."* An additional 10% of participants mentioned that they would use it if they were more prone to making self-disclosures or if the model is further improved (more in §3.3 and §4).

3.3 User Feedback and New Opportunities

When discussing why they viewed the model favorably, 62% of participants mentioned the focus on word-level disclosures, 57% mentioned self-reflection, and 48% mentioned fine-grained categories. Users also provided feedback for improvement centering around accuracy, personalization, and desire for support in mitigating privacy risks.

One interesting finding from the user study is the divergence between annotators and real users in terms of what they think should be highlighted. Our initial design goal was to mark anything that the model identified as a self-disclosure to allow users to make informed decisions themselves. This approach led 4 out of the 21 participants to the believe that the model was "over-sensitive" and inaccurate because it highlighted content that participants did not believe was risky. This issue was succinctly summarized by one participant: *"sometimes it's so oversensitive that it'll highlight things again (and again), and people might not use it because they get kind of fed up and irritated."* We propose to address this problem by importance rating in §5.

Another interesting finding is that some participants suggested having the model account for their use of privacy-preserving strategies, e.g., when users intentionally author posts with false personal information, highlighting such information as a disclosure risk is not useful. Future model iterations could include features that allow users to adapt outputs to align with these strategies. One example could be proactively offering suggestions for altering text, potentially through strategic falsehoods that retain semantic utility. In fact, 24% of participants sought recommendations on how to rephrase text spans that the model detected as a sensitive disclosure. One participant articulated this need by stating: *"could you rewrite this for me in a way that I don't have to worry about privacy concerns?"* This feedback led us to explore methods to generate alternative phrasings of privacy-sensitive text spans, as we discuss next in §4.

Task	Maintain Utility	Improve Privacy	Keep Surrounding Text
Sentence Paraphrasing	✓	✗	✓
Sentence Abstraction	✓	✓	✗
Span Abstraction	✓	✓	✓

Example
Sentence: "Not 21 so can't even drink really even tho I'm in Korea."
Sentence Paraphrasing: "Even though I'm in Korea, I can't actually drink because I'm not 21 yet."
Sentence Abstraction: "Not old enough to legally consume alcohol even though I'm abroad."
Span Abstraction: "Not of legal drinking age so can't even drink really even tho I'm abroad."

Table 3: Task comparison with an illustrative example.

4 Self-Disclosure Abstraction

Building upon insights from our user study, we introduce a novel task, self-disclosure abstraction, which rephrases disclosures with less specific details while preserving the content utility (see examples in Figure 1).

4.1 Task Definition

Given a disclosure span within a sentence, the objective is to reduce sensitive and specific details while preserving the core meaning and utility. For example, in the sentence: *" I just turned 32 last month* and have been really ...", the highlighted disclosure span can be abstracted to *"I recently entered my early 30s"*. This task operates at the span level, functioning similarly to a text editing tool such as Grammarly. In practice, we envision that abstraction will work with the detection model in a pipeline–first identifying self-disclosures, then users can select which disclosures to abstract. Abstracted spans must fit seamlessly into the original sentence without changing the rest of the text.

Comparison with Sentence-level Tasks. Table 3 illustrates the differences between sentential paraphrasing, sentence- and span-level abstraction. Sentential abstraction generalizes the entire sentence with or without disclosure spans provided. This approach generally modifies non-disclosure text as well as disclosures, affecting the original writing style, and potentially introducing unintended abstractions or hallucinations (Zhang et al., 2023), which may be undesirable. To assess the effectiveness of each task, we randomly sample 100 test sentences and apply each method by zero-shot prompting GPT-4. Two annotators are asked to rank and rate model outputs on a scale from 0 to 100, with

Methods	Automatic Evaluation (Matching Metric)				Human Evaluation			
	BLEU	ROUGE-2	ROUGE-L	Diversity	Rank Dist. (1→9)	Rank ▽	Raw	Z-Score
Sampling								
Special Token	17.40	18.81	38.42	2,576		5.37	79.22	-0.10
Instruction	16.95	18.78	38.35	2,564		5.80	77.17	-0.22
Instruction (w/ thought)	17.06	18.47	38.43	2,882		5.32	78.13	-0.13
End-to-end training								
Special Token	17.22	19.67	38.38	2,911		4.65	79.03	-0.07
Instruction	17.99	19.60	**39.57**	2,992		**3.77**	**81.53**	**0.21**
Instruction (w/ thought)	16.53	19.24	38.81	2,801		4.48	79.76	0.04
Iterative generation								
Special Token	**18.13**	19.74	38.71	2,913		4.12	80.72	0.11
Instruction	17.80	**19.81**	39.56	**3,067**		5.10	78.46	-0.03
Instruction (w/ thought)	16.89	18.84	38.31	2,914		4.18	81.27	0.17

Table 4: Test results on generating three alternative abstracted spans. End-to-end instruction tuning and iterative instruction training with thought achieve the top two performances under human evaluation. *Rank Dist.* presents the histograms of the rank distribution, where 1 is the best and 9 the worst. *Diversity* presents # unique bigrams.

Task	Rating	Rank 1	Rank 2	Rank 3	Rank 4
Sentence Paraphrasing	72.52	4%	14%	16%	66%
Sentence Abstraction:					
w/o providing spans	80.33	38%	**34%**	22%	6%
w/ providing spans	**86.16**	**54%**	24%	18%	**4%**
Span Abstraction	85.62	50%	30%	**14%**	6%

Table 5: Average human rating and rank distribution across four tasks, evaluating overall effectiveness, which considers both utility preservation and privacy increase.

an average 0.52 Kendall's τ for agreement. We aggregate the annotations by re-ranking the sums. Table 5 shows that when disclosures spans are provided, sentence-level abstraction achieves similar high effectiveness as span-level abstraction, with scores over 85. Prompts are listed in Appendix H.

4.2 Automatic Generation of Training Data

We use the most recent 10% of posts (plus associated comments) from our corpus (§2.1), which are further divided into training (159 posts), dev (25 posts), and test sets (50 posts). While creating diverse abstractions is challenging for human annotators, LLMs are adept at this task. We use chain-of-thought (Wei et al., 2022) prompting with few-shot demonstrations, which asks the model to first generate a rationale on why the disclosure span needs abstraction, and then generate three diverse abstractions, aiming to accommodate varied user preferences. For training and dev sets, we choose GPT-3.5 (06-13) to balance cost and performance. We use GPT-4 (06-13) for the test set, given its more advanced capabilities. We list the prompt in Appendix H.

4.3 Abstraction Models

We fine-tune Llama-2-7B (Touvron et al., 2023) with LoRA (Hu et al., 2021) for generating abstractions. We experiment with three different methods: sampling three times from an abstraction model (see Appendix D) that generates only one abstraction at a time (**sampling**), training an abstraction model that generates three abstractions all at once (input → A, B, C) (**end-to-end training**), or using an iterative approach that breaks the three abstractions (A, B, C) into three separate training instances: input → A, input+A → B, input+A+B → C (**iterative generation**). For each method, we consider the top three input-output setups identified in Appendix D, where we train models to generate only a single abstraction. These setups include formatting the input with special tokens and calculating loss on abstraction, formatting input with instruction and calculating loss either on abstraction or on rationale plus abstraction. Overall, we evaluate a total of 9 models.

4.4 Results

Automatic Evaluation. Table 4 presents the test set performance for each model. We adopt matching BLEU and ROUGE metrics proposed by (Dou et al., 2021) to encourage diversity in generated abstractions. These matching metrics use the Hungarian algorithm (Kuhn, 1955) to calculate the highest matching scores among one-to-one pairings be-

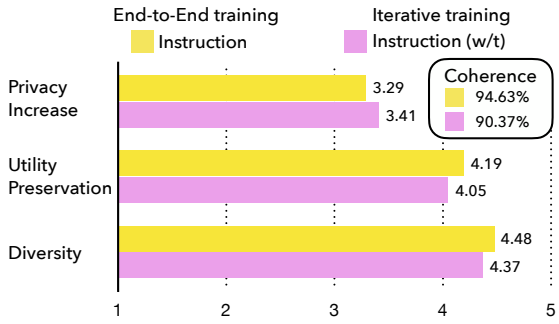

Figure 2: Human evaluation with Likert-scale (1-5) of the top two models. The best model shows moderate privacy increase, high utility preservation, and very high diversity in abstractions. w/t denotes *with thought*.

