VU VRIJE UNIVERSITEIT AMSTERDAM

VU Research Portal

ToxiREX

Schouten, Stefan F.; Markov, Ilia; Vossen, Piek

2026

Link to publication in VU Research Portal

citation for published version (APA)
Schouten, S. F., Markov, I., & Vossen, P. (2026). *ToxiREX: A Dataset on Toxic REasoning in ConteXt*.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
vuresearchportal.ub@vu.nl

Download date: 16. Nov. 2026

arXiv:2606.27981v1 [cs.CL] 26 Jun 2026

ARTICLE

ToxiREX: A Dataset on Toxic REasoning in ConteXt

Stefan F. Schouten, Ilia Markov, Piek Vossen

Vrije Universiteit Amsterdam, the Netherlands
Corresponding author: Stefan F. Schouten; Email: s.f.schouten@vu.nl

Abstract

We introduce a new, contextual, multilingual dataset called ToxiREX: **Toxi**c **RE**asoning in Conte**X**t. The dataset consists of threads of Reddit comments and structured characterizations of what the comments imply, following a systematic toxic reasoning schema developed in a previous paper (Schouten et al. 2026). Using the schema allows us to capture and explain implicit and context-dependent toxicity, while supporting mappings to existing toxicity taxonomies. The dataset includes comments in six languages (English, Arabic, Turkish, Spanish, German, and Dutch), collected from posts connected to specific major events (e.g. the 2023 Turkey earthquakes; the Russian invasion of Ukraine). We describe the context-preserving preprocessing of the threads. We create a training set of 125 thousand comments which is annotated by a commercially available LLM, and a test set of just under three thousand comments that is annotated by native speakers. We show that apparent disagreements in the test set annotations often reflect defensible alternative interpretations rather than noise. Finally, we provide baseline results by prompting and fine-tuning language models. To produce these results, we develop evaluation strategies for our hierarchical, schema-based predictions. While models perform better than random, there remains a lot of room for improvement, showing the task to be challenging. ToxiREX is the first dataset to simultaneously incorporate multiple languages, conversational context, and implicit toxicity, while using the toxic reasoning schema for rich, structured annotations. Dataset available at: `https://github.com/cltl/toxirex`

Warning: This paper contains examples of toxic language.

Keywords: Language Resources, Toxic Language, Multilinguality, Reasoning, Language Models

1. Introduction

Toxic language, defined here broadly as any language that conveys hateful, derogatory, or offensive ideas (Garg et al. 2023), is widely regarded as undesirable and must be identified and addressed to foster healthier online environments (Fortuna and Nunes 2018). While explicit expressions of toxicity are often easier to recognize, toxicity conveyed implicitly (Waseem et al. 2017) presents a greater challenge (Caselli et al. 2020). Determining whether speech is implicitly toxic can be subjective, as individuals may interpret the same text in different ways. Moreover, understanding the author's intended implications may depend on access to the relevant context, including cultural, situational, and discourse context.

The widespread adoption of social media has significantly amplified the need to automate the detection of toxic language. Such language can inflict harm on individuals and communities, reinforce harmful stereotypes, and can contribute to violence (Madriaza et al. 2025). As a result, identifying and mitigating toxic content is essential for maintaining inclusive and healthy online environments.

In this work, we introduce ToxiREX, a dataset of social media comments taken from Reddit. These comments have been annotated for toxic implications with the original conversational context present. For the annotation, we adopt the *toxic reasoning schema* (Schouten et al. 2026). This

schema systematizes categories of (implicit) toxic language by breaking down implications into independent traits. In doing so, we separate the propositional content of an implication from the attitudes held toward it by various stakeholders. For three traits in the schema, we also produce span-level annotations to identify the relevant parts of a comment. We use GPT4o to produce silver labels for each comment in the training data, and organized an annotation campaign to create the test set. To establish its quality, we also performed an analysis of apparent disagreements in the annotations of the test set.

The schema supports mapping to and from existing taxonomies of toxic language. This allows us to envision at least two strategies: (1) a bottom-up approach, where data is annotated according to the schema directly and then mapped onto categories of toxic language, or (2) a top-down approach, where data is first annotated according to conventional taxonomies, and then mapped onto the schema. These methods are complementary and both come with their own up- and downsides. The top-down approach is straightforward to execute, and annotation campaigns remain relatively simple. However, the bottom-up approach, while labor-intensive, allows for an open-ended exploration where annotations cover all kinds of potentially toxic content, even that which may not be covered by existing taxonomies.

We adopt the bottom-up approach to develop the ToxiREX dataset, with the goal of investigating: (1) the feasibility of directly annotating comments according to the toxic reasoning schema; (2) if the toxic reasoning schema is expressive enough to cover most potentially toxic implications; and finally, (3) whether existing approaches, like prompting and fine-tuning, can be used to detect and categorize implications according to the toxic reasoning schema.

We find that annotating according to the toxic reasoning is feasible, albeit with some unique challenges. Because of the hierarchical nature of the schema, the annotations are more sparse, making it harder to calculate to what degree annotators agreed. Similarly, this also complicates evaluating model predictions. The findings of the annotation campaign suggest the toxic reasoning schema is able to characterize a wide variety of potentially toxic implications expressed in many different ways in different languages.

Finally, our experiments show that detecting and characterizing the implications of potentially toxic comments is a very challenging task. While LLMs perform reasonably well when evaluated in a zero-shot setting, there remains substantial room for improvement. Fine-tuning a smaller language model on LLM predictions retains much of the LLM's performance at a much smaller computational budget.

2. Related Work

Toxic language detection has been studied extensively in the last decade (Jahan and Oussalah 2023; Fortuna and Nunes 2018). Here we compare and contrast our contribution to previous work along various dimensions.

Fine-grained classification. Toxic language detection is often approached as a binary classification problem, but others have broken down the problem by introducing fine-grained categorizations. For example, Vidgen et al. (2021), ElSherief et al. (2021), Kirk et al. (2023), and Pachinger et al. (2025) all introduce (multi-level) taxonomies for variants of toxic language detection. We also break down the problem, but rather than a hierarchy, we use the toxic reasoning schema, which has many independent traits. Simultaneously, specific combinations of these traits can still be mapped onto existing categories of toxic language, when desired.

Span-level and Structured annotation. Various datasets have introduced span-level annotations. Some providing general 'toxic spans', which indicate a particular span has a higher degree of responsibility towards the toxicity than the rest of the comment (Pavlopoulos et al. 2021; Mathew et al. 2021). Others include spans that indicate more specific aspects, such as the target of the

toxicity (Barbarestani et al. 2022; Jafari et al. 2024). Our work similarly includes spans that mark the target, and optionally spans marking a group or individual to which the target is compared.

Previous work has pointed out that human moderators use moderation guidelines to decide if posts are problematic (Calabrese et al. 2022). They argue that many cases depend on the exact rules/norms that should be applied. Motivated by this they use intent classification and slot filling (ICSF), which requires parts of the text (spans) beyond merely the target to be assigned to sets of possible slots. A similar work is that of Pachinger et al. (2025), which similarly annotates multiple kinds of spans. Our work also addresses the dependency of toxic language detection on guidelines, and similarly uses structured multiple-span annotation. But, instead of designing slots for a specific guideline, we use the toxic reasoning schema, which is intended to generalize across guidelines.

Multilinguality. Many existing toxic language datasets are monolingual, with most of the resources being in English. However, previous work has pointed out the importance of having resources that cover multiple languages (Basile et al. 2019). Existing multilingual resources include Ousidhoum et al. (2019) and Yadav et al. (2022). Our work incorporates six languages: Arabic, Dutch, English, German, Spanish, and Turkish. Our dataset further distinguishes itself by connecting comments and posts to major world event across different language-cultural groups. Because of this connection, the data uniquely enables the investigation of variation in toxic language reasoning and expression across these communities, as well as the capacity of LLMs to reflect these.

Implicit Toxicity. In the literature, a distinction is often made between explicit and implicit toxicity. The most explicit forms of toxicity involve the use of words which are almost always toxic. Implicit toxicity often comes in two degrees: (1) people plainly asserting something harmful or offensive, but without using toxic words; and (2) people stating one thing while implying another. Recent work primarily targeting the latter forms of implicit toxicity has done so by including free-text annotations of implied statements (Sap et al. 2020; ElSherief et al. 2021). We also include free-text annotations of implied statements, but further characterize the implications using the traits of the toxic reasoning schema.

In-Context Annotation. Many existing datasets of toxic language include social media posts on their own, with no context (e.g. Basile et al. 2019; Pavlopoulos et al. 2021). However, often posts are part of conversations, and cannot be understood independently. Previous work has firmly established that conversational context is important to understand if and why a post is toxic (Pavlopoulos et al. 2020; Yu et al. 2022; Xenos et al. 2022; Ljubešić et al. 2023). The comments in our dataset are all annotated within their original conversational context i.e. the Reddit thread they were posted in.

While there are existing datasets that have one or some of the features mentioned above, to the best of our knowledge, ToxiREX is the first to combine all of them.

3. Toxic Reasoning

In Schouten et al. (2026), we defined toxic reasoning as "toxic language detection that further requires making explicit the reasons for why something should be considered toxic" and proposed that it may be approached with the toxic reasoning schema. In this schema, what is conveyed by a toxic message is characterized in two parts: (1) a proposition predicating something of a subject, and (2) the author (implicitly) reporting an attitude (of belief, desire, or intent) towards this proposition. The subject will often be the target of the toxicity (when expressing negativity) , but can also be the author's in-group (when expressing exclusionary or discriminatory positivity).

Besides the author's attitudes, we also include the attitudes of the general public and experts[a] toward the proposition, which allows us to further differentiate between specific kinds of toxicity. The schema is used along with mappings from the schema to existing taxonomies, like that of the Implicit Hate Corpus (IHC, ElSherief et al. 2021). Such mappings allow for schema-based annotations and predictions to be converted back to (fine-grained) classifications.

3.1 Toxic Reasoning Schema

Here we will briefly summarize the toxic reasoning schema. The schema's first trait categorizes what is at the core of a message's implication: the message content characterizes the proposition that is central to what the text conveys. The schema distinguishes between three high-level **categories** of content.

- *Situation*(`subject`, [`other`]): a situation (e.g. environment, condition, circumstance, etc.) applies to the `subject`.
- *Quality*(`subject`, [`other`]): the `subject` possesses a given inherent quality, or has a certain nature.
- *Behavior*(`subject`, [`other`]): the `subject` behaves in a particular way.

Each can occur with *Negative* or *Positive* **polarity**, and can optionally apply relative to a an 'other' (group or individual).

A number of other traits are included in the schema. For example, IHC's 'Threatening and Intimidation' is mapped onto the schema as follows. ElSherief et al. (2021) state that this class "convey[s] a speaker commitment to a target's pain, injury, damage, loss, or violation of rights.", explicitly choosing a wider definition that does not focus solely on threatening 'life or limb'. Thus, we say that the class applies if and only if it expresses an intention to subject (*future*) a person or group of people to a *Negative Situation*. The *future*-oriented nature of threats is covered by **temporality** trait.

Now consider 'Inferiority language', which can, for example, include instances where someone states or implies that a *group* has a *Positive Quality* or *Behavior* (e.g. "Japanese people have higher IQ", or "white people invented civilization"). In this case, in order for it to be considered toxic, we also require that an expert has an attitude of disbelief towards the message's content. The fact that these instances necessarily involve a group, is covered by the **specificity** trait.

The schema includes a few other traits, including the **group-type**, and the **subject-role / other-role**, which specify the role of subject in the conversation, i.e. is the subject or other: the author, another conversation participant, someone else, or a group that includes them. For a complete description of the schema, see Schouten et al. (2026).

Finally, the schema supports identifying spans that clarify which part of a message is responsible for each of these traits. In ToxiREX, we include spans highlighting the `subject`, the `other`, and the content **category**.

4. Data Selection & Pre-processing

In this section, we give details on the following: (1) the source data, (2) context-preserving pre-processing of the data; and, (3) selection of a subset which contains sufficient and diverse toxicity. See Figure 2 for an overview.

[a] A person with the expertise to assess the truth of the relevant proposition.

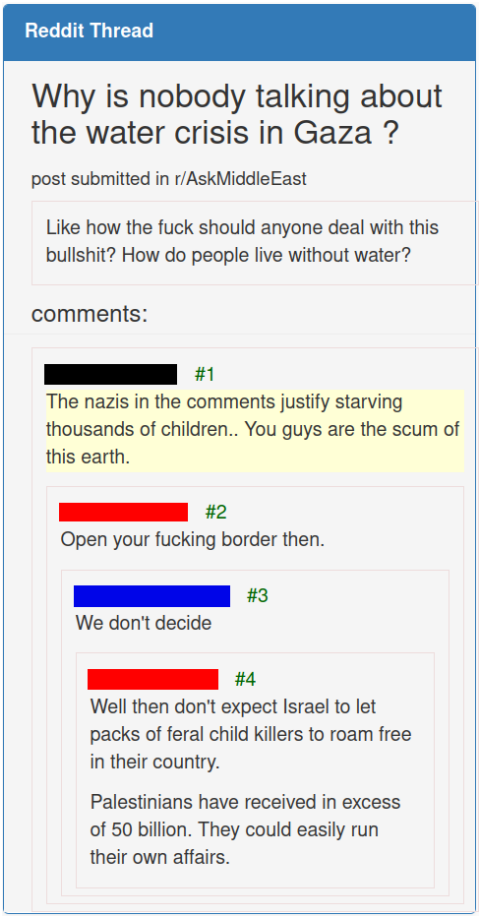

Figure 1: Thread of comments as presented to annotators. User names have been redacted by replacing each name with a colored box.

