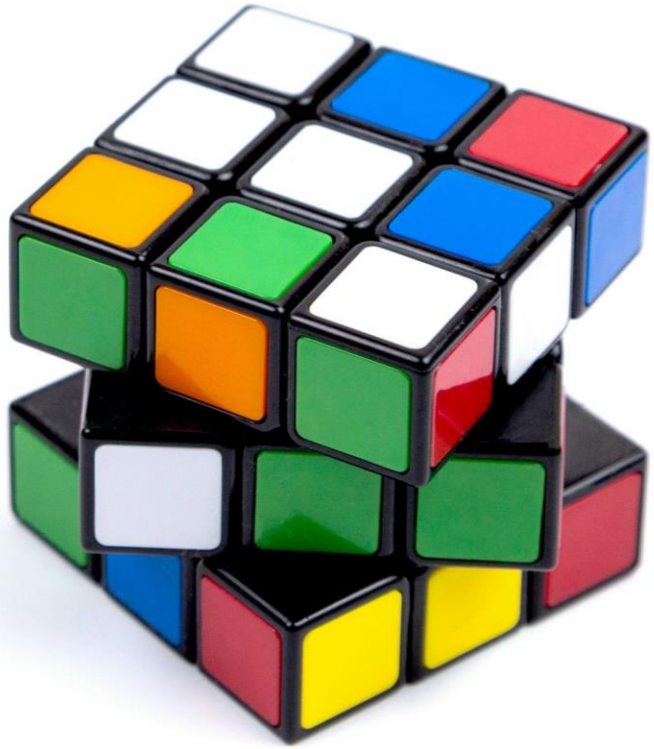
Trigger Warnings in Fanfiction

An Analysis of Usage Consistency and the Effect of Prescriptive Annotation Guidelines

Sebastian Heineking

04.07.2024

Motivation



NLP Tasks with room for subjective judgements are difficult to label

Davani et al. (2021) found personal biases of annotators and societal stereotypes to influence labeling behavior

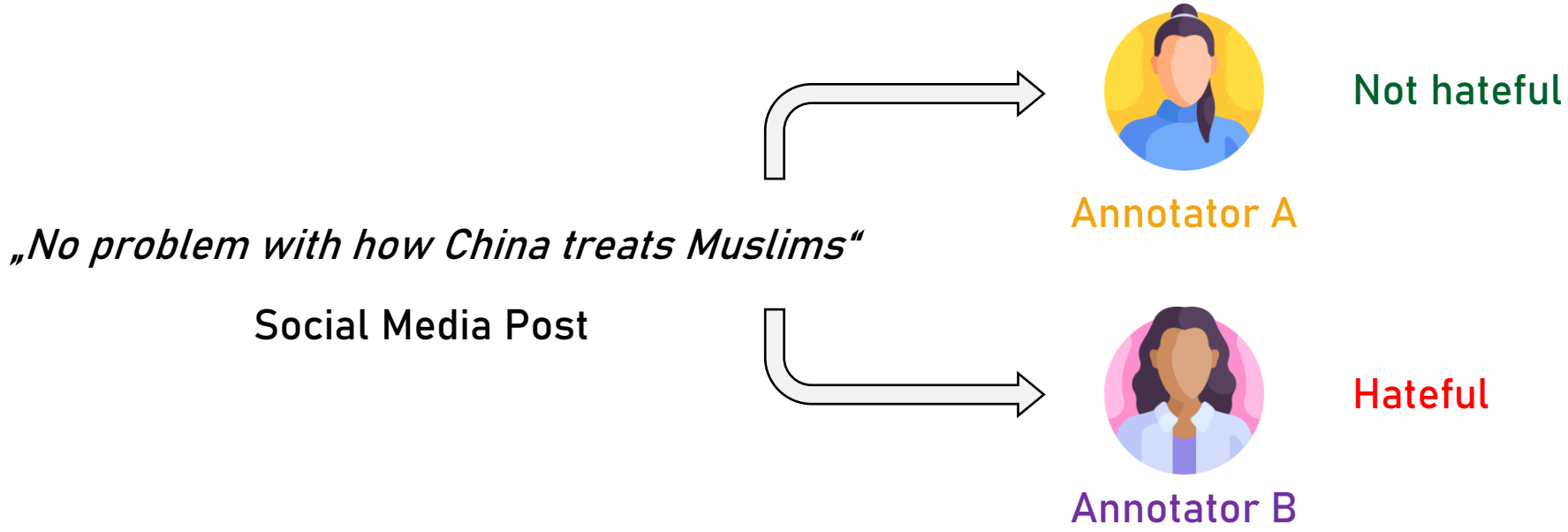
„No problem with how China treats Muslims“

Social Media Post

Röttger et al. (2022), Davani et al. (2021)

NLP Tasks with room for subjective judgements are difficult to label

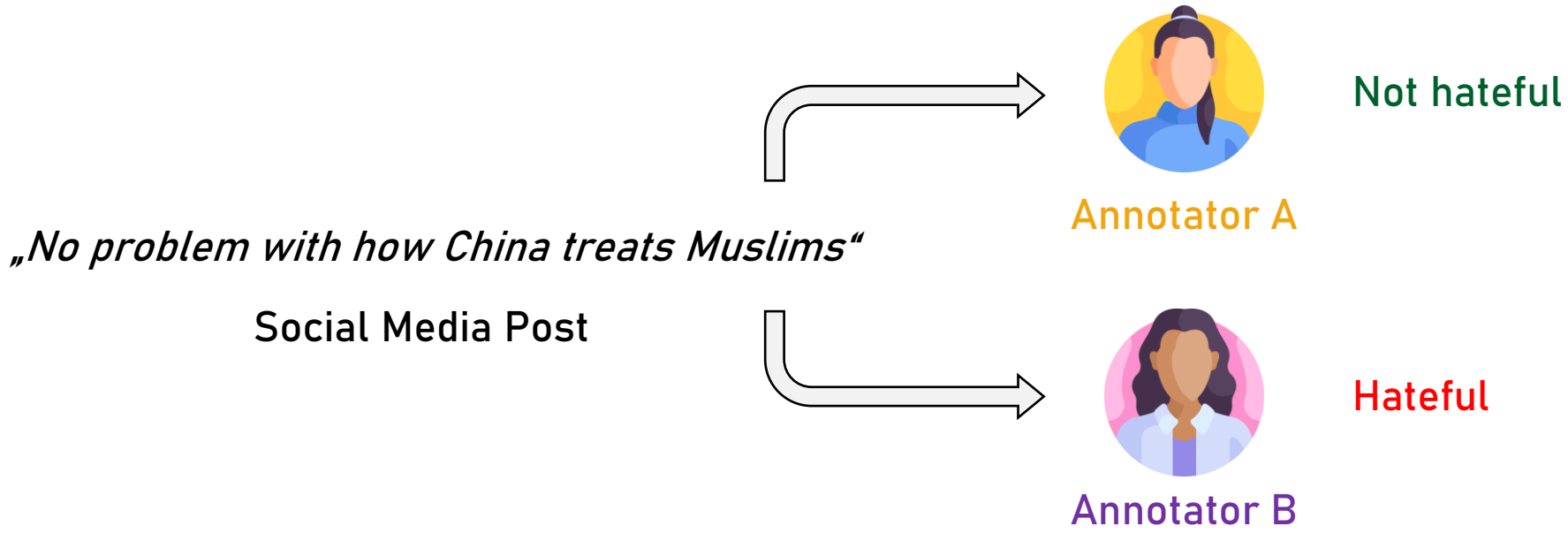
Davani et al. (2021) found personal biases of annotators and societal stereotypes to influence labeling behavior



Röttger et al. (2022), Davani et al. (2021)

NLP Tasks with room for subjective judgements are difficult to label

Davani et al. (2021) found personal biases of annotators and societal stereotypes to influence labeling behavior



Problem: Subjective labeling can produce inconsistent datasets which are unsuited for machine learning

Röttger et al. (2022), Davani et al. (2021)

Röttger et al. (2022) propose two contrasting annotation paradigms

The descriptive and prescriptive paradigm are suited for different goals in dataset creation

Imagine you come across the post below on social media. Do you personally feel this post is hateful?

Descriptive Paradigm

Röttger et al. (2022)

Röttger et al. (2022) propose two contrasting annotation paradigms

The descriptive and prescriptive paradigm are suited for different goals in dataset creation

Imagine you come across the post below on social media. Do you personally feel this post is hateful?

Descriptive Paradigm

- Ask annotators to make subjective judgements
- Encode different, personal beliefs in the dataset

→ Study annotator beliefs

Röttger et al. (2022)

Röttger et al. (2022) propose two contrasting annotation paradigms

The descriptive and prescriptive paradigm are suited for different goals in dataset creation

Imagine you come across the post below on social media. Do you personally feel this post is hateful?

Descriptive Paradigm

- Ask annotators to make subjective judgements
- Encode different, personal beliefs in the dataset

→ Study annotator beliefs

Imagine you come across the post below on social media. Does this post meet the criteria for hate speech?

Prescriptive Paradigm

Röttger et al. (2022)

Röttger et al. (2022) propose two contrasting annotation paradigms

The descriptive and prescriptive paradigm are suited for different goals in dataset creation

Imagine you come across the post below on social media. Do you personally feel this post is hateful?

Descriptive Paradigm

- Ask annotators to make subjective judgements
- Encode different, personal beliefs in the dataset

→ Study annotator beliefs

Imagine you come across the post below on social media. Does this post meet the criteria for hate speech?

Prescriptive Paradigm

- Define annotation guidelines for one specific belief
- Encode one belief consistently in the dataset

→ Train a model to apply the belief

Röttger et al. (2022)

Röttger et al. (2022) propose two contrasting annotation paradigms

The descriptive and prescriptive paradigm are suited for different goals in dataset creation

Imagine you come across the post below on social media. Do you personally feel this post is hateful?

Descriptive Paradigm

- Ask annotators to make subjective judgements
- Encode different, personal beliefs in the dataset

→ Study annotator beliefs

Imagine you come across the post below on social media. Does this post meet the criteria for hate speech?

Prescriptive Paradigm

- Define annotation guidelines for one specific belief
- Encode one belief consistently in the dataset

→ Train a model to apply the belief

Röttger et al. (2022)

Central Question of my Thesis:

Can prescriptive annotation guidelines increase annotator agreements when labeling texts for trigger warnings?

Part 1:
Do authors on Archive of Our Own apply warning tags in a consistent fashion?

Logo source: https://upload.wikimedia.org/wikipedia/commons/8/88/Archive_of_Our_Own_logo.png;

Central Question of my Thesis:

Can prescriptive annotation guidelines increase annotator agreements when labeling texts for trigger warnings?

Part 1:
Do authors on Archive of Our Own apply warning tags in a consistent fashion?

Part 2:
What is the effect of prescriptive annotation guidelines in trigger warning annotation?

Logo source: https://upload.wikimedia.org/wikipedia/commons/8/88/Archive_of_Our_Own_logo.png;

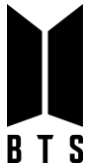
Wiegmann et al. (2023) collected stories on Archive of our Own (AO3)

The authors on Archive of our Own (AO3) assign a range of different tags to their stories

618,138 works

470,326 works

303,570 works



206,231 works

On AO3, people write and share stories based on popular media such as movies, books, anime, or music.

Logo sources (top to bottom): https://upload.wikimedia.org/wikipedia/commons/b/b9/Marvel_Logo.svg; https://upload.wikimedia.org/wikipedia/commons/c/c9/Naruto_logo.svg; https://upload.wikimedia.org/wikipedia/commons/6/6e/Harry_Potter_wordmark.svg; https://upload.wikimedia.org/wikipedia/commons/e/ef/BTS_logo.svg

Wiegmann et al. (2023) collected stories on Archive of our Own (AO3)

The authors on Archive of our Own (AO3) assign a range of different tags to their stories

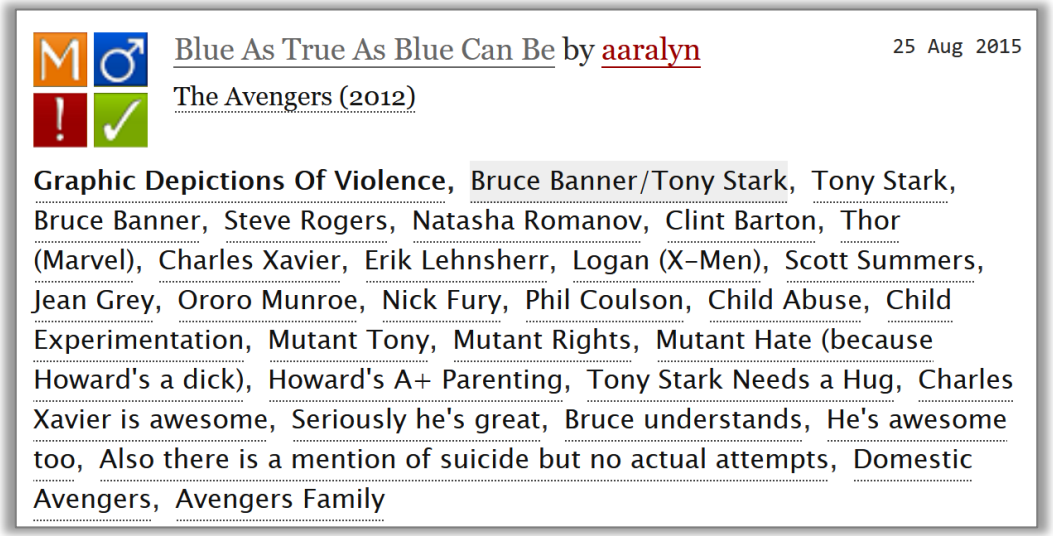
618,138 works

470,326 works

303,570 works

206,231 works

On AO3, people write and share stories based on popular media such as movies, books, anime, or music.



Blue As True As Blue Can Be by aaralyn

25 Aug 2015

The Avengers (2012)

Graphic Depictions Of Violence, Bruce Banner/Tony Stark, Tony Stark, Bruce Banner, Steve Rogers, Natasha Romanov, Clint Barton, Thor (Marvel), Charles Xavier, Erik Lehnsherr, Logan (X-Men), Scott Summers, Jean Grey, Ororo Munroe, Nick Fury, Phil Coulson, Child Abuse, Child Experimentation, Mutant Tony, Mutant Rights, Mutant Hate (because Howard's a dick), Howard's A+ Parenting, Tony Stark Needs a Hug, Charles Xavier is awesome, Seriously he's great, Bruce understands, He's awesome too, Also there is a mention of suicide but no actual attempts, Domestic Avengers, Avengers Family

https://archiveofourown.org/works/648087

Logo sources (top to bottom): https://upload.wikimedia.org/wikipedia/commons/b/b9/Marvel_Logo.svg; https://upload.wikimedia.org/wikipedia/commons/c/c9/Naruto_logo.svg; https://upload.wikimedia.org/wikipedia/commons/6/6e/Harry_Potter_wordmark.svg; https://upload.wikimedia.org/wikipedia/commons/e/ef/BTS_logo.svg

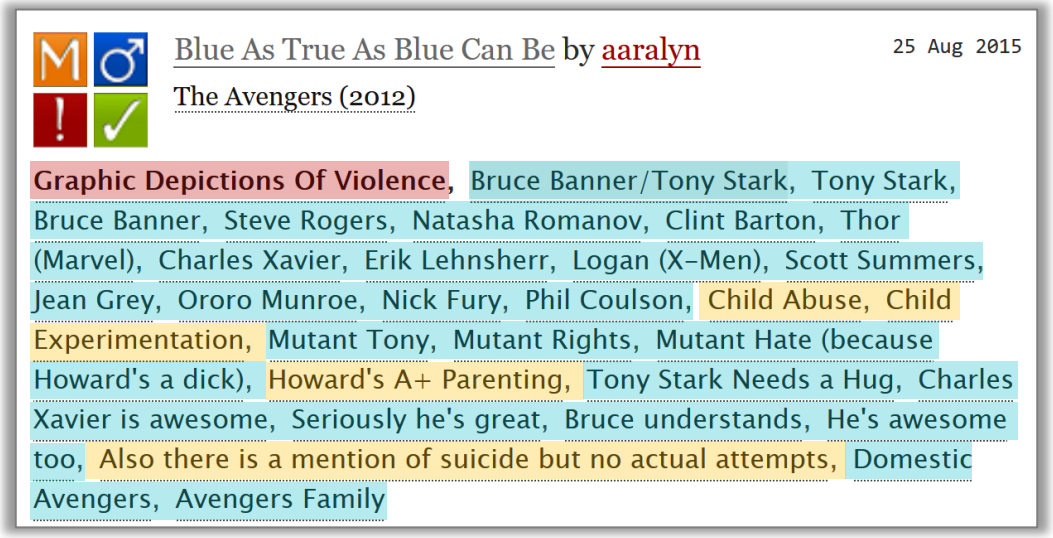
Wiegmann et al. (2023) collected stories on Archive of our Own (AO3)

The authors on Archive of our Own (AO3) assign a range of different tags to their stories

Archive Warnings – Set of 5 “official” warnings

Freeform tags about the fandom/characters

Freeform tags that contain a warning



Blue As True As Blue Can Be by aaralyn

25 Aug 2015

The Avengers (2012)

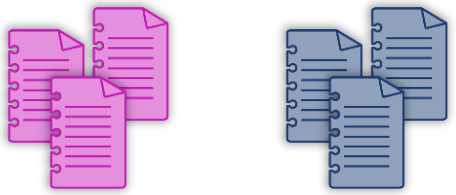
Graphic Depictions Of Violence, Bruce Banner/Tony Stark, Tony Stark, Bruce Banner, Steve Rogers, Natasha Romanov, Clint Barton, Thor (Marvel), Charles Xavier, Erik Lehnsherr, Logan (X-Men), Scott Summers, Jean Grey, Ororo Munroe, Nick Fury, Phil Coulson, Child Abuse, Child Experimentation, Mutant Tony, Mutant Rights, Mutant Hate (because Howard's a dick), Howard's A+ Parenting, Tony Stark Needs a Hug, Charles Xavier is awesome, Seriously he's great, Bruce understands, He's awesome too, Also there is a mention of suicide but no actual attempts, Domestic Avengers, Avengers Family

https://archiveofourown.org/works/648087

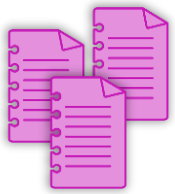
Usage Consistency by Authors

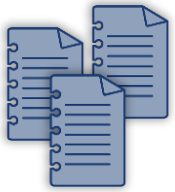
Hypothesis 1:

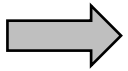
The authors on AO3 apply warning tags in a way that is consistent with common language understanding

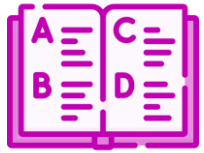


1. Collect Corpora of Documents
Documents tagged for a warning
vs. Baseline Documents



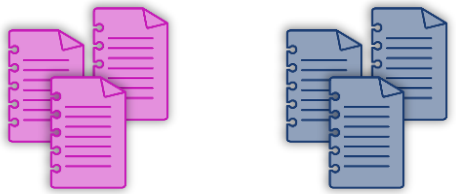


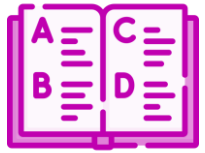


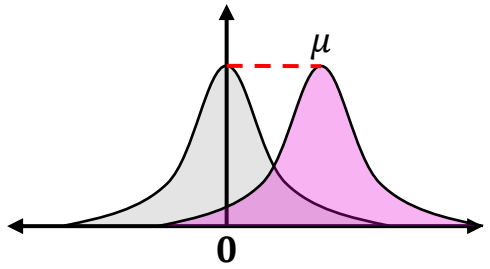


1. Collect Corpora of Documents
Documents tagged for a warning
vs. Baseline Documents

2. Define a Vocabulary
Expected Terms for a Warning







1. Collect Corpora of Documents
Documents tagged for a warning
vs. Baseline Documents

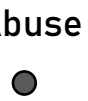
2. Define a Vocabulary
Expected Terms for a Warning

3. Apply Statistical Tests
Test the Term Frequencies

Usage consistency was analyzed for three categories of abuse

1. Collect 5,654 tags associated with abuse

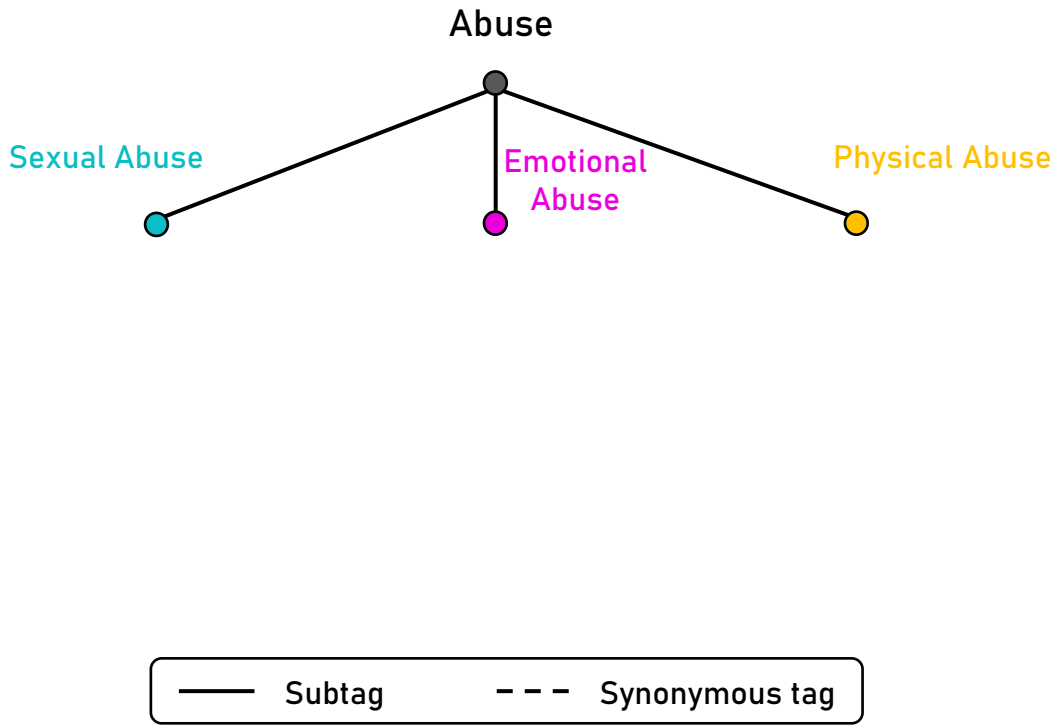
Tag graph from Wiegmann et al. (2023) based on relations recorded by the AO3 community



Usage consistency was analyzed for three categories of abuse

1. Collect 5,654 tags associated with abuse

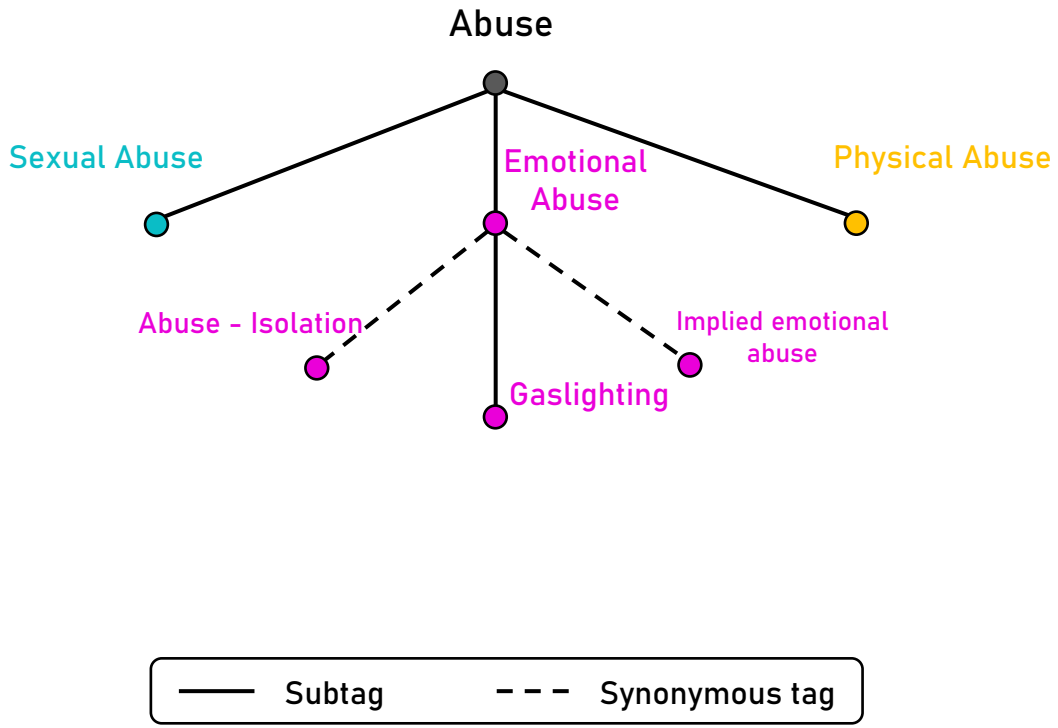
Tag graph from Wiegmann et al. (2023) based on relations recorded by the AO3 community



Usage consistency was analyzed for three categories of abuse

1. Collect 5,654 tags associated with abuse

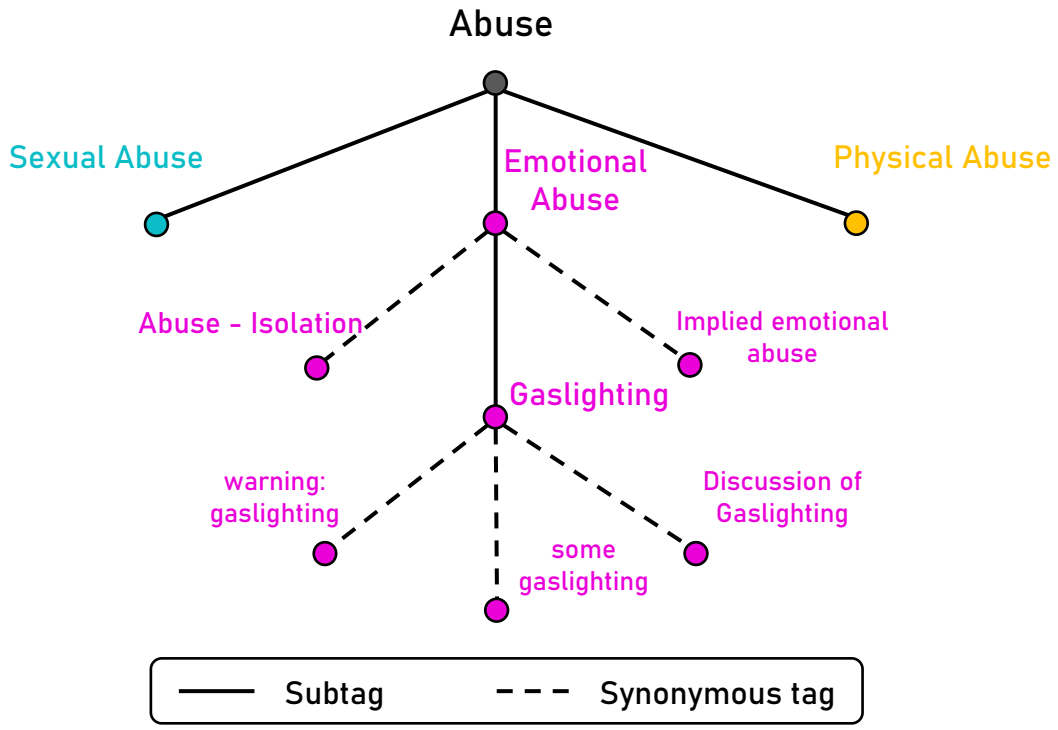
Tag graph from Wiegmann et al. (2023) based on relations recorded by the AO3 community



Usage consistency was analyzed for three categories of abuse

1. Collect 5,654 tags associated with abuse

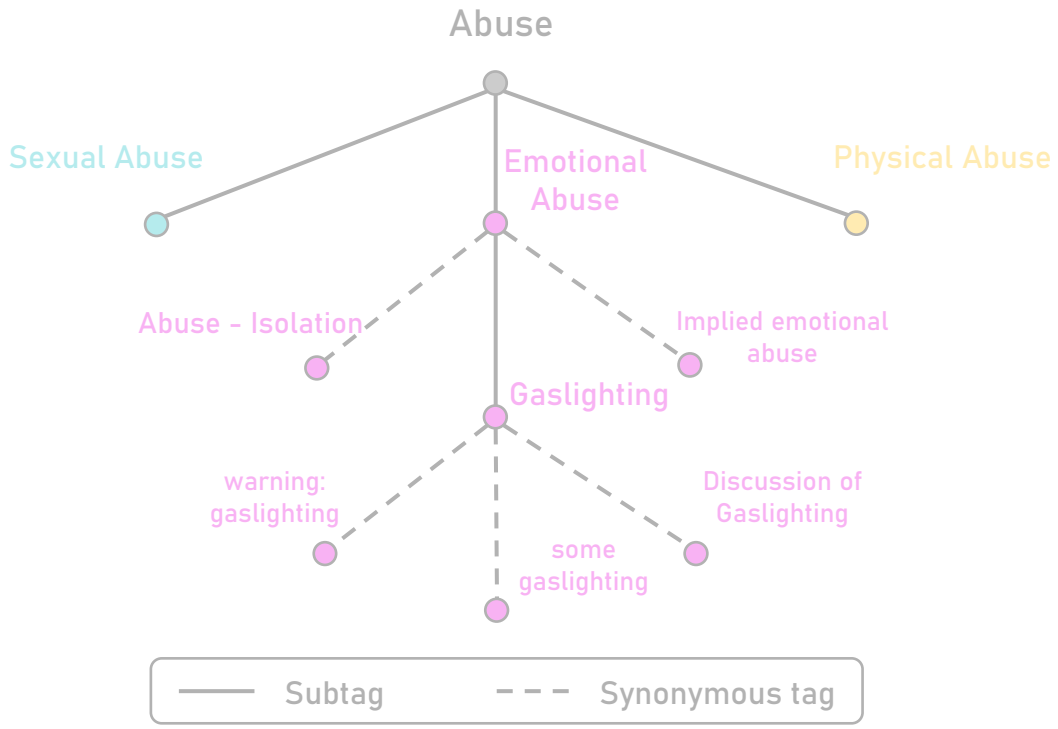
Tag graph from Wiegmann et al. (2023) based on relations recorded by the AO3 community



Usage consistency was analyzed for three categories of abuse

1. Collect 5,654 tags associated with abuse

Tag graph from Wiegmann et al. (2023) based on relations recorded by the AO3 community



2. Label tags manually for contained categories

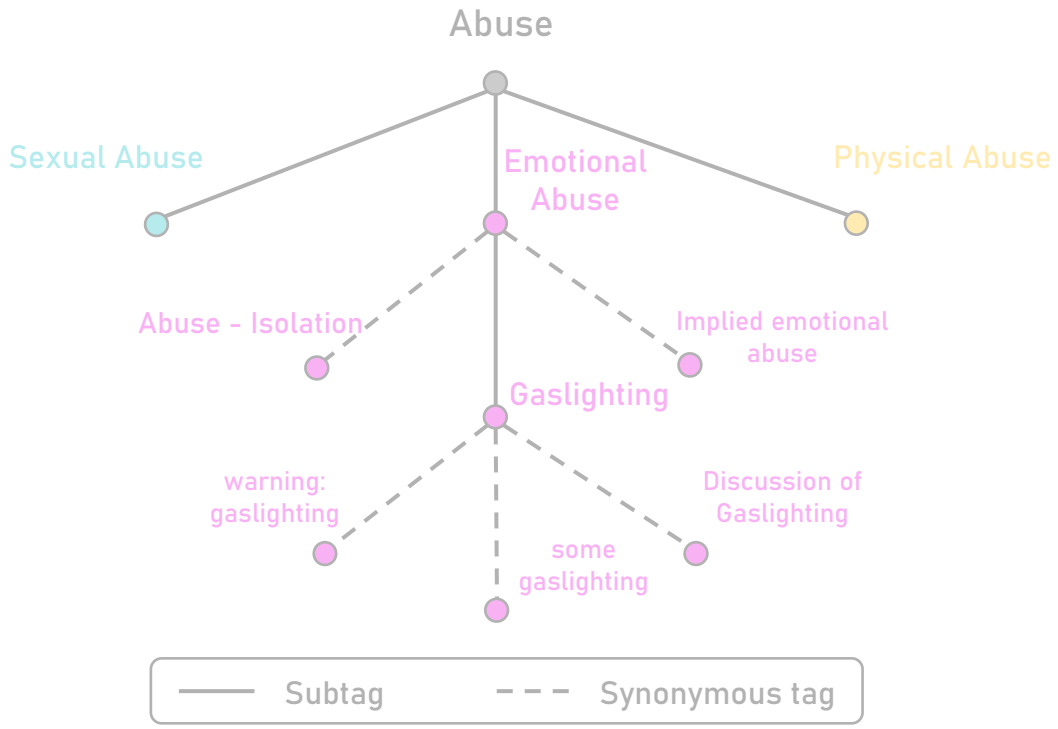
Tags can cover multiple categories

CW: Psychological abuse by a parent

Usage consistency was analyzed for three categories of abuse

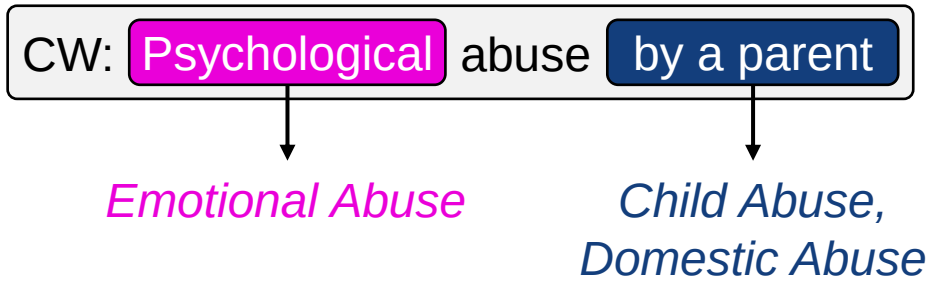
1. Collect 5,654 tags associated with abuse

Tag graph from Wiegmann et al. (2023) based on relations recorded by the AO3 community



2. Label tags manually for contained categories

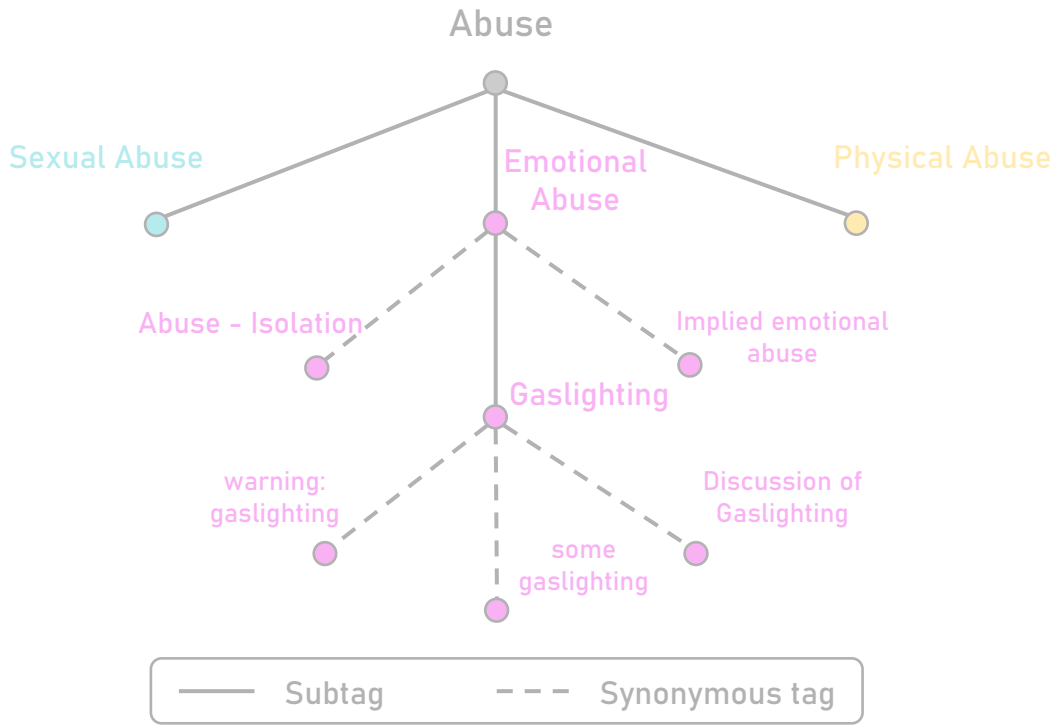
Tags can cover multiple categories



Usage consistency was analyzed for three categories of abuse

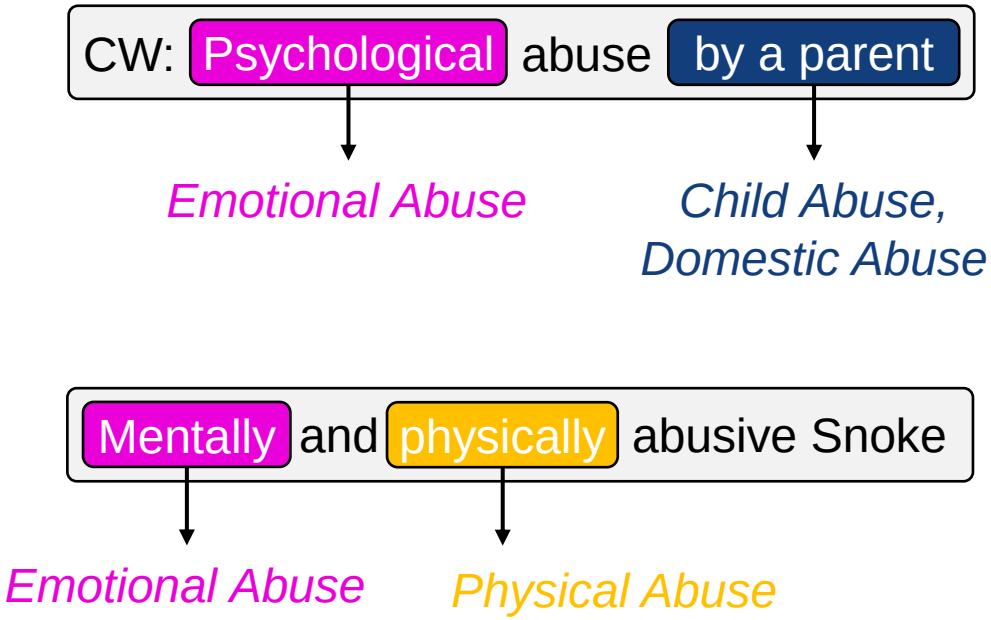
1. Collect 5,654 tags associated with abuse

