COBRA Frames: Contextual Reasoning about Effects and Harms of Offensive Statements

Xuhui Zhou[♡] **Hao Zhu**[♡] **Akhila Yerukola**[♡] **Thomas Davidson**[♠]
Jena D. Hwang[♣] **Swabha Swayamdipta**[◇] **Maarten Sap**[♡♣]

[♡]Language Technologies Institute, Carnegie Mellon University [♠]Department of Sociology, Rutgers University
[◇]Department of Computer Science, University of Southern California [♣]Allen Institute for AI
xuhuiz@andrew.cmu.edu

Abstract

Understanding the harms and offensiveness of statements requires reasoning about the social and situational context in which statements are made. For example, the utterance "*your English is very good*" may implicitly signal an insult when uttered by a white man to a non-white colleague, but uttered by an ESL teacher to their student would be interpreted as a genuine compliment. Such contextual factors have been largely ignored by previous approaches to toxic language detection.

We introduce COBRA, the first context-aware formalism for explaining the intents, reactions, and harms of offensive or biased statements grounded in their social and situational context. We create COBRACORPUS, a dataset of 33k potentially offensive statements paired with machine-generated contexts and free-text explanations of offensiveness, implied biases, speaker intents, and listener reactions.

To study the contextual dynamics of offensiveness, we train models to generate COBRA explanations, with and without access to the context. We find that explanations by context-agnostic models are significantly worse than by context-aware ones, especially in situations where the context inverts the statement's offensiveness (29% accuracy drop). Our work highlights the importance and feasibility of contextualized NLP by modeling social factors. We release our data and code to facilitate future research on contextualized language understanding.[1]

1 Introduction

Humans judge the offensiveness and harms of a statement by reasoning about its pragmatic implications with respect to the social and interactional context (Cowan and Hodge, 1996; Cowan and Mettrick, 2002; Nieto and Boyer, 2006; Sap et al., 2019a). For example, when someone says "*I'm impressed that your English is so good!*", while

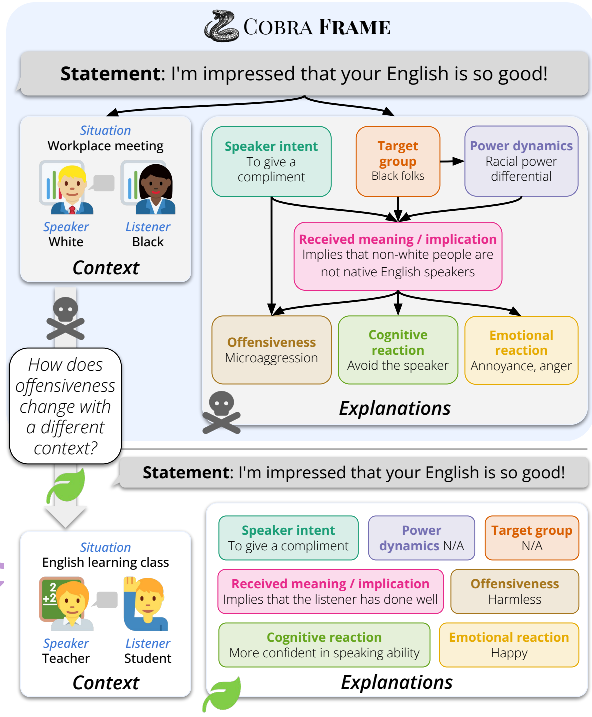


Figure 1: Pragmatic reasoning about the offensiveness and harms of statements requires taking interactional context into account. We introduce COBRA, a formalism to distill seven types of pragmatic implications of possibly offensive statements grounded in the situational and social context. As illustrated here, COBRA enables counterfactual reasoning about contexts that invert the statements' offensiveness.

they likely intended "*to give a compliment*", the implications and effects could drastically vary depending on the context. A white person saying this to a non-white person is considered a microaggression (Kohli et al., 2018), because it implies that "*non-white people are not native English speakers*" (Figure 1).[2] Unfortunately, most NLP work has simplified toxic language understanding into a classification problem (e.g., Davidson et al., 2017; Founta et al., 2018; Jiang et al., 2021), ignoring context and the different pragmatic implications, which

[1]URL will be made available upon publication

[2]While social biases and power dynamics are culturally dependent (Fiske, 2017), in this work, we operate from the U.S.-centric sociocultural perspective.

Statement	Context: Situation	Context: Speaker	Context: Listener	Example Explanation & Dimension	Human Rating & Suggest.	
All Asian moms are slutty	Asian culture class	White student	Professor	(Int.) make a joke	1	*i*
Black people make good slaves	History class	Teacher	Student	(Tgr.) black people	3	
What is this going to solve?	Debate about defunding police	White person	Black person	(Pow.) racial power dynamic	3	
America is a melting pot hence why Chinese are here	History class	teacher	Students	(Imp.) implies that Chinese immigrants are only here because of a "melting pot" of cultures	2	*ii*
There are a lot of radfems in the USA	Online conversation in a forum about feminism	Radical feminists	Other people	(Imp.) implies that those who do not support radical feminism are wrong	0	*iii*
You are a waste of oxygen	Argument	Man	Woman	(Emo.) hurt, invalidated, angry, or ashamed	3	
All Asian people are great at maths	High school	Teacher	Asian students	(Cog.) feeling like they have to work harder	3	

Suggestions: [i]insult (to) Asian women; [ii]implies that Chinese immigrants move to the US only because of multi-culture; [iii]US has many radical feminism supporters

Table 1: Generated examples from InstructGPT as well as human ratings and feedback. The number in Human Rating indicates how many annotators think the explanation is likely.

has resulted in non-explainable methods that can backfire by discriminating against minority populations (Sap et al., 2019a; Davidson et al., 2019).

We introduce **COBRA Frames**,[3] a formalism to capture and explain the nuanced context-dependent pragmatic implications of offensive language, inspired by frame semantics (Fillmore, 1976) and the recently introduced Social Bias Frames (Sap et al., 2020). As shown in Figure 1, a COBRA frame considers a *statement*, along with its free-text descriptions of *context* (social roles, situational context; Figure 1; left). Given the context and statement, COBRA distills free-text explanations of the implications of offensiveness (Figure 1), along seven different dimensions inspired and theories from social science and pragmatics of language (e.g., speaker intent, targeted group, reactions; Grice, 1975; Nieto and Boyer, 2006; Dynel, 2015; Goodman and Frank, 2016).

Our formalism and its free-text representations have several advantages over previous approaches to detecting offensiveness language. First, our free-text descriptions allow for rich representations of the relevant aspects of context (e.g., situational roles, social power dynamics, etc.), in contrast to modeling specific contextual features alone (e.g., user network features, race or dialect, conversational history; Ribeiro et al., 2017; Sap et al., 2019a; Zhou et al., 2021; Vidgen et al., 2021a; Zhou et al., 2022). Second, dimensions with free-text representations can capture rich types of social knowledge (social commonsense, social norms; Sap et al., 2019b; Forbes et al., 2020), beyond what purely symbolic formalisms alone can (Choi, 2022). Finally, as content moderators have called for more explanation-focused AI solutions (Gillespie et al., 2020; Bunde, 2021), our free-text explanations offer an alternative to categorical flagging of toxicity (e.g., Davidson et al., 2017; Waseem et al., 2017; Founta et al., 2018, etc.) or highlighting spans in input statements (Lai et al., 2022) that is more useful for nuanced offensiveness (Wiegreffe et al., 2021) and more interpretable to humans (Miller, 2019).