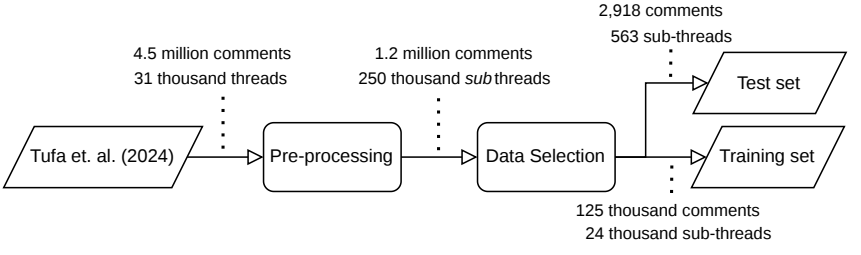

Figure 2: Flow chart showing the data processing steps.

4.1 Source Data

Rather than collect data from scratch, we use the Reddit data scraped by Tufa et al. (2024), but with our own pre-processing. This data was selected for its relation to one of fifteen major real-world events spanning four years from 2020 to 2023. Events cover a wide range of topics, for example: the January 6th U.S. capitol attack; the 2023 Turkey earthquakes; the Russian invasion of Ukraine, etc. (See Appendix E for the full list.) The raw data consists of just under 4.5 million *comments*

from 31 thousand *posts* in six languages: English (EN), Dutch (NL), Arabic (AR), Turkish (TR), Spanish (ES), and German (DE).

On Reddit, comments can be left directly in response to posts, or comments can reply to another comment, forming a tree structure. An example of some comments can be seen in Figure 1.

4.2 Pre-processing

Our data pre-processing is designed to allow studying toxicity in context, while maintaining annotation feasibility. While it would be optimal to annotate all comments in their full context, in practice asking annotators to read up to dozens of comments is not practical. So, we need to select data where the context is of a manageable size. A naive approach could be to select only comments up to a certain depth, but that would involve throwing away lots of potentially interesting data. Thus, we make a compromise by splitting up the comments below a post into what we call *sub-threads*: sequences of comments where each comment is a reply to the previous.

The sub-thread creation process for a given post proceeds as follows:

(1) Each comment is associated with its 'ancestry', i.e. its parent,[b] its parent's parent, and so on until the top-level.
(2) The candidate sub-threads consist of every leaf comment's ancestry.[c]
(3) Leaf comments whose ancestry is shorter than the minimum sub-thread length of three comments are dropped from consideration.
(4) From among the remaining candidate sub-threads, we now look for a candidate C to be confirmed as a sub-thread. This selection is done by heuristic. We look for the candidate that minimizes the following formula: $|L_C - L_{\text{ideal}}| + B_C$, where L_C is the candidate's length, $L_{\text{ideal}} = 5$ is the ideal length we choose, and B_C is the number of branching points in C, i.e. the number of comments that have more than one reply. The goal is to prioritize candidates close to the ideal length, that also leave the tree as intact as possible.
(5) The confirmed sub-thread's comments are now removed from further consideration and removed from the ancestries of the remaining comments. We now return to step 2, and the process is repeated until no candidates remain.

Figure 3 depicts a small example comment hierarchy that demonstrates the motivation behind the heuristic.

4.3 Data Selection

We use an LLM to select data in two steps.

Step 1. We suspect that the prevalence of toxic language in the full set of sub-threads is fairly low. Therefore, we use an LLM to clasify sub-threads by their toxicity, and GPT4o finds around 10% of the sub-threads to be toxic. So, rather than annotate everything, we use this subset for the dataset. Knowing that LLMs tend to be biased toward classifying text as toxic (Kumar et al. 2024), we assume that doing so will keep most toxicity, including implicit toxicity, while removing sub-threads that are clearly non-toxic. The prompt we used can be seen in Prompt 1.

Step 2. The goal of the human test set annotation is validate the expressivity of the toxic reasoning schema, and to produce a high-quality test set. For this purpose, we first perform another selection, this time to create a stratified sample. Specifically, we ask an LLM to categorize

[b] A comment's parent is the comment it replies to.
[c] A leaf comment is a comment without replies.

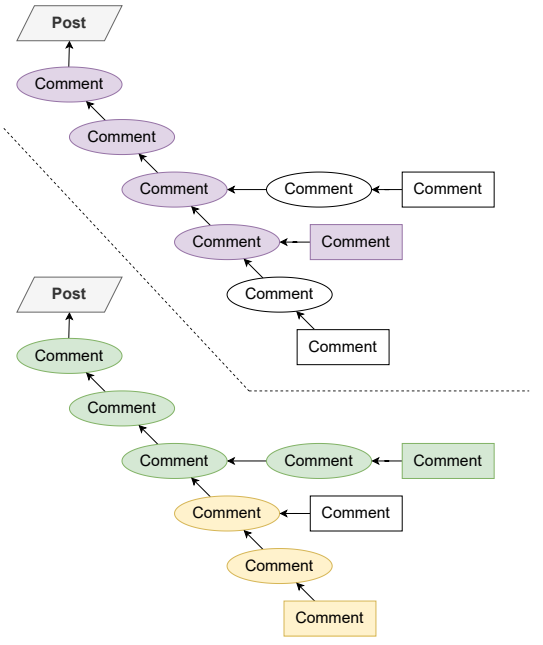

Figure 3: Example of sub-thread creation from a post's comment hierarchy. On the top, it shows an example of creating a sub-thread (purple) with two branching points, it leaves four unusable comments. On the bottom, it shows how selecting the candidate with fewer branching points (green), leaves another viable sub-thread (yellow).

```
Read the title: {title}
Read the post text: {post_text}
Read the full conversation: {context}
Note that the name of the author of each comment in the conversation is written at the start of the comment in brackets []
Read the comment: {comment}
Note that this comment is one of the comments in the conversation.
Given all provided information, is the comment an instance of hate speech? Say True if it is an instance of hate speech. Say False if it is not an instance of hate speech.
```

Prompt 1: Used for selecting comments likely to be toxic.

comments into a few high-level categories: Harmed, Inferior, Negative, NonHuman, Stereotype, Superior.[d]

See Table 1 for the statistics of how often each category was assigned to the data by the LLM, and exactly how many we sampled from each. Note that we did not sample the exact same number of comments from each combination of category and language. This is because categories apply on the comment level, but we sample entire sub-threads, which also contain comments of the other categories. This makes an exact uniform sampling over the categories difficult to realize.

The data which was not selected for human annotation was annotated by OpenAI's GPT4o and used as training data in our experiments.

[d]These categories are based on an early version of the toxic reasoning schema (TRS0). See Appendix D and F for details.

Table 1. : Number of sub-threads per language and TRS0 category in pre-annotation selection. Each cell shows nr. of sub-threads in test set out of total nr. of sub-threads.

	English	Spanish	German	Turkish	Dutch	Arabic
Harmed	20 / 5337	23 / 804	34 / 449	23 / 298	25 / 210	35 / 130
NotHuman	24 / 2185	31 / 444	22 / 207	21 / 147	21 / 101	30 / 81
Superior	24 / 1129	33 / 245	36 / 111	21 / 68	7 / 50	26 / 46
Inferior	25 / 3230	12 / 875	19 / 293	19 / 237	14 / 131	34 / 117
KnownStereotype	24 / 3018	27 / 641	28 / 340	35 / 91	29 / 138	18 / 32
SomethingNegative	14 / 10863	21 / 2120	27 / 934	24 / 720	23 / 480	44 / 290
SomethingPositive	7 / 1514	6 / 326	7 / 178	19 / 146	6 / 81	20 / 93
Nr. Sub-threads	102 / 17935	95 / 3431	98 / 1518	97 / 1026	79 / 691	92 / 388
Nr. Comments	499 / 92139	504 / 17600	497 / 7635	503 / 5352	419 / 3668	496 / 2106

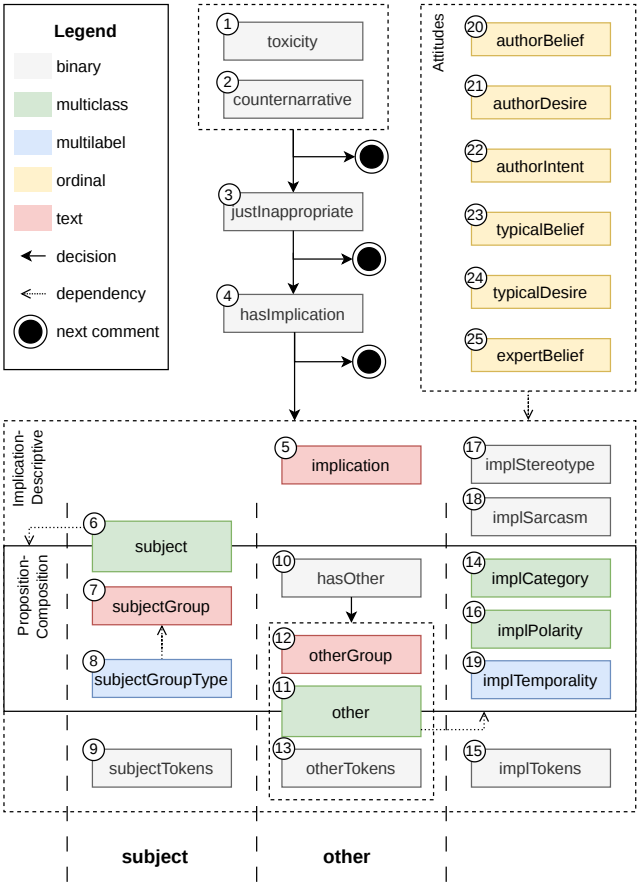

Figure 4: Schematic overview of variables in the annotation. The numbers in the top-left corners of each box indicate the order in which the variables were asked about. Arrows show decisions made during the annotation (solid) and dependencies between variables (dotted).

5. Test Data Annotation

Here, we will present the results of annotating data according to the toxic reasoning schema.

Annotators were shown a reproduction of the data as they could have seen it on Reddit (see Figure 1). This interface was created using Lingoturk.[e]

First, they would be shown the post text, if available.[f] Next, they would be shown a selection of comments left by Reddit users in response to that post. Specifically, they would be shown one sub-thread, constructed as described in 4.2. One of the comments in that sub-thread would be highlighted, indicating the comment that is to be annotated. Next, the annotator was shown a number of questions about the comment that was to be annotated.

5.1 Annotation Details

Variables. Each question asked of the annotator corresponds to a variable in Figure 4. The purpose of the first few variables is to prevent annotators from having to do a full annotation for comments that are either clearly and uncontroversially non-toxic, or toxic but merely 'inappropriate'. The answer given to these first questions determines whether or not to fully annotate the comment. This is indicated in Figure 4 with a solid black arrow.

The binary `toxicity` variable should equal 'Yes' if the annotator believes the "comment might be (perceived as) potentially toxic, polarizing, or otherwise negative toward a group or individual". The `counternarrative` variable should equal 'Yes' if the comment is "an argument-based reply aimed at the previous comment, providing an alternative perspective".[g] If the answer is 'No' for both of the above, the annotation proceeds with the next comment.

The `justInappropriate` variable must be 'Yes' if the comment is "only toxic because it uses inappropriate words, e.g. use of slurs, name-calling, etc.", with no toxic implication whatsoever. If the answer is 'Yes', the annotation proceeds with the next comment.

Finally, the `hasImplication` field asks annotators to confirm whether they have identified a toxic implication. If, despite making it past the previous variables, the annotator cannot identify what toxicity the comment is communicating (explicitly or implicitly), the annotator can answer 'No' and the annotation proceeds with the next comment.

We added `implication`, which is a textual description of the main implication. Annotators are asked to identify what they believe to be the main toxic implication of the comment. The rest of the questions are asked with respect to whatever implication is chosen by the annotator. The primary aim of including the `implication` is to encourage the annotator to actually specify an implication before proceeding. For the rest of the annotation, the `implication` remained at the top of the screen for reference.

The rest of the variables are based on traits of the toxic reasoning schema. Typically, the correspondence is one-to-one. The only exceptions are the `subject` and `other` variables, which include aspects of both the *role* and *specificity* traits. See Appendix A for an overview of all questions.

In addition to those variables, we also added the `subjectTokens`, `otherTokens` and `implTokens`, which asked annotators to indicate which tokens in the comment (or its context) corresponded to the `subject`, `other` and `implTopic`, respectively.

Summary sentence. Another part of the annotation tool was the *summary sentence*. This is a sentence generated with a simple template from the labels chosen by the annotator. Looking at Figure 4, this sentence was shown after the variables that are part of the box labeled 'Implication', but before the 'Attitudes'. The summary sentence uses the variables under 'Proposition-Composition' to fill out a sentence template that creates an English sentence.

---

[e] `https://github.com/FlorianPusse/Lingoturk/`

[f] Posts on Reddit can also be what are known as 'Link Posts', which provide a topic of conversation purely by linking to another webpage, rather than through a written post.

[g] Definition in line with previous work (Carthy et al. 2020).

For example, for the following comment from Yu et al. (2022):

> That chick$_{\text{SUBJ}}$ needs a high-five in the face with a chair. Damn her$_{\text{SUBJ}}$ for making us look bad!

For an annotation of: `Harm`, temporality of *future*, specificity *personal*, and subject role *other*; the summary sentence would be: "An individual ['That', 'chick', 'her'] will have harm come to them."

These sentences were presented to the annotators, who were asked to annotate the attitudes with respect to this sentence. For example, for the *author_desire* trait, we asked: "Based on the comment, to what degree does the author seem to wish (or prefer) that: SUMMARY_SENTENCE?". See Appendix B for pseudocode for how such sentences can be generated.