Tag graph from Wiegmann et al. (2023) based on relations recorded by the AO3 community



2. Label tags manually for contained categories

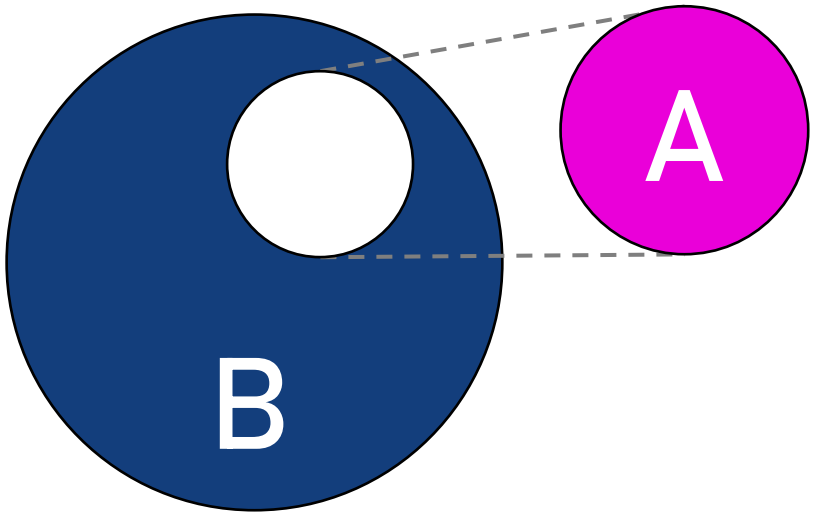
Tags can cover multiple categories



The consistency of each category was tested on the tagged chapters

3. Create a category corpus and a baseline

The corpora contain chapters with specified tags



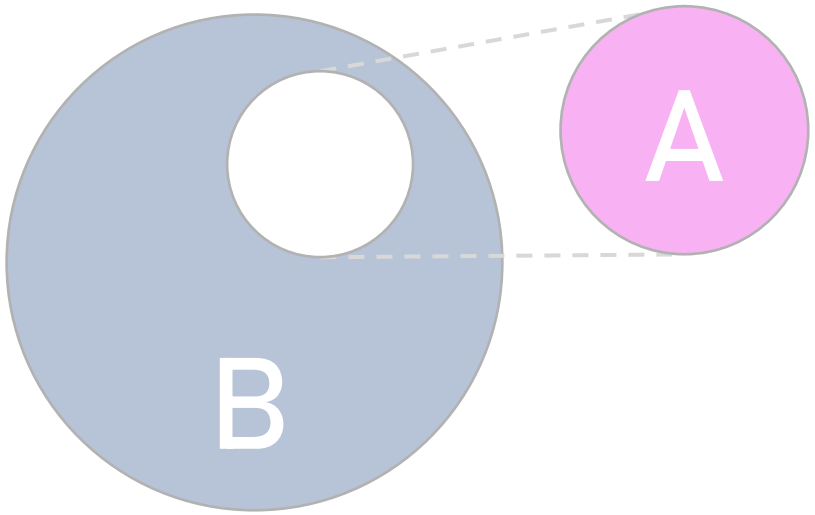
A: Chapters with an emotional abuse tag

B: Chapters with an abuse tag != emotional abuse

The consistency of each category was tested on the tagged chapters

3. Create a category corpus and a baseline

The corpora contain chapters with specified tags



A: Chapters with an emotional abuse tag

B: Chapters with an abuse tag != emotional abuse

4. Collect vocabularies of expected terms

$$T_{Emotional\ Abuse} = \{\text{hurt (V), angry (A), guilt (N), ...}\}$$

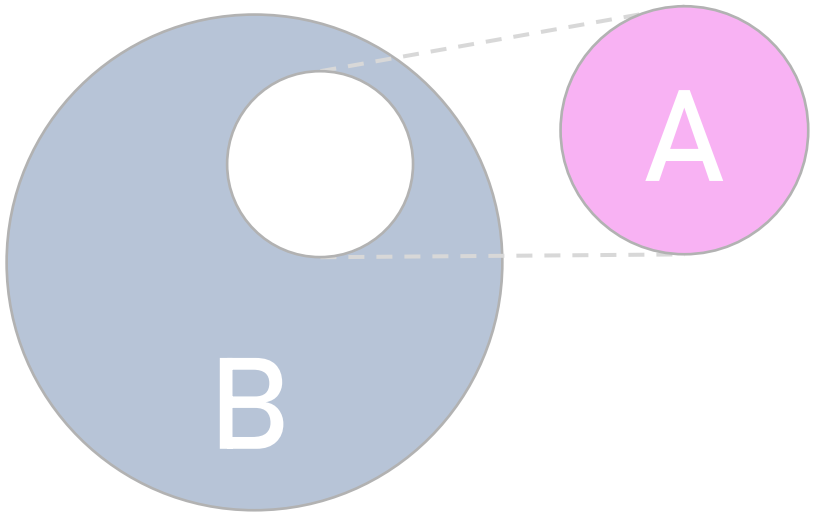
$$T_{Physical\ Abuse} = \{\text{beat (V), broken (A), bruise (N), ...}\}$$

$$T_{Sexual\ Abuse} = \{\text{molest (V), sexual (A), consent (N), ...}\}$$

The consistency of each category was tested on the tagged chapters

3. Create a category corpus and a baseline

The corpora contain chapters with specified tags



A: Chapters with an emotional abuse tag

B: Chapters with an abuse tag != emotional abuse

4. Collect vocabularies of expected terms

$$T_{Emotional\ Abuse} = \{\text{hurt (V), angry (A), guilt (N), ...}\}$$

$$T_{Physical\ Abuse} = \{\text{beat (V), broken (A), bruise (N), ...}\}$$

$$T_{Sexual\ Abuse} = \{\text{molest (V), sexual (A), consent (N), ...}\}$$

Sources for initial vocabulary

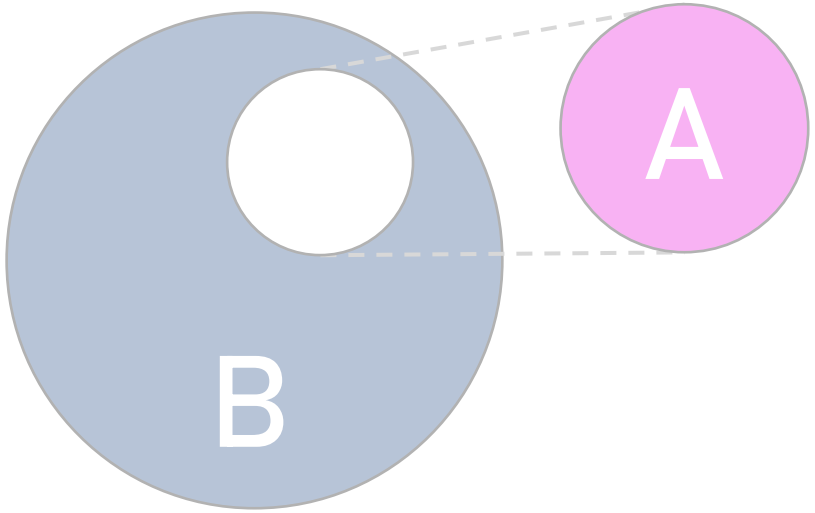
- Social Care Institute for Excellence[1]
- Washington State Department of Social and Health Services[2]

1: https://www.scie.org.uk/safeguarding/adults/introduction/types-and-indicators-of-abuse
2: https://www.dshs.wa.gov/altsa/home-and-community-services/types-and-signs-abuse

The consistency of each category was tested on the tagged chapters

3. Create a category corpus and a baseline

The corpora contain chapters with specified tags



A: Chapters with an emotional abuse tag

B: Chapters with an abuse tag != emotional abuse

4. Collect vocabularies of expected terms

$$T_{Emotional\ Abuse} = \{\text{hurt (V), angry (A), guilt (N), ...}\}$$

$$T_{Physical\ Abuse} = \{\text{beat (V), broken (A), bruise (N), ...}\}$$

$$T_{Sexual\ Abuse} = \{\text{molest (V), sexual (A), consent (N), ...}\}$$

Sources for initial vocabulary

- Social Care Institute for Excellence[1]
- Washington State Department of Social and Health Services[2]

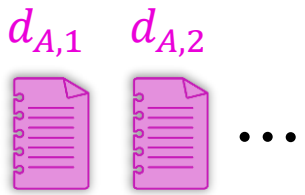
Expansion of vocabulary

- Synonyms using thesauri and GPT-4
- Syntactic categories (Adjectives, Nouns, Verbs) with the same word stem

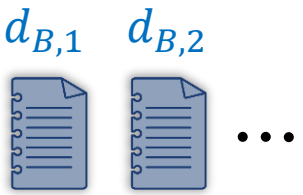
1: https://www.scie.org.uk/safeguarding/adults/introduction/types-and-indicators-of-abuse
2: https://www.dshs.wa.gov/altsa/home-and-community-services/types-and-signs-abuse

1. Method: Mann-Whitney U-Test on chapter frequencies

Calculate term frequencies for t for all chapters



$\mathrm{tf}(t, d_{A,1}), \mathrm{tf}(t, d_{A,2}), \ldots$



$\mathrm{tf}(t, d_{B,1}), \mathrm{tf}(t, d_{B,2}), \ldots$

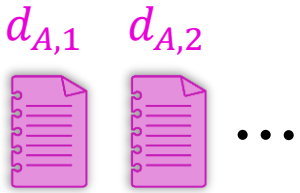
$$\mathrm{tf}(t, d_A) = \frac{\mathrm{f}_{t,d_A}}{\sum_{t' \in d_A} \mathrm{f}_{t',d_A}}, \qquad d_A \in A$$

- Term t with raw count f_{t,d_A} in chapter d_A
- Chapter d_A from corpus A
- Chapter d_B from corpus B

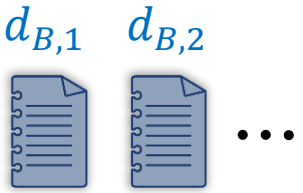
Lijffijt et al. (2014); Icons from Flaticon.com

1. Method: Mann-Whitney U-Test on chapter frequencies

Calculate term frequencies for t for all chapters



$\mathrm{tf}(t, d_{A,1}), \mathrm{tf}(t, d_{A,2}), \ldots$



$\mathrm{tf}(t, d_{B,1}), \mathrm{tf}(t, d_{B,2}), \ldots$

$$\mathrm{tf}(t, d_A) = \frac{\mathrm{f}_{t,d_A}}{\sum_{t' \in d_A} \mathrm{f}_{t',d_A}}, \qquad d_A \in A$$

- Term t with raw count f_{t,d_A} in chapter d_A
- Chapter d_A from corpus A
- Chapter d_B from corpus B

Rank chapters by the term frequencies (ascending)

Conclusion:
Term t is less likely in corpus A

Statistics:

- U_1/U_2: No. of pairs in which d_A/ d_B has a lower term frequency
- z: Standardized U (for large samples)

Lijffijt et al. (2014); Icons from Flaticon.com

1. Method: Mann-Whitney U-Test on chapter frequencies

Calculate term frequencies for t for all chapters

$d_{A,1}$ $d_{A,2}$ …

$\mathrm{tf}(t, d_{A,1}), \mathrm{tf}(t, d_{A,2}), \ldots$

$d_{B,1}$ $d_{B,2}$ …

$\mathrm{tf}(t, d_{B,1}), \mathrm{tf}(t, d_{B,2}), \ldots$

$$\mathrm{tf}(t, d_A) = \frac{\mathrm{f}_{t,d_A}}{\sum_{t' \in d_A} \mathrm{f}_{t',d_A}}, \qquad d_A \in A$$

- Term t with raw count f_{t,d_A} in chapter d_A
- Chapter d_A from corpus A
- Chapter d_B from corpus B

Rank chapters by the term frequencies (ascending)

Conclusion:
Term t is less likely in corpus A

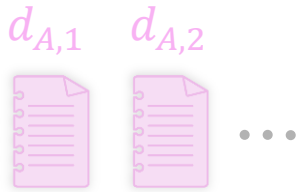
Conclusion:
Term t is equally likely for both

Statistics:
- U_1/U_2: No. of pairs in which d_A/ d_B has a lower term frequency
- z: Standardized U (for large samples)

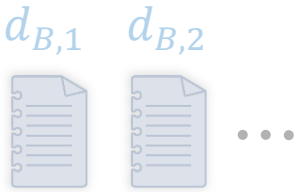
Lijffijt et al. (2014); Icons from Flaticon.com

1. Method: Mann-Whitney U-Test on chapter frequencies

Calculate term frequencies for t for all chapters



$\mathrm{tf}(t, d_{A,1}), \mathrm{tf}(t, d_{A,2}), \ldots$



$\mathrm{tf}(t, d_{B,1}), \mathrm{tf}(t, d_{B,2}), \ldots$

Rank chapters by the term frequencies (ascending)

Conclusion:
Term t is less likely in corpus A

Conclusion:
Term t is equally likely for both

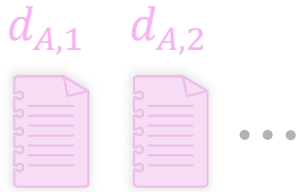
Strengths

- Shows which words occur significantly more
- Does not overestimate significance by assuming word independence

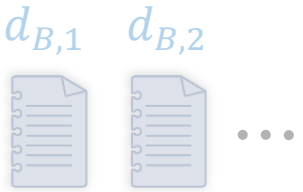
Lijffijt et al. (2014); Icons from Flaticon.com

1. Method: Mann-Whitney U-Test on chapter frequencies

Calculate term frequencies for t for all chapters



$\mathrm{tf}(t, d_{A,1}), \mathrm{tf}(t, d_{A,2}), \ldots$



$\mathrm{tf}(t, d_{B,1}), \mathrm{tf}(t, d_{B,2}), \ldots$

Rank chapters by the term frequencies (ascending)

Conclusion:
Term t is less likely in corpus A

Conclusion:
Term t is equally likely for both

Strengths

- Shows which words occur significantly more
- Does not overestimate significance by assuming word independence

Weaknesses

- Prefers “common” words that occur in a lot of chapters (esp. on larger corpora)
- Does not yield effect size, i.e. “how much more“ frequent a word is

Lijffijt et al. (2014); Icons from Flaticon.com

2. Method: Log ratio on corpus frequencies

Calculate corpus-level term frequencies TF for t

$$\mathrm{TF}(t, A) = \frac{\sum_{d_A \in A} \mathrm{f}_{t,d_A}}{\sum_{d_A \in A} \sum_{t' \in d_A} \mathrm{f}_{t',d_A}} \qquad \mathrm{TF}(t, B) = \frac{\sum_{d_B \in B} \mathrm{f}_{t,d_B}}{\sum_{d_B \in B} \sum_{t' \in d_B} \mathrm{f}_{t',d_B}}$$

Hardie (2014)

2. Method: Log ratio on corpus frequencies

Calculate corpus-level term frequencies TF for t

$$\mathrm{TF}(t, A) = \frac{\sum_{d_A \in A} \mathrm{f}_{t,d_A}}{\sum_{d_A \in A} \sum_{t\prime \in d_A} \mathrm{f}_{t\prime,d_A}} \qquad \mathrm{TF}(t, B) = \frac{\sum_{d_B \in B} \mathrm{f}_{t,d_B}}{\sum_{d_B \in B} \sum_{t\prime \in d_B} \mathrm{f}_{t\prime,d_B}}$$

Calculate the binary logarithm of the ratio of frequencies

$$lr(t) = \log_2 \left(\frac{\mathrm{TF}(t, A)}{\mathrm{TF}(t, B)} \right)$$

- $lr(t) = 0$ t is equally likely for A and B
- $lr(t) = 1$ t is twice as likely for A
- $lr(t) = 2$ t is four times as likely for A

Hardie (2014)

2. Method: Log ratio on corpus frequencies

Calculate corpus-level term frequencies TF for t

$$\mathrm{TF}(t, A) = \frac{\sum_{d_A \in A} \mathrm{f}_{t,d_A}}{\sum_{d_A \in A} \sum_{t\prime \in d_A} \mathrm{f}_{t\prime,d_A}} \qquad \mathrm{TF}(t, B) = \frac{\sum_{d_B \in B} \mathrm{f}_{t,d_B}}{\sum_{d_B \in B} \sum_{t\prime \in d_B} \mathrm{f}_{t\prime,d_B}}$$