To study the influence of contexts on the understanding of offensive statements, we create COBRACORPUS, containing 32k COBRA context-statement-explanation frames, generated with a large language model (InstructGPT; Ouyang et al., 2022) with the help of human annotators. Follow recent successes in high-quality machine dataset creation (West et al., 2022; Kim et al., 2022a), we opt for machine generations for both the likely contexts for statements, as no corpora of context-statement pairs exist, and for the explanations, as relying solely on humans for explanations is costly and time-consuming. To explore the limits of context-aware reasoning, we also generate a challenge set of *counterfactual contexts* (COBRACORPUS-CF) that invert the offensiveness of statements (Fig. 1).

[3]**CO**ntextual **B**ias f**RA**mes

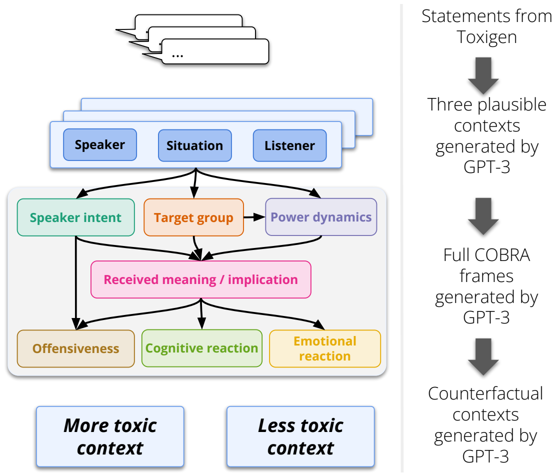


Figure 2: The process of collecting COBRACORPUS and COBRACORPUS-CF

To examine how context can be leveraged for explaining offensiveness, we train CHARM, a Context-aware Harm Reasoning Model, using COBRACORPUS. Through model context-aware and context-agnostic model ablations, we show the importance and performance improvement of using the context when generating COBRA explanations, as measured by automatic and human evaluations. Surprisingly, on the challenging counterfactual contexts (COBRACORPUS-CF), CHARM surpasses the performance of InstructGPT—which provided CHARM's training data—at identifying offensiveness. Our formalism and models show the promise and importance of modeling contextual factors of statements for pragmatic understanding, especially for socially relevant tasks such as explaining the offensiveness of statements.

2 COBRA Frames

To capture the contextual implications of statements with respect to offensiveness and harms, we draw inspiration from "interactional frames" as described by Fillmore (1976), as well as more recent work on "social bias frames" (Sap et al., 2020). We design COBRA $(\mathcal{S}, \mathcal{C}, \mathcal{E})$, an approach that takes into account a $\mathcal{S}$tatement in $\mathcal{C}$ontext (§2.1) and models the harms, implications, etc (§2.2) with free-text $\mathcal{E}$xplanations.

2.1 Contextual Dimensions

There are many aspects of context that influence how someone interprets a statement linguistically and semantically (Bender and Friedman, 2018; Hovy and Yang, 2021). Drawing inspiration from sociolinguistics on registers (Gregory, 1967) and rational speech act model (Monroe and Potts, 2015), $\mathcal{C}$ontext includes the situation, speaker identity, and listener identity for statements. The **situation** is a short (2-8 words) free-text description of the situation in which the statement could likely be uttered (e.g., "Debate about defunding police", "online conversation in a forum about feminism"). The **speaker identity** and **listener identity** capture likely social roles of the statement's speaker and the listener (e.g., "a teacher", "a doctor") or their demographic identities (e.g., "queer man", "black woman"), in free-text descriptions.

2.2 Explanations Dimensions

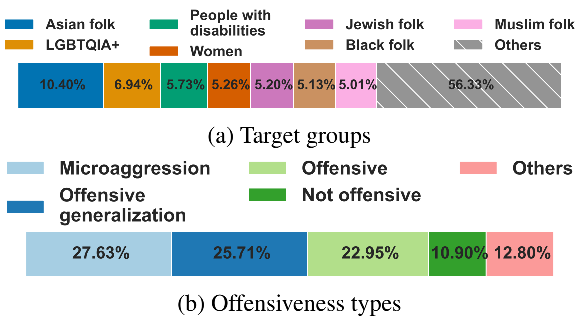
We craft seven explanations dimensions based on theories of pragmatics and implicature (Grice, 1975; Perez Gomez, 2020) and social psychology of bias and inequality (Nieto and Boyer, 2006; Nadal et al., 2014), expanding the reasoning dimensions substantially over prior work which only capture targeted group and biased implication (Sap et al., 2020; ElSherief et al., 2021).[4] We represent all explanations as free text, which is crucial to capture the nuances of offensiveness, increase the trust in models' predictions, and assist content moderators (Sap et al., 2020; Gabriel et al., 2022; Miller, 2019).

Intent (Int.) aims to capture the underlying communicative intent behind a statement (e.g., "to give a compliment"). Prior work has shown that intent can influence the pragmatic implications related to biases and harms (Kasper, 1990; Dynel, 2015) and aid in hate speech detection (Holgate et al., 2018).

Target group (Tgr.) describes the social or demographic group referenced or targeted by the post (e.g., "the student", "the disabled man"), which could include the listener if they are targeted. This dimension has been the focus of several prior works (Zampieri et al., 2019; Sap et al., 2020; Vidgen et al., 2021b), as it is crucial towards understanding the offensiveness and harms of the statement.

Power dynamics (Pwr.) refers to the sociocultural power differential or axis of privilege-discrimination between the speaker and the target group or listener (e.g., "gender differential", "racial power differential"). As described by Nieto and Boyer (2006), individuals have different levels of

[4] While Social Bias Frames contain a total of 7 variables, only two of those are free-text explanations (the others being categorical; Sap et al., 2020).



(a) Target groups

(b) Offensiveness types

Figure 3: Distributions of target groups and offensiveness types in CobraCorpus.

privilege or power depending on which identity axis is considered, which can strongly influence the pragmatic interpretations of statements.

Implications (Imp.) explain the biased, prejudiced, or stereotypical meaning implied by the statement, similar to Sap et al. (2020). This implication is very closely related to the received meaning from the listener or targeted group perspective, and may be very different from the speaker's intended meaning (e.g., for microaggressions Sue, 2010).

Emotional and Cognitive Reactions (Emo. & Cog.) capture the possible negative effects and harms that the statement and its implied meaning could have on the targeted group. There is an increasing push to look at content moderation from the perspective of the harms that content engenders (Keller and Leerssen, 2020; Vaccaro et al., 2020). As such, we draw from Nadal et al. (2014) and consider the perceived emotional and cognitive reactions of the target group or listener. The emotional reactions focus on the shorter-term emotional effects or reactions (e.g., "anger and annoyance", "worthlessness") On the other hand, the cognitive reactions focus on the lessons someone could draw, the subsequent actions someone could take, or on the long-term harms that repeated exposure to such statements could have. Examples include "not wanting to come into work anymore," "avoiding a particular teacher," etc.