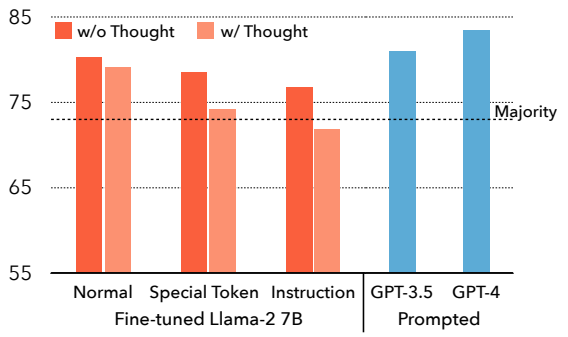

Figure 3: Test results of importance rating, measured in accuracy. Fine-tuning directly on output is better than on thought. The best fine-tuned model achieves comparable performance with GPT-3.5.

tween generations and references. According to the automatic metrics, all three methods show similarly high performance, while generating three abstractions in iterative steps yields slightly better results.

Human Evaluation. We further conduct a human evaluation on 60 sampled self-disclosure test instances. We first use Rank & Rate (Maddela et al., 2023) to rate the abstractions generated from each of the nine models in Table 4 on a scale of 0-100. Each instance receives three ratings from three in-house annotators. We also report the z-scores, which normalize the raw scores by the mean and standard deviation for each annotator to reduce individual bias. The rank is based on the average z-score. For inter-annotator agreement, we calculate an interval Krippendorff's alpha (Krippendorff, 2004) of 0.37 on z-score, and an ordinal Krippendorff's alpha of 0.41 on rank. These values indicate a fair level of agreement, given these methods performing closely, in line with Maddela et al. (2023).

We then evaluate the top two models on four aspects: privacy increase, utility preservation, and diversity, all rated on a 1-5 Likert scale, along with a binary assessment of coherence, evaluating whether each abstraction integrates seamlessly into the sentence. Detailed definitions for each aspect and Likert scale are provided in Appendix G.2. For inter-annotator agreement, we calculate an interval Krippendorff's alpha, which are 0.46 for Privacy Increase, 0.28 for Utility Preservation, and 0.60 for Diversity, as well as two agree% (% of instances where at least two annotators agreed), which are 61.7%, 80.0%, 96.7% for each category respectively, showing a good level of agreement.

Human evaluation results in Table 4 reveal that end-to-end training using the instruction and iterative generation with thought and instruction achieve the highest z-scores, with 0.21 and 0.17 respectively. The additional aspect-based human evaluation conducted on these two models is shown in Figure 2. Both models are capable of generating high-quality abstractions. The end-to-end training model performs slightly better in utility preservation, diversity, and coherence, while the iterative generation model is better at increasing privacy.

5 Importance Rating of Self-Disclosures

To highlight and abstract self-disclosure more selectively, as users suggested in §3, we consider an additional task that rates the importance of each disclosure within context.

Task Definition. Given the disclosure span and its surrounding context, the task is to estimate how important this disclosure is for others to understand the user's message and communication goals. We consider three levels: `low`, `moderate`, and `high`, corresponding to disclosures that can be removed, essential but can be abstracted, and have to be kept as it is (see App. G for details). For disclosures that appear in a post's title or body, we consider both the title and body of the post as context. For disclosures in comments (i.e., replies to the main post), the context extends to the entire comment and its parent comment in the reply chain, if existing.

Training Data. Compared to abstraction, humans perform more effectively than LLMs on importance rating. We use the same train/dev/test split as the abstraction experiment and have each instance annotated by three in-house annotators. 24% instances reach a consensus and 65% have agreement between two annotators. Only 11% exhibit complete disagreement. We also calculate Krippendorff's α as 0.29. This fair level of agreement is anticipated given the task's subjective nature.

For example, people have different opinions about whether details like age in a dating post should be retained (*"32"*) or abstracted (*"early 30s"*). There often isn't a clear cut between low and moderate, moderate and high; yet, they provide useful signals to users. Further discussions are in Appendix E. For training labels, we take the majority vote or `moderate` if the annotators choose all three levels.

Model. We fine-tune Llama-2-7B with various input-output formats as in the abstraction experiment (§4). We use GPT-3.5 to generate the reasoning that leads to the human-assigned ratings.

Evaluation Results. Given the task subjectivity, we measure accuracy by considering a prediction as correct if it matches any one of the three annotations. Figure 3 shows the accuracy of each method in comparison with GPT-3.5 and GPT-4. We find that fine-tuning on thought process degrades performance across all input format performance. The top performing fine-tuned model achieves 80.37%, on par with GPT-3.5's 80.98%.

6 Related Work

There is an excellent survey-position paper by Lison et al. (2021), which has provided a comprehensive review of literature in the NLP for Privacy area. It identified one key research challenge as: "*Most importantly, they (NLP approaches) are limited to predefined categories of entities and ignore how less conspicuous text elements may also play a role in re-identifying the individual. For instance, the family status or physical appearance of a person may lead to re-identification but will rarely be considered as categories to detect.*" — which motivated this very work of ours. We discuss some related works, including the newer ones, below.

Online Self-Disclosure Detection. Most existing research addressed self-disclosure detection in social media as sentence or document classification, which could not accurately pinpoint the specific disclosure spans. Many prior work focused on one specific kind of self-disclosure, such as medical and mental health conditions (De Choudhury et al., 2016; Yates et al., 2017; Benton et al., 2017; Klein et al., 2017; Zhao et al., 2019; Valizadeh et al., 2021a, 2023), sexual harassment (Schrading et al., 2015; Andalibi et al., 2016; Chowdhury et al., 2019), personal opinions or sentiments (Cho et al., 2022), and employment history (Preoţiuc-Pietro et al., 2015; Tonneau et al., 2022). Other research considered all types of personal information that a user can reveal as a single category (Mao et al., 2011; Caliskan Islam et al., 2014; Bak et al., 2014; Balani and De Choudhury, 2015; Wang et al., 2016; Yang et al., 2017; Blose et al., 2020; Reuel et al., 2022). A few works considered self-disclosures of multiple classes, but covered only a small number of categories (Lee et al., 2023; Akiti et al., 2020), or relied on dictionary/rule-based methods (Guarino et al., 2022), or limited to a particular context (e.g., bragging (Jin et al., 2022) and news comments (Umar et al., 2019)). To address these limitations, we emphasize detecting self-disclosure at the span level and broadening the coverage to include 19 distinct categories. This allows for a more fine-grained detection of privacy leaks for users.

PII Identification and Anonymization. Personal Identifiable Information (PII) is closely related to self-disclosures, but with a focus on highly sensitive attributes, such as full names, social security numbers, dates of birth, etc. (Regulation, 2016; Morris et al., 2022; Adams et al., 2019; Lukas et al., 2023; Hathurusinghe et al., 2021). Such sensitive data is more commonly encountered in legal (Pilán et al., 2022; Mansfield et al., 2022) and medical text (Yue and Zhou, 2020; Dernoncourt et al., 2017), as opposed to in social media or online community sources. Existing tools for PII identification such as Microsoft's Presidio [5] use regular expressions (Subramani et al., 2023; Mouhammad et al., 2023), and Named Entity Recognition (NER) detectors (Honnibal et al., 2020). However, such approaches indiscriminately mark entities (e.g., a business phone number on an advertisement) without considering whether the information is self-disclosed. PII anonymization (Azure, 2023; AWS, 2023), widely used in healthcare records management and machine learning training pipelines, replaces sensitive data with masked tokens (e.g., *[xxx]*) or weaker labels (e.g., *[Location]*). This aggressive approach hurts the utility of the message and is not suited for online self-disclosures, which are often voluntary and serve specific functions. To address this, we introduce a novel task of **self-disclosure abstraction** that strikes a balance of utility and privacy.