Calculate the binary logarithm of the ratio of frequencies

$$lr(t) = \log_2 \left(\frac{\mathrm{TF}(t, A)}{\mathrm{TF}(t, B)} \right)$$

- $lr(t) = 0$ t is equally likely for A and B
- $lr(t) = 1$ t is twice as likely for A
- $lr(t) = 2$ t is four times as likely for A

Strengths

- Shows the effect size for a word
- Able to identify rare words with high difference in frequencies

Hardie (2014)

2. Method: Log ratio on corpus frequencies

Calculate corpus-level term frequencies TF for t

$$\mathrm{TF}(t, A) = \frac{\sum_{d_A \in A} \mathrm{f}_{t,d_A}}{\sum_{d_A \in A} \sum_{t\prime \in d_A} \mathrm{f}_{t\prime,d_A}} \qquad \mathrm{TF}(t, B) = \frac{\sum_{d_B \in B} \mathrm{f}_{t,d_B}}{\sum_{d_B \in B} \sum_{t\prime \in d_B} \mathrm{f}_{t\prime,d_B}}$$

Calculate the binary logarithm of the ratio of frequencies

$$lr(t) = \log_2 \left(\frac{\mathrm{TF}(t, A)}{\mathrm{TF}(t, B)} \right)$$

- $lr(t) = 0$ t is equally likely for A and B
- $lr(t) = 1$ t is twice as likely for A
- $lr(t) = 2$ t is four times as likely for A

Strengths

- Shows the effect size for a word
- Able to identify rare words with high difference in frequencies

Weaknesses

- Does not yield significance
- Prefers rare words; High log ratios can result from a few documents

Hardie (2014)

Test hypotheses for both approaches using the vocabulary

Mann-Whitney U-Test
For each term $t \in T_{C_i}$, test the following hypotheses:

$$H_0: F_A(x) = F_B(x), \qquad H_A: F_A(x) \neq F_B(x)$$

- $F_A(x), F_B(x)$: CDF of term frequencies $tf(t,d)$ for documents $d_A \in A$, $d_B \in B$

Test hypotheses for both approaches using the vocabulary

Mann-Whitney U-Test
For each term $t \in T_{C_i}$, test the following hypotheses:

$$H_0: F_A(x) = F_B(x), \qquad H_A: F_A(x) \neq F_B(x)$$

- $F_A(x), F_B(x)$: CDF of term frequencies $tf(t,d)$ for documents $d_A \in A$, $d_B \in B$

1. Calculate p-values from z-scores
2. Define significance level $\alpha = 0.05$
3. Bonferroni-correction to account for multiplicity

$$p_t \leq \frac{\alpha}{|T_{C_i}|}, \qquad t \in T_{C_i}$$

Test hypotheses for both approaches using the vocabulary

Mann-Whitney U-Test
For each term $t \in T_{C_i}$, test the following hypotheses:

$$H_0: F_A(x) = F_B(x), \qquad H_A: F_A(x) \neq F_B(x)$$

- $F_A(x), F_B(x)$: CDF of term frequencies $tf(t,d)$ for documents $d_A \in A$, $d_B \in B$

1. Calculate p-values from z-scores
2. Define significance level α
3. Bonferroni-correction to account for multiplicity

$$p_t \leq \frac{\alpha}{|T_{C_i}|}, \qquad t \in T_{C_i}$$

Log Ratio:
Test if the mean log ratio $\overline{lr}$ for T_{C_i} is different from 0:

$$H_0: \overline{lr} = 0, \qquad H_A: \overline{lr} \neq 0$$

- Under the null hypothesis, the distribution of log ratios LR is normally distributed → Test mean with t-test

Test hypotheses for both approaches using the vocabulary

Mann-Whitney U-Test
For each term $t \in T_{C_i}$, test the following hypotheses:

$$H_0: F_A(x) = F_B(x), \qquad H_A: F_A(x) \neq F_B(x)$$

- $F_A(x), F_B(x)$: CDF of term frequencies $tf(t,d)$ for documents $d_A \in A$, $d_B \in B$

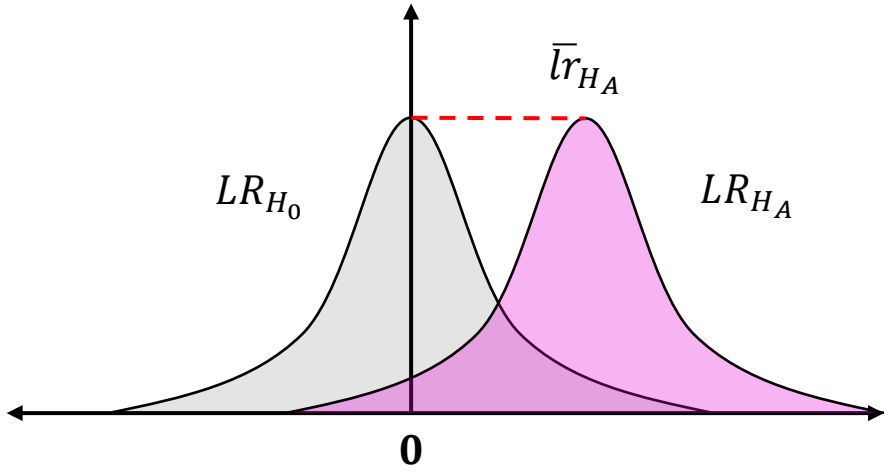
1. Calculate p-values from z-scores
2. Define significance level α
3. Bonferroni-correction to account for multiplicity

$$p_t \leq \frac{\alpha}{|T_{C_i}|}, \qquad t \in T_{C_i}$$

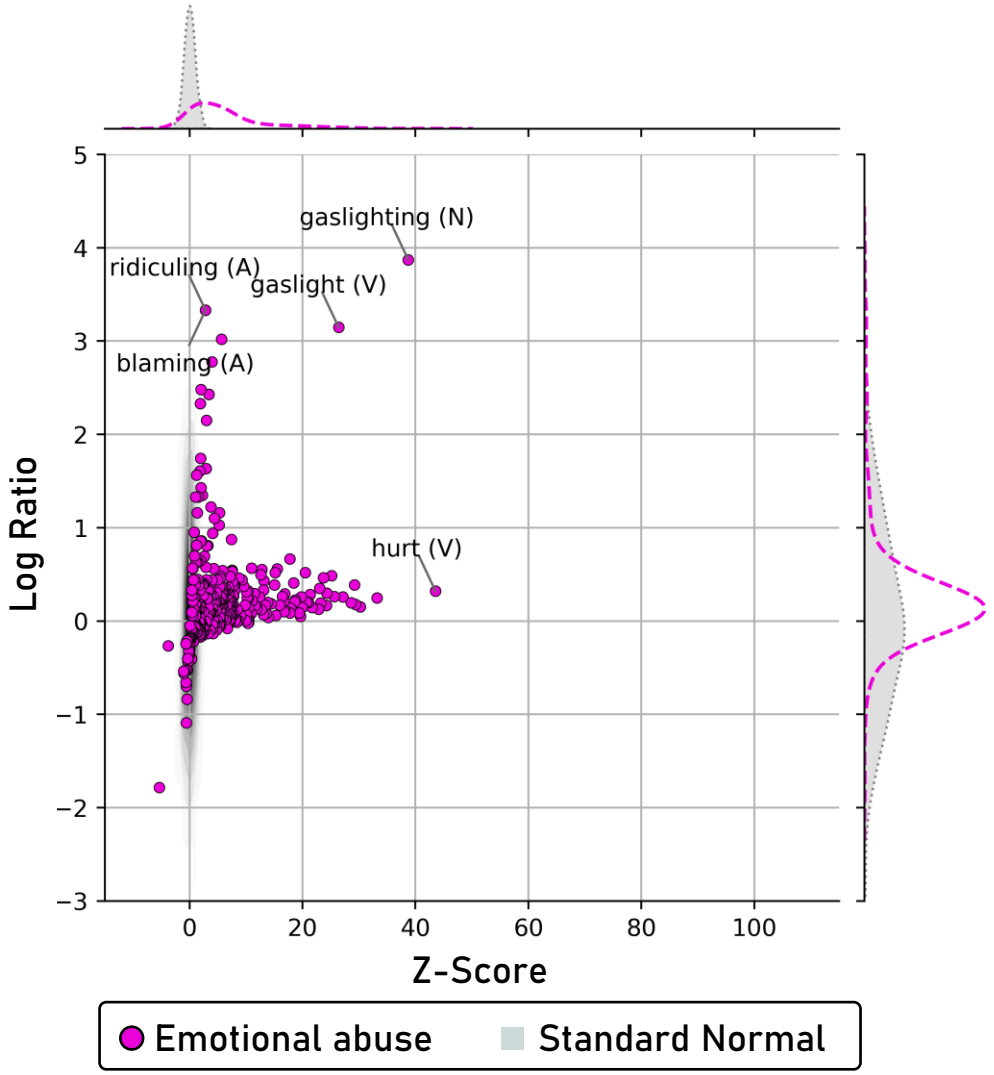
Log Ratio:
Test if the mean log ratio $\overline{lr}$ for T_{C_i} is different from 0:

$$H_0: \overline{lr} = 0, \qquad H_A: \overline{lr} \neq 0$$

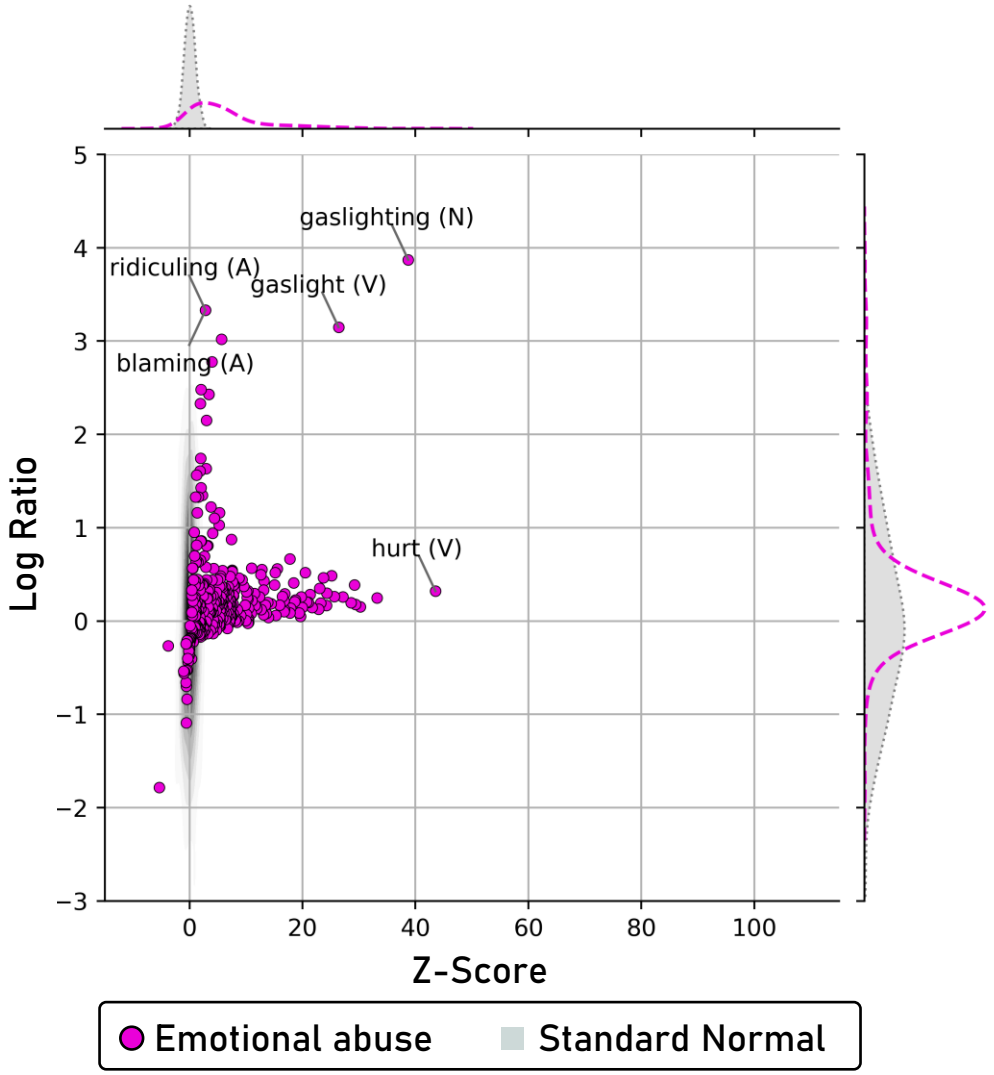
- Under the null hypothesis, the distribution of log ratios LR is normally distributed → Test mean with t-test



Emotional abuse: Vocabulary as a whole is significantly more frequent



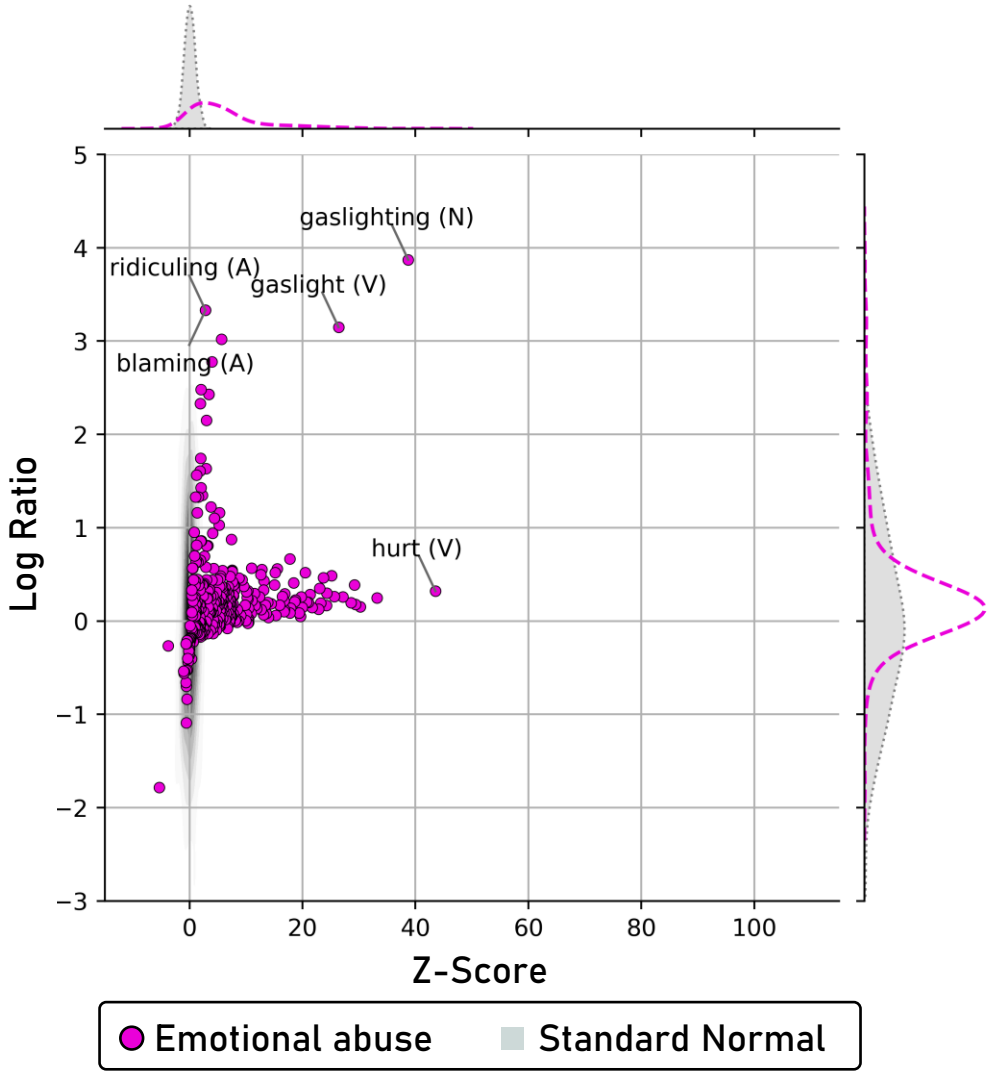
Emotional abuse: Vocabulary as a whole is significantly more frequent



Mann-Whitney U-Test:

- 190 terms significantly more frequent $H_A: F_A(x) > F_B(x)$
- 152 terms are not significant $H_0: F_A(x) = F_B(x)$
- 1 term significantly less frequent $H_A: F_A(x) < F_B(x)$

Emotional abuse: Vocabulary as a whole is significantly more frequent



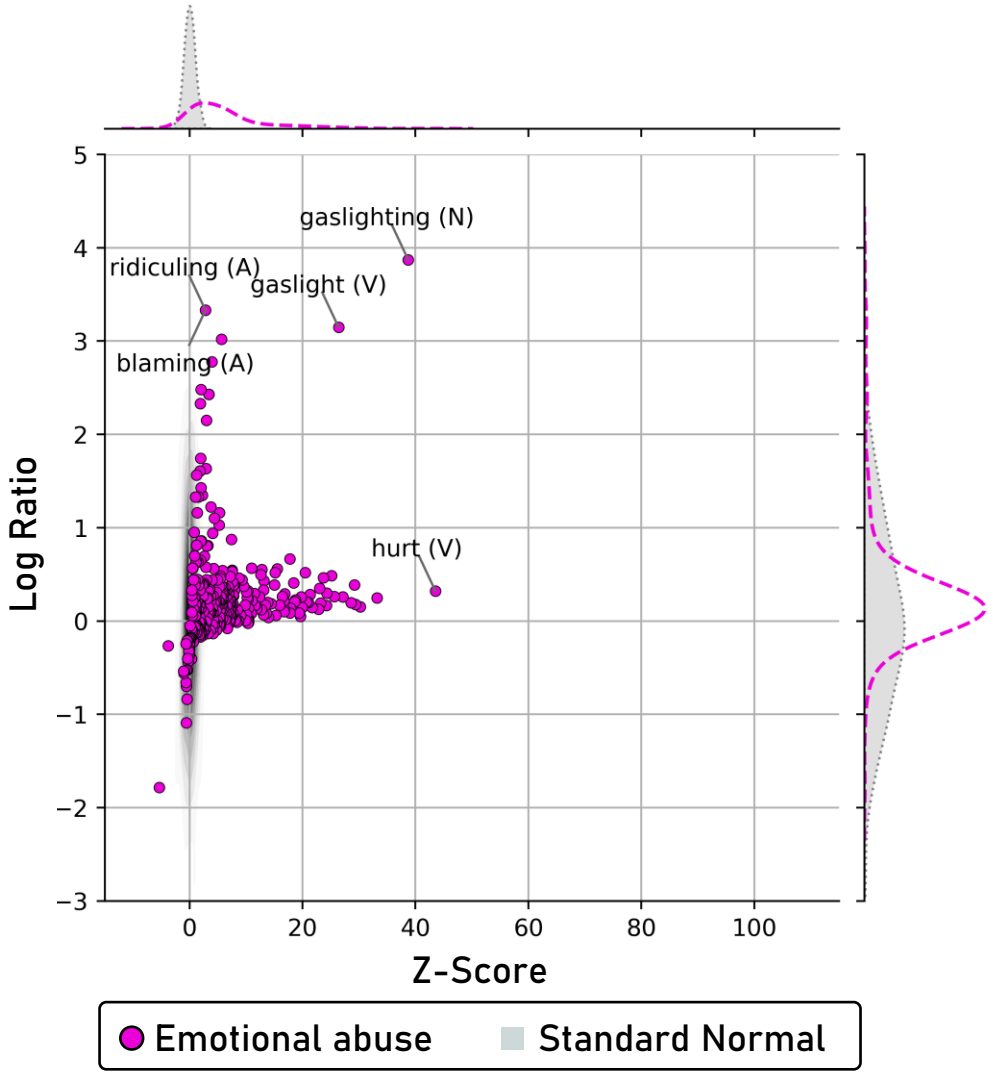
Mann-Whitney U-Test:

- 190 terms significantly more frequent $H_A: F_A(x) > F_B(x)$
- 152 terms are not significant $H_0: F_A(x) = F_B(x)$
- 1 term significantly less frequent $H_A: F_A(x) < F_B(x)$

Distribution of Log Ratios:

- p-value for t-test: $1.93 * 10^{-14}$; $H_A: \overline{lr} > 0$
- Cohen's d: 0.32 (Small to medium effect)

Emotional abuse: Vocabulary as a whole is significantly more frequent



Mann-Whitney U-Test:

- 190 terms significantly more frequent $H_A: F_A(x) > F_B(x)$
- 152 terms are not significant $H_0: F_A(x) = F_B(x)$
- 1 term significantly less frequent $H_A: F_A(x) < F_B(x)$

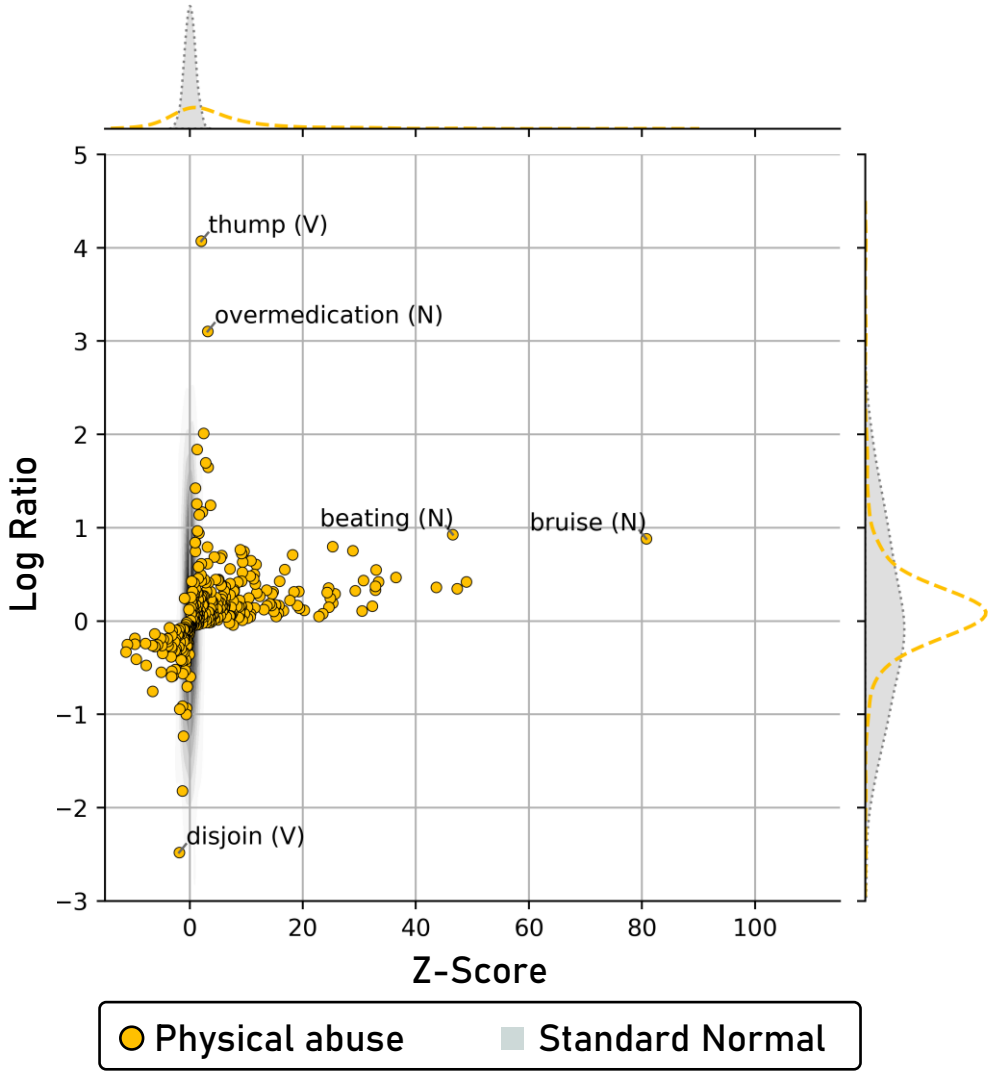
Distribution of Log Ratios:

- p-value for t-test: $1.93 * 10^{-14}$; $H_A: \overline{lr} > 0$
- Cohen's d: 0.32 (Small to medium effect)

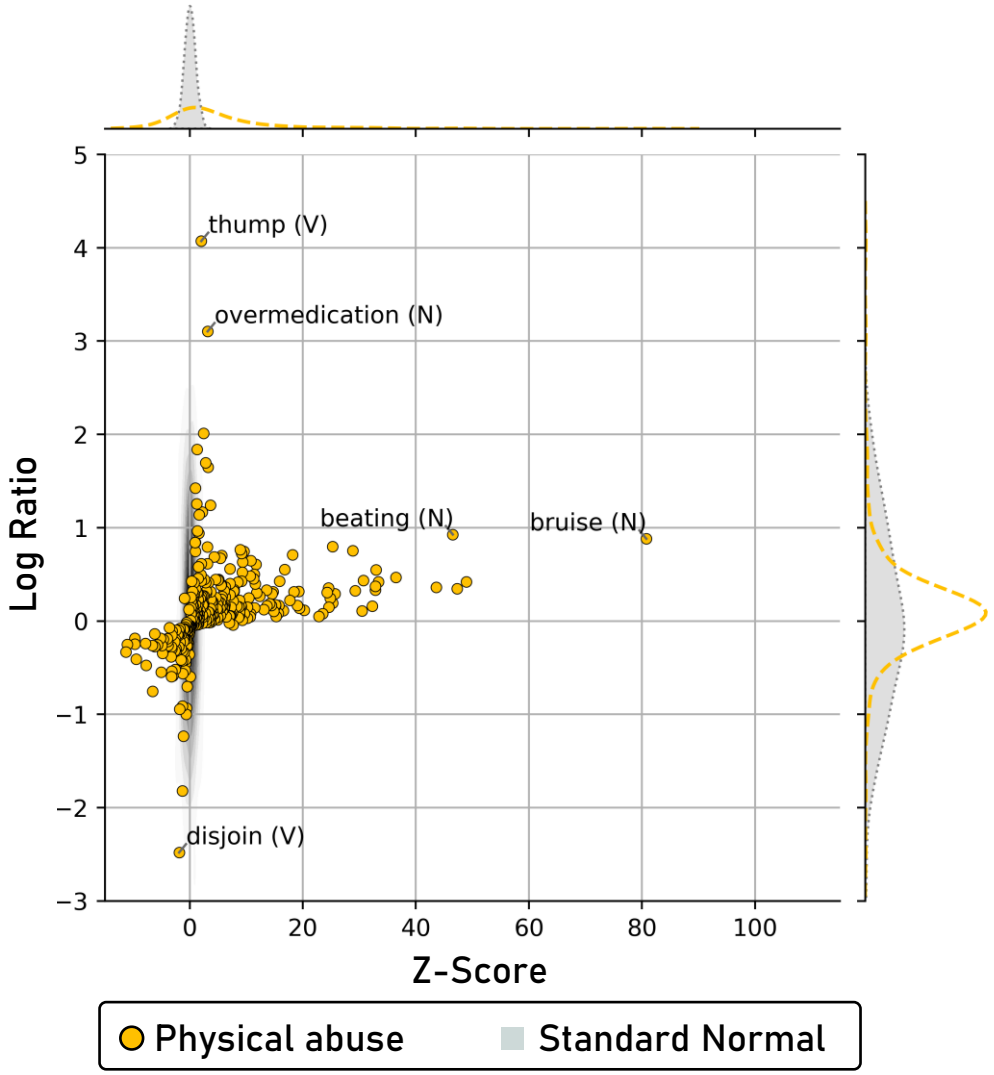
Conclusion:

- On term level, the H_0 can be rejected in 190 cases
- For the vocabulary as a whole, the H_0 can be rejected

Physical abuse: Vocabulary as a whole is significantly more frequent



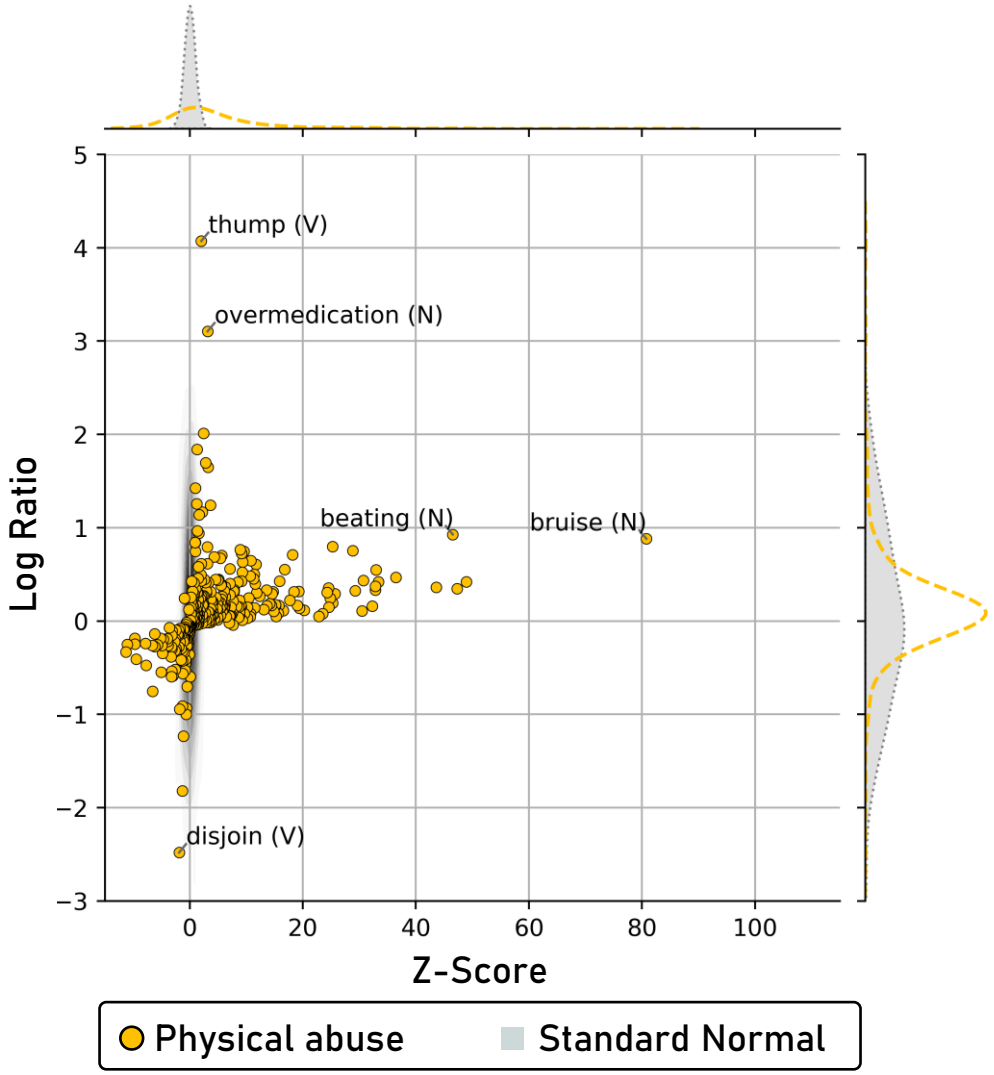
Physical abuse: Vocabulary as a whole is significantly more frequent



Mann-Whitney U-Test:

- 130 terms significantly more frequent $H_A: F_A(x) > F_B(x)$
- 196 terms are not significant $H_0: F_A(x) = F_B(x)$
- 20 terms significantly less frequent $H_A: F_A(x) < F_B(x)$

Physical abuse: Vocabulary as a whole is significantly more frequent



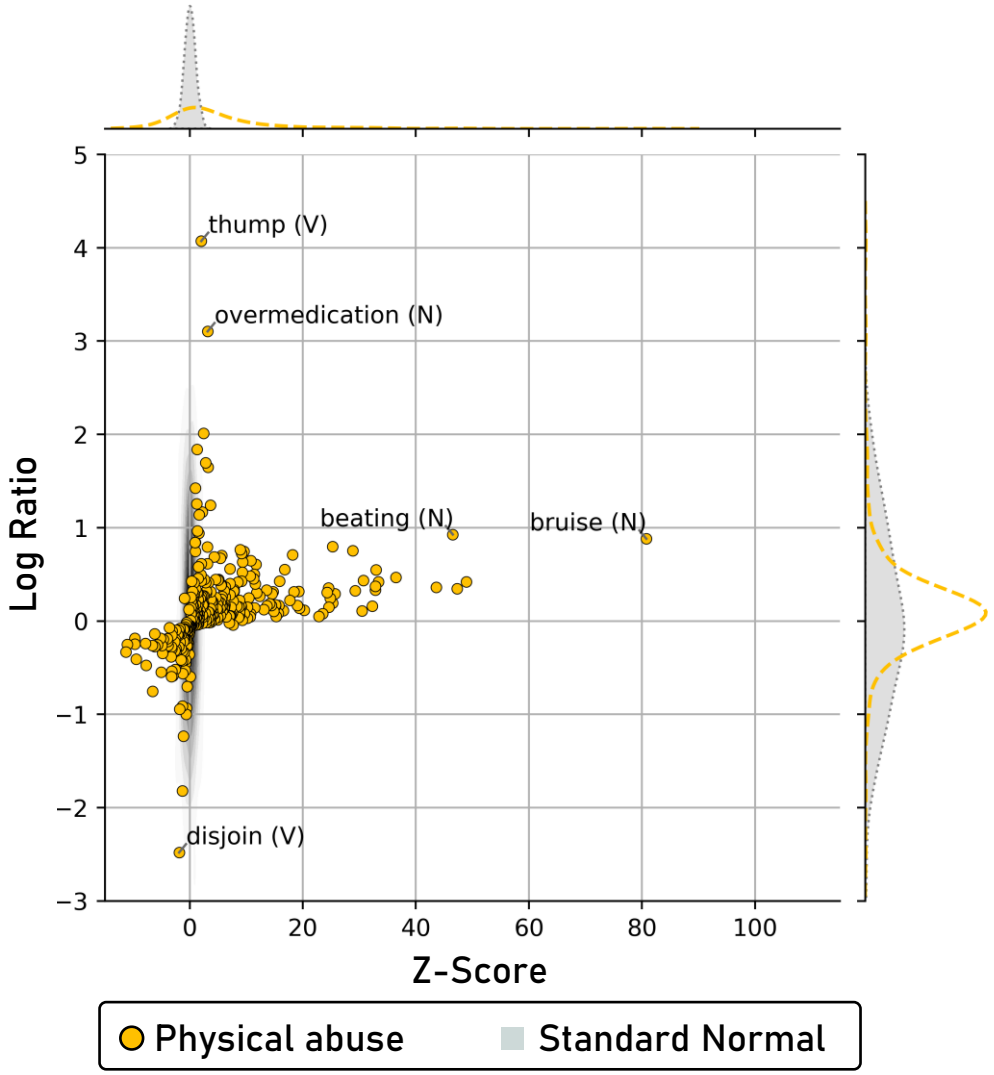
Mann-Whitney U-Test:

- 130 terms significantly more frequent $H_A: F_A(x) > F_B(x)$
- 196 terms are not significant $H_0: F_A(x) = F_B(x)$
- 20 terms significantly less frequent $H_A: F_A(x) < F_B(x)$

Distribution of Log Ratios:

- p-value for t-test: $3.70 * 10^{-6}$; $H_A: \overline{lr} > 0$
- Cohen's d: 0.16 (Small effect)

Physical abuse: Vocabulary as a whole is significantly more frequent



Mann-Whitney U-Test:

- 130 terms significantly more frequent $H_A: F_A(x) > F_B(x)$
- 196 terms are not significant $H_0: F_A(x) = F_B(x)$
- 20 terms significantly less frequent $H_A: F_A(x) < F_B(x)$