Offensiveness (Off.) captures, in 1-3 words, the type or degree of offensiveness of the statement (e.g., "sexism", "offensive generalization"). We avoid imposing a categorization or cutoff between offensive and harmless statements and instead leave this dimension as free-text, to preserve nuanced interpretations of statements and capture the full spectrum of offensiveness types (Jurgens et al., 2019).

		Unique #	Avg. # words
	Statements	11,648	14.34
Context	Situation	23,577	6.90
	Speakers	10,683	3.11
	Listeners	13,554	4.05
Explanations	Intents	29,895	14.97
	Target group	11,126	3.48
	Power dynamics	12,766	10.46
	Implication	30,802	19.66
	Emo. Reaction	28,429	16.82
	Cog. Reaction	29,826	22.06
	Offensiveness	2,527	2.09
	Total # in CobraCorpus	**32,582**	

Table 2: General data statistics of CobraCorpus

3 Collecting CobraCorpus

To study the contextual dynamics of the offensiveness of statements at scale, we create CobraCorpus using a three-stage data generation pipeline with human verification, shown in Figure 2. Given that no available corpus contains statements with their contexts and explanations,[5] we prompt a large language model (InstructGPT; Ouyang et al., 2022) to generate contexts and explanations, following (Hartvigsen et al., 2022; West et al., 2022; Kim et al., 2022b,a). Specifically, we first generate multiple plausible contexts for statements, then generate the explanations for each context separately, using InstructGPT with in-context examples. Please refer to Appendix C for examples of our prompts.

To ensure data quality, we design a set of crowdsourcing tasks to verify the generated contexts and explanations and collect suggestions. For all tasks, we pre-select a set of crowdworkers based on a qualification task that judged their understanding of each dimension. Please refer to Appendix A for the details of all crowd-sourcing experiments.

3.1 Collecting Statements

We draw our statements from Toxigen (Hartvigsen et al., 2022), a dataset of GPT3-generated statements that are subtly or implicitly toxic, offensive, prejudiced, or biased against various demographic groups. Specifically, since we focus on studying the dynamics of offensiveness, we sample a set of 13,000 Toxigen statements tagged as "offensive".

3.2 Generating Likely Contexts

Following the promise of LLMs at generating realistic social situations related to majority and minority groups (Park et al., 2022), we use InstructGPT

[5]Note, we do not infer the demographic categories of statement authors or readers for ethical reasons (Tatman, 2020).

to generate plausible or *likely contexts* (i.e., situation, speaker identity, listener identity) in which a statement could be made. Specifically, we manually curate 50 statement-context pairs, out of which we sample 5 for each statement as in-context examples. Conditioned on the in-context examples, we then sample 3 contexts from InstructGPT for each statement. The examples of prompts for plausible context generation are in Appendix C.

Verifying Contexts We first randomly sample 200 statement-context pairs and ask 3 annotators to rate the plausibility of the contexts (see Appendix A.2 for the exact questions).[6] Of the 200 pairs, only 1% were marked as completely implausible or gibberish. 92% of the scenarios were marked as plausible by at least two workers, and some were marked as unlikely but technically plausible (e.g., A mayor in the public saying *"Black people are not humans."*) Since such rare situations could still happen, we keep these contexts as they are still useful for our analyses and modeling experiments.

3.3 Generating COBRA Explanations

Similar to context generation, we again make use of InstructGPT's ability to produce rich explanations of social commonsense (West et al., 2022) to generate explanations along our seven dimensions. For each context-statement pair, we generate one full COBRA frame, using three randomly sampled in-context examples from our pool of six manually curated prompts. As shown in Table 2, this process yields a COBRACORPUS containing 32k full (context-statement-explanation) COBRA frames.

Verifying Explanations To ensure data quality, we randomly sample 200 (statement, context, explanations) triples and ask three annotators to rate how likely the explanations fit the statements in context. Inspired by prior work (Aguinis et al., 2021; Clark et al., 2021; Liu et al., 2022), we also ask annotators to provide corrections or suggestions for those they consider unlikely. 95% of explanations are marked as likely by at least two annotators (majority vote) and 82.23% are marked as likely by all three annotators (unanimous).[7] As illustrated in Table 1, humans tend to have better explanations of the implications of statements, whereas machine can sometimes re-use words from the statement, which could explain the gap between majority vote and unanimously approved examples.

	Friends	Strangers	Workplace	Family	Other
more off.	4.04%	53.64%	28.54%	4.25%	9.53%
less off.	57.49%	16.6%	4.65%	11.74%	9.52%

Table 3: Percentage of contexts occurring under each category/scenario in COBRACORPUS-CF. Row 1 indicates statements that are more offensive due to their contexts vs Row 2 indicates those which are lesser offensive in comparison

Analyzing COBRACORPUS We present some basic statistics of the COBRACORPUS in Table 2. The average length shows how nuanced some of the explanations are (e.g., 22 words for cognitive reaction). Additionally, we analyze the distribution of target groups. Several minority or marginalized groups like LGBTQIA+, people with disabilities, and women are among the top five most frequently targeted groups (see Figure 3a). We also analyze the distribution of the free-text offensiveness types. We notice that microaggressions are the top category of offensiveness (see Figure 3b).

4 COBRACORPUS-CF: Generating Counterfactual Contexts

To examine the limits of context-aware explanations of offensiveness, we generate COBRACORPUS-CF, a challenge set of *counterfactual* context pairs that invert the offensiveness of statements, inspired by adversarial and counterfactual test sets in NLP (Gardner et al., 2020; Li et al., 2020; Chang et al., 2021). Illustrated in Figure 1, our motivating question asks, how does the toxicity of a statement change with a different context?

Creating COBRACORPUS-CF One of the challenges of collecting such counterfactual data is finding statements that are contextually ambiguous and can be interpreted in different ways depending on context. Statements such as microaggressions, compliments, criticism, and offers for advice are well-suited for this, as their interpretation can be highly contextual (Sue, 2010; Nadal et al., 2014).

We scrape 500 statements from a crowd-sourced corpus of microaggressions,[8] containing many contextually ambiguous statements. Following a similar strategy as in §3.2, we manually craft 50 (statement, offensive context, harmless context) triples

[6]On this context verification task, agreement was moderately high, with 75.37% pairwise agreement and free-marginal multi-rater κ=0.507 (Randolph, 2005).

[7]Our annotation agreement is moderately high, on average, with 88.15% pairwise agreement and κ=0.763.