Privacy Leakage in Language Models. Recent work has shown however that LLMs are prone to leaking personal information (Sun et al., 2023; Kim et al., 2023; Huang et al., 2022; Lukas et al., 2023)

[5] `https://microsoft.github.io/presidio/`

and lack the ability to reason about privacy compared with humans (Mireshghallah et al., 2023). This phenomenon of user privacy violation is due to the issue of memorization (Carlini et al., 2022, 2021), where LMs recall individual sequences from their pre-training corpora. Recent efforts in solving this leakage problem include differentially-private training (Ponomareva et al., 2023; Li et al., 2021), decoding methods that prevent generation of memorized sequences (Ippolito et al., 2022), and prompt self-moderation (Chen et al., 2023). We refer interested readers to the recent surveys of Smith et al. (2023), Ishihara (2023), and Klymenko et al. (2022) for additional information on privacy leakage and memorization. Our work takes a different angle by providing a user-centered approach that tackles a root cause of privacy leakage; helping users make more informed decisions when posting online through self-disclosure identification and abstraction. This in turn can help decrease the chance of personal information ending up in pre-training corpora and reduce potential privacy violations.

7 Conclusion

We push the first steps in protecting user-side privacy in online self-disclosures. Our disclosure detection model trained on our new fine-grained corpus with span-level annotations achieves over 65% of partial span F_1 and is further validated through an HCI user study, highlighting its real-world applicability. Responding to the need from participants for balancing privacy risk reduction with message utility, we propose a novel task of self-disclosure abstraction, and explore various fine-tuning methods to generate three diverse abstractions. Our human evaluation shows that the best model can provide diverse abstractions that reduce privacy risks while highly preserving utility. We further fine-tune a model to rate the importance of the self-disclosure on understanding user's perspective and context. This model reaches 80% accuracy, matching the performance of GPT-3.5, thereby helping users to decide which disclosures to abstract. Overall, we believe our work paves the way for a new direction of using LLMs to protect user-side privacy.

Limitations

Our user study (§3.3) has revealed some additional limitations and research directions, including personalization, explainability, and contextual awareness. More specifically, 5 participants suggested that the tool should consider subreddit-specific norms. For example, the `r/diabetes` subreddit inherently expects that users may discuss their medical condition, rendering some model-predicted highlights redundant. Participants also expressed a desire for more transparency in the model's decision-making process, such as more explanations as to *why* certain highlights were marked as disclosures. Future work could expand our research to include other social platforms to provide broader insights and applicability in diverse social environments. In this work, we evaluate self-disclosure detection and abstraction individually for accurate assessment. As these tasks work consecutively as a pipeline in practice, future work could conduct user studies on the whole pipeline. Future research could also investigate the model's performance after quantization, which will allow deploying Llama-7B completely on personal devices for better privacy protection.

Ethics Statement

This research was approved by the Institutional Review Board (IRB) at Georgia Institute of Technology. We take the following measures to safeguard the personal information in our corpus before the annotation process. First, all personal identification information (PII), such as names and emails, is replaced with synthetic data. Second, we hired in-house student annotators ($18 per hour) instead of crowd workers for annotation. Every annotator was informed that their annotations were being used to create a dataset for online self-disclosure detection. All examples, except for those generated by the model, shown in the paper are synthetic but accurately reflect the real data. Our user study was approved by IRB at Carnegie Mellon University. The primary data collected from user interviews (including the Reddit posts run through the model) was self-disclosed and gathered in a survey with the participants' consent. In accordance with IRB policy, we anonymized all data collected during the study by removing any PII. The primary purpose of our models is to provide users with a tool to mitigate the privacy risks associated with online self-disclosures. In cases where the models fail, they do not pose additional risks but rather tend towards overprotection, either by identifying more spans of text or by overly abstracting the disclosed information. We identified no potential harms that

would disproportionately impact marginalized or otherwise vulnerable populations. To prevent misuse, we will not release our dataset and models to the public. Instead, we will share them upon request with researchers who agree to adhere to these ethical guidelines:

1. Use of the corpus is limited to research purposes only.
2. Redistribution of the corpus without the authors' permission is prohibited.
3. Compliance with Reddit's policy is mandatory.

Annotations of posts that have been deleted by users will be excluded at the time of the data request. To request access, please email the authors.

Acknowledgments

The authors would like to thank Srushti Nandu, Chase Perry, and Shuheng Liu for conducting pilot studies; Piranava Abeyakaran, Nour Allah El Senary, Vishnesh Jayanthi Ramanathan, Ian Ligon, Govind Ramesh, Ayush Panda, and Grace Kim for their help with data annotation and evaluation; Yang Chen, Duong Minh Le, and Nghia T. Le for their helpful feedback. We would also like to thank Azure's Accelerate Foundation Models Research Program graciously providing access to API-based models, such as GPT-4. This research is supported in part by NSF awards IIS-2144493, IIS-2052498 and IIS-2112633, ODNI and IARPA via the ODNI and IARPA via the HIATUS program (contract 2022-22072200004). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Allison Adams, Eric Aili, Daniel Aioanei, Rebecca Jonsson, Lina Mickelsson, Dagmar Mikmekova, Fred Roberts, Javier Fernandez Valencia, and Roger Wechsler. 2019. AnonyMate: A toolkit for anonymizing unstructured chat data. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 1–7.

Chandan Akiti, Anna Squicciarini, and Sarah Rajtmajer. 2020. A semantics-based approach to disclosure classification in user-generated online content. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3490–3499.

Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. 2016. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 3906–3918.

Dhananjay Ashok and Zachary C Lipton. 2023. Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*.

AWS. 2023. Amazon comprehend.

Azure. 2023. Azure AI language.

JinYeong Bak, Chin-Yew Lin, and Alice Oh. 2014. Self-disclosure topic model for classifying and analyzing twitter conversations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1986–1996.

Sairam Balani and Munmun De Choudhury. 2015. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1373–1378.

Adrian Benton, Margaret Mitchell, Dirk Hovy, et al. 2017. Multitask learning for mental health conditions with limited social media data. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017. Proceedings of Conference*. Association for Computational Linguistics.

Taylor Blose, Prasanna Umar, Anna Squicciarini, and Sarah Rajtmajer. 2020. Privacy in crisis: A study of self-disclosure during the coronavirus pandemic. *arXiv preprint arXiv:2004.09717*.

Aylin Caliskan Islam, Jonathan Walsh, and Rachel Greenstadt. 2014. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, pages 35–46.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Yang Chen, Ethan Mendes, Sauvik Das, Wei Xu, and Alan Ritter. 2023. Can language models be instructed to protect personal information? *arXiv preprint arXiv:2310.02224*.

Won Ik Cho, Soomin Kim, Eujeong Choi, and Younghoon Jeong. 2022. Assessing how users display self-disclosure and authenticity in conversation with human-like agents: A case study of luda lee. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 145–152.

Arijit Ghosh Chowdhury, Ramit Sawhney, Puneet Mathur, Debanjan Mahata, and Rajiv Ratn Shah. 2019. Speak up, fight back! detection of social media disclosures of sexual harassment. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Student research workshop*, pages 136–146.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Paul C. Cozby. 1973. Self-disclosure: a literature review. *Psychological bulletin*, 79 2:73–91.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Yao Dou, Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2021. Multitalk: A highly-branching dialog testbed for diverse conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12760–12767.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Alfonso Guarino, Delfina Malandrino, and Rocco Zaccagnino. 2022. An automatic mechanism to provide privacy awareness and control over unwittingly dissemination of online private information. *Computer Networks*, 202:108614.

Rajitha Hathurusinghe, Isar Nejadgholi, and Miodrag Bolic. 2021. A privacy-preserving approach to extraction of personal information through automatic annotation and federated learning. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pages 36–45.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

David Heineman, Yao Dou, Mounica Maddela, and Wei Xu. 2023. Dancing between success and failure: Edit-level simplification evaluation using SALSA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. 2020. spaCy: Industrial-strength natural language processing in python.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047.

Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A Choquette-Choo, and Nicholas Carlini. 2022. Preventing verbatim memorization in language models gives a false sense of privacy. *arXiv preprint arXiv:2210.17546*.

Shotaro Ishihara. 2023. Training data extraction from pre-trained language models: A survey. *arXiv preprint arXiv:2305.16157*.