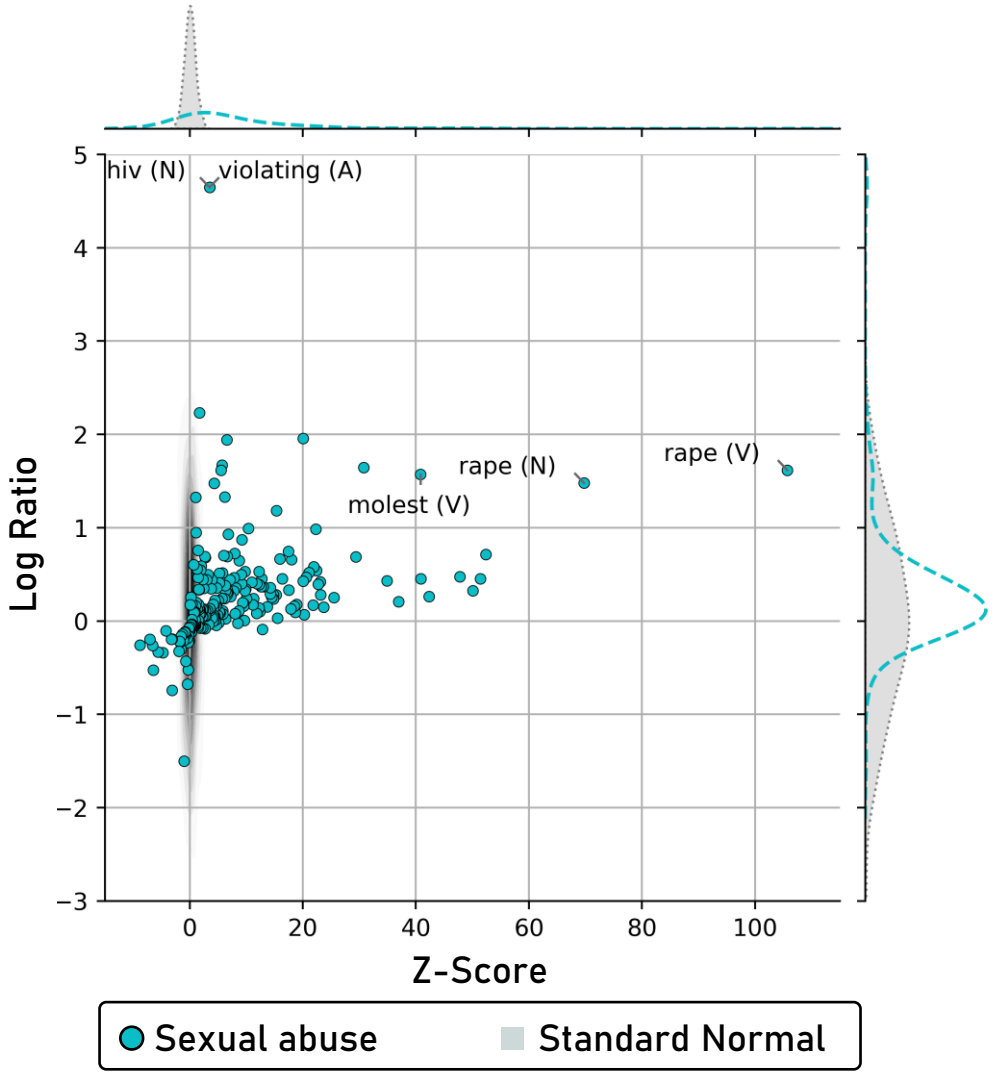
Distribution of Log Ratios:

- p-value for t-test: $3.70 * 10^{-6}$; $H_A: \overline{lr} > 0$
- Cohen's d: 0.16 (Small effect)

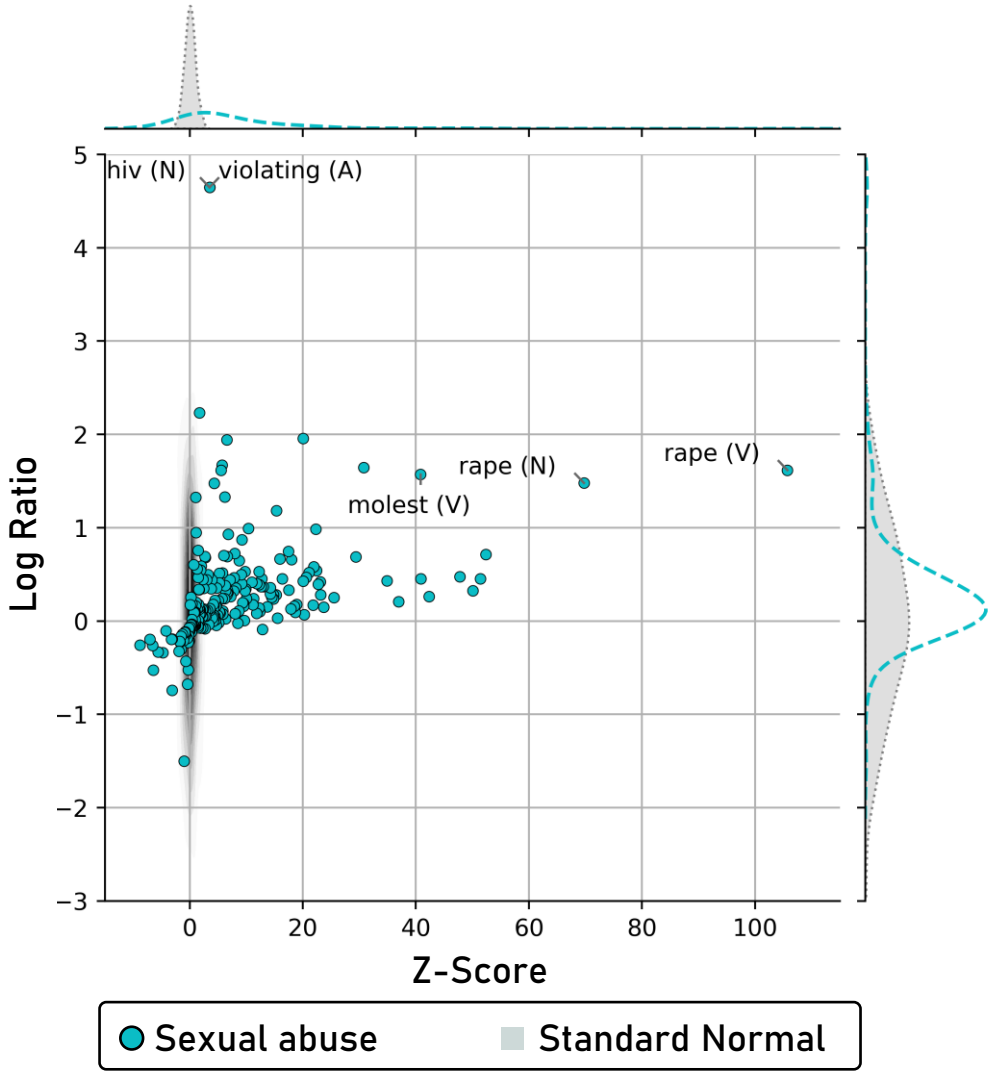
Conclusion:

- On term level, the H_0 can be rejected in 130 cases
- For the vocabulary as a whole, the H_0 can be rejected

Sexual abuse: Vocabulary as a whole is significantly more frequent



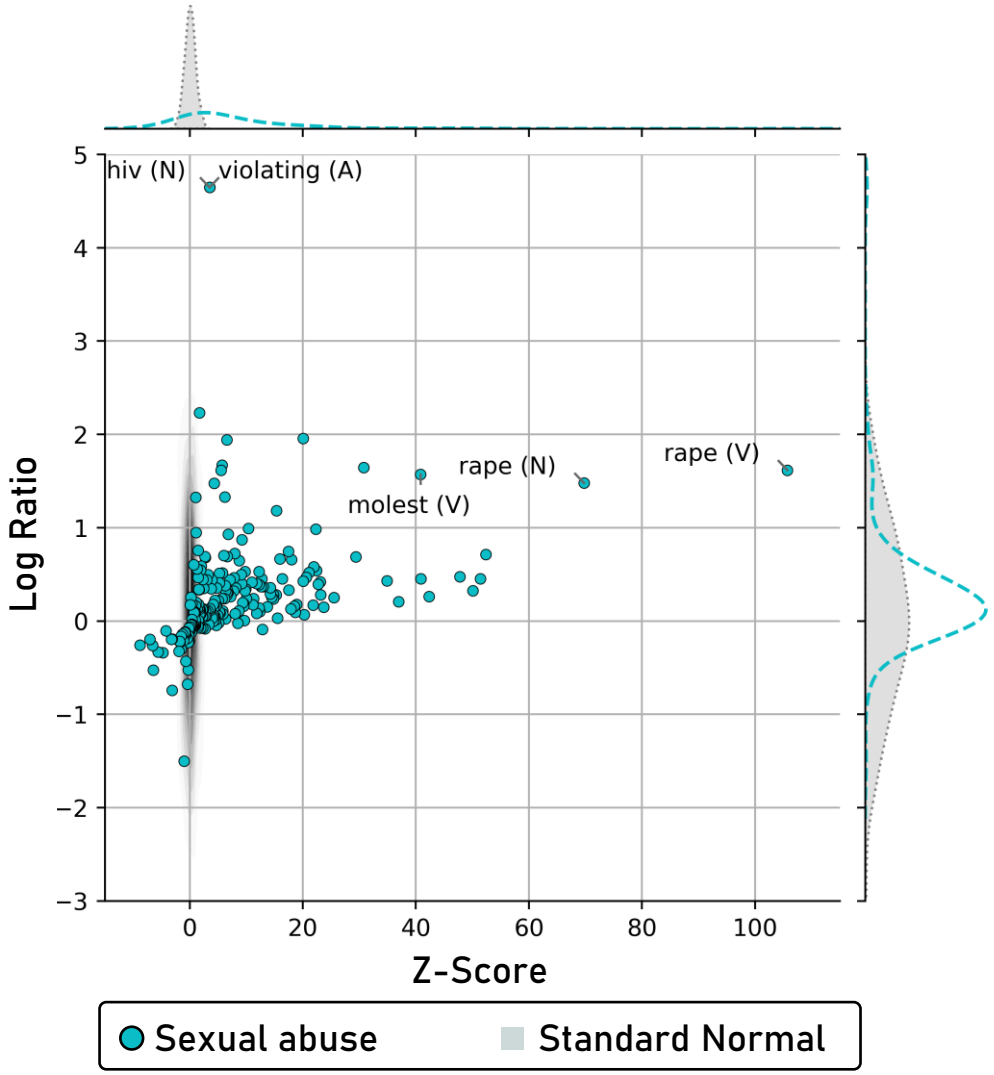
Sexual abuse: Vocabulary as a whole is significantly more frequent



Mann-Whitney U-Test:

- 124 terms significantly more frequent $H_A: F_A(x) > F_B(x)$
- 101 terms are not significant $H_0: F_A(x) = F_B(x)$
- 7 terms significantly less frequent $H_A: F_A(x) < F_B(x)$

Sexual abuse: Vocabulary as a whole is significantly more frequent



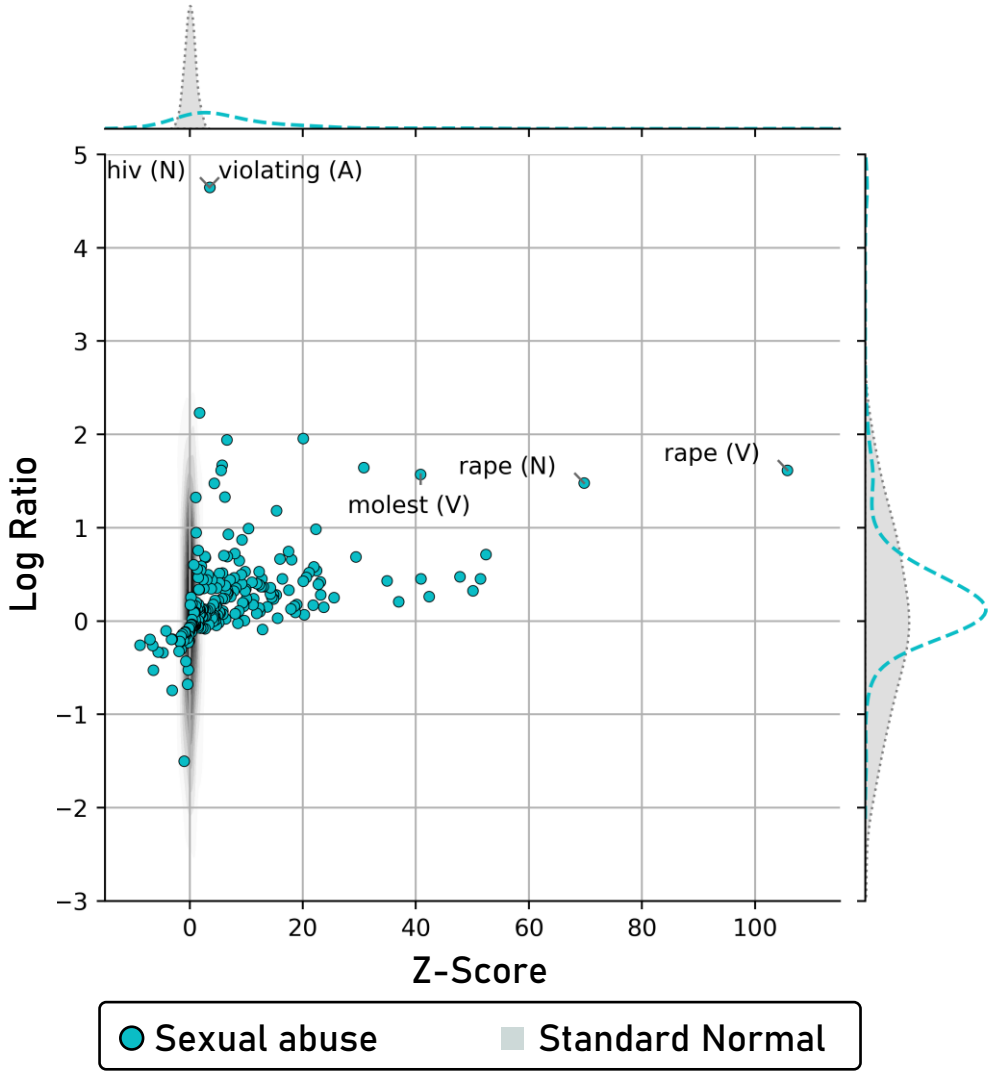
Mann-Whitney U-Test:

- 124 terms significantly more frequent $H_A: F_A(x) > F_B(x)$
- 101 terms are not significant $H_0: F_A(x) = F_B(x)$
- 7 terms significantly less frequent $H_A: F_A(x) < F_B(x)$

Distribution of Log Ratios:

- p-value for t-test: $1.74 * 10^{-11}$; $H_A: \overline{lr} > 0$
- Cohen's d: 0.34 (Small to medium effect)

Sexual abuse: Vocabulary as a whole is significantly more frequent



Mann-Whitney U-Test:

- 124 terms significantly more frequent $H_A: F_A(x) > F_B(x)$
- 101 terms are not significant $H_0: F_A(x) = F_B(x)$
- 7 terms significantly less frequent $H_A: F_A(x) < F_B(x)$

Distribution of Log Ratios:

- p-value for t-test: $1.74 * 10^{-11}$; $H_A: \overline{lr} > 0$
- Cohen's d: 0.34 (Small to medium effect)

Conclusion:

- On term level, the H_0 can be rejected in 124 cases
- For the vocabulary as a whole, the H_0 can be rejected

Hypothesis 1:

The authors on AO3 apply warning tags in a way that is consistent with common language understanding

For all three categories, H_0 can be rejected on vocabulary level

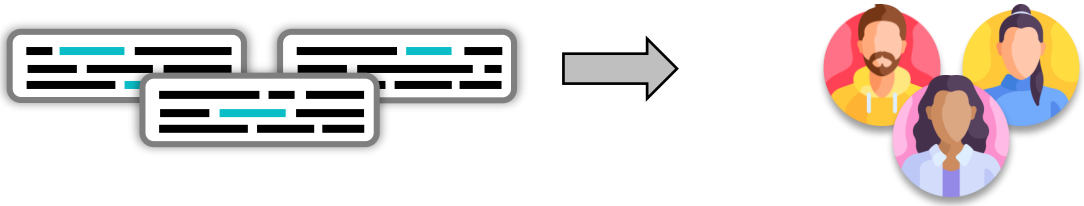
On term level, H_0 can be rejected in 38%–55% cases

Prescriptive Annotation Guidelines

Hypothesis 2:

Prescriptive annotation guidelines for trigger warnings lead to higher annotator agreements than descriptive guidelines

1. Passages for Labeling
Retrieve Passages based on
Significant Terms

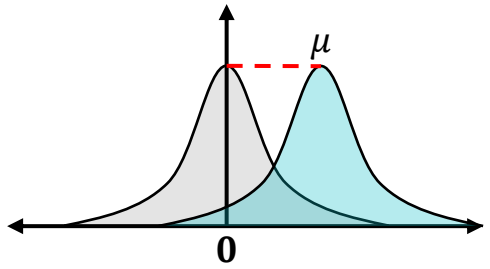


1. Passages for Labeling
Retrieve Passages based on
Significant Terms

2. Collect Annotations
Prompt LLM for Annotations with
Sociodemographic Prompting

1. Passages for Labeling
Retrieve Passages based on
Significant Terms

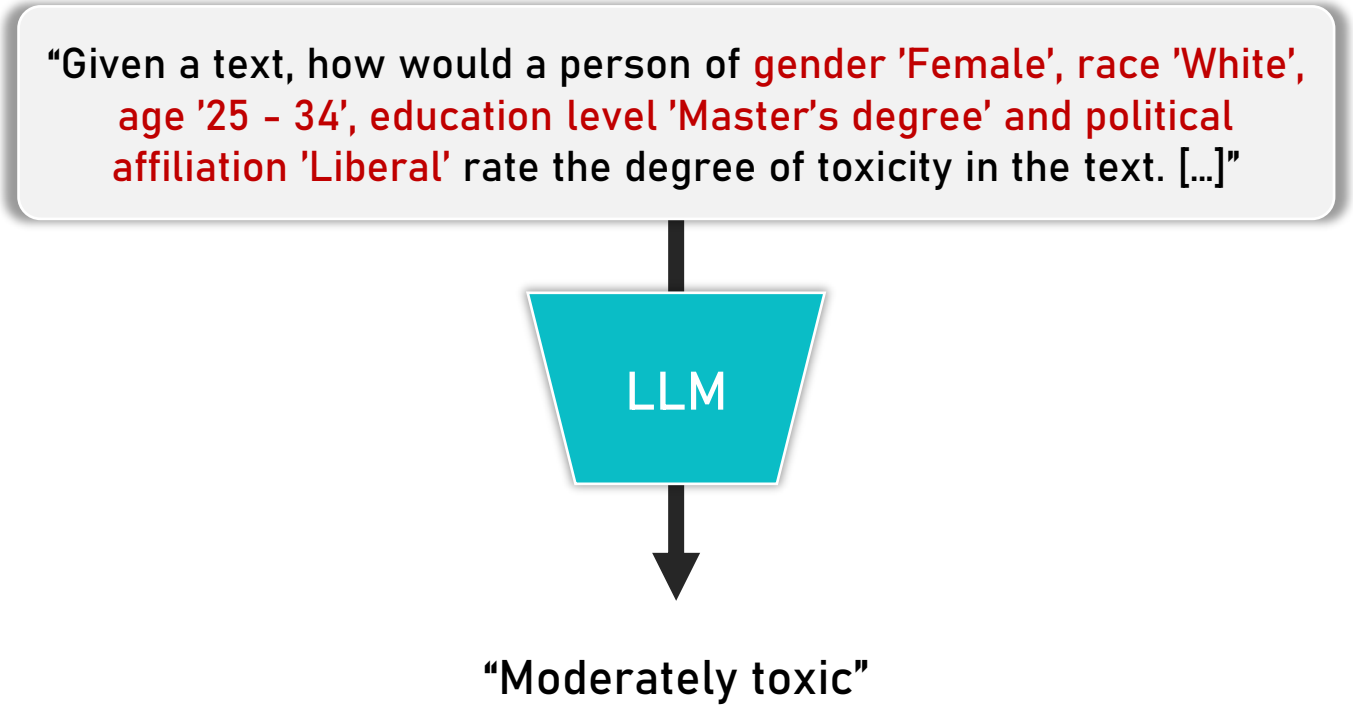
2. Collect Annotations
Prompt LLM for Annotations with
Sociodemographic Prompting