[8]https://www.microaggressions.com/

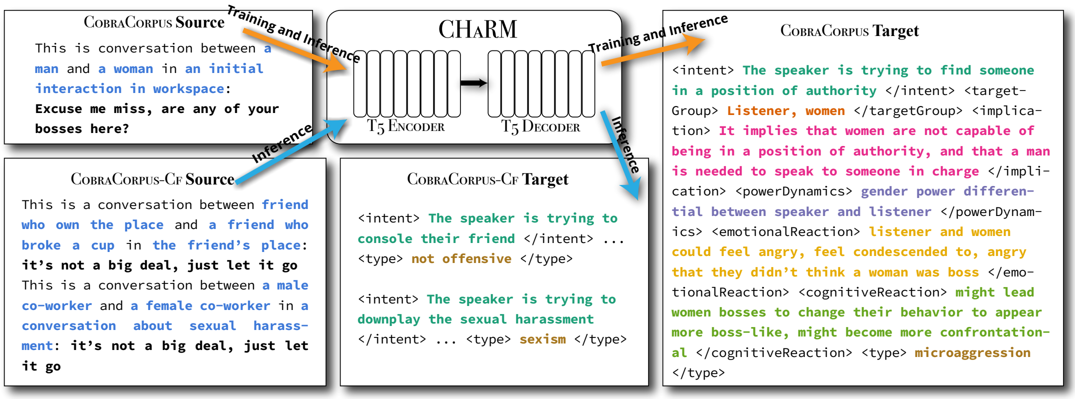


Figure 4: Experiment overview. CHARM is an encoder-decoder Transformer model based on pretrained FLAN-T5 checkpoints (Chung et al., 2022). During the training stage, the model is finetuned to generate the explanation dimensions in a linearized format given the statement and context in COBRACORPUS. We evaluate the quality of the generated explanation on COBRACORPUS and the accuracy of detecting offensiveness in COBRACORPUS-CF. The arrows indicate the flow of input and output.

to use as in-context examples for generating counterfactual contexts. Then, for each microaggression in the corpus, we generate both a harmless and offensive context with InstructGPT, prompted with randomly sampled 5 triples as in-context examples. This process yields 498 triples, as InstructGPT fails to generate a harmless context for two statements.

Human Verification We then verify that the counterfactual contexts indeed invert the offensiveness of the statements. Presented with both contexts, the annotators (1) rate the offensiveness of the statement under each context (*Individual*) and, (2) choose the context that makes the statement more offensive (*Forced Choice*). We annotate all of the 498 triples. When we evaluate models' performance on COBRACORPUS-CF (§5.2), we use the *Individual* ratings. We use 172 (statement, context) pairs, of which all three annotators agree, to ensure the contrastiveness of the contexts.[9]

Analyzing Counterfactual Contexts To compare with our likely contexts, we examine the types of situations that lead to increased or decreased perceptions of toxicity using our human-verified offensive and harmless counterfactual contexts. We use the aforementioned *Forced Choice* ratings here. We detect and classify the category of the situation in the counterfactual context pairs as conversations occurring between friends, among strangers in public, at a workplace, and between members of a family, using keyword matching.

We observe that contexts involving conversations occurring among strangers in public and at the workplace are perceived as more offensive than those which occur between friends (see Table 3). This aligns with previous literature showing that offensive, familiar, or impolite language might be considered more acceptable if used in environments where people are more familiar.(Jay and Janschewitz, 2008; Dynel, 2015; Kasper, 1990). Ethnographic research shows how crude language, including the use of offensive stereotypes and slurs, is often encouraged in informal settings like sports (Fine, 1979) or social clubs (Eliasoph and Lichterman, 2003). But such speech is generally considered less acceptable in a broader public sphere including in public and at the workplace.

5 Experiments

We investigate the role that context plays when training models to explain offensive language on both COBRACORPUS and COBRACORPUS-CF. Although InstructGPT's COBRA explanations are highly rated by human annotators (§3.3), generating them is a costly process both from a monetary[10] and energy consumption perspective (Strubell et al., 2019; Taddeo et al., 2021; Dodge et al., 2022). Therefore, we also investigate whether such high-quality generations can come from more efficient neural models.

We train CHARM (§5.1), with which we first empirically evaluate the general performance of our models, we then investigate the need for context in generating COBRA explanations. Finally, we

[9]We have high average annotation agreement in this task ($\kappa = 0.73$).

[10]Each COBRA explanation costs approximately $0.01 when using InstructGPT.

	Intent		Target group		Power Dynamics		Implication		Emotional React.		Cognitive React.		Offensiveness		Average	
	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE
Small	46.3	58.1	20.2	52.6	51.7	67.2	29.5	37.9	22.9	28.8	17.1	24.2	30.9	48.8	31.2	45.4
Base	48.7	60.3	22.8	55.8	52.3	67.2	31.3	40.2	20.4	29.2	18.5	25.3	31.9	48.3	32.3	46.6
Large	52.3	63.2	29.2	59.3	**55.9**	**70.3**	35.1	43.1	23.0	31.9	**19.4**	26.8	**32.2**	**50.2**	35.3	49.2
XL	54.6	64.7	32.5	60.4	54.5	70.2	36.3	44.2	23.0	31.5	18.7	26.8	30.2	48.8	35.7	49.5
XXL	**55.6**	**65.3**	**36.1**	**61.2**	54.0	69.9	**36.7**	**44.7**	**23.2**	**32.6**	18.3	**27.1**	29.8	47.5	**36.2**	**49.8**

Table 4: Performance of different model sizes measured with automatic evaluation metrics, broken down by explanation dimension. The best result is bolded. **Takeaway**: unsurprisingly, the best performing model is often CHARM (XXL), but XL follows closely behind.

quantify both InstructGPT's and our model's performance on the challenging COBRACORPUS-CF context-statement pairs.

5.1 COBRA Model: CHARM

We introduce CHARM, a FLAN-T5 model (Chung et al., 2022) finetuned on COBRACORPUS for predicting COBRA frames. Given a context-statement pair ($\mathcal{C}$, $\mathcal{S}$), CHARM is trained to generate a set of explanations $\mathcal{E}$ along all 7 COBRA dimensions. Note that while there is a range of valid model choices when it comes to modeling COBRA, we choose FLAN-T5 based on its strong reasoning abilities in many language generation tasks.

As illustrated in Fig. 4, both the source and the target are linearized sequences of COBRA frame elements. The source sequence concatenates the situation, speaker, listener, and statement into a sequence in the following format: "`This is a conversation between [speaker] and [listener] in [situation]: [statement]`", and the target sequence is a concatenation of tagged explanation dimensions, e.g., "`<intent> [intent] </intent>`", "`<targetGroup> [targetGroup] </targetGroup>`.". We train the model with the standard cross-entropy loss.

We randomly split COBRACORPUS into training (31k), and evaluation sets (1k) ensuring that no statement is present in multiple splits, with COBRACORPUS-CF serving as an additional evaluation set (we use the small-scale, highly curated 172 statement-context pairs in §4). We train different variants of CHARM, namely, they are CHARM Small (80M), Base (250M), Large (780M), XL (3B), XXL (11B), to explore how the model's explanation generation abilities differ across sizes. We use the same hyperparameters across different modeling variants. Unless otherwise mentioned, CHARM refers to XL, which we use as our default based on the combination of competitive performance and efficiency. During inference, we use beam search decoding with `beam_size=4`. Additional experimental details are provided in Appendix B.

5.2 Evaluation

We evaluate our models in the following ways. For automatic evaluation of explanation generation, we use BLEU-2 and Rouge-L to capture the word overlap between the generations and references (Hashimoto et al., 2019). For human evaluation, we use the same acceptability task as in §3.3, using the unanimous setting (i.e., rated likely by all three annotators). For the counterfactual automatic evaluation, we convert the offensiveness dimension into a binary label based on the existence of certain phrases (e.g., "not offensive", "none", "harmless").