Annotators. For each language, three annotators who spoke the language fluently or natively completed the annotation for each comment. The English and Dutch comments were annotated by experts, while for the other four languages, students were recruited to perform the annotation. The expert annotators did various small trial annotations of five threads each, which were also used to refine the software with which the annotation was performed.

Two batches of five English threads were selected to train the student annotators in two stages. Each stage was followed by a meeting where feedback was presented and opportunity provided for the students to ask questions. At every point in the process, the students were given the opportunity to stop annotating if the exposure to toxic content was too much.

To get an indication of the biases annotators may have, we asked them to share some information about themselves. Following the recommendations of Vidgen and Derczynski (2020), we report demographics, as well as information on (level of) expertise and personal experiences with toxic language. This information can be seen in Appendix C, Table 14 and 15.

5.2 Example of Data and Annotation

To get an idea of what kinds of comments make up ToxiREX, we present a full thread in Table 2. We can see that the sub-thread contains implications, counterspeech, and clearly non-toxic comments. The thread consists of comments that were left on a post titled: "Ukrainians were among tens of thousands of LGBTQ activists who took part in the Polish capital's Pride parade Saturday, in a country hosting tens of thousands of Ukrainians who have fled the Russian invasion". It features two users discussing whether refugees should "participate in a protest while they're in a country that isnt theirs".

The discussion starts with user A asking if Ukrainian refugees are engaging in bad behavior by participating. User B challenges this, and uses a (somewhat extreme) hypothetical scenario to demonstrate their point. After some back and forth, the exchange ends with both users accusing each other of bad behavior, such as "justifying rape", and being "homophobic".

Our dataset contains many examples of such political conversations. They often include users making controversial claims that could be interpreted as toxic, and then devolve into creative name calling. We believe this type of data is underexplored compared to more obviously harmful text.

5.3 Initial Findings

Statistics. Sub-thread lengths, the number of potentially toxic comments, and the frequencies of implication properties and polarities can be seen in Table 3. The number of comments considered potentially toxic is close to or over half, depending on the language. This is higher than levels of toxicity reported for other datasets (Vidgen and Derczynski 2020). There are two reasons for this. First, we specifically ask annotators if a comment *might* be considered toxic. And second, we used an LLM to pre-select sub-threads based on whether it contained any toxicity.

The testset contains diverse implications, although situational implications are underrepresented compared to qualitative and behavioral implications. Problematic annotations for the

Table 2. : A full (sub-)thread from ToxiREX.

(a) Comments and consensus annotation (based on value for first 4 variables)

#	comment	annotation
A1.	Why do the Ukrainians think it's okay to participate in a protest while they're in a country that isnt theirs? If they have Polish citizenship its one thing but if they're only there while the war is going on why should they be protesting anything?	Impl.
B1.	Maybe they think human rights are for all humans	Countersp.
A2.	Why does a foreigner have a right to tell a country that is not his own how it should be run, as long as it isn't a threat to his own?	Impl.
B2.	So if you let me into your house and I saw you raping your sister I should just keep my mouth shut because it's your house and your sister?	Countersp.
A3.	Ah yes, stopping what is universally recognized as a crime is imperialism	Non-toxic
B3.	Don't intentionally miss the point. Should I keep my mouth shut because it's your house or should I stand against something that is wrong?	Countersp.
A4.	I answered you. One situation is an obvious universally accepted crime against humanity. The other is cultural imperialism	Impl.
B4.	Hate to break it to ya bud but rape isn't universally accepted as a crime against humanity. In fact, there are instances of prepubescent girls being raped by much, much older men only to then have those girls murdered by way of honor killing. Still want to sit on your hands and cry *cultural imperialism*?	Countersp.
A5.	You seem very hellbent on justifying rape and incest. But I'll give it to you, sometimes you are justified in stopping practices that are accepted in another culture. But this isn't one of those times. Gay marriage is hotly contested in many countries, and it isn't the place of a foreigner to try and influence the decisions another country makes on some issues	Impl.
B5.	Sounds like you actually don't give a shit about human rights or maybe you're just homophobic and are trying to weasel out of coming across as such. This will be my last response because you're just not someone worth interacting with.	Impl.

implication category (`implCategory`) and polarity (`implPolarity`) variables are rare. These include implications being labeled under the 'NonSpecific' category, as well as those missing a polarity. From these results, we can conclude that the implication properties included in the schema cover a large majority of the potentially toxic implications present in our data.

Inter-annotator Agreement. The agreement, calculated independently for the categorical variables, can be seen in Table 4. Going by Landis and Koch (1977), agreement is fair to moderate (between 0.2 and 0.6) for the `toxicity` and `justInappropriate` label. Agreement for the variables related to the implication are often lower (ranging from -0.3 to 0.5). However, given the hierarchical nature of the annotation task, disagreements are not immediately indicative of a

(b) Comment implications. Square brackets contain `implTokens` and, if an indivual is targeted, `subjTokens`.

#	implication	author attitude
A1.	Ukrainian refugees in Poland have/will have Negative behavior ["participate in a protest while they're in a country that isnt theirs"] compared to Polish citizens.	*belief=yes* *desire=no*
A2.	"	"
A4.	Ukrainians have Negative behavior ["cultural imperialism"].	*belief=yes* *desire=no*
A5.	An individual ["You", "you"] has Negative behavior ["justifying rape and incest"].	*belief=yes*
B5.	An individual ["you", "you're"] has Negative behavior ["don't give a shit about human rights", "homophobic"].	*belief=yes*

problem. As mentioned previously, annotators were instructed to annotate with respect to what they believed to be the main implication of a comment. Because annotators could have different implications in mind, a degree of disagreement is expected for these variables.

Furthermore, different languages were annotated by different annotators, making agreement vary by language. While Arabic, German, Spanish and Turkish were annotated by students, the agreement does not appear to be systematically lower than for English and Dutch, which were annotated by experts. For some languages the annotators came from more diverse backgrounds, for example for Arabic, annotators came from Morocco, Egypt, and Jordan, this could be another contributor to variation in agreement.

It is also likely that agreement would be higher if measured by sub-thread, since different annotators might differ on which comments they believe to be the toxic ones, while agreeing on the presence and type of toxicity.

5.4 Analysis of Annotator Disagreements

Given that not all disagreements are problematic in our case, we seek to better understand what is behind the disagreements in our annotations. Thus, we set out to categorize each disagreement by what we believe to be the source of the disagreement. The disagreements were categorized as:

- **error**, indicating the disagreement is due to an annotation error by one or more annotators;
- **orthogonal**, signifying that there is a disagreement in name only, because the other answers are also correct;
- **conditional**, meaning the disagreement is a direct consequence of an upstream disagreement, not a disagreement in and of itself;
- **subjective**, for disagreements that results from a difference in perspective; or,
- **disjunct**, indicating we were unable to reduce the disagreement to one of the above cases.

Because of the labor-intensive nature of this analysis, it was done only for a limited subset of the annotation. First, it was done only for the English part of the data. Second, we restricted the analysis to the subset of comments for which annotators agreed on the first four annotation variables. Last, we only categorized disagreements for those variables essential for capturing what is implied by a comment. These variables are those (partially) inside the box labeled 'Proposition-Composition' in Figure 4. This leaves 47 out of the 499 English comments in the pre-selection, coming from 38 out of 102 sub-threads.

Table 3. : Test set statistics, including label statistics for the human annotations. The first three rows give sub-thread lengths. The 'Nr. of non-toxic' vs. 'possibly toxic' are based on a majority vote. The rest of the rows show the raw number of annotations which has that category or polarity.

		AR	DE	EN	ES	NL	TR
	Avg. sub-thread length	5.50	5.22	5.10	5.44	5.30	5.23
	Min. sub-thread length	4	4	4	4	4	4
	Max. sub-thread length	10	10	10	10	9	10
	Nr. of non-toxic	293	246	309	232	265	233
	Nr. of possibly toxic	213	260	211	263	136	217
Category	Situational	14	20	43	45	16	9
	Harm	11	11	30	29	23	
	Qualitative	54	64	65	118	24	40
	Dehumanization	8	4	3	26	3	13
	Behavioral	122	96	201	141	149	78
	NonSpecific†	7	2	9	3	3	5
	UnclearOrNone†	11	9	6	20	3	3
Polarity	Positive	10	8	22	8	12	
	Neutral	13	6	74	21	10	6
	Negative	157	160	219	275	169	115
	None†			2	3	2	
	†Nr. of problematic	18	11	17	26	8	8
Polarity + Category	Positive-Situational	1	2	5	1	5	
	Neutral-Situational	1	2	9	3		
	Negative-Situational	12	17	29	41	11	9
	Harm	11	11	30	29	23	
	Positive-Qualitative	2		5	2		
	Neutral-Qualitative	1	2	18	5		4
	Negative-Qualitative	51	62	45	110	24	37
	Dehumanization	8	4	3	26	3	13
	Positive-Behavioral	4	5	10	5	7	
	Neutral-Behavioral	6	2	38	5	5	2
	Negative-Behavioral	116	90	170	135	139	76

The first stage of this analysis consisted of the annotators going over their annotations again, to correct any errors they observed with the benefit of both hindsight and access to the other annotators' choices. Simultaneously, while checking for errors, the annotators indicated for each disagreement which of the categories given above was most applicable.

Table 4. : The agreement scores (Fleiss's Kappa), computed naively.

	AR	DE	EN	ES	NL	TR	Average
toxicity	0.42	0.32	0.35	0.53	0.32	0.60	0.42
countnernarrative	0.08	0.08	0.32	0.24	0.25	0.28	0.19
Below is reported only on cases where there was agreement on toxicity and counternarrative.							
SUPPORT	84	57	61	114	29	91	
justInappropriate	0.20	0.35	0.29	0.44	0.57	0.55	0.41
Below is reported only on cases where each annotator agreed there was some toxic implication present.							
SUPPORT	28	21	46	62	29	19	
subject	0.11	0.38	0.50	0.41	0.81	0.03	0.36
implPolarity	0.22	-0.02	0.11	0.29	-0.05	-0.02	0.14
implCategory	0.28	0.29	0.06	0.37	0.30	0.19	0.24
implSarcasm	0.15	0.35	0.13	0.15	0.46	-0.07	0.19
implStereotype	0.07	0.35	-0.15	-0.01	0.17	-0.33	0.03

In the second stage, the variable-level categorizations of each annotator are considered to decide which category applies best for each variable of each comment. This was partially done by majority vote, and partially based on discussions between the annotators.

We noticed that disagreements between individual variables often cannot be understood independently, often making it difficult to categorize the disagreement. So, in the third stage, we supplemented the analysis with proposition-level categorizations. Taking the corrected errors into account, we generated the summary sentences again. Then, for each comment, the variable-level categorizations of disagreements are considered alongside the summary sentences to come to a proposition-level categorization. When the proposition-level process gave additional insight into the cause of the disagreement, this was also used to adjust categorizations on the variable-level categorizations.

Findings. The statistics of how many times each category was chosen can be seen in Table 5. A selection of examples is given in Table 6.

When annotators categorize an apparent disagreement as orthogonal, they do not truly disagree with the other annotators. While they chose to annotate differently, they acknowledge that the other annotations are equally and simultaneously correct. The percentage of comments where the annotators chose the exact same answers before error correction ranges from 34% to 82%. When including the comments marked orthogonal and the error correction, the percentage is at least 72% for all variables. This shows that, as expected, a large portion of apparent disagreements, were not truly disagreements at all.

Looking at the remaining disagreements, most are subjective. When the annotators marked something subjective, it is usually due to a different interpretation of the comment. For these cases, the annotators do disagree with each other, but still recognize that the other annotations rest on a valid understanding of the comment and annotation schema.

Table 5. : The statistics of the categories that were found to be most descriptive of why a given disagreement between annotators occurred. Disagreement categories are in bold, aggregates are italicized. Rows where the categories are accompanied by a '+' show the frequency of that category occurring in combination with one or more of the categories that occur above it in the table. For example, the row labeled subjective+ presents frequencies of subjective occurring together with orthogonal and/or error. The first row marked '#' shows the number of comments that were not marked as 'conditional'. The statistics are computed over the comments not marked 'conditional'.

	subject	*subjectGroup*	*hasOther*	*other*	*otherGroup*	*implPolarity*	*implCategory*	*implTemporality*	*proposition*
#	47	46	47	22	31	47	47	47	47
agreement	68%	47%	36%	82%	65%	55%	38%	34%	4%
error	6%	2%	13%	18%	10%	21%	11%	4%	6%
sum	*74%*	*49%*	*49%*	*100%*	*74%*	*77%*	*49%*	*38%*	*11%*
orthogonal	11%	28%	23%	0%	3%	0%	23%	34%	17%
orthogonal+	*0%*	*2%*	*0%*	*0%*	*0%*	*2%*	*0%*	*4%*	*15%*
sum	*85%*	*79%*	*72%*	*100%*	*77%*	*79%*	*72%*	*77%*	*43%*
subjective	13%	15%	17%	0%	6%	4%	15%	19%	21%
subjective+	0%	2%	2%	0%	13%	2%	0%	0%	15%
sum	*98%*	*96%*	*91%*	*100%*	*97%*	*85%*	*87%*	*96%*	*79%*
disjunct	2%	4%	9%	0%	0%	9%	9%	4%	9%
disjunct+	0%	0%	0%	0%	3%	6%	4%	0%	13%
sum	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*

Only a small fraction of the disagreements did not fall into one of the other categories (ranging from 0% to 9%), and were thus marked disjunct. These are usually cases where the annotators fundamentally disagreed with another annotator's interpretation. Given these results, we believe the data to be of high quality.