3. Apply Statistical Tests
Test the Annotator Agreements

Beck et al. (2024) analyzed sociodemographic prompting of LLMs

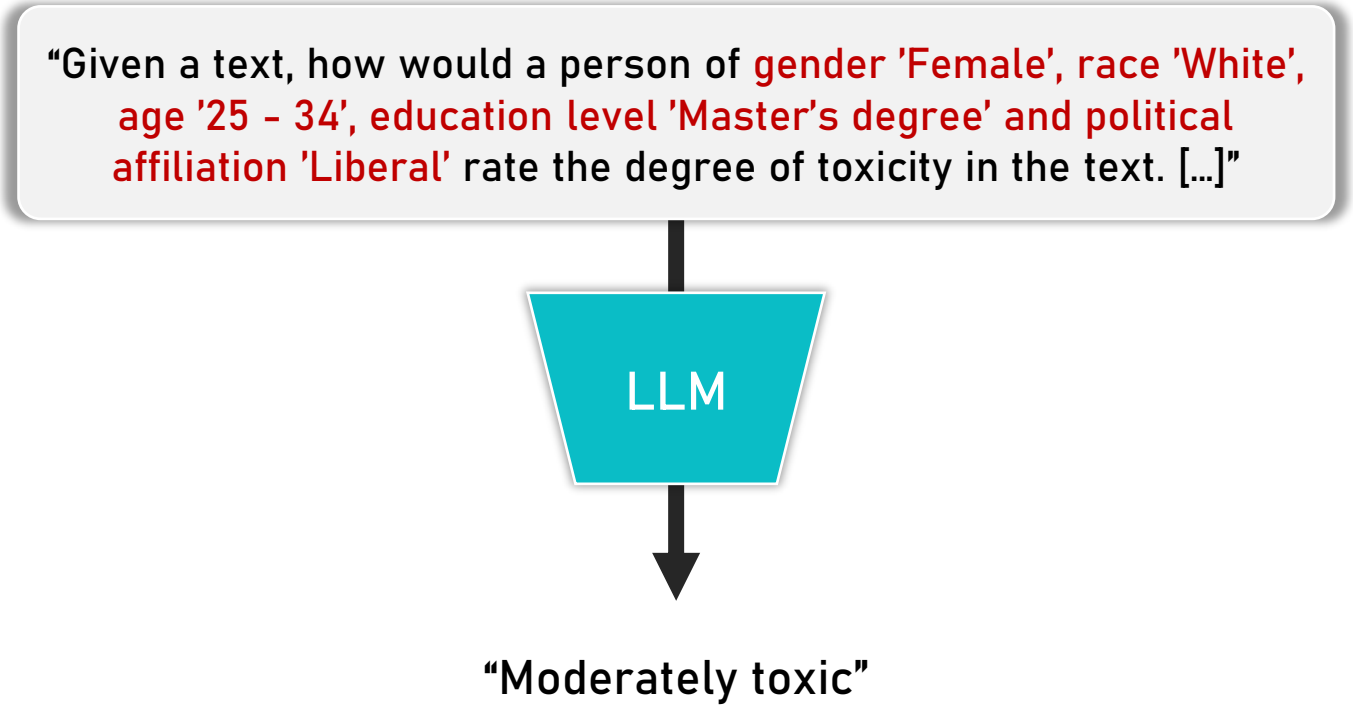
Enrich the prompt with sociodemographic properties to try to align LLM output with the described population



Beck et al. (2024)

Beck et al. (2024) analyzed sociodemographic prompting of LLMs

Enrich the prompt with sociodemographic properties to try to align LLM output with the described population



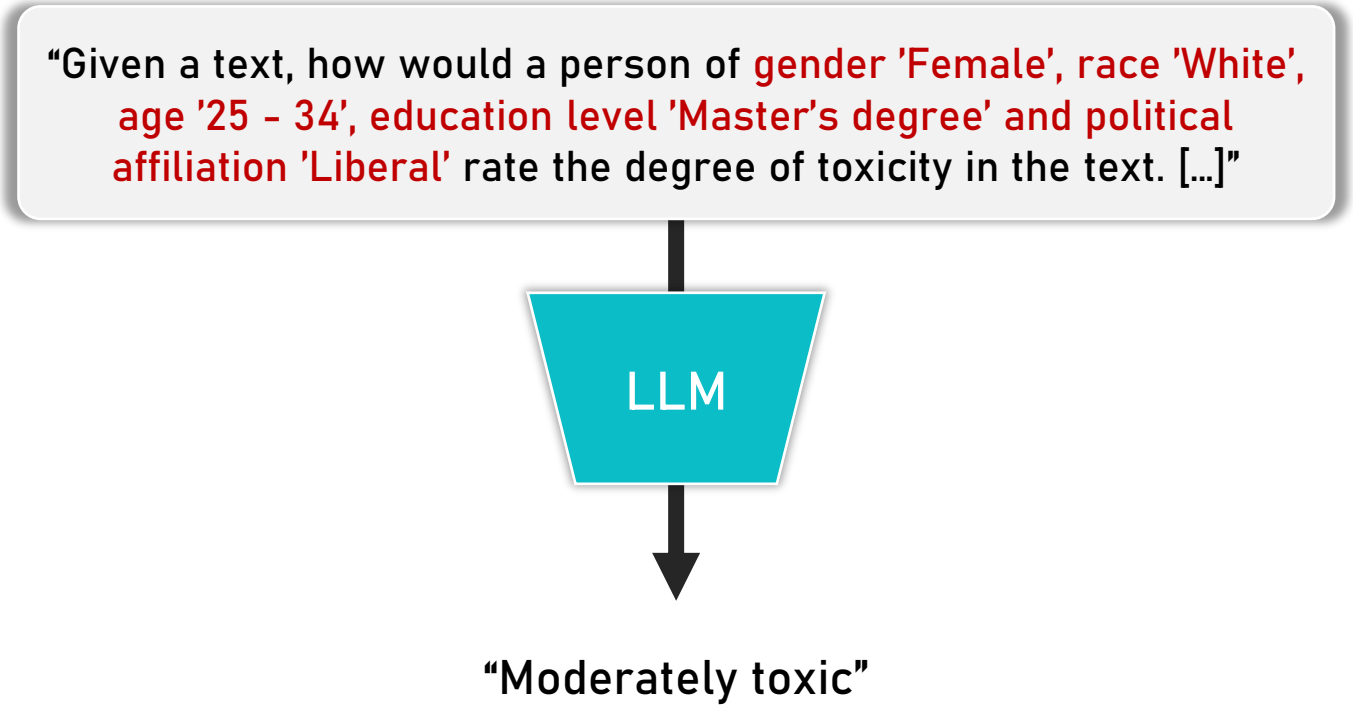
Findings

- Instruction-tuned models based on T5 were affected the most
  - Mean prediction change of >40% across the seven datasets in comparison with no sociodemographic prompting

Beck et al. (2024)

Beck et al. (2024) analyzed sociodemographic prompting of LLMs

Enrich the prompt with sociodemographic properties to try to align LLM output with the described population



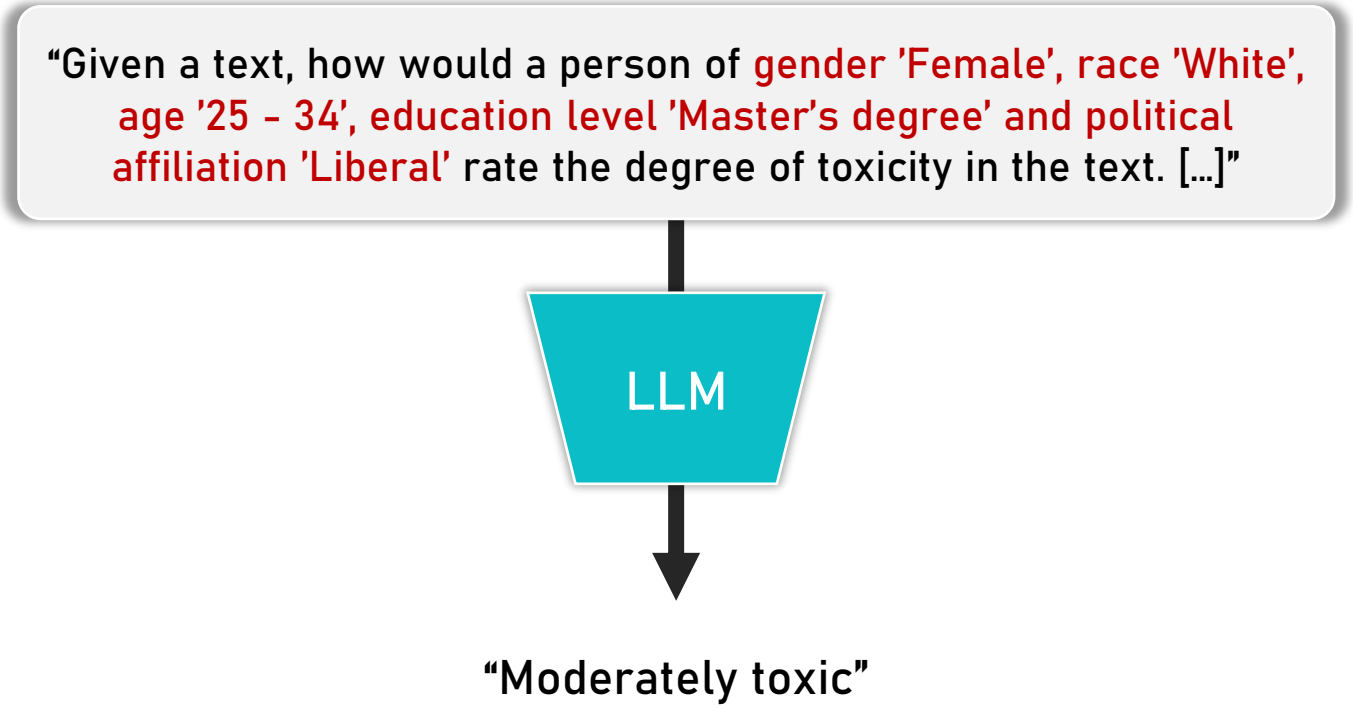
Findings

- Instruction-tuned models based on T5 were affected the most
  - Mean prediction change of >40% across the seven datasets in comparison with no sociodemographic prompting
- Choice of model more influential than the choice of text

Beck et al. (2024)

Beck et al. (2024) analyzed sociodemographic prompting of LLMs

Enrich the prompt with sociodemographic properties to try to align LLM output with the described population



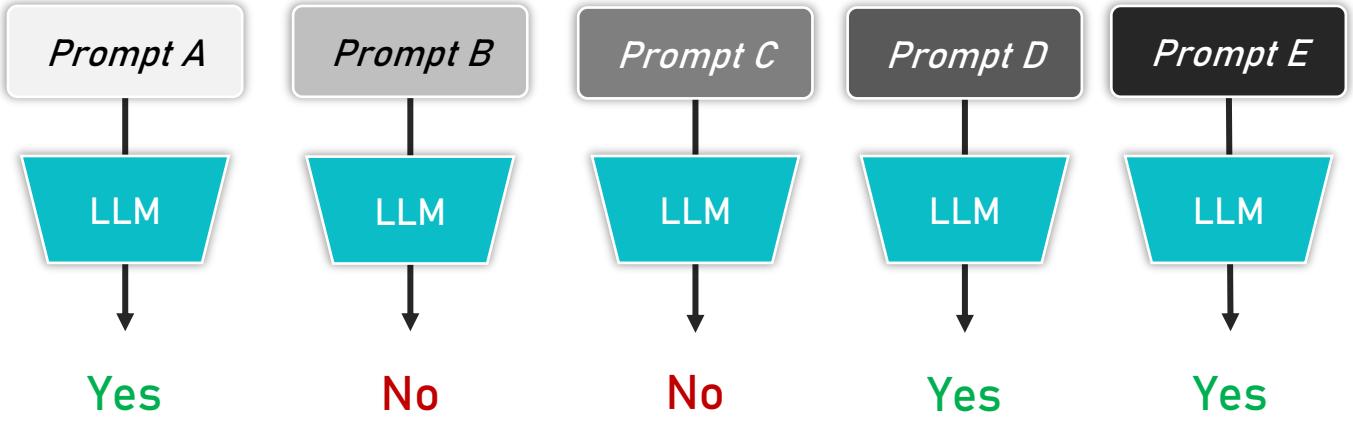
Findings

- Instruction-tuned models based on T5 were affected the most
  - Mean prediction change of >40% across the seven datasets in comparison with no sociodemographic prompting
- Choice of model more influential than the choice of text
- Authors were not able to consistently reproduce human annotations based on sociodemographic profiles

Beck et al. (2024)

Sociodemographic prompting can be used to predict disagreement

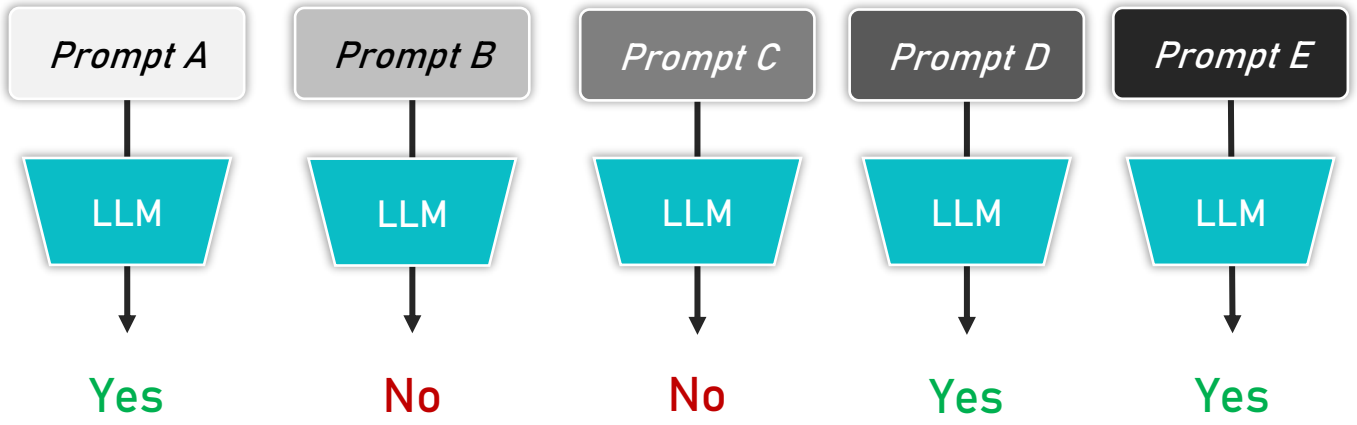
Model annotator disagreement using different prompts



Beck et al. (2024)

Sociodemographic prompting can be used to predict disagreement

Model annotator disagreement using different prompts



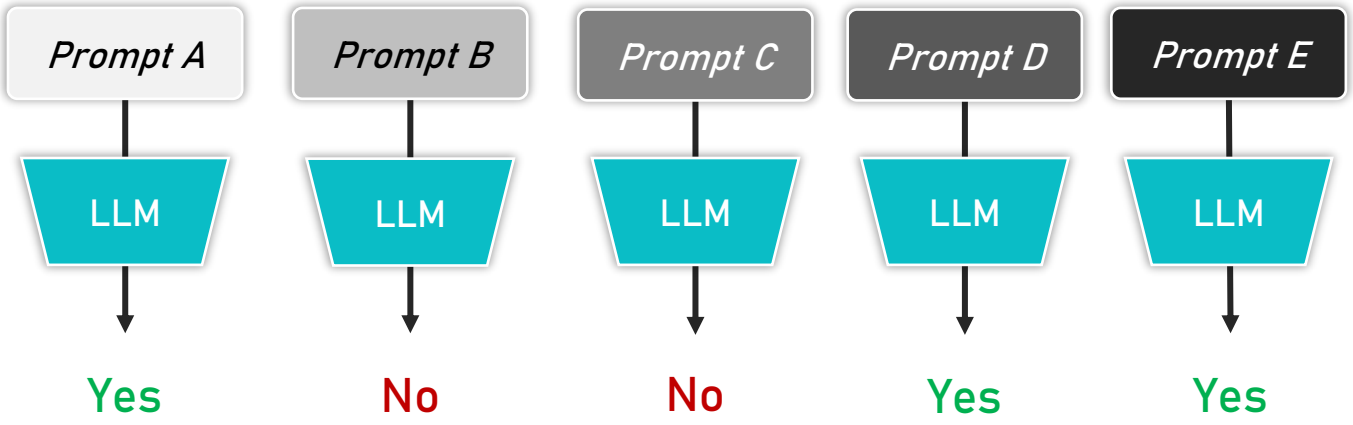
Treat disagreement prediction as binary classification

- Disagreement between annotators and LLM output differs by prompt → True Positive
- All annotators agree and the LLM output is the same for all prompts → True Negative

Beck et al. (2024)