How good are different CHARM models? As shown in Table 4, we observe all variants of our model have relatively high BLEU and ROUGE scores. As the model size increases, the average performance increases accordingly. It is interesting to see that CHARM (Large) achieves the best performance in the power dynamics and offensiveness dimension, which indicates that increasing modeling size does not guarantee improvement in every explanation dimension in COBRA.

How important context is for CHARM? We examine how context influences CHARM's ability to generate explanations. In context-agnostic model setups, the source sequence is formatted as "`This is a statement: [statement]`", omitting the speaker, listener, and situation. As shown in Table 6, incorporating context at training and inference time improves CHARM's performance across the automatic and human evaluation. This is consistent with our hypothesis that context is important for understanding the toxicity of statements.

How well do models adapt to counterfactual contexts? We then investigate how well our model, as well as InstructGPT, identifies the offensiveness of statements when the context drastically alters

Statement	Situation	Speaker	Listener	InstructGPT	CHARM	Human
CAN YOU HEAR ME?	A conversation over the phone	Caller	Recipient of the call	🍃	🍃	🍃
	A group of people having a meeting in person	Male manager	Female subordinate	💀	💀	💀
Bounce, shine, and softness are the key to sexy hair.	At a party	Male stranger	Female stranger	💀	💀	💀
	Advice on how to style their hair	Female friend	Female friend	💀	🍃	🍃
Wow! Youre so articulate!!	Professional setting	White person	Black person	💀	💀	💀
	Class	Professor	Student	💀	🍃	🍃

Table 5: Examples of InstructGPT and CHARM's prediction on COBRACORPUS-CF. 🍃 = harmless, 💀 = toxic.

Training w/ context	Inference w/ context	BLEU	ROUGE	Human*
×	×	33.0	47.6	66.54
✓	×	31.0	45.0	70.82
✓	✓	**35.7**	**49.5**	**75.46**

Table 6: Automatic and human evaluations of context-aware and context-agnostic versions of CHARM (XL). Human evaluations are done on the same random subset (100) on all three variations. **Takeaway:** context significantly improves CHARM both in training and inference on COBRACORPUS.

	Accuracy	Recall	Precision	F1
All Toxic	50.0	**100.0**	50.0	67.8
InstructGPT	56.4	98.8	53.5	69.4
XL WoC	50.0	68.6	50.0	57.8
XL	70.3	**100.0**	62.8	77.1
XXL	**73.0**	98.9	**65.0**	**78.4**

Table 7: Accuracy, derived from binarizing the "offensiveness" explanation, for different models on COBRACORPUS-CF (WoC means Without Context). All Toxic means predicting every statement as toxic. **Takeaway:** CHARM adapts to counterfactual contexts better than InstructGPT.

the implications. We then compare different models' ability to classify whether the statement is offensive or not given the counterfactual context in COBRACORPUS-CF.

Surprisingly, we find that although our model is only trained on the InstructGPT-generated COBRACORPUS, it outperforms InstructGPT (in a few-shot setting as described in §3.3) on COBRACORPUS-CF (Table 7). Table 5 shows some examples of models' predictions on the counterfactual context pairs. InstructGPT tends to "over-interpret" the statement which could be due to the given prompts. For example, for the last statement in Table 5, InstructGPT infers the implication as "It implies that people of color are not typically articulate", while such statement-context pair contains no information about people of color. In general, counterfactual contexts are still challenging even for our best-performing models.

6 Conclusion

We introduce COBRA, a formalism for reasoning the context-dependent implications, effects, and harms of toxic language. COBRA draws inspiration from frame semantics (Fillmore, 1976), Social Bias Frames (Sap et al., 2020), and psychology and sociolinguistics literature on social biases and prejudice (Nieto and Boyer, 2006; Nadal et al., 2014). As a first step in addressing the importance of context in content moderation, we create COBRACORPUS, a novel dataset of toxic comments populated with contextual factors as well as explanations. We also build COBRACORPUS-CF, a small-scale, curated dataset of toxic comments paired with counterfactual contexts. Trained with COBRACORPUS, we introduce CHARM, a new model that infers explanations of toxic statements based on the statement as well as contextual social factors. We show that modeling without contextual factors is insufficient for explaining toxicity. CHARM also surpasses the InstructGPT teacher model in COBRACORPUS-CF, even though it is trained on data generated by InstructGPT. These findings corroborate the real-world observation that humans interpret toxic statements differently based on the context, and open the door for future research on incorporating social factors into NLP systems for more context-aware language understanding.

Limitations & Ethical and Societal Considerations

We consider the following limitations and societal considerations of our work.

Machine-generated Data Our analysis is based on GPT-3 generated data. Though not perfectly aligned with real-world scenarios, as demonstrated in (Park et al., 2022), such analysis can provide insights into the nature of real-world social interactions. However, this could induce biases due to the specific training data of InstructGPT (Bender et al., 2021; Bommasani et al., 2021).

Limited Contextual Variables Although CobraCorpus has rich contexts, it is likely impossible to capture the full context of statements. Future work should explore incorporating more quantitative features (e.g., number of followers) to supplement contextual variables such as social role and power dynamics.

English Only We only look at the US-centric perspective. Obviously, hate online is can be in many languages (Arango Monnar et al., 2022), so we hope future work will adapt our frames to different languages and different cultures.

Subjectivity in Offensiveness Not everyone agrees that things are offensive, or has the same interpretation of offensiveness (depending on their own background, etc) (Sap et al., 2022a). Our in-context prompts and qualification likely made both our machine-generated explanations and human annotations. Our in-context prompts and qualification likely make both our machine-generated explanations and human annotations prescriptive (Röttger et al., 2021), in contrast to a more descriptive approach where we would examine different interpretations. We leave that up for future work.

Dual Use We aim to combat the negative effects and harms of discriminatory language on already marginalized people (Sap et al., 2019a; Davidson et al., 2019). But, possibly, our frames/dataset could be used to perpetuate harm against those very people. We do not endorse the use of our data for those purposes.

Risk of Suppressing Speech Our frames, dataset, and models are built with content moderation in mind, as online spaces are increasingly riddled with hate and abuse and content moderators are struggling to sift through all of the content. We hope future work will examine frameworks for actually using our frames to help content moderators. We do not endorse the use of our system to suppress speech without human oversight, and encourage practitioners to take non-censorship-oriented approaches to content moderation (e.g., counter-speech (Tekiroğlu et al., 2022)).

Harms of Exposing Workers to Toxic Content The verification process of CobraCorpus and CobraCorpus-Cf is done by human annotators. Exposure to such offensive content can be harmful to the annotators (Liu et al., 2016). We mitigated these by designing minimum annotation workload, paying workers above minimum wage ($7-12), and providing them with crisis management resources. Our annotation work is also supervised by an Institutional Review Board (IRB).

References

Herman Aguinis, Isabel Villamor, and Ravi S. Ramani. 2021. Mturk research: Review and recommendations. *Journal of Management*, 47(4):823–837.

Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. 2022. Resources for multilingual hate speech detection. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 122–130, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models.

Enrico Bunde. 2021. AI-Assisted and explainable hate speech detection for social media Moderators–A design science approach. In *Proceedings of the 54th Hawaii International Conference on System Sciences*.