6. Training Data Annotation

To produce silver-standard annotations for the training data, we use OpenAI's GPT4o model. The model is used with the 'Structured Outputs' feature which uses the pydantic model described below in 6.1. Sub-threads are presented to the model as shown in Prompt 2. If a sub-thread has no post text, it is instantiated with the value "EMPTY".

Table 6. : Examples of each of the types of disagreement. The answers given by annotators are represented by the *Summary Sentences*. Newline characters in the the comments were replaced by the '↩' character.

(a) Orthogonal.

Comment	**Category**	**Summary Sentences**
The rise of Islam becoming a state religion in Turkey is the reason why I think the "God is punishing them' comments are going on. Edrogan has abandoned secularism to try to turn Turkey into an Islamic state and now this earthquake. Is it God punishing them? I HIGHLY doubt it (agnostic here) but hopefully the Turkish people think it is so they don't descend into a full religious dictatorship.	orthogonal	• Muslim Turks had/have harm come to them. • An individual [President Erdogan] has negative behavior. • Erdogan government had negative behavior.
All jokes aside.. This protest showed most controversial police brutality we've seen in recent years in Belgium, with multiple claims of unlawful enforcement including but not limited to: ↩ ↩ Unprovoked use of the water cannon.↩ Unprovoked use of teargas canisters on protestors (including woman and children within the crowd) during a speech.↩ Police refusing to grand medical aid to wounded unconscious protester, who was dragged by fellow protestors towards them, refusing said protestor's passage past a barbed wire barricade.↩ ↩ Footage of this all over the place.. yet there was a serious lack of media coverage on the entire event, which was expecting 10k visitors but ended up around 35k.↩ ↩ Not looking to start an pro/anti argument, simply providing some background to the event.	orthogonal	• the Belgian police had negative behavior compared to protesters. • Belgian police has negative behavior. • Belgian police had negative behavior.

(b) Subjective & disjunct.

Comment	**Category**	**Summary Sentences**
Yes, it's not a legal case, it's a political decision. It doesn't matter if all 10k rioters say that Trump incited them. We already know he provably did and if this were a real jury, he'd be convicted. We need to put the stains on the GOP's "blue lives matter" dress force them to bleach themselves.	subjective	• far right Republicans supporting Trump had/have/will have negative behavior. • The Republican Party (GOP) has negative behavior. • Republicans had negative behavior.
But they're really bad at it. They just lie and immediately get mocked for it. It's just a terrible look for Israel. If you can't be honest how will anyone take you seriously or get people to be sympathetic to you stealing land.	subjective	• people supporting Israel including the people of Israel themselves have/will have negative behavior. • Israel has negative behavior. • Israel had negative behavior.
My argument? ↩ ↩ Antifa = proud boys, put them in the same jail cell and let them fight it out there. ↩ ↩ r/law: dOwNvOTe!	disjunct	• Antifa has harm come to them compared to proud boys. • Antifa and proud boys will have negative behavior. • Antifa and the proud boys will have harm come to them.
And we all know what the Russians are gonna do:↩ ↩ Go to Moldova and kill the Russians first. Then forcibly conscript the rest of the men and rape the Russian women remaining.	disjunct	• Russians will have negative nature. • Russians will have negative behavior. • Russia will have negative behavior.

```
From a thread in r/{subreddit}

Post Title: `{title}`
Post Text: ```
{text}
```

Message {i} by {author_name} on {creation_time}:
```
{comment_body}
```

Please indicate for each of the messages in this thread whether they include toxic language, using the specified schema.
```

Prompt 2: Template string containing messages in a sub-thread and instruction. Lines 8 through 12 are repeated and instantiated independently for each message.

### *6.1 Structured Outputs for Toxic Reasoning*

To make an LLM produce outputs in accordance with the toxic reasoning schema, we encode a version of the schema as a pydantic[h] model.

The pydantic model is converted into a JSON Schema[i]. The role of the JSON schema is two-fold. First, with this JSON schema, models can be forced to produce responses which are valid instances of the toxic reasoning schema. To make this work, a sampling strategy is used that masks tokens which would produce invalid outputs (Willard and Louf 2023). Second, the JSON schema is provided in the input to the model. For an appropriately trained model, this largely replaces the role of the instructions provided to a model. As such, the pydantic model also includes textual descriptions of each field, which guide the LLM on how to fill out those fields.

In total, the pydantic model consists of 9 Python classes. The main class is the `ToxicReasoning` class, every sub-thread is provided to the LLM as one input, and the model is asked to produce one `ToxicReasoning` instance per comment that is deemed toxic. The full pydantic code can be found in Appendix G.

### *6.2 Statistics*

Statistics of the training set can be seen in Table 7. Roughly 30%-40% of comments were considered potentially toxic by GPT4o, varying by language. This is lower than the amount considered potentially toxic by human annotators (closer to 50%). Compared to human annotators, GPT4o is also somewhat more likely to annotate negative polarity and qualitative category implications.

Note that the test set data was selected based on a stratified sample, but the train set was not. Thus, differences in these statistics cannot be attributed purely to differences between GPT4o and the human annotators.

## 7. Baseline Experiments

In this section, we provide baseline results on our dataset. We evaluate OpenAI's GPT4o, a commercially available LLM, in a zero-shot setup. We also fine-tune XLM-RoBERTa (Conneau et al. 2020) on the training set using a simple baseline setup.

[h] https://docs.pydantic.dev/
[i] https://json-schema.org/
```

Table 7. : Train set statistics, including label statistics. The first three rows give sub-thread lengths. The rest of the rows show the number of annotations which has that category or polarity.

		AR	DE	EN	ES	NL	TR
	Nr. sub-threads	296	1,421	17,833	3,340	612	940
	Avg. sub-thread length	5.51	5.16	5.20	5.21	5.39	5.43
	Min. sub-thread length	4	4	4	4	4	4
	Max. sub-thread length	10	10	10	10	10	10
	Nr. of non-toxic	1,034	5,389	62,224	10,533	2,249	3,097
	Nr. of possibly toxic	597	1,944	30,493	6,858	1,049	2,005
Category	Situational	33	149	1,950	477	51	81
	Harm	28	176	2,957	571	97	148
	Qualitative	223	583	8,705	2,496	369	603
	Dehumanization	23	83	763	333	25	148
	Behavioral	240	799	13,436	2,412	423	574
	NonSpecific	19	80	1,623	230	39	36
	UnclearOrNone	0	1	17	3	3	1
Polarity	Positive	9	6	102	20	23	
	Neutral	2	16	189	51	12	10
	Negative	555	1,849	29,160	6,451	993	1,578
Polarity + Category	Positive-Situational	1		7	1		
	Neutral-Situational		4	13	5		
	Negative-Situational	32	145	1,930	471	51	77
	Harm	28	176	2,957	571	97	148
	Positive-Qualitative			12	2		
	Neutral-Qualitative		1	26	13	2	2
	Negative-Qualitative	223	582	8,667	2,481	367	601
	Dehumanization	23	83	763	333	25	148
	Positive-Behavioral	7	6	53	14	1	
	Neutral-Behavioral	2	8	87	22	7	1
	Negative-Behavioral	231	785	13,296	2,376	415	573

7.1 A Fine-tuning Baseline

The main approach we evaluate works by adding classification heads to a pre-trained language model. We use this strategy in combination with XLM-RoBERTa-large (Conneau et al. 2020), and EuroBERT (Boizard et al. 2025). A sub-thread is provided to the model as input, the same way as shown in Prompt 2. However, one change is made: we add a special '`<COMMENT>`' token to the vocabulary and randomly initialize its embedding. Then, after each comment in the sub-thread we add this token. It is the representations for these '`<COMMENT>`' tokens that we use as input to the

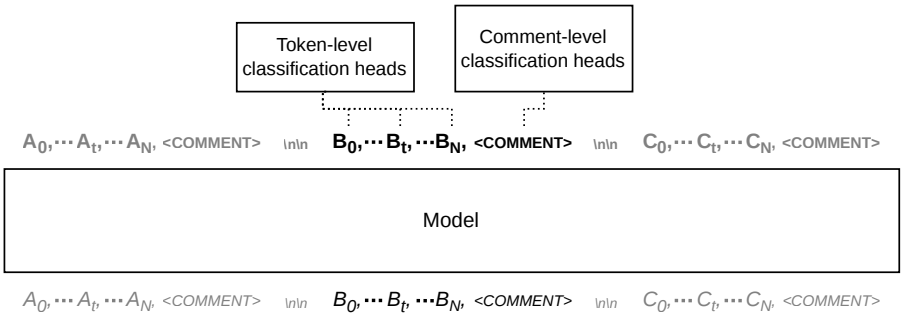

Figure 5: Schematic view of approach for fine-tuning baseline. From bottom to top, we see: input tokens for comments A, B and C; the model; output hidden states; and finally, the classification heads.

classification heads. The only exception are the `subjectTokens`, `otherTokens`, and `implTokens` variables, which are predicted using binary classification heads on the tokens of each comment.

This approach is a relatively straightforward way of getting a model to make predictions according to the toxic reasoning schema. Unlike GPT4o, which predicts each field in sequence, this model outputs predictions on all variables 'simultaneously'. Given two variables A and B, where A precedes B in the hierarchy, this setup does not allow a model to adjust its prediction of variable B to the its prediction on variable A.

Alternative Strategy. We also explored fine-tuning a model for producing instances of the pydantic model in the same way that GPT4o does. Preliminary experiments using this strategy did not result in better performance than the strategy given above (see Appendix H). Since our primary goal was to produce a simple baseline, we leave such strategies for future work.

7.2 Evaluation

In our setup, every annotator potentially has a different implication in mind when producing their annotations. This presents two problems for the use of ordinary classification metrics, like F1 scores.

Problem 1. For a given variable V, comparing model predictions against annotations only makes sense if the annotation and prediction agree on the variables which precede V in the hierarchy of Figure 4. Disagreements on earlier variables can cause predictions and annotations to disagree on whether there is an implication at all. Thus, the model may produce a value for a given variable, while the annotator did not, or vice versa. To address this, we consider the following variants of the metrics.

- **propagated** — If a model does not produce any implication when it should have, or does when it should not have, then propagate this error to all subsequent variables as well.
- **conditional** — Report the metric on the subset of comments/annotations for which both the model and annotator agreed that the variable should be annotated (i.e. that none of the preceding variables had a value that should terminate the prediction/annotation).
- **optimistic** — Models like that described in subsection 7.1 can produce predictions for all variables even if the prediction for an earlier variable would require ignoring later predictions (e.g. `hasImplication` being false). This allows us to evaluate later predictions as though the prediction for the earlier variable was correct, even if it was not.

Table 8. : Weighted F1, recall (Rec.), and precision (Prec.) scores, as well as the 'Prior' baseline, calculated per language. Left column shows English, and the right shows average over all languages. Labels obtained by majority vote amongst annotators. F1 scores that do not beat the 'Prior' baseline are underlined. Darker shades of blue correspond to higher scores.

		English				All			
		F1	Prec.	Rec.	Prior	F1	Prec.	Rec.	Prior
GPT4o (prop.)	toxicity	.74	.75	.75	.52	.77	.79	.77	.50
	counternarrative	.95	.94	.96	.94	.88	.84	.92	.85
	justInappropriate	.87	.86	.90	.83	.78	.82	.83	.71
	hasImplication	.78	.79	.78	.60	.76	.80	.74	.68
	hasOther	.38	.77	.28	.65	.43	.80	.32	.72
XLM-R (prop.)	toxicity	.67	.69	.69	.52	.67	.72	.69	.50
	counternarrative	.95	.94	.97	.94	.87	.84	.91	.85
	justInappropriate	.87	.92	.91	.83	.76	.85	.82	.71
	hasImplication	.75	.75	.76	.60	.77	.77	.76	.68
	hasOther	.33	.70	.22	.65	.34	.78	.22	.72
XLM-R (opt.)	toxicity	.67	.69	.69	.52	.67	.72	.69	.50
	counternarrative	.95	.94	.97	.94	.87	.84	.91	.85
	justInappropriate	.87	.92	.91	.83	.76	.85	.82	.71
	hasImplication	.75	.75	.76	.60	.77	.77	.76	.68
	hasOther	.69	.73	.67	.65	.71	.77	.67	.72

Problem 2. Combining annotations to produce a single label, often done by majority vote, similarly requires that annotations agree on the variables which precede it in the hierachy. Rather than combine the annotations, we compare predictions against individual annotations. Various strategies have been used to evaluate on datasets that contain multiple annotations per item (Uma et al. 2021). We use three approaches that allow the use of ordinary classification metrics:

- **all**, do not select one annotation per comment, but repeat comments such that each annotation is included independently;
- **random**, sample one annotation per comment, repeating N times, then report the median; and,
- **max**, choose the annotation for which another (set of) variable is most similar to the prediction.

Metrics. For the evaluation, we report the average precision, recall, and F1, with each being calculated per class first. The number we report is the weighted average over the classes, where the weights are proportional to the frequency of the class.

We compare these scores against what we call the 'Prior' baseline. This is the score obtained by guessing classes proportional to their frequency in the label distribution of the test set.

Table 9. : The Jaccard index between GPT4o and human annotations for the token-level variables.

	implTokens		otherTokens		subjectTokens	
	Jaccard	Supp.	Jaccard	Supp.	Jaccard	Supp.
AR	.26	161	.11	17	.23	147
DE	.22	163	.23	54	.19	144
EN	.18	263	.13	91	.19	246
ES	.25	280	.20	45	.20	249
NL	.19	172	.09	37	.23	149
TR	.29	117	.19	26	.23	116

Table 10. : F1 scores compared to prior baseline for PropositionComposition variables, calculated across all languages. Evaluated in 3 scenarios, propagated, conditional on the model predicting there is an implication, and conditional on the relevant _Tokens variable.