Mali Jin, Daniel Preoţiuc-Pietro, A Seza Doğruöz, and Nikolaos Aletras. 2022. Automatic identification and classification of bragging in social media. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3945–3959.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Sidney M Jourard. 1971. Self-disclosure: An experimental analysis of the transparent self.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. ProPILE: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

Ari Klein, Abeed Sarker, Masoud Rouhizadeh, Karen O'Connor, and Graciela Gonzalez. 2017. Detecting personal medication intake in twitter: an annotated corpus and baseline classification system. In *BioNLP 2017*, pages 136–142.

Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. 2022. Differential privacy in natural language processing the story so far. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 1–11.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage Publications.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Jooyoung Lee, Sarah Rajtmajer, Eesha Srivatsavaya, and Shomir Wilson. 2023. Online self-disclosure, social support, and user engagement during the covid-19 pandemic. *ACM Transactions on Social Computing*.

Ada Lerner, Helen Yuxun He, Anna Kawakami, Silvia Catherine Zeamer, and Roberto Hoyle. 2020. Privacy and activism in the transgender community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2021. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. *arXiv preprint arXiv:2302.00539*.

Mufan Luo and Jeffrey T Hancock. 2020. Self-disclosure and social media: motivations, mechanisms and psychological well-being. *Current opinion in psychology*, 31:110–115.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Courtney Mansfield, Amandalynne Paullada, and Kristen Howell. 2022. Behind the mask: Demographic bias in name detection for pii masking. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 76–89.

Huina Mao, Xin Shuai, and Apu Kapadia. 2011. Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, pages 1–12.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*.

John Morris, Justin Chiu, Ramin Zabih, and Alexander M Rush. 2022. Unsupervised text deidentification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4777–4788.

Nina Mouhammad, Johannes Daxenberger, Benjamin Schiller, and Ivan Habernal. 2023. Crowdsourcing on sensitive data with privacy-preserving text rewriting. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 73–84, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for

text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. 2023. How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1754–1764.

Protection Regulation. 2016. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679:2016.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ann-Katrin Reuel, Sebastian Peralta, João Sedoc, Garrick Sherman, and Lyle Ungar. 2022. Measuring the language of self-disclosure across corpora. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1035–1047.

Nicolas Schrading, Cecilia Ovesdotter Alm, Raymond Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on reddit. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2577–2583.

Manya Sleeper. 2016. *Everyday online sharing*. Ph.D. thesis, Carnegie Mellon University.

Victoria Smith, Ali Shahin Shamsabadi, Carolyn Ashurst, and Adrian Weller. 2023. Identifying and mitigating privacy risks stemming from language models: A survey. *arXiv preprint arXiv:2310.01424*.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

Nishant Subramani, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. 2023. Detecting personal information in training corpora: an analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 208–220.

Albert Yu Sun, Eliott Zemour, Arushi Saxena, Udith Vaidyanathan, Eric Lin, Christian Lau, and Vaikkunth Mugunthan. 2023. Does fine-tuning gpt-3 with the openai api leak personally-identifiable information? *arXiv preprint arXiv:2307.16382*.

Manuel Tonneau, Dhaval Adjodah, Joao Palotti, Nir Grinberg, and Samuel Fraiberger. 2022. Multilingual detection of personal employment status on twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6564–6587.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Prasanna Umar, Anna Squicciarini, and Sarah Rajtmajer. 2019. Detection and analysis of self-disclosure in online news commentaries. In *The World Wide Web Conference*, pages 3272–3278.

Mina Valizadeh, Xing Qian, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2023. What clued the ai doctor in? on the influence of data source and quality for transformer-based medical self-disclosure detection. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1193–1208.

Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2021a. Identifying medical self-disclosure in online communities. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4398–4408.

Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2021b. Identifying medical self-disclosure in online communities. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4398–4408, Online. Association for Computational Linguistics.

Yi-Chia Wang, Moira Burke, and Robert Kraut. 2016. Modeling self-disclosure in social networking sites. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pages 74–85.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,

et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Rachel Wicks and Matt Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3995–4007, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Diyi Yang, Zheng Yao, and Robert Kraut. 2017. Self-disclosure and channel difference in online health support groups. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 704–707.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Conference on Empirical Methods in Natural Language Processing*, pages 2958–2968. ACL.

Xiang Yue and Shuang Zhou. 2020. PHICON: Improving generalization of clinical text de-identification models via data augmentation. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 209–214.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Xinyan Zhao, Deahan Yu, and VG Vinod Vydiswaran. 2019. Identifying adverse drug events mentions in tweets using attentive, collocated, and aggregated medical representation. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 62–70.

A Self-Disclosure Corpus

A.1 Quality Control

Recent works have shown that crowdsourcing often leads to lower quality annotations (Clark et al., 2021; Gilardi et al., 2023). To ensure the quality, we hire seven in-house annotators, who are undergraduate students at a US university and native English speakers, and compensate them at a rate of $18 per hour. Each annotator first undergoes a training that includes tutorials and 20 exercise examples. During the annotation, we release the data in ten batches, allowing us to constantly monitor performance and provide feedback as needed. We use the BRAT interface for the annotation process.[6]

Category	Krippendorff's α	Two Agree (%)
AGE	0.80	73.1
AGE&GENDER	0.88	81.6
RACE/NATIONALITY	0.88	82.2
GENDER	0.79	74.0
LOCATION	0.71	65.4
APPEARANCE	0.68	57.1
WIFE/GF	0.70	62.1
FINANCE	0.65	57.6
OCCUPATION	0.66	58.4
FAMILY	0.70	61.5
HEALTH	0.53	46.7
MENTAL HEALTH	0.45	37.0
HUSBAND/BF	0.78	70.8
EDUCATION	0.72	64.9
PET	0.48	37.7
RELATIONSHIP STATUS	0.48	37.3
SEXUAL ORIENTATION	0.58	49.9
OVERALL	0.54	45.7

Table 6: Inter-annotator agreement for each self-disclosure category.

A.2 Inter-annotator Agreement

Table 6 shows the inter-annotator agreement per self-disclosure category, measured by Krippendorff's α (Krippendorff, 2018) and *Two Agree*. Due to computational constraints, we calculate Krippendorff's α per post and report the average across the dataset. Two agree is the percentage of words labeled as *disclosure* by both annotators $a1$ and $a2$: $\frac{|\text{Words}_{a1} \cap \text{Words}_{a2}|}{|\text{Words}_{a1} \cup \text{Words}_{a2}|}$. The numbers are higher than or similar to those in other span-level annotation work (Dou et al., 2022; Heineman et al., 2023).

B User Study Recruitment

We first launched a pre-study survey on Prolific, targeting participants who met the following criteria: had an active Reddit account, made at least

[6] https://brat.nlplab.org/

Input	Span F_1		Partial F_1		Token F_1	
	Multi	*Binary*	*Multi*	*Binary*	*Multi*	*Binary*
Normal	42.41	43.93	58.99	60.50	66.73	72.23
256	43.17	45.01	60.49	60.64	67.24	71.35
128	42.86	45.25	60.39	60.56	69.22	71.78
64	43.22	43.91	59.51	62.27	68.18	73.95
Sentence	**48.88**	**52.92**	**65.71**	**72.29**	**74.17**	**83.26**

Table 7: Test performance of DeBERTaV3-large fine-tuned on various data setups, with training on sentence level achieving the best results. For *Multi*-class models, the results are averaged over all classes, excluding label "O", while *Binary* models report F_1 for "disclosure".

three posts on Reddit, resided in the U.S., and were 18 years of age or older. The survey asked them to provide posts they had authored and had privacy concerns over. A total of 158 individuals participated in the survey. We kept the participants whose posts are majority text-based. We then invited 21 Reddit users to participate in the interview based on their availability. We make sure of the diverse demographics among the 21 participants where 12 identified as female, 16 were below the age of 50, and 15 held a bachelor's degree or higher.

C Further Detection Results

In this section, we describe the details of our self-disclosure detection model and show the performance of the binary-class model.

C.1 Experiment Details

We fine-tune RoBERTa-large (Liu et al., 2019), a transformer-based encoder with 355M parameters, and DeBERTaV3-large (He et al., 2021) with 435M parameters on our dataset by minimizing the cross-entropy loss for each token's label. As some words are tokenized into multiple subword tokens, during inference, we use the hidden states of the first token to get the label (Rei et al., 2022).