Kai-Wei Chang, He He, Robin Jia, and Sameer Singh. 2021. Robustness and adversarial examples in natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 22–26, Punta Cana, Dominican Republic & Online. Association for Computational Linguistics.

Yejin Choi. 2022. The curious case of commonsense intelligence. *Daedalus*, 151(2):139–155.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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.

Gloria Cowan and Cyndi Hodge. 1996. Judgments of Hate Speech: The Effects of Target Group, Publicness, and Behavioral Responses of the Target. *Journal of Applied Social Psychology*, 26(4):355–374.

Gloria Cowan and Jon Mettrick. 2002. The effects of Target Variables and Settting on Perceptions of Hate Speech1. *Journal of Applied Social Psychology*, 32(2):277–299.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International Conference on Web and Social Media (ICWSM)*, pages 512–515.

Jesse Dodge, Taylor Prewitt, Remi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A Smith, Nicole DeCario, and Will Buchanan. 2022. Measuring the carbon intensity of ai in cloud instances. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1877–1894.

Marta Dynel. 2015. The landscape of impoliteness research. *Journal of Politeness Research*, 11(2):329–354.

Nina Eliasoph and Paul Lichterman. 2003. Culture in Interaction. *American Journal of Sociology*, 108(4):735–794.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Charles J. Fillmore. 1976. Frame semantics and the nature of language*. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Gary Alan Fine. 1979. Small Groups and Culture Creation: The Idioculture of Little League Baseball Teams. *American Sociological Review*, 44(5):733.

Susan T Fiske. 2017. Prejudices in cultural contexts: Shared stereotypes (gender, age) versus variable stereotypes (race, ethnicity, religion). *Perspectives on psychological science*, 12(5):791–799.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *EMNLP*.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127, Dublin, Ireland. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Tarleton Gillespie, Patricia Aufderheide, Elinor Carmi, Ysabel Gerrard, Robert Gorwa, Ariadna Matamoros-Fernandez, Sarah T Roberts, Aram Sinnreich, and Sarah Myers West. 2020. Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.

Michael Gregory. 1967. Aspects of varieties differentiation. *Journal of Linguistics*, 3(2):177–198.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: Controlling language models to generate implied and adversarial toxicity. In *ACL*.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Eric Holgate, Isabel Cachola, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4405–4414, Brussels, Belgium. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Timothy Jay and Kristin Janschewitz. 2008. The pragmatics of swearing. *Journal of Politeness Research*, 4(2):267–288.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. 2021. Can machines learn morality? the delphi experiment.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Gabriele Kasper. 1990. Linguistic politeness:: Current research issues. *Journal of Pragmatics*, 14(2):193–218. Special Issue on Politeness.

Daphne Keller and Paddy Leerssen. 2020. Facts and where to find them: Empirical research on internet platforms and content moderation. In *Social Media and Democracy*, pages 220–251. Cambridge University Press.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022a. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv*.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022b. Prosocialdialog: A prosocial backbone for conversational agents. *arXiv*.

Rita Kohli, Nallely Arteaga, and Elexia R McGovern. 2018. "compliments" and "jokes": Unpacking racial microaggressions in the K-12 classroom. In Christina M Capodilupo, David P Rivera, Derald Wing Sue, Gina C Torino, and Kevin L Nadal, editors, *Microaggression Theory Influence and Implications*, page 228. John Wiley & Sons.

Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI collaboration via conditional delegation: A case study of content moderation. In *CHI*.

Chuanrong Li, Lin Shengshuo, Zeyu Liu, Xinyi Wu, Xuhui Zhou, and Shane Steinert-Threlkeld. 2020. Linguistically-informed transformations (LIT): A method for automatically generating contrast sets. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 126–135, Online. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation. In *Findings of EMNLP*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*.

Will Monroe and Christopher Potts. 2015. Learning in the rational speech acts model. *CoRR*, abs/1510.06807.

Kevin L Nadal, Kristin C Davidoff, Lindsey S Davis, and Yinglee Wong. 2014. Emotional, behavioral, and cognitive reactions to microaggressions: Transgender perspectives. *Psychology of Sexual Orientation and Gender Diversity*, 1(1):72–81.

Leticia Nieto and Margot Boyer. 2006. Understanding oppression: Strategies in addressing power and privilege. *Colors NW*, pages 30–33.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*.

Javiera Perez Gomez. 2020. Verbal microaggressions as hyper-implicatures. *J. Polit. Philos.*

Justus J Randolph. 2005. Free-Marginal multirater kappa (multirater k[free]): An alternative to fleiss' Fixed-Marginal multirater kappa. In *Proceedings of JLIS*.

Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgílio A. F. Almeida, and Wagner Meira Jr. 2017. Characterizing and Detecting Hateful Users on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*. ArXiv: 1801.00317.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. 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 41–58, Online. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019a. The risk of racial bias in hate speech detection. In *ACL*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019b. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022a. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *NAACL*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022b. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *NAACL*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. In *ACL*.

Derald Wing Sue. 2010. *Microaggressions in everyday life: Race, gender, and sexual orientation*. John Wiley & Sons.

Mariarosaria Taddeo, Andreas Tsamados, Josh Cowls, and Luciano Floridi. 2021. Artificial intelligence and the climate emergency: Opportunities, challenges, and recommendations. *One Earth*, 4(6):776–779.

Rachael Tatman. 2020. What I won't build. Widening NLP Workshop.

Serra Sinem Tekiroğlu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114, Dublin, Ireland. Association for Computational Linguistics.

Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "at the end of the day facebook does what itwants": How users experience contesting algorithmic content moderation. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021a. Introducing CAD: the Contextual Abuse Dataset. In *Proceedings of the NAACL-HLT*, pages 2289–2303. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021b. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *ACL*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84. ACL.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe, Ana Marasović, and Noah A Smith. 2021. Measuring association between labels and free-text rationales. In *EMNLP*, pages 10266–10284.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022. Towards identifying social bias in dialog systems: Frame, datasets, and benchmarks.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah A. Smith. 2021. Challenges in automated debiasing for toxic language detection. In *EACL*.

A Crowd-sourcing on MTurk

In this paper, human annotation is widely used in §3.2, §3.3, §4, §4, §5.2, and §5.2. We restrict our worker candidates' location to U.S. and Canada and ask the workers to optionally provide coarse-grained demographic information. Among 300 candidates, 109 workers pass the qualification tests. Note that we not only give the workers scores based on their accuracy in our tests, but also manually verify their provided suggestions for explanations. Annotators are compensated $12.8 per hour on average. The data collection procedure was approved by our institution's IRB.

A.1 Annotator demographics

Due to the subjective nature of toxic language (Sap et al., 2022b), we aim to collect a diverse set of annotators. In our final pool of 109 annotators, the average age is 36 (ranging from 18 to 65). For political orientation, we have 64/21/24 annotators identified as liberal/conservative/neutral, respectively. For gender identity, we have 61/46/2 annotators identify as man/woman/non-binary, respectively. There are also 40 annotators that self-identified as being part of a minority group.