		GPT4o			XLM-RoBERTa		
		F1	Prior	Supp.	F1	Prior	Supp.
propagated	subject	.54	.49	1051	.39	.49	1051
	subjectGroupType	.29	.12	1070	.13	.12	1070
	other	.32	.62	301	.27	.62	301
	implPolarity	.61	.79	1169	.44	.79	1169
	implTemporality	.55	.55	1180	.39	.55	1180
	implCategory	.34	.38	1170	.24	.38	1170
cond-impl	subject	.76	.50	583	.74	.50	372
	subjectGroupType	.41	.12	636	.26	.12	419
	other	.48	.58	171	.74	.65	68
	implPolarity	.87	.84	627	.89	.88	397
	implTemporality	.79	.56	652	.77	.56	419
	implCategory	.49	.36	628	.48	.37	399
cond-tokens	subject	.87	.57	286	.83	.56	15
	subjectGroupType	.48	.11	297	.35	.14	17
	other	.58	.59	46	-	-	0
	implPolarity	.85	.81	340	.75	.72	70
	implTemporality	.84	.60	311	.85	.57	62
	implCategory	.50	.37	341	.53	.36	70

8. Results

We report F1, Precision, and Recall against the 'Prior' baseline for the binary, multiclass and multilabel variables. There are two exceptions. First, for the _Tokens variables we report the Jaccard

index,[j] because there is no fixed set of labels, as the tokens are different for each comment. Second, for the attitudes we do not report classification scores because the values are from an ordinal scale. To assess the quality of predictions on the level of implications, including the attitudes, we also report another statistic. Specifically, we convert the predictions to summary sentences, as described previously, and then count how often these sentences match at least one of the human annotations.

The scores obtained by GPT4o and XLM-RoBERTa can be seen in Tables 8, 9, 10 and 11. We omit the scores obtained by EuroBERT, since they consistently underperformed XLM-R by a small margin. In Table 8, we can see the classification scores obtained by GPT4o and RoBERTa for the variables we evaluate against a majority vote amongst the annotations. For token-level variables, Table 9 shows the Jaccard index between GPT4o and human annotations. In Table 10, we can see F1 scores for variables which together describe the proposition at the heart of the implication being annotated.

In general, we found that the choice between **all**, **random**, and **max** was of little impact. This may be because a large share of comments were only found to have potentially toxic implications by one or two annotators (see Table 16). Thus, we only report **random** for all variables.

8.1 GPT4o and XLM-R Performance

Starting with Table 8, we can see that both models beat the Prior baseline for the first four variables. With GPT4o being evaluated in zero-shot setting, it has had no exposure at all to instances of our dataset, making this performance quite impressive. Simultaneously, there is still room for improvement, with F1 scores for toxicity (.74), justInappropriate (.87), and hasImplication (.78) all remaining well below a perfect score. Performance for GPT4o leads on the toxicity variable, but is otherwise very similar to XLM-R. The performance for the `hasOther` variable is poorer, primarily due to poor recall. For XLM-R, we also report an optimistic evaluation, where errors in the first four variables are not propagated. In that setting, we can see that performance for `hasOther` gets much closer to the Prior baseline. GPT4o is not evaluated in the **optimistic** variant, because similar to the human annotation, the model only produces a full prediction if it did not encounter one of stopping conditions in the first few variables.

Looking at Table 10, we can see that in the **propagated** setting, F1 scores for both models often do not outperform the Prior baseline. However, when looking at **cond-impl**, which includes only the instances for which the model predicted an implication, we see that F1 scores are much higher. Of course, just because the model predicted an implication, does not mean that this is the same implication the human annotator had in mind. Thus, we also include **cond-tokens**. In that evaluation setting, we only include instances where the tokens selected by the model and the annotation have a Jaccard score of at least 0.5. We use the most relevant `_Tokens` variable, that is: `subjectTokens` for the variables related to the subject, `otherTokens` for `other`, and `implTokens` for the variables related to the implication. In this setting, we see the performance of both models rise further. This supports the idea that the regular classification scores are too pessimistic. At the same time, it should be noted that by selecting samples where the model agrees with human annotators, we might be selecting for the instances that are most straightforward to predict. For XLM-R, the number of instances in this last setting is also low to very low, so those particular F1 scores may be unreliable.

The results show that XLM-RoBERTa keeps up in performance when it comes to characterizing implications. However, when it comes to identifying whether a comment may contain toxicity in the first place, or identifying spans of tokens corresponding to elements of the toxicity, GPT4o clearly outperforms XLM-RoBERTa.

[j]The Jaccard index is defined between two sets as the size of their intersection divided by the size of their union.

Table 11. : F1 scores for variables that describe the implication.

		GPT4o			RoBERTa		
		F1	Prior	Supp.	F1	Prior	Supp.
prop.	implSarcasm	.55	.62	1169	.38	.62	1169
	implStereotype	.45	.57	1169	.31	.57	1169
cond-impl	implSarcasm	.79	.65	628	.75	.67	399
	implStereotype	.64	.55	628	.61	.58	399

Table 12. : GPT4o stakeholder attitude evaluation. Rows corresponds to differences in ordinal values assigned by GPT4o vs. Human annotators. Row marked '#' shows number of samples where that attitude was annotated.

	Author			Typical		Expert
	Belief	Prefer	Account	Belief	Prefer	Belief
#	20	9	1	21	8	17
0	45%	0%	0%	19%	25%	12%
1	45%	56%	100%	62%	75%	47%
2	5%	33%	0%	19%	0%	41%
3	5%	11%	0%	0%	0%	0%

8.1.1 Implication-level Performance

Here we investigate the performance of the models on the level of implications. We do this to get a better sense of how well the models perform at the full task, rather than just the individual variables.

Included in the implication-level performance is are also the variables that are annotated w.r.t. the earlier variables which characterize component parts of the implication. This includes the stakeholder attitudes, as well as the `implSarcasm` and `implStereotype` variables.

For `implSarcasm` and `implStereotype`, we give F1 scores in Table 11. In the cond-impl setting, GPT4o outperforms the Prior baseline on both of these variables. XLM-R does as well, although with a slim margin. For these variables, we do not include a 'cond-tokens' evaluation setting, because there is no single _Tokens variable that would be suitable.

Summary Sentence Accuracy. We perform another evaluation for GPT4o where we count the number of samples for which it produces the same annotation as at least one of the annotators. This evaluation is performed on the same subset of English sentences used for the disagreement analysis.

GPT4o's predictions are *identical* to one of the annotators for 10 out of 47 samples. For an additional 13 samples, we see a *match*, but with a small difference. Differences are: GPT4o predicts only a subset of the temporalities that the annotator included (3 samples); GPT4o is imprecise in its prediction for `subjectGroup` (5 samples), for example predicting "The group mentioned" or "Certain individuals"; GPT4o predicted `hasOther`, when annotators did not (2 samples); or, GPT4o predicted nature instead of behavior or vice versa (3 samples). For 14 samples, GPT4o predicted *no implication*. Finally, for 10 samples, there was *no match* with any annotator's implication.

Attitudes. To evaluate the stakeholder attitudes, we look at samples for which the summary sentences were either a *match* or *identical* to one of the annotators. For each attitude, the annotators chose between the ordinal values "Very low", "Low", "Medium", "High", "Very high", or "Not applicable". The last option is there to indicate when there is too little information in a comment and its context to tell the author's attitude.

In Table 12, we report the number of GPT4o's predictions that were off by 0, 1, 2, or 3 steps in the ordinal scale. What we can see is that `authorPrefer` is the attitude with which GPT4o struggles most. For all attitudes, over half of the predictions are 0 or 1 steps away.

8.2 Discussion

The results show that toxic reasoning as a task, and the ToxiREX test set in particular, present a challenging problem. Neither GPT4o nor the fine-tuned XLM-RoBERTa are full solutions to this problem.

XLM-RoBERTa generally underperforms GPT4o. This is expected as it was fine-tuned on labels produced by GPT4o.

For the part of the data we analyzed, GPT4o gives characterizations of toxic implications that match or are identical to one of the human annotations in 23 out of 47 cases. In many of the remaining cases, GPT4o does not match a human's annotation. However, looking at GPT4o's free-text description of the implication it (presumably) attempted to characterize, we can see that often those implications do still match (one of) the human annotations. This suggests that the model struggles with correctly using the toxic reasoning schema as presented in the Pydantic model. However, for 14 out of 47 instances, GPT4o did not predict any implication. Thus, the model still misses out on many of the toxic implications spotted by human annotators. And, more deficiencies can be seen in the results on the attitudes and the span-level predictions.

From the results it is clear that there is still much to improve with regards to models' ability to exhaustively list and characterize the toxic implications of social media comments.

9. Conclusion

We have released ToxiREX, a contextual, multilingual dataset annotated for toxic reasoning on Reddit comments, where annotations provide structured characterizations of what a comment implies. The annotations capture and explain implicit and context-dependent toxicity, and can be mapped to existing toxicity taxonomies. It includes a test set of almost 3,000 comments, split between six languages, and annotated by native speakers. The ToxiREX training set consists of roughly 125,000 comments with silver-standard labels generated by GPT4o. The dataset's conversational threads, often about controversial topics, include toxic implications of various kinds. In the process of annotating the test set, we have shown how the toxic reasoning schema has, with few exceptions, proven expressive enough for the human annotators.

To produce baseline results and to show how to evaluate models on the test set, we evaluate GPT4o and a fine-tuned XLM-RoBERTa. We have shown how GPT4o-based annotation (and prediction) is made possible by using structured outputs. When GPT4o is evaluated, its performance is strong for a zero-shot setup, but also shows that automatically detecting and characterizing toxic implications is still an open challenge.

Both the toxic reasoning approach and the ToxiREX dataset present various opportunities for future work. For example, the multilinguality of our dataset presents an opportunity to better understand toxic language across cultures. Characterizing toxicity in different languages and different cultures using the same schema allows for direct comparisons. Do different cultures appeal to different kinds of negative situations, behaviors, and qualties? Another direction is to pursue

strategies that improve performance on the toxic reasoning tasks, i.e. detecting and characterizing toxic implications.

For ToxiREX we used a 'bottom-up' approach to construct our dataset. We asked annotators to become familiar with the full toxic reasoning schema, and to annotate according to it. Afterwards, we can map those annotations to existing taxonomies of toxic language. In future work, we can also identify any instances that do not cleanly map onto existing categories, or identify if instances mapped to the same existing category still differ along some other axis, potentially giving cause to recognize new subcategories.

Further opportunities lie in 'top-down' data annotations, where annotators select familiar categories of toxic language, and then verify that their understanding comports with how it maps onto the toxic reasoning schema. For example, when an annotation of 'Intimidation' is chosen, we ask if the author indeed desires someone be subjected to a negative situation, and if so, what spans correspond to the subject and the negative situation. This would be substantially less labor-intensive than the bottom-up approach, which requires a full characterization of all aspects, rather than just those already indicated as relevant by the mapping. Such a top-down annotation campaign will be complementary by allowing a larger scale, in exchange for trading in some expressivity, as the categories will be fixed beforehand.

Funding Statement. This research was supported by Huawei Finland through the DreamsLab project. All content represented the opinions of the authors, which were not necessarily shared or endorsed by their respective employers and/or sponsors.

Competing Interests. The authors declare none.

Data Availability. The ToxiREX dataset can be found on Github at `https://github.com/cltl/toxirex/`.

Ethics. Ethical approval was not required.

References

Barbarestani, B., **Maks, I.**, **and Vossen, P.** (2022). Annotating Targets of Toxic Language at the Span Level. In **Kumar, R.**, **Ojha, A. K.**, **Zampieri, M.**, **Malmasi, S.**, **and Kadar, D.**, editors, *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pp. 43–51, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Basile, V., **Bosco, C.**, **Fersini, E.**, **Nozza, D.**, **Patti, V.**, **Rangel Pardo, F. M.**, **Rosso, P.**, **and Sanguinetti, M.** (2019). SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Boizard, N., **Gisserot-Boukhlef, H.**, **Alves, D. M.**, **Martins, A.**, **Hammal, A.**, **Corro, C.**, **Hudelot, C.**, **Malherbe, E.**, **Malaboeuf, E.**, **Jourdan, F.**, **Hautreux, G.**, **Alves, J.**, **El-Haddad, K.**, **Faysse, M.**, **Peyrard, M.**, **Guerreiro, N. M.**, **Fernandes, P.**, **Rei, R.**, **and Colombo, P.** (2025). EuroBERT: Scaling Multilingual Encoders for European Languages. arXiv:2503.05500 [cs].

Calabrese, A., **Ross, B.**, **and Lapata, M.** (2022). Explainable Abuse Detection as Intent Classification and Slot Filling. *Transactions of the Association for Computational Linguistics*, 10:1440–1454. Place: Cambridge, MA.

Carthy, S. L., **Doody, C. B.**, **Cox, K.**, **O'Hora, D.**, **and Sarma, K. M.** (2020). Counter-narratives for the prevention of violent radicalisation: A systematic review of targeted interventions. *Campbell Systematic Reviews*, 16(3):e1106. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cl2.1106.