We experiment with various data processing methods during fine-tuning, given that Reddit posts and comments can significantly vary in length and disclosure spans do not always require extended contexts. Specifically, we segment comments or posts into shorter chunks of 64, 128, and 256 words, as well as individual sentences using Ersatz (Wicks and Post, 2021).

For self-disclosure classes which are infrequent in Reddit data (specifically `name` and `contact` disclosures) and thus insufficient to train models from scratch, we turn to existing models and tools. We use the state-of-the-art NER model, LUKE (Yamada et al., 2020) to identify person names and Microsoft Presidio[7] to recognize contact information such as phone numbers and social media usernames. To specifically identify self-disclosures as opposed to generic names (e.g., Taylor Swift), we further train a sentence classifier using RoBERTa-large, which achieves 84.4 test F_1, to first determine whether a sentence contains a self-disclosure.

C.2 Results

Table 7 presents the average test-set performance for both binary and multi-class models when fine-tuned under different data configurations. Due to the inherent simplicity, binary models typically outperform their multi-class counterparts. In addition, we find that dividing the long Reddit posts and comments into shorter pieces generally improves performance. The most significant gain is achieved by segmenting the data at the sentence level, leading to an increase of over 6 span-level F1 points in both binary and multi-class settings, compared to the normal baseline.

Table 8, a detailed version of Table 2 from Section 2.2, shows span, partial span, and token-level F_1 of fine-tuned RoBERTa-large, DeBERTaV3-large, and prompted GPT-4 Turbo. For GPT-4 Turbo, we iteratively refined the prompt, listed in Appendix H, incorporating definitions, guidelines, and chain-of-thought. We find that generating thought before outputing results leads to an increase of 2.43 partial span F_1. We also discover that when breaking down categories into multiple groups performs worse than detecting all categories at once, as GPT-4 Turbo tend to predict unspecified categories or over-classifiy like "40 yo man" as `appearance`. For our user study, we use RoBERTa-large model.

C.3 ShareGPT

We describe the process of collecting data and annotating self-disclosures within ShareGPT conversations [8]. This process is conducted in four steps: **1)** we filter out conversations turns that are AI-generated, have over 500 tokens, or don't contain "I" or "my", resulting in 59,533 human-authored turns, **2)** we randomly select 1,600 conversation turns from them, **3)** five in-house annotators then identify 105 turns containing self-disclosures, **4)**

[7] https://microsoft.github.io/presidio/

[8] https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered

Class (#Spans)	RoBERTa-large			DeBERTaV3-large			GPT-4		
	Span F_1	Partial F_1	Token F_1	Span F_1	Partial F_1	Token F_1	Span F_1	Partial F_1	Token F_1
AGE (35)	60.87	72.46	87.04	64.62	70.77	84.97	*65.71*	**80.0**	80.54
AGE&GENDER (17)	*73.68*	**84.21**	81.63	54.05	70.27	75.47	69.77	74.42	70.97
RACE/NATIONALITY (8)	60.0	60.0	64.94	*82.35*	**82.35**	81.25	70.59	70.59	71.64
GENDER (17)	61.11	61.11	56.47	*72.73*	**72.73**	61.73	57.14	57.14	62.50
LOCATION (41)	52.87	71.26	76.60	*57.78*	**73.33**	83.61	41.30	54.35	69.63
APPEARANCE (31)	*44.07*	64.41	78.14	35.48	**67.74**	76.27	12.77	42.55	54.71
WIFE/GF (30)	53.33	66.67	74.19	*65.52*	**75.86**	78.36	48.39	64.52	74.32
FINANCE (33)	35.82	68.66	76.57	*40.0*	**71.43**	77.68	31.17	54.55	69.49
OCCUPATION (44)	40.0	64.44	72.39	*45.65*	**65.22**	75.08	39.56	52.75	68.21
FAMILY (44)	*50.0*	**58.70**	66.52	35.29	49.02	69.23	46.60	58.25	64.68
HEALTH (40)	44.21	56.84	76.54	*45.10*	**58.82**	75.94	26.45	38.02	50.17
MENTAL HEALTH (46)	39.22	**64.71**	81.18	*40.0*	63.16	75.75	36.36	52.73	66.51
HUSBAND/BF (14)	*68.75*	**75.0**	71.91	64.71	70.59	71.11	55.17	68.97	68.49
EDUCATION (21)	55.32	68.09	85.23	*65.38*	**69.23**	86.73	42.55	51.06	67.37
PET (15)	46.15	46.15	56.88	*48.28*	**55.17**	61.54	*48.28*	48.28	56.25
RELATIONSHIP STATUS (31)	*38.36*	41.10	58.80	36.92	**43.08**	60.11	34.29	42.86	57.58
SEXUAL ORIENTATION (12)	28.57	**76.19**	68.0	25.0	58.33	66.0	*52.17*	69.57	71.29
AVERAGE	48.32	64.71	72.53	*48.88*	**65.71**	74.17	41.93	57.68	66.14

Table 8: Test performance per class for fine-tuned models and prompted `GPT-4-0125-preview`. *Italic* highlights the best model for each class on span F_1. **Bold** on partial span F_1. Underline on token F_1.

Class (#spans)	Span F_1	Partial F_1	Token F_1
OCCUPATION (75)	38.89 (-6.8)	61.11 (-4.1)	64.14 (-10.9)
LOCATION (31)	56.0 (-1.8)	66.67 (-6.7)	77.35 (-6.3)
EDUCATION (10)	57.14 (-8.2)	66.67 (-2.6)	91.80 (+5.1)
RELATION. STATUS (6)	42.11 (+5.2)	42.11 (-1.0)	63.16 (+3.1)
FAMILY (6)	66.67 (+31.4)	66.67 (+17.7)	74.47 (+5.2)
AVERAGE	52.16 (+4.0)	60.64 (+0.7)	74.18 (-0.8)

Table 9: Per-class performance of the fine-tuned DeBERTa in detecting self-disclosure within ShareGPT conversations. Differences compared to in-domain performance are shown in parentheses ().

these turns undergo a two-round annotation–initial annotation followed by adjudication. In the end, there are 105 human-written turns with annotated self-disclosure spans.

Table 9 presents the performance of DeBERTa, fine-tuned on Reddit data, for each class in detecting self-disclosure within ShareGPT conversations. The model demonstrates good generalizability, performing comparably to its in-domain results.

D Further Abstraction Results

D.1 Generate a Single Abstraction

Besides generating three diverse abstractions given a self-disclosure span, we also fine-tune Lllama-2-7B (Touvron et al., 2023) with LoRA (Hu et al., 2021) for generating a single abstraction. We consider three different **input formats**, that use standard input, special token, and natural language instructions:

BLEU / ROUGE-2		Output	
		w/o Thought	w/ Thought
Input	Standard	15.3 / 25.1	14.4 / 22.9
	Special Token	17.0 / 25.4	12.9 / 21.7
	Instruction	17.9 / 24.8	18.3 / 25.5

Table 10: Test results on self-disclosure abstraction task. Training with special token and instruction with thought lead to the best performance.

STANDARD/NORMAL INPUT:
```
Sentence:{s}\nDisclosure Span:{d}\nAbstraction Span:
```

SPECIAL TOKEN:
```
<SENTENCE>{s}<SPAN>{d}<ABSTRACTION>
```

INSTRUCTION:
```
Your task is to abstract the given disclosure...
Sentence:{s}\nDisclosure Span:{d}
```

For **output formats**, which is the text that the model is trained to generate and where the loss is calculated, we explore two options. One is solely the desired output, which is the abstraction, and another is a rationale plus the abstraction, also known as chain-of-thought (Wei et al., 2022) training.

Table 10 reports BLEU (Papineni et al., 2002) and ROUGE-2 (Lin, 2004) for comparing each input and output configurations. We find that using special tokens and instruction with thoughts helps improve performance over the standard method.