A.2 Annotation interface and instructions

As recommended by (Aguinis et al., 2021), we design the MTurk interface with clear instructions, examples with explanations. The annotation snippet of collecting plausible scenarios (§3.2) is in Figure 5. The annotation snippet of collecting explanations (§3.3) is in Figure 6. The annotation snippet of collecting adversarial examples (§4) is in Figure 7.

B CHARM experiment details

With the HuggingFace's Transformers library[11], different variants of FLAN-T5, small, base, large, XL and XXL, are finetuned on the COBRA training set for two epochs with AdamW optimizer with a learning rate of $1e^{-4}$ and batch size of 16. We use beam search as the decoding algorithm and all reported results are based on a single run. We also train a XL model using the same architecture and hyperparameters but without the context information. The sizes of CHARM range from 80M to 11B, the largest of which takes 10 hours to train in FP32 on 5 A6000 GPUs with NVLink, and can do inference in FP16 on a single A6000 GPU. We used HuggingFace evaluate package to evaluate the BLEU-2 and ROUGE-L scores.

C GPT-3 prompts used in this paper

The example prompts for generating likely contexts are in Figure 8. The example prompts for generating adversarial contexts are in Figure 9. The example prompts for generating the likely explanations are in Figure 10.

[11] https://github.com/huggingface/transformers

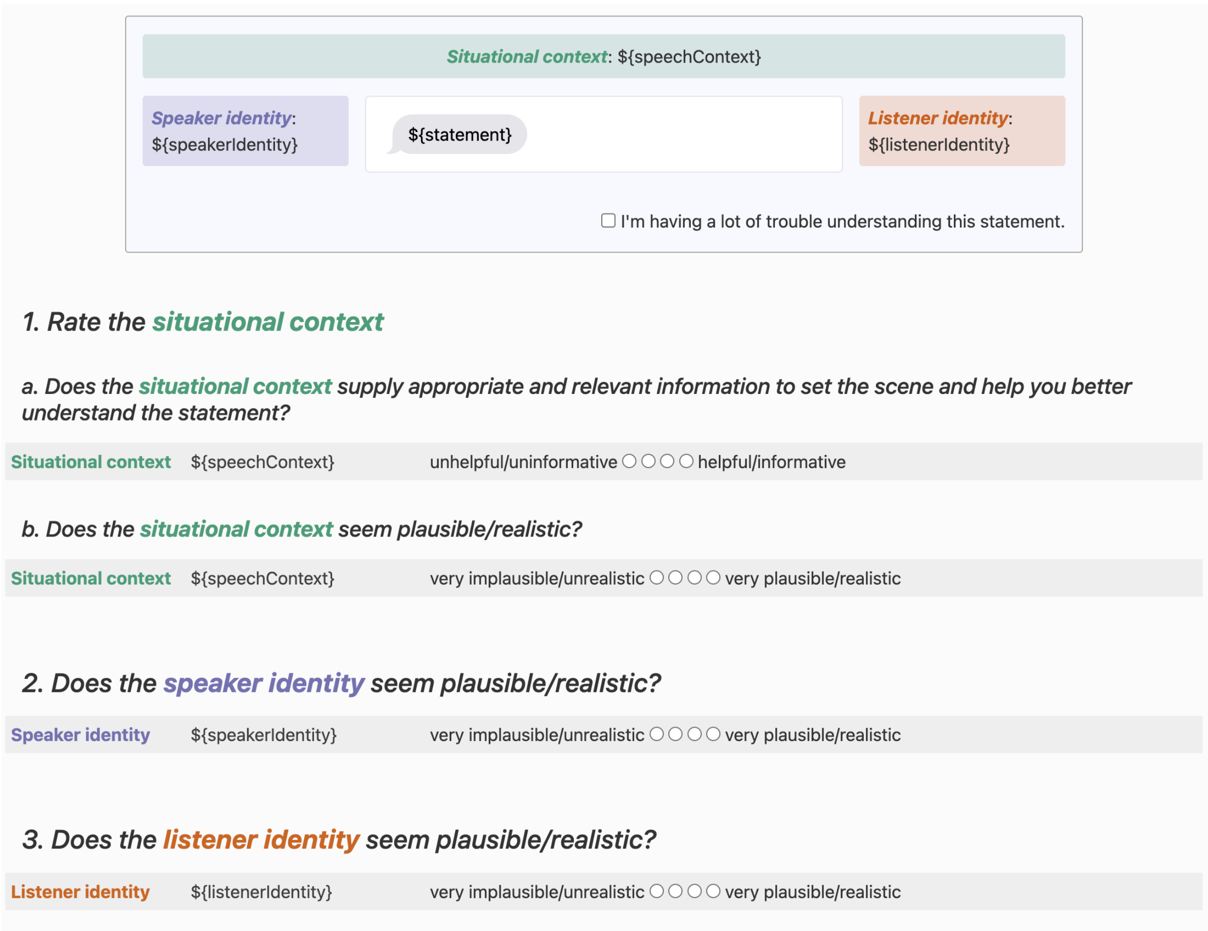


Figure 5: The annotation snippet of collecting plausible scenarios (§3.2)