Caselli, T., **Basile, V.**, **Mitrović, J.**, **Kartoziya, I.**, **and Granitzer, M.** (2020). I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language. In **Calzolari, N.**, **Béchet, F.**, **Blache, P.**, **Choukri, K.**, **Cieri, C.**, **Declerck, T.**, **Goggi, S.**, **Isahara, H.**, **Maegaard, B.**, **Mariani, J.**, **Mazo, H.**, **Moreno, A.**, **Odijk, J.**, **and Piperidis, S.**, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 6193–6202, Marseille, France. European Language Resources Association.

Conneau, A., **Khandelwal, K.**, **Goyal, N.**, **Chaudhary, V.**, **Wenzek, G.**, **Guzmán, F.**, **Grave, E.**, **Ott, M.**, **Zettlemoyer, L.**, **and Stoyanov, V.** (2020). Unsupervised Cross-lingual Representation Learning at Scale. arXiv:1911.02116 [cs].

ElSherief, M., **Ziems, C.**, **Muchlinski, D.**, **Anupindi, V.**, **Seybolt, J.**, **De Choudhury, M.**, **and Yang, D.** (2021). Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical*

Methods in Natural Language Processing, pp. 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fortuna, P. and Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):1–30.

Garg, T., Masud, S., Suresh, T., and Chakraborty, T. (2023). Handling Bias in Toxic Speech Detection: A Survey. *ACM Comput. Surv.*, 55(13s):264:1–264:32.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Ding, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Chen, J., Yuan, J., Tu, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., You, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Zhou, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. In *10th International Conference on Learning Representations (ICLR 2022)*, pp. 12513–12525. Curran Associates, Inc.

Jafari, N., Allan, J., and Sarwar, S. M. (2024). Target Span Detection for Implicit Harmful Content. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '24, pp. 117–122, New York, NY, USA. Association for Computing Machinery.

Jahan, M. S. and Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Kirk, H., Yin, W., Vidgen, B., and Röttger, P. (2023). SemEval-2023 Task 10: Explainable Detection of Online Sexism. In **Ojha, A. K., Doğruöz, A. S., Da San Martino, G., Tayyar Madabushi, H., Kumar, R., and Sartori, E.**, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp. 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Kumar, D., AbuHashem, Y. A., and Durumeric, Z. (2024). Watch Your Language: Investigating Content Moderation with Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:865–878.

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Ljubešić, N., Mozetič, I., and Novak, P. K. (2023). Quantifying the impact of context on the quality of manual hate speech annotation. *Natural Language Engineering*, 29(6):1481–1494.

Madriaza, P., Hassan, G., Brouillette-Alarie, S., Mounchingam, A. N., Durocher-Corfa, L., Borokhovski, E., Pickup, D., and Paillé, S. (2025). Exposure to hate in online and traditional media: A systematic review and meta-analysis of the impact of this exposure on individuals and communities. *Campbell Systematic Reviews*, 21(1):e70018.

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., and Mukherjee, A. (2021). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14867–14875. Number: 17.

OpenAI, Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., Barak, B., Bennett, A., Bertao, T., Brett, N., Brevdo, E., Brockman, G., Bubeck, S., Chang, C., Chen, K., Chen, M., Cheung, E., Clark, A., Cook, D., Dukhan, M., Dvorak, C., Fives, K., Fomenko, V., Garipov, T., Georgiev, K., Glaese, M., Gogineni, T., Goucher, A., Gross, L., Guzman, K. G., Hallman, J., Hehir, J., Heidecke, J., Helyar, A., Hu, H., Huet, R., Huh, J., Jain, S., Johnson, Z., Koch, C., Kofman, I., Kundel, D., Kwon, J., Kyrylov, V., Le, E. Y., Leclerc, G., Lennon, J. P., Lessans, S., Lezcano-Casado, M., Li, Y., Li, Z., Lin, J., Liss, J., Lily, Liu, Liu, J., Lu, K., Lu, C., Martinovic, Z., McCallum, L., McGrath, J., McKinney, S., McLaughlin, A., Mei, S., Mostovoy, S., Mu, T., Myles, G., Neitz, A., Nichol, A., Pachocki, J., Paino, A., Palmie, D., Pantuliano, A., Parascandolo, G., Park, J., Pathak, L., Paz, C., Peran, L., Pimenov, D., Pokrass, M., Proehl, E., Qiu, H., Raila, G., Raso, F., Ren, H., Richardson, K., Robinson, D., Rotsted, B., Salman, H., Sanjeev, S., Schwarzer, M., Sculley, D., Sikchi, H., Simon, K., Singhal, K., Song, Y., Stuckey, D., Sun, Z., Tillet, P., Toizer, S., Tsimpourlas, F., Vyas, N., Wallace, E., Wang, X., Wang, M., Watkins, O., Weil, K., Wendling, A., Whinnery, K., Whitney, C., Wong, H., Yang, L., Yang, Y., Yasunaga, M., Ying,

K., **Zaremba, W.**, **Zhan, W.**, **Zhang, C.**, **Zhang, B.**, **Zhang, E.**, **and Zhao, S.** (2025). gpt-oss-120b & gpt-oss-20b Model Card. arXiv:2508.10925 [cs].

Ousidhoum, N., **Lin, Z.**, **Zhang, H.**, **Song, Y.**, **and Yeung, D.-Y.** (2019). Multilingual and Multi-Aspect Hate Speech Analysis. In **Inui, K.**, **Jiang, J.**, **Ng, V.**, **and Wan, X.**, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Pachinger, P., **Goldzycher, J.**, **Planitzer, A. M.**, **Neidhardt, J.**, **and Hanbury, A.** (2025). A Disaggregated Dataset on English Offensiveness Containing Spans. In **Abercrombie, G.**, **Basile, V.**, **Frenda, S.**, **Tonelli, S.**, **and Dudy, S.**, editors, *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pp. 1–14, Suzhou, China. Association for Computational Linguistics.

Pavlopoulos, J., **Sorensen, J.**, **Dixon, L.**, **Thain, N.**, **and Androutsopoulos, I.** (2020). Toxicity Detection: Does Context Really Matter? In **Jurafsky, D.**, **Chai, J.**, **Schluter, N.**, **and Tetreault, J.**, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4296–4305, Online. Association for Computational Linguistics.

Pavlopoulos, J., **Sorensen, J.**, **Laugier, L.**, **and Androutsopoulos, I.** (2021). SemEval-2021 Task 5: Toxic Spans Detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pp. 59–69, Online. Association for Computational Linguistics.

Sap, M., **Gabriel, S.**, **Qin, L.**, **Jurafsky, D.**, **Smith, N. A.**, **and Choi, Y.** (2020). Social Bias Frames: Reasoning about Social and Power Implications of Language. In **Jurafsky, D.**, **Chai, J.**, **Schluter, N.**, **and Tetreault, J.**, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5477–5490, Online. Association for Computational Linguistics.

Schouten, S. F., **Markov, I.**, **and Vossen, P.** (2026). A Position Paper on Toxic Reasoning: Grounding Categories of Toxic Language in Implications and Attitudes. In **Barnes, J.**, **Barriere, V.**, **De Clercq, O.**, **Klinger, R.**, **Nouri, C.**, **Nozza, D.**, **and Singh, P.**, editors, *The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA 2026)*, pp. 134–145, Rabat, Morocco. Association for Computational Linguistics.

Team, G., **Kamath, A.**, **Ferret, J.**, **Pathak, S.**, **Vieillard, N.**, **Merhej, R.**, **Perrin, S.**, **Matejovicova, T.**, **Ramé, A.**, **Rivière, M.**, **Rouillard, L.**, **Mesnard, T.**, **Cideron, G.**, **Grill, J.-b.**, **Ramos, S.**, **Yvinec, E.**, **Casbon, M.**, **Pot, E.**, **Penchev, I.**, **Liu, G.**, **Visin, F.**, **Kenealy, K.**, **Beyer, L.**, **Zhai, X.**, **Tsitsulin, A.**, **Busa-Fekete, R.**, **Feng, A.**, **Sachdeva, N.**, **Coleman, B.**, **Gao, Y.**, **Mustafa, B.**, **Barr, I.**, **Parisotto, E.**, **Tian, D.**, **Eyal, M.**, **Cherry, C.**, **Peter, J.-T.**, **Sinopalnikov, D.**, **Bhupatiraju, S.**, **Agarwal, R.**, **Kazemi, M.**, **Malkin, D.**, **Kumar, R.**, **Vilar, D.**, **Brusilovsky, I.**, **Luo, J.**, **Steiner, A.**, **Friesen, A.**, **Sharma, A.**, **Sharma, A.**, **Gilady, A. M.**, **Goedeckemeyer, A.**, **Saade, A.**, **Feng, A.**, **Kolesnikov, A.**, **Bendebury, A.**, **Abdagic, A.**, **Vadi, A.**, **György, A.**, **Pinto, A. S.**, **Das, A.**, **Bapna, A.**, **Miech, A.**, **Yang, A.**, **Paterson, A.**, **Shenoy, A.**, **Chakrabarti, A.**, **Piot, B.**, **Wu, B.**, **Shahriari, B.**, **Petrini, B.**, **Chen, C.**, **Lan, C. L.**, **Choquette-Choo, C. A.**, **Carey, C. J.**, **Brick, C.**, **Deutsch, D.**, **Eisenbud, D.**, **Cattle, D.**, **Cheng, D.**, **Paparas, D.**, **Sreepathihalli, D. S.**, **Reid, D.**, **Tran, D.**, **Zelle, D.**, **Noland, E.**, **Huizenga, E.**, **Kharitonov, E.**, **Liu, F.**, **Amirkhanyan, G.**, **Cameron, G.**, **Hashemi, H.**, **Klimczak-Plucińska, H.**, **Singh, H.**, **Mehta, H.**, **Lehri, H. T.**, **Hazimeh, H.**, **Ballantyne, I.**, **Szpektor, I.**, **Nardini, I.**, **Pouget-Abadie, J.**, **Chan, J.**, **Stanton, J.**, **Wieting, J.**, **Lai, J.**, **Orbay, J.**, **Fernandez, J.**, **Newlan, J.**, **Ji, J.-y.**, **Singh, J.**, **Black, K.**, **Yu, K.**, **Hui, K.**, **Vodrahalli, K.**, **Greff, K.**, **Qiu, L.**, **Valentine, M.**, **Coelho, M.**, **Ritter, M.**, **Hoffman, M.**, **Watson, M.**, **Chaturvedi, M.**, **Moynihan, M.**, **Ma, M.**, **Babar, N.**, **Noy, N.**, **Byrd, N.**, **Roy, N.**, **Momchev, N.**, **Chauhan, N.**, **Sachdeva, N.**, **Bunyan, O.**, **Botarda, P.**, **Caron, P.**, **Rubenstein, P. K.**, **Culliton, P.**, **Schmid, P.**, **Sessa, P. G.**, **Xu, P.**, **Stanczyk, P.**, **Tafti, P.**, **Shivanna, R.**, **Wu, R.**, **Pan, R.**, **Rokni, R.**, **Willoughby, R.**, **Vallu, R.**, **Mullins, R.**, **Jerome, S.**, **Smoot, S.**, **Girgin, S.**, **Iqbal, S.**, **Reddy, S.**, **Sheth, S.**, **Põder, S.**, **Bhatnagar, S.**, **Panyam, S. R.**, **Eiger, S.**, **Zhang, S.**, **Liu, T.**, **Yacovone, T.**, **Liechty, T.**, **Kalra, U.**, **Evci, U.**, **Misra, V.**, **Roseberry, V.**, **Feinberg, V.**, **Kolesnikov, V.**, **Han, W.**, **Kwon, W.**, **Chen, X.**, **Chow, Y.**, **Zhu, Y.**, **Wei, Z.**, **Egyed, Z.**, **Cotruta, V.**, **Giang, M.**, **Kirk, P.**, **Rao, A.**, **Black, K.**, **Babar, N.**, **Lo, J.**, **Moreira, E.**, **Martins, L. G.**, **Sanseviero, O.**, **Gonzalez, L.**, **Gleicher, Z.**, **Warkentin, T.**, **Mirrokni, V.**, **Senter, E.**, **Collins, E.**, **Barral, J.**, **Ghahramani, Z.**, **Hadsell, R.**, **Matias, Y.**, **Sculley, D.**, **Petrov, S.**, **Fiedel, N.**, **Shazeer, N.**, **Vinyals, O.**, **Dean, J.**, **Hassabis, D.**, **Kavukcuoglu, K.**, **Farabet, C.**, **Buchatskaya, E.**, **Alayrac, J.-B.**, **Anil, R.**, **Dmitry**, **Lepikhin**, **Borgeaud, S.**, **Bachem, O.**, **Joulin, A.**, **Andreev, A.**, **Hardin, C.**, **Dadashi, R.**, **and Hussenot, L.** (2025). Gemma 3 Technical Report. arXiv:2503.19786 [cs].

Tufa, W. T., **Markov, I.**, **and Vossen, P.** (2024). Grounding Toxicity in Real-World Events Across Languages. In **Rapp, A.**, **Di Caro, L.**, **Meziane, F.**, **and Sugumaran, V.**, editors, *Natural Language Processing and Information Systems*, pp. 197–210, Cham. Springer Nature Switzerland.

Uma, A. N., **Fornaciari, T.**, **Hovy, D.**, **Paun, S.**, **Plank, B.**, **and Poesio, M.** (2021). Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Vidgen, B. and Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

Vidgen, B., **Nguyen, D.**, **Margetts, H.**, **Rossini, P.**, **and Tromble, R.** (2021). Introducing CAD: the Contextual Abuse Dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2289–2303, Online. Association for Computational Linguistics.