Example 1 **Sentence:** [22M] [21F] **My girlfriend** cheated on me with a coworker, it's a little messy. **Abstractions:** "My partner", "The individual involved", "The person I was in a relationship with"
Example 2 **Sentence:** Did what happened at work make **my PTSD** from the military worse, or is it a new PTSD? **Abstractions:** "my previous trauma", "my mental health condition", "my past emotional challenges"
Example 3 **Sentence:** my friend then asked how I could even consider liking eminem cuz of his lyrics, and that **I am gay**. **Abstractions:** "I identify as LGBTQIA+.", "I am part of the rainbow community.", "I belong to the LGBTQ+ community."
Example 4 **Sentence:** ...then stop and **I took amitriptyline 10mg once and stop,** then some other sleeping pills... **Abstractions:** "I tried a prescribed medication for a short period of time,", "I experimented with a medication for a limited time,", "I explored the use of a prescribed medication for a brief period,"
Example 5 **Sentence:** Now I stole and lied to **my dad and sister** the only two people who go the extra way for me. **Abstractions:** "my closest family members", "the people who care about me the most", "the individuals who are always there for me"
Example 6 **Sentence:** I've been on antidepressants for 6 months and **I want to kill myself everyday.** **Abstractions:** "I'm feeling a profound sense of despair daily.", "I'm experiencing intense feelings of hopelessness every day.", "I'm struggling with extreme sadness and discomfort each day."
Example 7 **Sentence: I am an International Student** and i am new to this place. **Abstractions:** "I am a student from another country", "I am a foreign student adjusting to my new surroundings", "I am a student coming from another part of the world"
Example 8 **Sentence: I was open about my abusive childhood** and still received a rating for PTSD. **Abstractions:** "I shared my difficult childhood experiences", "I spoke openly about my traumatic childhood", "I talked publicly about my challenging upbringing"

Table 11: 8 randomly sampled examples with abstractions generated by the fine-tuned Llama-2-7B. The self-disclosure span to abstract is marked in **bold**.

D.2 Examples

Table 11 displays 8 randomly sampled examples with three abstractions generated by the best fine-tuned Llama-2-7B from Section 4.

E Further Importance Rating Discussions

To further illustrate the subjectivity, here is an example where each of three annotators assigned different labels:

Post: "At what age in your life did you want to settle down?"

Comment: "When I was 23 after *I finished my master degree*, I married my best friend."

Disclosure in the comment: "I finished my master degree,".

In this example, the annotators have different opinions on whether to keep, abstract or remove the self-disclosure. Each choice reflects a valid perspective: keeping it provides a clear milestone that may resonate with readers; abstracting to "a life milestone" omits specific education details; deleting it as the post only asks for age.

F Implementation Details

F.1 GPT-3.5 and GPT-4

We use gpt-3.5-turbo-0613 as GPT-3.5 and gpt-4-0613 as GPT-4.

F.2 Evaluation Metrics

For BLEU (Papineni et al., 2002), we use SacreBLEU (Post, 2018). For ROUGE (Lin, 2004), we use the one from torchmetrics.[9]

F.3 Experiments

We implemented our models using Huggingface 4.33.2 (Wolf et al., 2019) and PyTorch 2.0.1 (Paszke et al., 2019). All results are from single runs.

[9] https://torchmetrics.readthedocs.io/en/stable/text/rouge_score.html

Self-disclosure Detection. We train RoBERTa-large (Liu et al., 2019) and DeBERTaV3-large (He et al., 2021) on 2 A40 GPUs. We first train on data with single annotations for 10 epochs with batch size of 32, which takes around 16 minutes per run. We evaluate on the dev set every 50 steps, saving the checkpoint with the highest partial span-level F_1. We then fine-tune on a batch of double annotated data for another 10 epochs, which takes about 5 minutes per run. We evaluate every 20 steps on the dev set, and save the checkpoint with the highest partial span-level F1 for final evaluation. We perform a learning rate sweep over 1e-5, 2e-5, 3e-5, 5e-5, 8e-5 on our evaluation set, with 5e-5 being the best for RoBERTa and 3e-5 for DeBERTa. We use AdamW (Loshchilov and Hutter, 2017) as the optimizer with weight decay of 0.01. Additionally, we use a cosine learning rate schedule with a warmup ratio of 0.06.

Self-disclosure Abstraction. We train Llama-2 7B (Touvron et al., 2023) with LoRA (Hu et al., 2021) on 8 A40 GPUs for 5 epochs, with a total batch size of 32, which takes at most 2 hours per run. We use a learning rate of 1e-4, which is the standard when fine-tuning LLMs with LoRA. We use AdamW (Loshchilov and Hutter, 2017) as the optimizer with a weight decay of 0.01. Additionally, we use a cosine learning rate schedule with a warmup ratio of 0.03. For LoRA hyperparameters, we set rank=8, alpha=16, dropout=0.05, target modules=Q and V attention matrices. Additionally, we also update the embedding layer during fine-tuning.

Importance Rating. We use the same setup and hyperparameters as the abstraction experiment. It takes around 25 minutes per training run.

G Annotation Guidelines

G.1 Self-disclosure Annotation

Annotators were instructed to annotate explicit self-disclosures that concern the user based on our defined list of categories (Table 1). Most of the categories under "attributes" are straightforward to annotate such as the user's age, location, gender, etc. We considered instances where Reddit users revealing both their age and gender in one word, such as *"M24"*, under a specific category AGE/-GENDER that is different from AGE and GENDER which are for disclosures of age and gender individually. For tricky categories, most of which are under "experiences", we provided exact definitions to follow which help the annotators make decisions and ensure consistent labeling. Those definitions are as follows:

- *Appearance* self-disclosures are defined as descriptions of bodily features of the user, such as their height, weight, eye or hair color, or any other specific features.
- *Health* self-disclosures are defined as the disclosure of a specific disease or illness the user has, specific medications they take, or medical tests they perform.
- *Mental Health* self-disclosures are defined as situations where users discuss their feelings, state of mind, or suicidal thoughts.
- *Finance* self-disclosures are defined as mentions of specific personal financial details such as details about one's salary, recent transactions, affordability of items, choice of bank, and similar specifics.
- *Education* self-disclosures are defined as mentions what the user is currently or planning on studying, or degrees they hold.
- *Occupation* self-disclosures are defined as mentions of the current or past occupations of the user.
- *Family* self-disclosures are defined as any disclosure that fits within our specified attributes and experiences but concern a family member of the user, such as their parents, siblings, or extended family members.

G.2 Human Evaluation for Abstraction

The following are the 1-5 Likert scales for the aspects: privacy increase, utility preservation, and diversity, used in human evaluation for three-span abstraction (§4).

Privacy Increase:

1 – No Privacy Increase: The abstractions are the same or paraphrases to the disclosure span.

2 – Low Privacy Increase: The abstractions slightly obscure sensitive details but are still quite similar to the original.

3 – Moderate Privacy Increase: The abstractions moderately obscure sensitive details.

4 – High Privacy Increase: The abstractions significantly obscure sensitive details and remove details.

5 – Maximum Privacy Increase: The abstractions eliminate nearly all specific details.

Utility preservation:

1 – No Utility Preserved: The abstractions remove or significantly change the disclosure span, losing all the utility.

2 – Low Utility Preserved: The abstractions preserve a small amount of the disclosure span, but major aspects are lost or altered.

3 – Moderate Utility Preserved: The abstractions maintain a part of the disclosure span's utility.

4 – High Utility Preserved: The abstractions maintain most of the disclosure span's utility, with only minor aspects lost.

5 – Full Utility Preserved: The abstractions maintain the complete utility of the disclosure span, effectively conveying the intended function.

Diversity:

1 - Identical Abstractions: All three abstractions are essentially the same, exhibiting no diversity in wording or style.

2 - Minimal Diversity: Two of the three abstractions are identical, with only one offering a different expression.

3 - Low Diversity: All three abstractions are different, yet they exhibit similar styles and only minor variations in wording.

4 - Moderate Diversity: Each abstraction differs significantly in wording, with about half of the words unique to each. The styles are somewhat varied but maintain a degree of similarity.

5 - High Diversity: Each abstraction is distinctly unique, both in wording and in expression style, demonstrating a broad range of diversity.

G.3 Importance Rating Annotation

Low – Can be removed without compromising the understanding of the user's perspective and context.
Moderate – Adds a meaningful layer to the context but is not essential to the user's overall message. Can be abstracted.
High – Essential for an accurate understanding of the user's perspective and context, should be kept as is.