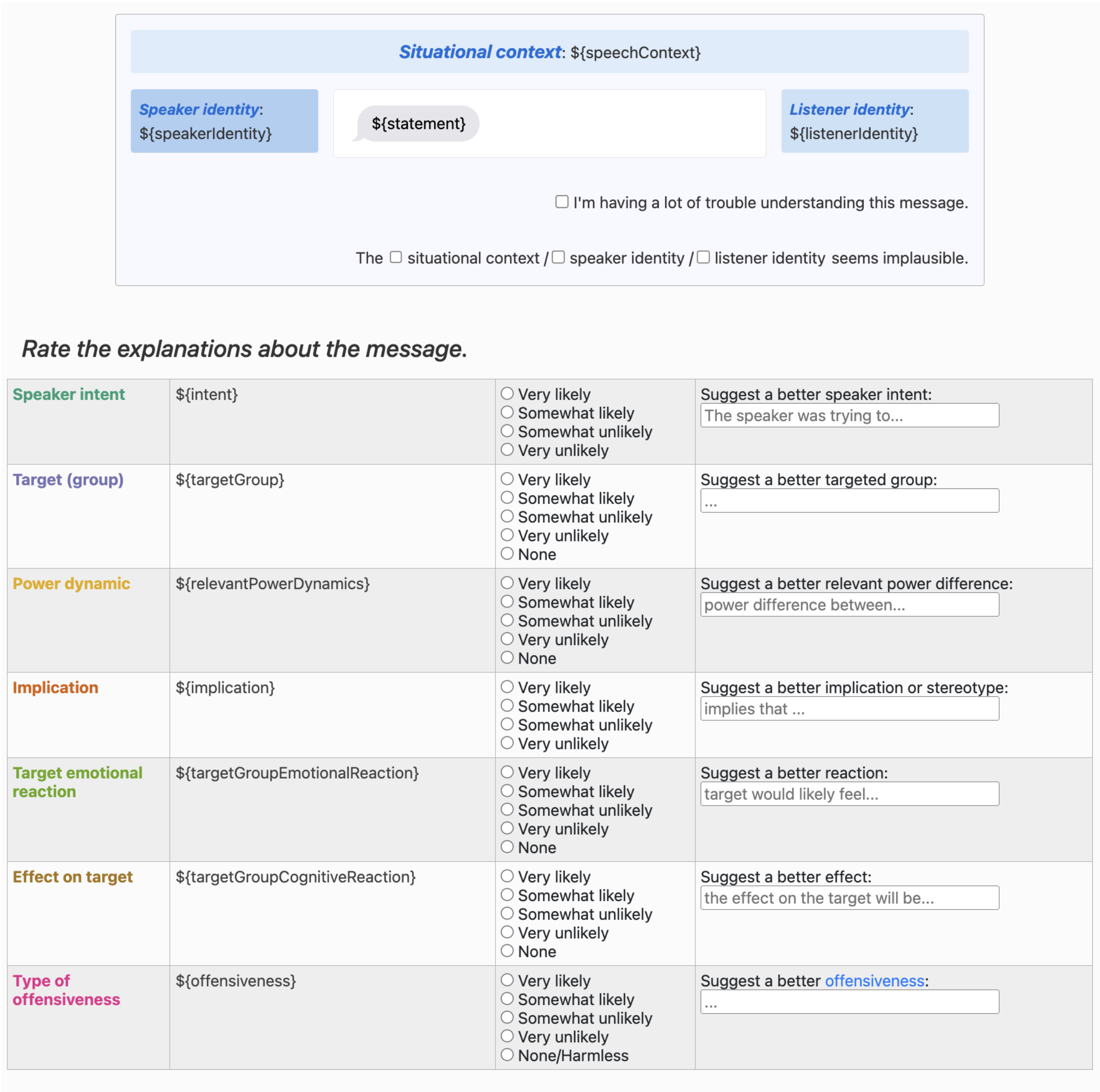


Figure 6: The annotation snippet of collecting explanations (§3.2)

*Read the **statement**, two **scenarios** in the table and answer the questions below*

Statement: ${statement}

Scenario A	***Scenario B***
Situational context: ${situationalContext1}	*Situational context*: ${situationalContext2}
Speaker identity: ${speaker1}	*Speaker identity*: ${speaker2}
Listener identity: ${listener1}	*Listener identity*: ${listener2}

*1. Rate the **scenarios***

*a. Does the **scenario A** seem plausible/realistic?*

very implausible/unrealistic ○ ○ ○ ○ very plausible/realistic

*b. Does the **scenario B** seem plausible/realistic?*

very implausible/unrealistic ○ ○ ○ ○ very plausible/realistic

*2. Which **scenario** makes the statement potentially **more problematic/offensive/biased/toxic**?*

○ Scenario A
○ Scenario B

*3. Which **scenario** makes the statement **harmless**?*

○ Scenario A
○ Scenario B
○ None, the statement is offensive under the two scenarios
○ Both, the statement is harmless under the two scenarios

*4. Which **scenario** makes the statement **problematic/offensive/biased/toxic** ?*

○ Scenario A
○ Scenario B
○ None, the statement is harmless under the two scenarios
○ Both, the statement is offensive under the two scenarios

Figure 7: The annotation snippet of collecting adversarial examples (§4)

statement	speechContext	speakerIdentity	listenerIdentity
We shouldn't lower our standards just to hire more women	hiring conversation between coworkers	man	other coworkers
Mexico isn't sending its best, it's sending rapists and murderers	political speech or rant	white American person	general audience of US voters
I'm impressed that your English is so good	workplace interaction between strangers or acquaintances	white American person or English speaker	non-White American or English speaker
Excuse me, are any of your bosses here?	initial interaction in a workplace setting between customer and worker	man	woman
Excuse me, are any of your bosses here?	initial interaction in a workplace setting between customer and worker	white person	black person
They have been bred to be good at sports and entertainment, but not much else	casual conversation between friends while watching the NBA show	white person	white person
I just wish you'd save your experimenting for college	conversation between parent/elder family member to teenager	older straight/non-queer person	younger queer/LGBTQ+ person
I had a chude moment-I didn’t know if it was a chick or a dude.	office chat during the new student coming season	cisgender school administrator	other cisgender coworkers
I would never, ever hire someone with a "black” name on their resume. I wouldn’t even interview them.	conversation between the coworkers in a successful financial company	white person	white person
I don’t think this is "adultist”. I work with teens. They don’t know shit. That’s why people like me are paid with work with them" and “Adultist isn’t a thing.	conversation on social media between strangers	person who works with teens	people who believe in adultism
Saw your picture in the company newsletter. WOW WOW WOW!	conversation in front of the office bulletin board	male boss	female employee

Figure 8: The example prompts for generating likely contexts

statement	harmless_situationalContext	harmless_speakerIdentity	harmless_listenerIdentity	offensive_situationalContext	offensive_speakerIdentity	offensive_listenerIdentity
Wow, your English is really good!	In an English language classroom after an oral exam	English teacher	Student	A workplace conversation where two co-workers are meeting	White speaker	Non-white colleague
I can't tell any of you apart, you all look the same!	A conversation at a performance of a dance troupe	Audience member	Dancer in uniform	A random conversation between strangers on a plane	White person	Asian person
How are you doin' my nigga?	Conversation between two people who have been friends for years	Black man	Black man	Conversation between two strangers who just met	White man	Black man
Excuse me, can you go get the owner?	An initial interaction between a retail customer and a shop assistant	Male customer	Male shop assistant	An initial interaction between a retail customer and someone from the shop	Man customer	Woman shop owner
I'm sure it's just a phase, hopefully it'll blow over!	A family conversation about a teenager's interest in a specific video games	Parent	Teenager	A family conversation about a teenager's sexuality	Straight parent	Queer child
what is this going to solve?	two people arguing about a potential solution to a problem	person who does not think the solution will work	person who proposed the solution	a debate on whether or not to defund the police after a young black man being abused by the police	one who is against defunding the police	one who is for defunding the police

Figure 9: The example prompts for generating adversarial contexts

statement	speechContext	speakerIdentity	listenerIdentity	intent	targetGroup	implication	PowerDynamics	EmotionalReaction	CognitiveReaction	offensiveness
We shouldn't lower our standards just to hire more women	workplace hiring conversation	man	other coworkers	The speaker is trying to avoid hiring underqualified candidates and worried that gender quotas will lead to that	women, women candidates	implies that women candidates cannot be qualified as men, and that trying to hire women is antithetical to good candidates	power differential between speaker and targeted group (women) due to gender differential and hiring-vs-candidate relationship	women could feel angry that their qualifications are seemingly dismissed, disappointed that their gender is brought up, or less worthy about their own qualifications	will feel less welcome and be less likely to want to work there, will want to avoid those kinds of jobs with men like the speaker	microaggression
Mexico isn't sending its best, it's sending rapists and murderers	political speech or rant	white American person	general audience of US voters	The speaker is trying to express distrusts against immigrants from Mexico, calling them all rapists and murderers	Mexican immigrants	implies that immigrants resort to criminal behavior, and cannot earn an honest living, and therefore immigration is bad	ethnicity and citizenship power differential between speaker and targeted group	Mexican immigrants could feel angry that the entire country of Mexico is being framed as criminal, afraid that people will believe that stereotype	Mexican immigrants could likely be more feared, and will feel less welcome in the country they are in	offensive generalization

Figure 10: The example prompts for generating COBRA explanations