Waseem, Z., **Davidson, T.**, **Warmsley, D.**, **and Weber, I.** (2017). Understanding Abuse: A Typology of Abusive Language

Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pp. 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Willard, B. T. and Louf, R. (2023). Efficient Guided Generation for Large Language Models. arXiv:2307.09702 [cs].

Xenos, A., **Pavlopoulos, J.**, **Androutsopoulos, I.**, **Dixon, L.**, **Sorensen, J.**, **and Laugier, L.** (2022). Toxicity detection sensitive to conversational context. *First Monday*.

Yadav, A., **Chandel, S.**, **Chatufale, S.**, **and Bandhakavi, A.** (2022). LAHM : Large Annotated Dataset for Multilingual & Multi-Domain Hate Speech Identification. In **Das, A.**, **Chakraborty, T.**, **Ekbal, A.**, **and Sheth, A.**, editors, *Proceedings of the Workshop on Multi-Modal Fake News and Hate-Speech Detection (DE-FACTIFY 2022)*, volume 3199 of *CEUR Workshop Proceedings*, Virtual Event, Vancouver. CEUR.

Yu, X., **Blanco, E.**, **and Hong, L.** (2022). Hate Speech and Counter Speech Detection: Conversational Context Does Matter. In **Carpuat, M.**, **de Marneffe, M.-C.**, **and Meza Ruiz, I. V.**, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5918–5930, Seattle, United States. Association for Computational Linguistics.

Appendix A. Annotation Guidelines & Questions

```
Your task is to review a selection of Reddit comments. You will be asked a number of questions about each comment, focusing on that comment's potential toxicity.

We ask you to first read the context (see 'Reddit Thread').  The context consists of the following parts:
 - a post title (shown at the top in a large font);
 - optionally, the text of the post itself; and finally,
 - the comments section, where people are talking about the post.

After reading the context, please proceed to answer the questions about the current (highlighted in yellow) comment.  The current comment will be repeated as well, and will remain at the top of the page after you scroll past it. NOTE: The box that shows the current comment at the top of the page can be resized by dragging the bottom right corner.

The first few questions will determine if this comment is relevant (toxic in the right way). If it is not relevant you will be asked to continue to the next comment. If it is relevant, you will be asked what you think the main toxic implication is.

The rest of the questions should be answered with respect to that main implication. The main toxic implication you write down will also remain at the top of the page, underneath the current comment, in case you need to refer back to it.
```

Table 13. : Annotation Questions

(a) Questions 1–7

#	Question & Answer options	Variable
1	**In your view, do think this comment might be (perceived as) potentially toxic, polarizing, or otherwise negative towards a group or individual?** — "Yes/Maybe", "No"	
2	**Is it counter-speech? (An argument-based reply aimed at the previous comment, providing an alternative perspective.)** — CHECKBOX	toxicity
-	*If 'No' or 'No'+'Counter-speech', skip the following questions and go to the next comment.*	
3	**Is the comment only toxic because it uses inappropriate words? e.g. use of slurs, name-calling, etc.** — "Yes", "No"	justInappropriate
-	*If 'Yes', skip the following questions and go to the next comment.*	
4&5	**If the toxicity stems from what the text is either explicitly or implicitly communicating, please write one sentence that conveys the main toxic implication.** — TEXTBOX	implication
	If not (i.e. the toxicity cannot be captured in an implication), choose 'Different kind of toxicity', and continue to the next comment. — CHECKBOX	hasImplication
6	**Who is the subject of the implication?** — "the author themselves and/or their ingroup", "another participant in the conversation and/or the group they belong to", "an individual outside of the conversation", "another group", "none of the above"	subject
-	*If 'none of the above', continue to the next comment.*	
7	**If the subject is a group, or if the subject is a member of a group, which group?** — TEXTBOX	subjectGroup

(b) Questions 8–15

#	Question & Answer options	Variable
8	**What kind of group is it, i.e. what characteristic(s) define the group?** — "Sexual orientation", "Gender", "Disability", "Race/Ethnicity", "Age", "Religion", "Famous individual", "Political affiliation", "Social belief", "Body image", "Addiction", "Socioeconomic status", "Profession", "Nationality", "other", "Not applicable"	subjectGroupType
9	**Please mark tokens that are part of references to the subject, including references by name, by pronoun, and by description.** — "None" / TOKENS IN COMMENT	subjectTokens
10&11	**Is the subject implicitly or explicitly compared or contrasted to another person or group? If so, who?** — "No other", "the author themselves and/or their ingroup", "another participant in the conversation and/or the group they belong to", "an individual outside of the conversation", "another group", "none of the above"	hasOther & other
12	**If the other is a group, or if the other is a member of a group, which group?** — TEXTBOX	otherGroup
13	**Please mark tokens that are part of references to the other, including references by name, by pronoun, and by description.** — "None" / TOKENS IN COMMENT	otherTokens
14	**What is the main implication about?** — "(a) the subject's circumstances, living conditions, physical condition or health, general wellbeing, access to resources, etc.", "... (a.1) some kind of harm coming to the subject", "(b) the subject's nature, inherent qualities or abilities, etc.", "... (b.1) dehumanisation of the subject", "(c) the subject's choices/decisions, lifestyle, beliefs, etc.", "(d) a non-specific comparison (does not fall under (a), (b), or (c)) between the subject and the other", "(e) unclear or none of the above"	implCategory
15	**Please mark tokens that were relevant to determining your answer for the previous question.** — "None" / TOKENS IN COMMENT	implTokens

(c) Questions 16–25

#	Question & Answer options	Variable
16	**Does the implication say something positive, negative, or neither (neutral)?** — "Positive", "Neutral", "Negative"	implPolarity
17	**In your view, does the implication play into a known stereotype?** — "Yes", "No"	implStereotype
18	**Is the implication conveyed through sarcasm?** — "Yes", "No"	implSarcasm
19	**At what point in time is the implication meant to hold?** — "Past", "Present", "Future"	implTemporality
20	**Based on the comment, to what degree does the author seem to believe that: SUMMARY_SENTENCE?** — "Not applicable", "Very low", "Low", "Medium", "High", "Very high"	authorBelief
21	**Based on the comment, to what degree does the author seem to wish (or prefer) that: SUMMARY_SENTENCE?** — "Not applicable", "Very low", "Low", "Medium", "High", "Very high"	authorPrefer
22	**Based on the comment, to what degree does the comment author seem personally committed to, or accountable/responsible for, the truth of the following statement: SUMMARY_SENTENCE?** — "Not applicable", "Very low", "Low", "Medium", "High", "Very high"	authorAccount
23	**In your view, to what degree would the general public believe that: SUMMARY_SENTENCE?** — "Not applicable", "Very low", "Low", "Medium", "High", "Very high"	typicalBelief
24	**In your view, to what degree would the general public wish (or prefer) that: SUMMARY_SENTENCE?** — "Not applicable", "Very low", "Low", "Medium", "High", "Very high"	typicalPrefer
25	**In your view, to what degree would an expert believe that: SUMMARY_SENTENCE?** — "Not applicable", "Very low", "Low", "Medium", "High", "Very high"	expertBelief

Appendix B. Summary sentence generation pseudocode

```
def summary_sentence(subject, subjectGroup, subjectTokens, hasOther, other,
    otherTokens, implTemporality, implPolarity, implTopic):

  match subject:
    case Personal:    subjectStr = "An␣Individual"
    case GroupMember: subjectStr = subjectGroup + "␣and␣a␣related␣individual"
    case GroupWhole:  subjectStr = subjectGroup
  subjectStr += f"␣({','.join(subjectTokens)})␣"

  match other:
    case Personal:    otherStr = "An␣Individual"
    case GroupMember: otherStr = otherGroup + "␣and␣a␣related␣individual"
    case GroupWhole:  otherStr = otherGroup
  otherStr += f"␣({','.join(otherTokens)})"

  verbStr = "had/" if Past in implTemporality else "./"
  isPresent = Present in implTemporality
  match (isPresent, other):
    case (true, Personal):  verbStr += "has/"
    case (true, _):         verbStr += "have/"
    case (false, _):        verbStr += "./"
  verbStr += "will␣have" if Future in implTemporality else "."

  polarityStr = implPolarity.toString()

  match implTopic:
    case Situational: topicStr = "circumstances"
    case Harm:        topicStr = "harm␣come␣to␣them"
    case Qualitative: topicStr = "nature"
    case SubHuman:    topicStr = "lack␣of␣humanity"
    case Behavioral:  topicStr = "behavior"
    default:          topicStr = "property"

  comparisonStr = "compared␣to␣" + otherStr if hasOther else ""

  return f"{subjectStr}␣{verbStr}␣{polarityStr}␣{topicStr}␣{comparisonStr}"
```

Pseudocode 1: Generation of summary sentence.

Appendix C. Annotator Information

Table 14. : Experience of annotators with toxic language. The scale for experience with toxicity is ‘None’, ‘Little’, ‘Average’, ‘Extensive’, or ‘Unsure’.

		Experience with Toxicity		
	ID	**Targeted By**	**Viewed**	**Position / Program**
Dutch	D1	None	Yes	Staff
	D2	None	Yes	Staff
English |	D3/E1	None	Yes	PhD Candidate
English	E2	-	-	PhD Candidate
	E3	None	Extensive	Staff
Spanish	S1	None	Little	BSc Artificial Intelligence
	S2	Extensive	Unsure	BSc Artificial Intelligence
	S3	None	Average	MSc Linguistics
German	G1	None	Little	MA Humanities Research
	G2	Little	Little	BSc Articifial Intelligence
	G3	Average	None	BSc Articifial Intelligence
Turkish	T1	None	Little	BSc Articifial Intelligence
	T2	Extensive	Extensive	BSc Articifial Intelligence
	T3	None	None	BSc Articifial Intelligence
Arabic	A1	None	Unsure	BSc Articifial Intelligence
	A2	None	Average	BSc Computer Science
	A3	Little	Extensive	BSc Computer Science

Table 15. : Demographic information of annotators. One person annotated both Dutch and English (D3/E1).

	ID	Age	Gender	Country	Native Language	Political Leaning	Religion	Sexuality
Dutch | English	D1	60-66	Female	Netherlands	Dutch	Left	Agnostic	Heterosexual
	D2	60-66	Male	Netherlands	Dutch	Leftwing Liberal	Atheist	Heterosexual
	D3/E1	25-31	Male	Netherlands	Dutch	Social Liberal	Atheist	Heterosexual
	E2	25-31	Female	Iran	Persian	-	-	-
	E3	39-45	Male	Russia	Russian	Liberal	Atheist	Heterosexual
Spanish	S1	25-31	Male	Mexico	English, Spanish	Conservative leaning	Agnostic	Heterosexual
	S2	18-24	Female	Venezuela	Spanish	Libertarian	Atheist	Bisexual
	S3	25-31	Female	Greece	Greek, Spanish	Socialist	Agnostic Atheist	Heterosexual
German	G1	25-31	Female	Germany	German	Left	Agnostic	Queer
	G2	25-31	Female	Germany	Russian, German	Left	None	Heterosexual
	G3	18-24	Male	Germany	German	Progressive conservatism	Atheist	Heterosexual
Turkish	T1	18-24	Male	Turkey	Turkish	Liberal	Islam	Heterosexual
	T2	18-24	Male	Turkey	English, Turkish	Non-political	Islam	Heterosexual
	T3	18-24	Male	Turkey	Turkish	Left-wing & libertarian	Islam	Heterosexual
Arabic	A1	18-24	Female	Morocco	Arabic	None	Islam	Heterosexual
	A2	18-24	Female	Egypt	Arabic	Moderate	Islam	Heterosexual
	A3	18-24	Male	Jordan	Arabic	Moderate / Centrist	Islam	Heterosexual

Table 16. : The nr. of comments for which $N \in \{0, 1, 2, 3\}$ annotators chose the option which continues the annotation, by language.

	N	AR	DE	EN	ES	NL	TR
toxicity	0	192	127	189	147	160	164
	1	101	119	116	85	123	69
	2	114	140	123	89	84	64
	3	99	120	92	174	52	153
justInappropriate	0	216	151	188	172	194	282
	1	135	153	148	109	130	104
	2	109	129	115	106	70	49
	3	46	73	69	108	25	15
hasImplication	0	337	340	254	207	239	331
	1	84	100	132	146	90	61
	2	57	45	87	80	61	39
	3	28	21	47	62	29	19

Appendix D. Early version of Toxic Reasoning Schema (TRS0)

Here we describe the traits that make up an older version (TRS0) of the toxic reasoning schema. It is very similar to the latest version, but less complete, and had a less general way of categorizing the implications. See Table 17 for how we envisioned mapping this version of the schema to the IHC (ElSherief et al. 2021).

The schema's first trait captures the **implication**(s) of a text. It characterizes the proposition that is central to the text's implication. We identify three high-level categories of implications:[k]

- [Negative/Positive]Condition(*group*): a condition (e.g. environment, circumstance, etc.) has, does, or will apply to *group*.
- [Negative/Positive]Quality(*group*): a *group* or a *group*-member possesses a particular quality;
- Inferior(*group1*, *group2*) / Superior(*group2*, *group1*): *group1* is inferior in some way to *group2*.

The first category contains statements about a group being affected by their environment. For example, *being evicted* would be a negative condition. The second category is about the group's inherent qualities, or nature. This could include statements describing groups of people as 'sub-human' or 'vermin'. Finally, the third category involves how groups relate to each other, e.g. suggesting that people of one ethnicity are more intelligent than another.

[k] By implication we mean the property (of a single group) or relation (between groups) that were either implicitly or explicitly applied to a group or groups.