H Prompt Templates

We first list the prompts used to prompt GPT-3.5 and GPT-4, and then the prompts for fine-tuning the Llama-2-7B model. For prompts used in the importance rating experiment (§5), we provide ones for disclosures in comments; for disclosures in other locations (title, post, subcomments), only the instance part of the prompt is modified. Please note that within these prompts, we use "generalization", "generalize", and "rationale" as equivalents to "abstraction", "abstract", and "thought" respectively.

H.1 GPT-3.5 and GPT-4 Prompts

Self-disclosure Detection. We use the following prompt to ask GPT-4 to detect self-disclosures.

```
## Task
Analyze the provided sentence to
    identify segments containing self-
    disclosure. Self-disclosure refers
    to personal information about the
    author or their close relations.

## Categories
There are 17 specific categories to
    consider:
* Age: "I am a 23-year-old"
* Gender: "I'm just a girl"
* Age_Gender: only when age and gender
    are combined in a *single* word,
    such as "20F" or "32M".
* Sexual_Orientation: "I'm a straight
    man"
* Race_Nationality: mentions of the user
    's nationality/race/ethnicity.
* Wife_GF: disclosures indicating the
    author has a wife, girlfriend, or
    fiancee, such as "My gf."
* Husband_BF: disclosures indicating the
     author has a husband, boyfriend, or
     fiance, such as "My bf".
* Relationship_Status: only includes
    mentions of marital status, being in
     a romantic relationship, or being
    single. For example, "my partner".
* Family: mentions of specific family
    members, as well as disclosures that
     related to themselves, such as "My
    child is 3 year old".
* Health: includes a wide range of
    health-related information, from
    discussing specific diseases or
    conditions to mentioning medications
    , medical tests, or treatments.
* Mental_Health: includes a broad range
    of emotional states and feelings,
    not necessarily limited to specific
    mental health diagnoses. Examples
    include feelings of sadness, anxiety
    , stress, or comments about
    struggling with emotional well-being
     in general.
* Location: captures specific
    geographical details disclosed by
    the author. It includes precise
    locations such as addresses, cities,
     countries, or distinctive landmarks
    , so vague spans like "this place"
    should not be categorized.
```

```
* Appearance: "I am 6'2".
* Pet: "I have two musk turtles"
* Occupation: "I'm a motorcycle tourer (
    by profession)", student should be
    categorized as Education.
* Education: "I got accepted to UCLA"
* Finance: any financial situations, not
     necessarily exact amounts.

## Sentence to Analyze
"{sentence}"

Please note:
1. Focus only on disclosures related to
    the author or their close relations.
     Avoid marking disclosures about
    others.
2. For disclosures about close relations
    , like family members, partners,
    pets, you should tag them with the
    subject category (e.g., "Family").
    For instance, "My sister is pregnant
    " should be tagged as "Family"
    rather than splitting into "Family"
    and "Health".
3. Always include subjects or pronouns
    that are adjacent to the disclosure
    in the span.
4. Never include temporal details in the
     span.
5. Avoid tagging hypothetical or non-
    factual information.
6. Do not include non-informational word
     like "also" in the span.
7. Health and Mental_Health disclosures
    can be general, do not have to be
    explicit.

Read the sentence carefully and identify
     all the segments that fit these
    categories. Write down a brief
    thought process first, with the 7
    notes in mind. Then finalize the
    self-disclosure spans in a JSON
    format: the key must be the self-
    disclosure span and the value must
    be its corresponding category.
```

Abstraction Comparison in Section 4. The following zero-shot prompts are used to instruct GPT-4 to compare sentence paraphrasing, sentence, and span-level abstraction.

* Sentence Paraphrasing:

```
Your task is to generate a paraphrase
    for the given sentence.

Sentence: "{sentence}"
Paraphrased Sentence:
```

* Sentence-level Abstraction (without disclosure spans provided):

```
Your task is to generate a paraphrase
    that conveys similar meaning and
    maintains the same utlity while
    removing any overly specific or
    sensitive information for the given
    sentence.

Sentence: "{sentence}"
Paraphrased Sentence (Reducing Private
    Information):
```

* Sentence-level Abstraction (with disclosure spans provided):

```
Your task is to generate a paraphrase
    that conveys similar meaning and
    maintains the same utlity while
    removing overly specific or
    sensitive information for the given
    sentence. Besides the sentence, I
    will give you a list of disclosures
    that need to be addressed.

Sentence: "{sentence}"
Disclosures: {disclosures}
Paraphrased Sentence (Reducing Private
    Information):
```

* Span-level Abstraction:

```
Your task is to generalize the given '
    disclosure span' in the sentence.
    Provide a generalized span that
    conveys similar meaning but remove
    any overly specific or sensitive
    information.

Remember the following criteria:
* Only the disclosure span should be
    generalized; the rest of the
    sentence should remain intact.
* Generalized span should retain the
    essence and utility of the original
    span.
* Make sure the generalized span fits
    seamlessly into the original
    sentence, maintaining proper syntax
    and grammar.

Sentence: "{sentence}"
Disclosure Span to Revise: "{span}"
Generalized Span:
```

Self-disclosure Abstraction. We use the following prompt to ask GPT-3.5 (train/dev sets) or GPT-4 (test) to generate three abstraction spans with thought process for each data instance.

```
Your task is to generalize the given '
    disclosure span' in the sentence.
    Provide three diverse generalized
    spans that convey similar meaning
    but remove any overly specific or
    sensitive information.

Example 1:
Sentence: "Should I submit a 1470 SAT
    score to Carnegie Mellon and Duke?"
Disclosure Span to Revise: "1470 SAT
    score"
Rationale: The original disclosure span
    explicitly states a specific SAT
    score, which may be sensitive
    information for some individuals. To
     retain the context that the score
    is competitively high while removing
```

```
     the exact number, we can use
    generalized terms.
Generalized Spans: {{"span 1": "a high
    1400-range SAT score", "span 2": "an
     SAT score in the upper 1400s", "
    span 3": "an SAT score above 1450"}}

Example 2:
Sentence: "My Top Spotify Artists of
    2022 were Charli XCX, Taylor Swift,
    Carly Rae Jepsen, Snow Strippers,
    and Danny L Harle."
Disclosure Span to Revise: "Top Spotify
    Artists of 2022 were Charli XCX,
    Taylor Swift, Carly Rae Jepsen, Snow
     Strippers, and Danny L Harle."
Rationale: The original span lists very
    specific artists that the person
    listened to in 2022, which can be
    considered personal and revealing.
    To maintain the general idea that
    the person has favorite artists from
     that year without giving away the
    exact names, generalized terms can
    be used.
Generalized Spans: {{"span 1": "favorite
     artists of 2022 included a mix of
    pop and electronic musicians", "span
     2": "top artists from 2022 covered
    a range of styles", "span 3": "
    favorite artists last year were from
     various genres"}}

Example 3:
Sentence: "I just turned 32 last month
    and have been really reflecting on
    my relationships with people from my
     relationship with my siblings to
    coworkers."
Disclosure Span to Revise: "just turned
    32 last month"
Rationale: The original span provides
    very specific details about the
    person's age and the timing of their
     birthday, which could be seen as
    personal information. To keep the
    essence of the span-that the
    individual is in their early 30s and
     recently had a birthday-while
    making it less specific, generalized
     terms can be used.
Generalized Spans: {{"span 1": "recently
     entered my early 30s", "span 2": "
    turned into my early thirties not
    long ago", "span 3": "just started
    my third decade a little while back
    "}}

First, provide a rationale explaining
    why the disclosure span needs
    generalization. Then, offer three
    diverse generalized alternatives in
    a JSON format like this: {{"span 1":
     "xxx", "span 2": "xxx", "span 3": "
    xxx"}}.

Criteria:
* Only the disclosure span should be
    generalized; the rest of the
    sentence should remain intact.
* Generalized spans should be diverse
    but should all retain the essence of
     the original span.
* Make sure the generalized span fits
    seamlessly into the original
    sentence, maintaining proper syntax
    and grammar.

Sentence: "{sentence}"
Disclosure Span to Revise: "{span}"
Rationale:
```

Importance Rating.

```
Rate the importance of the disclosure
    span in a Reddit comment within the
    context into three Likert-scale:
* Low Importance: Can be removed without
     compromising the understanding of
    the commenter's perspective and
    context.
* Moderate Importance: Adds a meaningful
     layer to the context but is not
    essential to the commenter's overall
     message. Can be generalized.
* High Importance: Essential for an
    accurate understanding of the
    commenter's perspective and context,
     should be kept as is.

Given instance:
* Title: {title}
* Post: {post}
* Comment: {comment}
* Disclosure Span in the Comment: {
    disclosure}
{post_empty_explaination}
Note: The disclosure span is marked
    between special tokens <disclosure
    ></disclosure> in the Comment.

Read the Title, Post, and Comment
    carefully to understand the context.
     Write down your thought process.
    And in the end, provide your
    importance rating of the disclosure
    span in a JSON format: {{"Importance
    ": "Low/Moderate/High"}}.
```

The following prompt is used to prompt GPT-3.5 to generate thought process for the human-annotated importance rating:

```
Provide the thought process of the
    importance rating of a disclosure
    span in a Reddit comment.

Here are the definitions of the three
    importance rating scales:
* Low Importance: Can be removed without
     compromising the understanding of
    the commenter's perspective and
    context.
* Moderate Importance: Adds a meaningful
     layer to the context but is not
    essential to the commenter's overall
     message. Can be generalized.
* High Importance: Essential for an
    accurate understanding of the
```

```
    commenter's perspective and context,
     should be kept as is.

Given instance:
* Title: {title}
* Post: {post}
* Comment: {comment}
* Disclosure Span in the Comment: {
    disclosure}
* Human Rating of Importance: {
    human_rating}
{post_empty_explaination}
Note: The disclosure span is marked
    between special tokens <disclosure
    ></disclosure> in the Comment.

Instructions:
1. Carefully read the Title, Post, and
    Comment to understand the context.
2. Based on the human rating, write a
    detailed thought process that leads
    to this rating, according to the
    definitions. Your thought process
    should build up the reasoning that
    culminates in the rating, rather
    than stating the rating and then
    explaining it.
3. Ensure that your thought process is
    clear and straight to the point,
    avoiding filler sentences or
    unnecessary elaboration.
4. Present your thought in a JSON format
     of {{"Thought": "xxx"}}.
```

H.2 One-Span Abstraction

Instruction:

```
Your task is to generalize the given '
    disclosure span' in the sentence,
    which is providing a generalized
    alternative that is less specific
    but retains the core meaning of the
    original span.

Remember the following criteria:
* Only the disclosure span should be
    generalized; the rest of the
    sentence should remain intact.
* Make sure the generalized span fits
    seamlessly into the original
    sentence, maintaining proper syntax
    and grammar.

Sentence: "{sentence}"
Disclosure Span to Revise: "{span}"
Generalized Span:
```

Instruction with thought:

```
Your task is to generalize the given '
    disclosure span' in the sentence.

Please follow these steps:
1. First, provide a rationale explaining
     why the disclosure span needs
    generalization.
2. Then, provide a generalized
    alternative that is less specific
    but retains the core meaning of the
    original span.

Remember the following criteria:
* Only the disclosure span should be
    generalized; the rest of the
    sentence should remain intact.
* Make sure the generalized span fits
    seamlessly into the original
    sentence, maintaining proper syntax
    and grammar.

Sentence: "{sentence}"
Disclosure Span to Revise: "{span}"
Rationale:
```

H.3 Three-Span Abstraction

End-to-end instruction:

```
Your task is to generalize the given '
    disclosure span' in the sentence.
    Provide three diverse generalized
    spans that convey similar meaning
    but remove any overly specific or
    sensitive information.

Remember the following criteria:
* Only the disclosure span should be
    generalized; the rest of the
    sentence should remain intact.
* Generalized spans should be diverse
    but should all retain the essence of
     the original span.
* Make sure the generalized span fits
    seamlessly into the original
    sentence, maintaining proper syntax
    and grammar.
* Provide three diverse generalized
    alternatives in a JSON format like
    this: {{"span 1": "xxx", "span 2": "
    xxx", "span 3": "xxx"}}.

Sentence: "{sentence}"
Disclosure Span to Revise: "{span}"
Generalized Spans:
```

End-to-end instruction with thought:

```
Your task is to generalize the given '
    disclosure span' in the sentence.
    Provide three diverse generalized
    spans that convey similar meaning
    but remove any overly specific or
    sensitive information.

Please follow these steps:
1. First, provide a rationale explaining
     why the disclosure span needs
    generalization.
2. Then, offer three diverse generalized
     alternatives in a JSON format like
    this: {{"span 1": "xxx", "span 2": "
    xxx", "span 3": "xxx"}}.

Remember the following criteria:
* Only the disclosure span should be
    generalized; the rest of the
    sentence should remain intact.
* Generalized spans should be diverse
    but should all retain the essence of
     the original span.
```

```
* Make sure the generalized span fits
    seamlessly into the original
    sentence, maintaining proper syntax
    and grammar.

Sentence: "{sentence}"
Disclosure Span to Revise: "{span}"
Rationale:
```

Iterative instruction:

```
Your task is to generalize the given '
    disclosure span' in the sentence,
    which is providing a generalized
    alternative that is less specific
    but retains the core meaning of the
    original span.

Remember the following criteria:
* Only the disclosure span should be
    generalized; the rest of the
    sentence should remain intact.
* The generalized span should be
    different from the example
    generalizations but should retain
    the essence of the original span.
* Make sure the generalized span fits
    seamlessly into the original
    sentence, maintaining proper syntax
    and grammar.

Sentence: "{sentence}"
Disclosure Span to Revise: "{span}"
Example Generalizations: {examples}
Generalized Span:
```

Iterative instruction with thought:

```
Your task is to generalize the given '
    disclosure span' in the sentence.

Please follow these steps:
1. First, provide a rationale explaining
     why the disclosure span needs
    generalization.
2. Then, offer one diverse generalized
    alternatives that is different from
    the example generalizations provided
    .

Remember the following criteria:
* Only the disclosure span should be
    generalized; the rest of the
    sentence should remain intact.
* The generalized span should be
    different from the examples but
    should retain the essence of the
    original span.
* Make sure the generalized span fits
    seamlessly into the original
    sentence, maintaining proper syntax
    and grammar.

Sentence: "{sentence}"
Disclosure Span to Revise: "{span}"
Example Generalizations: {examples}
Rationale:
```

H.4 Importance Rating

Instruction:

```
Rate the importance of the disclosure
    span in a Reddit comment within the
    context into three Likert-scale:
* Low Importance: Can be removed without
     compromising the understanding of
    the commenter's perspective and
    context.
* Moderate Importance: Adds a meaningful
     layer to the context but is not
    essential to the commenter's overall
     message. Can be generalized.
* High Importance: Essential for an
    accurate understanding of the
    commenter's perspective and context,
     should be kept as is.

Given instance:
* Title: {title}
* Post: {post}
* Comment: {comment}
* Disclosure Span in the Comment: {
    disclosure}
{post_empty_explaination}
Note: The disclosure span is marked
    between special tokens <disclosure
    ></disclosure> in the Comment.

Read the Title, Post, and Comment
    carefully to understand the context,
     and provide your importance rating
    of the disclosure span.
```

Instruction with thought:

```
Rate the importance of the disclosure
    span in a Reddit comment within the
    context into three Likert-scale:
* Low Importance: Can be removed without
     compromising the understanding of
    the commenter's perspective and
    context.
* Moderate Importance: Adds a meaningful
     layer to the context but is not
    essential to the commenter's overall
     message. Can be generalized.
* High Importance: Essential for an
    accurate understanding of the
    commenter's perspective and context,
     should be kept as is.

Given instance:
* Title: {title}
* Post: {post}
* Comment: {comment}
* Disclosure Span in the Comment: {
    disclosure}
{post_empty_explaination}
Note: The disclosure span is marked
    between special tokens <disclosure
    ></disclosure> in the Comment.

Read the Title, Post, and Comment
    carefully to understand the context.
     Write down your thought process.
    And in the end, provide your
    importance rating of the disclosure
    span in a JSON format: {{"Importance
    ": "Low/Moderate/High"}}.
```