		Threat / Intimidation	Incitement	Grievance	Inferiority		Misinformation	
					∗	Dehuman.	∗	Stereotype
Implications	Something(*group*)	·	·	·	·	·	✓	·
	SomethingNegative(*group*)	·	·	·	·	·	·	·
	NegativeCircumstance(*group*)	·	·	·	·	·	·	·
	Harmed(*target*)	✓	✓	·	·	·	·	·
	Harmed(*in_group*)	·	·	✓	·	·	·	·
	NegativeNature(*group*)	·	·	·	·	·	·	·
	NotHuman(*target*)	·	·	·	·	✓	·	·
	Inferior(*target*, *in_group*)	·	·	·	✓	·	·	·
	SomethingPositive(*group*)	·	·	·	·	·	·	·
	Superior(*in_group*, *target*)	·	·	·	✓	·	·	·
	KnownStereotype(*target*)	·	·	·	·	·	·	✓
	temporality	future	future	·	·	·	·	·
Attitudes	*author_belief*	+	·	+	+	+	+	·
	author_preference	·	+	−	·	·	·	·
	author_accountability	+	−	−	·	·	·	·
	typical_belief	·	·	−	·	·	·	·
	typical_preference	·	·	−	·	·	·	·
	expert_belief	·	·	−	·	·	−	−

Table 17. : Mapping between TRS0 and the IHC () (sub-)classes. Each column represents an IHC (sub-)class, it contains: (1) check marks for the implications that could be involved in a sample of that class; (2) the values for the temporality, specificity, and origin that are required by that class (or '·' for no requirements); and (3) constraints on the possible values for the attitudinal attributes for that class (with '+' indicating a minimum constraint: '> 0.5', '$-$' indicating a maximum constraint: '< 0.5', and '·' indicating no constraints).

Appendix E. ToxiREX Events

Table 18. : Events to which sub-threads in ToxiREX are linked. The training data includes sub-threads for each event. Whether an event is part of the test data is indicated in the 'Test?' column.

Event Title	Test?
2022_European_heatwaves	Yes
2022_FIFA_World_Cup	Yes
2023_Israel–Hamas_war	Yes
2023_Turkey–Syria_earthquakes	Yes
Baldur's_Gate_3	No
Chained_Echoes	No
Chicory:_A_Colorful_Tale	No
Cocoon	No
COVID-19_pandemic	Yes
Cuphead_in_the_Delicious_Last_Course	No
Cyberpunk_2077:_Phantom_Liberty	No
Dave_the_Diver	No
Death_and_state_funeral_of_Elizabeth_II	Yes
Diablo_IV	No
Disco_Elysium:_The_Final_Cut	Yes
Dwarf_Fortress	Yes
Elden_Ring	Yes
Eurovision_Song_Contest_2021	No
Eurovision_Song_Contest_2022	No
Eurovision_Song_Contest_2023	No
Fall_of_Kabul_(2021)	Yes
Final_Fantasy_XIV:_Endwalker	Yes
Forza_Horizon_5	Yes
God_of_War	Yes
God_of_War_Ragnarök	No
Hades	No
Honda_car_recall_2021	No
Honda_car_recall_2023	No
Hyundai_car_recall_2021	Yes
Hyundai_car_recall_2023	No
Jack_Jeanne	No
January_6_United_States_Capitol_attack	Yes
Jeep_car_recall_2023	Yes
Kia_car_recall_2023	Yes
Mahsa_Amini_protests	Yes
Mass_Effect_Legendary_Edition	Yes
Metroid_Prime_Remastered	No
Microsoft_Flight_Simulator	Yes
Neon_White	No
Persona_5_Royal	No
Portal_Companion_Collection	No
Psychonauts_2	Yes
Quake_II_-_Enhanced_Edition	No
Resident_Evil_4	No
Rogue_Legacy_2	Yes
Russian_invasion_of_Ukraine	Yes
Sea_of_Stars	Yes
Spider-Man_2	No
Super_Mario_Bros._Wonder	No
Tesla_car_recall_2023	No
Tetris_Effect:_Connected	Yes
The_House_in_Fata_Morgana_ -_Dreams_of_the_Revenants_Edition	Yes
The_Legend_of_Heroes:_Trails_from_Zero	No
The_Legend_of_Zelda:_Tears_of_the_Kingdom	Yes
The_Stanley_Parable:_Ultra_Deluxe	Yes
The_Witcher_3:_Wild_Hunt_-_Complete_Edition	No
Tony_Hawk's_Pro_Skater_1_+_2	No
Turbo_Overkill	No
Volkswagen_car_recall_2021	No
Walkabout_Mini_Golf	No
Xenoblade_Chronicles_3: _Expansion_Pass_Wave_4_-_Future_Redeemed	Yes

Appendix F. Pydantic Model TRS0

```
class Property(str, Enum):
    harmed = "Harmed(GROUP_A)"
    not_human = "NotHuman(GROUP_A)"
    known_stereotype = "KnownStereotype(GROUP_A)"
    inferior = "Inferior(GROUP_A, GROUP_B)"
    superior = "Superior(GROUP_B, GROUP_A)"
    something_positive = "SomethingPositive(GROUP_A)"
    something_negative = "SomethingNegative(GROUP_A)"
    nothing_implied = "SomethingNeutral(GROUP_A)"

class Temporality(str, Enum):
    past = "past"
    present = "present"
    future = "future"

class ToxicReasoning(BaseModel):
    message_nr: int = Field(description="The number of the message in the thread to which the
      reasoning applies.")

    implication: str = Field(description="If the message seems to imply something about a
     person or a group, give a short description. If there is no clear implication do not
     generate a ToxicReasoning.")

    prop: Property = Field(title="Property", description="The property that the author is
     implying has, will, or does apply to a group. Choose the most specific property that
     clearly applies.")
    prop_span: str = Field(title="Property span", description="The span in the original text
     most indicative of the property being implied.")

    group_a_description: str = Field(title="GROUP_A", description="The group about whom the
     property is asserted, or otherwise the first of two.")
    group_b_description: str = Field(title="GROUP_B", description="An optional second group
     of which something is implied, possibly used as a point of comparison to GROUP_A.")
    group_a_span: str = Field(description="The span in the original text most indicative of
     GROUP_A.")
    group_b_span: str = Field(description="The span in the original text most indicative of
     GROUP_B.")

    when: Temporality = Field(description="At what point in time the property is implied to
     apply to (one of) the groups.")

    author_belief: float = Field(description="[0-1]: Probability that the author believes
     their implication.")
    author_preference: float = Field(description="[0-1]: Probability that the author would
     prefer for their implication to be true.")
    author_responsibility: float = Field(description="[0-1]: Probability that the author is
     or feels personally responsible for the (future) truth of their implication.")
    typical_belief: float = Field(description="[0-1]: Probability that ordinary people would
     believe the author's implication.")
    typical_preference: float = Field(description="[0-1]: Probability that ordinary people
     would prefer for the implication to be true.")
    expert_belief: float = Field(description="[0-1]: Probability that experts would believe
     the author's implication.")

    confidence: float = Field(description="[0-1]: Confidence that the above is correct.")


class ThreadReasonings(BaseModel):
```

```
message_reasonings: list[ToxicReasoning] \
    = Field(description="A list of toxic reasonings describing something said or implied
 in one of the messages in the thread.")
```

Appendix G. Pydantic Model TRS1

```
class GroupRole(str, Enum):
    author = "the author themselves and/or their ingroup"
    participant = "another participant in the conversation and/or the group they belong to"
    outside = "an individual outside of the conversation"
    another = "another group"
    na = "none of the above"


class GroupCharacteristic(str, Enum):
    sexual = "Sexual orientation"
    gender = "Gender"
    disability = "Disability"
    race = "Race/Ethnicity"
    age = "Age"
    religion = "Religion"
    famous = "Famous individual"
    political = "Political affiliation"
    social = "Social belief"
    body_image = "Body image"
    addiction = "Addiction"
    socioeconomic = "Socioeconomic status"
    profession = "Profession"
    nationality = "Nationality"
    other = "Other"
    na = "Not applicable"


class ImplicationCategory(str, Enum):
    a = ("the subject's circumstances, living conditions, physical condition or health,
     general wellbeing, access to "
         "resources, etc.")
    a1 = "some kind of harm coming to the subject"
    b = "the subject's nature, inherent qualities or abilities, etc."
    b1 = "dehumanisation of the subject"
    c = "the subject's choices/decisions, lifestyle, beliefs, etc."
    d = "a non-specific comparison (does not fall under other categories) between the subject
      and the other"
    e = "unclear or none of the above"


class Polarity(str, Enum):
    positive = "Positive"
    neutral = "Neutral"
    negative = "Negative"


class Temporality(str, Enum):
    past = "Past"
    present = "Present"
    future = "Future"


class ToxicReasoning(BaseModel):
  # short description of implication
```

```
    implication: str = Field(description="If the message might imply something toxic about a
       person or a group, give the main implication as a single sentence. Do not describe the
       implication, but simply make explicit that which is implied; i.e. do not start with 'The
       comment/author implies ...', instead the sentence should generally start with the subject
        and continue with what is implied about the subject. If there is no clear implication do
        not generate a ToxicReasoning.")

    # SUBJECT
    subject_descr: str = Field(description="The subject (person or group) of the implication.")
    subject_role: GroupRole = Field(description="The role of the subject in the context of this
        thread.")
    subject_span: str = Field(description="The span in the original text most indicative of the
        subject.")
    subject_characteristic: GroupCharacteristic = Field(description="The characteristic that
       defines the subject group (if the subject is a group).")

    # OTHER
    has_other: bool = Field(description="If there is an 'other' to which the subject is (
       explicitly or implicitly) compared.")
    other_descr: str = Field(description="A description of the optional 'other' person or group
       .")
    other_role: GroupRole = Field(description="Who the other is in the context of this thread."
       )
    other_span: str = Field(description="The span in the original text most indicative of the
       other.")

    # IMPLICATION
    category: ImplicationCategory = Field(description="The category (topic) of the implication.
        Choose the most specific option that clearly applies.")
    impl_span: str = Field(description="The span in the original text most indicative of the
       Implication Category.")
    polarity: Polarity = Field(description="If positive, the main implication being true should
        be something for the subject to be happy about, or proud of. If negative, the main
       implication being true should be something for the subject to be sad about, or ashamed of
       .")
    stereotype: bool = Field(description="If the implication plays into a widely known
       stereotype.")
    sarcasm: bool = Field(description="If the implication is conveyed through sarcasm.")
    when: list[Temporality] = Field(description="At what point(s) in time the implication is
       meant to hold (at least 1).")

    # STAKEHOLDER ATTITUDES
    author_belief: float = Field(description="[0-1]: Probability that the author believes what
       their implication says was/is/will (or should [have] be[en]) the case.")
    author_preference: float = Field(description="[0-1]: Probability that the author prefers
       what their implication says was/is/will (or should [have] be[en]) the case.")
    author_responsibility: float = Field(description="[0-1]: Probability that the author is or
       feels personally responsible for the (past, present, or future) truth of what their
       implication says was/is/will (or should [have] be[en]) the case.")
    typical_belief: float = Field(description="[0-1]: Probability that ordinary people would
       believe what the implication says was/is/will (or should [have] be[en]) the case.")
    typical_preference: float = Field(description="[0-1]: Probability that ordinary people
       would prefer what the implication says was/is/will (or should [have] be[en]) the case.")
    expert_belief: float = Field(description="[0-1]: Probability that experts would believe
       what the implication says was/is/will (or should [have] be[en]) the case.")


class Toxicity(str, Enum):
    yes = "Yes/Maybe"
```

```
    no = "No"


class CommentAnnotation(BaseModel):
    message_nr: int = Field(description="The number of the message in the thread to which
     this annotation applies.")

    is_toxic: Toxicity = Field(description="If this comment might be perceived as toxic.")
    is_only_innapropriate: bool = Field(description="If the comment is only toxic because of
     an inappropriate or toxic word (rather than what is being said/implied).")
    is_counter_speech: bool = Field(description="If the comment is an argument-based counter
     of the previous comment, providing an alternative perspective.")
    toxic_reasoning: Optional[ToxicReasoning] = Field(
        description="If toxic and the toxicity stems from what the text is either explicitly
     or implicitly communicating this should contain a toxic reasoning. If toxic for a
     different reason, it should be null."
    )


class ThreadReasonings(BaseModel):
    comment_annotations: list[CommentAnnotation] = Field(
        description="A list of comment annotations specifying message toxicity."
    )
```

Appendix H. Alternative Fine-tuning Strategy

We explored an alternative fine-tuning strategy where a language model is expected to produce an instance of the pydantic model in the same way that GPT4o does.

We prepared supervised learning examples by adding sub-threads to a prompt and appending the silver-standard annotations of GPT4o. The annotations were presented as instances of the Pydantic model converted to JSON format, with the JSON schema of the pydantic model being prepended to the prompt. Thus, the model was expected to take its instructions from the JSON schema, and then produce a JSON output which adheres to it. During inference, we used structured outputs (Willard and Louf 2023) to force the model to produce valid instances of the schema. We used Gemma-3-270m-it (Team et al. 2025), DeepSeek-Qwen-2.5-1.5b (Guo et al. 2025), and GPT-OSS-20b (in 4-bit precision) (OpenAI et al. 2025). These models were trained with Low-rank Adaption (LoRA) (Hu et al. 2022) parameter-efficient fine-tuning. We also tried Gemma-3-270m-it model with ordinary all-parameter fine-tuning.

Unfortunately, these preliminary experiments did not result in a model with better performance than the baseline presented in the main text. We expect that a strategy like this one could still succeed. However, since our primary goal was to produce a simple baseline and our preliminary experiments did not succeed, we leave this for future work.