Sociodemographic prompting can be used to predict disagreement

Model annotator disagreement using different prompts



Treat disagreement prediction as binary classification

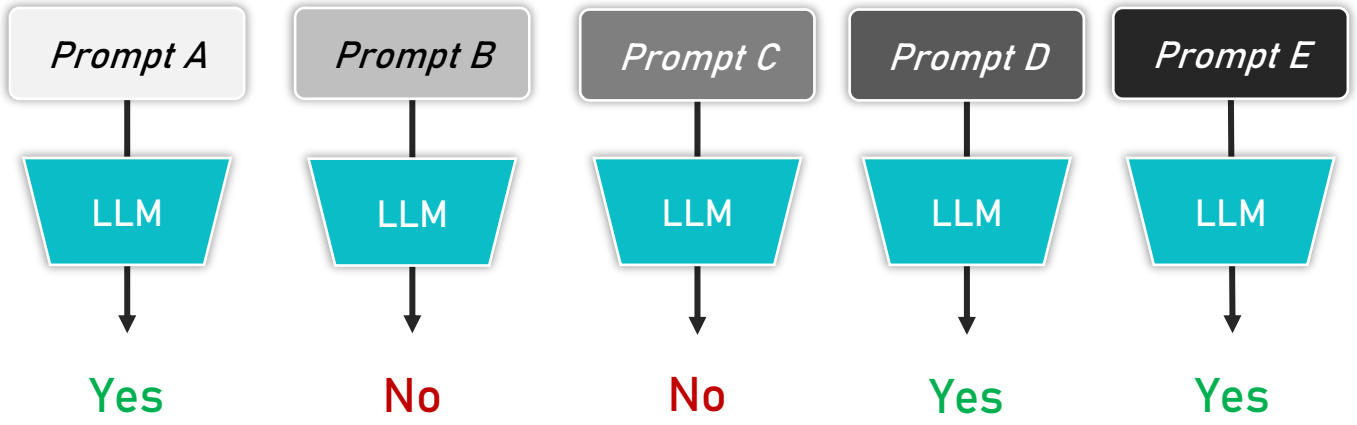
- Disagreement between annotators and LLM output differs by prompt → True Positive
- All annotators agree and the LLM output is the same for all prompts → True Negative

	DP	Jigsaw	GHC	H-Twitter	SE2016	GWSD	Diaz
InstructGPT (175B)	0.28	0.19	0.17	0.15	0.29	0.50	0.30
Flan-T5 (80M)	0.02	0.13	0.24	0.01	0.25	0.00	0.08
Flan-T5 (250M)	0.00	0.14	0.18	0.00	0.23	0.43	0.53
Flan-T5 (780M)	0.72	0.32	0.28	0.37	0.52	0.31	0.40
Flan-T5 (3B)	0.61	0.34	0.41	0.13	0.34	0.58	0.33
Flan-T5 (11B)	0.73	0.47	0.44	0.41	0.69	0.78	0.82
Flan-UL (20B)	0.28	0.16	0.34	0.41	0.34	0.52	0.63
Tk-Instruct (80M)	0.36	0.03	0.01	0.34	0.21	0.51	0.00
Tk-Instruct (250M)	0.38	0.07	0.16	0.28	0.11	0.04	0.21
Tk-Instruct (780M)	0.63	0.28	0.18	0.34	0.60	0.29	0.36
Tk-Instruct (3B)	0.50	0.07	0.25	0.30	0.66	0.47	0.36
Tk-Instruct (11B)	0.69	0.34	0.38	0.38	0.32	0.59	0.59
OPT-IML (1.3B)	0.44	0.28	0.33	0.39	0.12	0.44	0.71
OPT-IML (30B)	0.38	0.29	0.31	0.40	0.28	0.23	0.65
Dolly-V2 (2.8B)	0.56	0.18	0.07	0.00	0.44	0.02	0.72
Dolly-V2 (6.9B)	0.79	0.24	0.24	0.01	0.68	0.76	0.72
Dolly-V2 (12B)	0.88	0.36	0.26	0.17	0.24	0.43	0.56

Beck et al. (2024)

Sociodemographic prompting can be used to predict disagreement

Model annotator disagreement using different prompts



Treat disagreement prediction as binary classification

- Disagreement between annotators and LLM output differs by prompt → True Positive
- All annotators agree and the LLM output is the same for all prompts → True Negative

	DP	Jigsaw	GHC	H-Twitter	SE2016	GWSD	Diaz
InstructGPT (175B)	0.28	0.19	0.17	0.15	0.29	0.50	0.30
Flan-T5 (80M)	0.02	0.13	0.24	0.01	0.25	0.00	0.08
Flan-T5 (250M)	0.00	0.14	0.18	0.00	0.23	0.43	0.53
Flan-T5 (780M)	0.72	0.32	0.28	0.37	0.52	0.31	0.40
Flan-T5 (3B)	0.61	0.34	0.41	0.13	0.34	0.58	0.33
Flan-T5 (11B)	0.73	0.47	0.44	0.41	0.69	0.78	0.82
Flan-UL (20B)	0.28	0.16	0.34	0.41	0.34	0.52	0.63
Tk-Instruct (80M)	0.36	0.03	0.01	0.34	0.21	0.51	0.00
Tk-Instruct (250M)	0.38	0.07	0.16	0.28	0.11	0.04	0.21
Tk-Instruct (780M)	0.63	0.28	0.18	0.34	0.60	0.29	0.36
Tk-Instruct (3B)	0.50	0.07	0.25	0.30	0.66	0.47	0.36
Tk-Instruct (11B)	0.69	0.34	0.38	0.38	0.32	0.59	0.59
OPT-IML (1.3B)	0.44	0.28	0.33	0.39	0.12	0.44	0.71
OPT-IML (30B)	0.38	0.29	0.31	0.40	0.28	0.23	0.65
Dolly-V2 (2.8B)	0.56	0.18	0.07	0.00	0.44	0.02	0.72
Dolly-V2 (6.9B)	0.79	0.24	0.24	0.01	0.68	0.76	0.72
Dolly-V2 (12B)	0.88	0.36	0.26	0.17	0.24	0.43	0.56

Beck et al. (2024)

Estimate the effect of prescriptive guidelines using SD prompting

"Imagine you are a person of [SD Profile] that has experienced abuse themselves or of their relatives in the past. Is there anything in the given text that you could associate with this experience?"

Requires warning for abuse ('yes' or 'no'):

Descriptive Prompt (Underspecified)

- Asks to imagine having experienced trauma
- Question: Does the text require a warning?

Estimate the effect of prescriptive guidelines using SD prompting

"Imagine you are a person of [SD Profile] that has experienced abuse themselves or of their relatives in the past. Is there anything in the given text that you could associate with this experience?"

Requires warning for abuse ('yes' or 'no'):

Descriptive Prompt (Underspecified)

- Asks to imagine having experienced trauma
- Question: Does the text require a warning?

"Imagine you are a person of [SD Profile] that has experienced emotional abuse themselves or of their relatives in the past. Is there anything in the given text that you could associate with this experience?"

Requires warning for emotional abuse ('yes' or 'no'):

Descriptive Prompt (Category-Specific)

- Ask to imagine having experienced trauma
- Uses specific category (emotional abuse)
- Question: Does the text require a warning?

Estimate the effect of prescriptive guidelines using SD prompting

"Imagine you are a person of [SD Profile] that has experienced abuse themselves or of their relatives in the past. Is there anything in the given text that you could associate with this experience?"

Requires warning for abuse ('yes' or 'no'):

Descriptive Prompt (Underspecified)
- Asks to imagine having experienced trauma
- Question: Does the text require a warning?

"Imagine you are a person of [SD Profile] that has experienced emotional abuse themselves or of their relatives in the past. Is there anything in the given text that you could associate with this experience?"

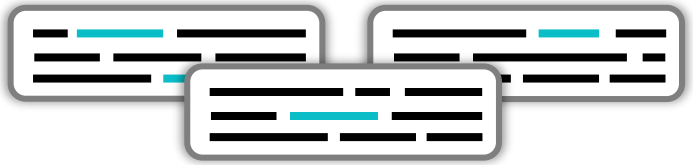
Requires warning for emotional abuse ('yes' or 'no'):

Descriptive Prompt (Category-Specific)
- Ask to imagine having experienced trauma
- Uses specific category (emotional abuse)
- Question: Does the text require a warning?

"Imagine you are a person of [SD Profile]. Does the following text contain acts/consequences of emotional abuse?
Examples are:
1) intimidation, harassment, humiliation, and other verbal abuse
2) gaslighting, lying and other forms of manipulation
3) socially isolating a person or preventing them from engaging in meaningful activities"
Contains acts/consequences of emotional abuse ('yes' or 'no'):

Prescriptive Prompt
- Does not ask to imagine trauma
- Gives lists of examples to check
- Question: Does the text contain examples from the list

Estimate the effect of prescriptive guidelines using SD prompting



1. Sample 10,000 passages uniformly for 60 significant and/or high log-ratio words

Icons from Flaticon.com

Estimate the effect of prescriptive guidelines using SD prompting

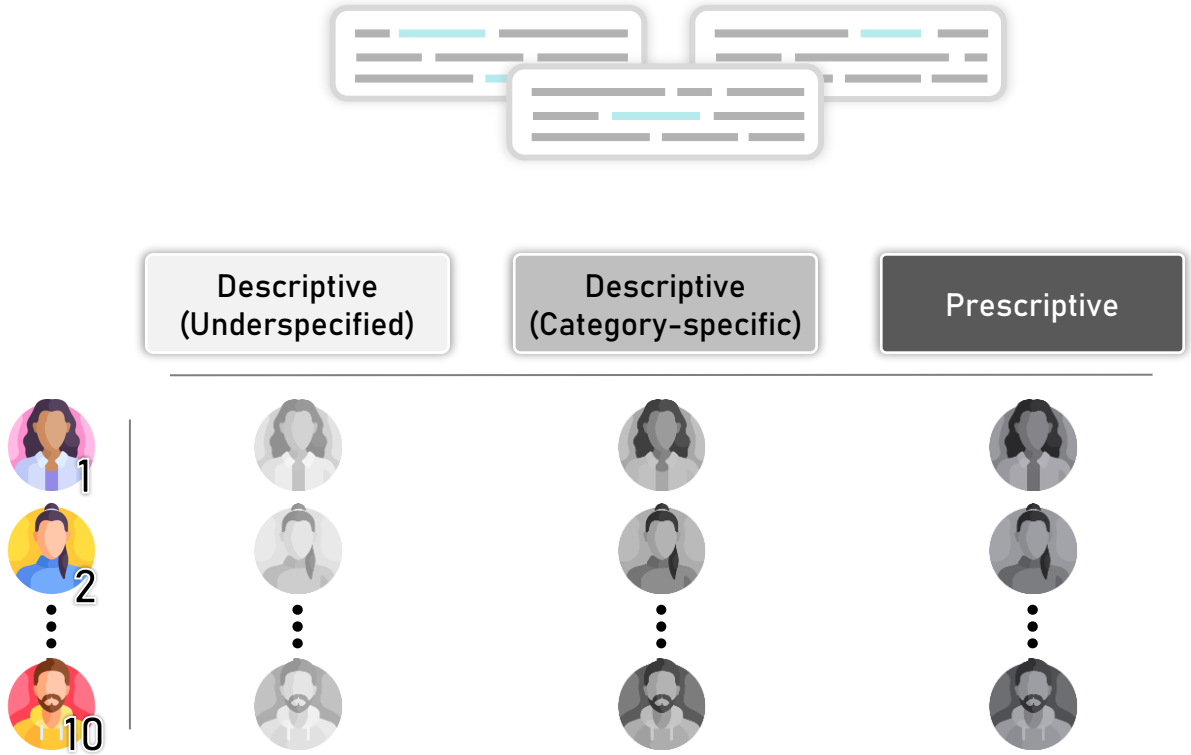
Descriptive (Underspecified) | Descriptive (Category-specific) | Prescriptive

1. Sample 10,000 passages uniformly for 60 significant and/or high log-ratio words

2. Prompt Flan-T5 11B with all three prompt types and ten different sociodemographic profiles

Icons from Flaticon.com

Estimate the effect of prescriptive guidelines using SD prompting



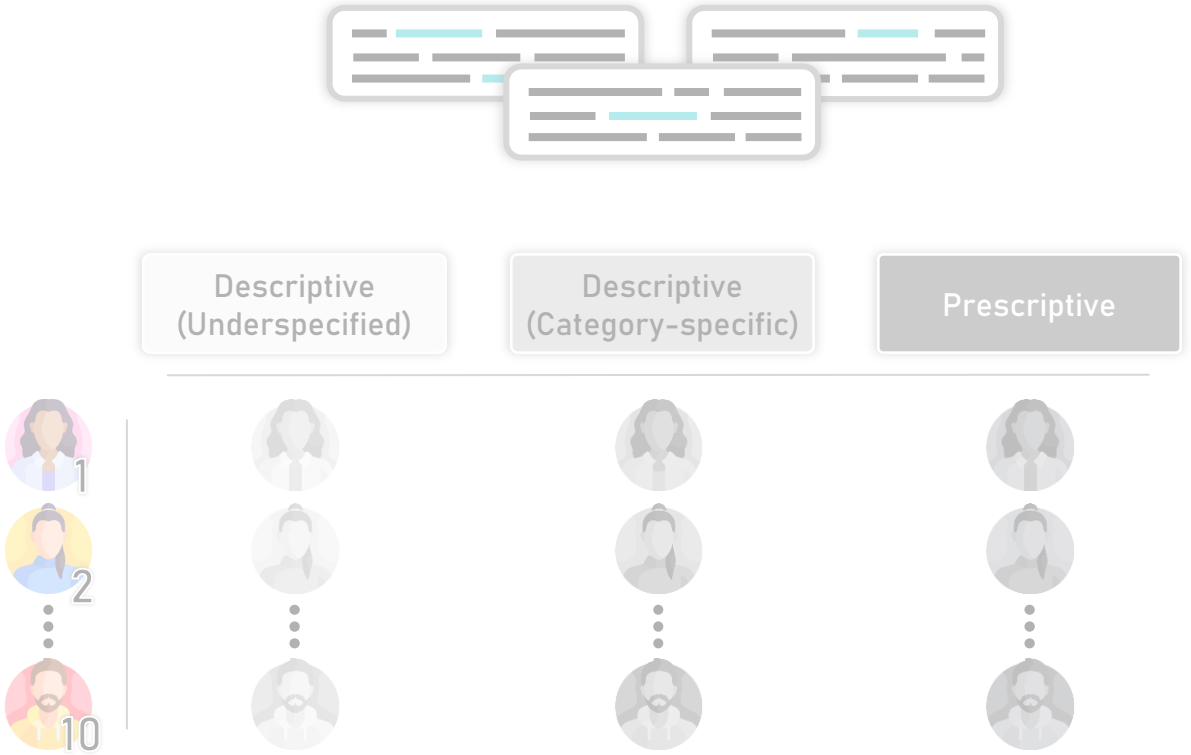
1. Sample 10,000 passages uniformly for 60 significant and/or high log-ratio words

2. Prompt Flan-T5 11B with all three prompt types and ten different sociodemographic profiles

- Each of the 30 prompt-profile combinations receives each of the 10,000 passages
- Make binary classification: “yes” or “no”

Icons from Flaticon.com

Estimate the effect of prescriptive guidelines using SD prompting



1. Sample 10,000 passages uniformly for 60 significant and/or high log-ratio words

2. Prompt Flan-T5 11B with all three prompt types and ten different sociodemographic profiles

- Each of the 30 prompt-profile combinations receives each of the 10,000 passages
- Make binary classification: “yes” or “no”

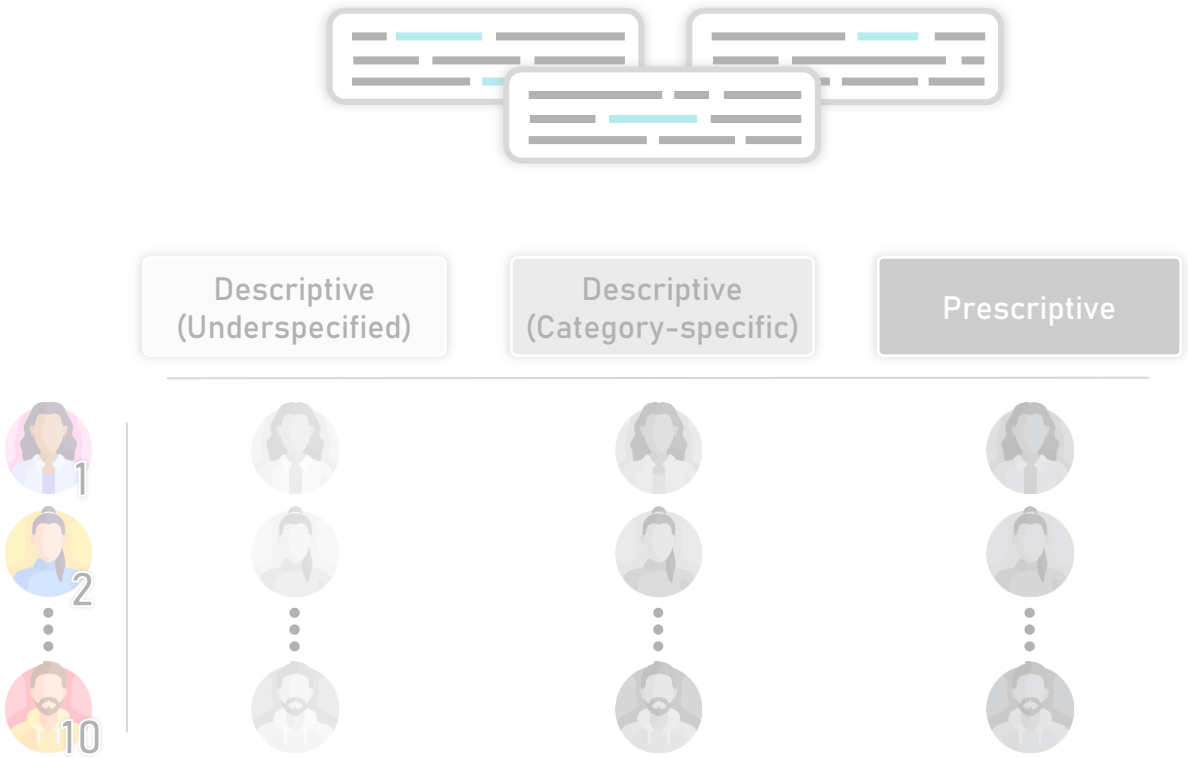
3. Calculate pairwise annotator agreements using Cohen’s Kappa

$$\kappa = P_o - \frac{P_e}{1 - P_e}, \kappa \in [-1, 1]$$

- P_o: P(Agreement)
- P_e: P(Random agreement)

Icons from Flaticon.com

Estimate the effect of prescriptive guidelines using SD prompting



1. Sample 10,000 passages uniformly for 60 significant and/or high log-ratio words

2. Prompt Flan-T5 11B with all three prompt types and ten different sociodemographic profiles

- Each of the 30 prompt-profile combinations receives each of the 10,000 passages
- Make binary classification: “yes” or “no”

3. Calculate pairwise annotator agreements using Cohen’s Kappa

- Null hypothesis: The mean Kappa is the same for descriptive and prescriptive prompts

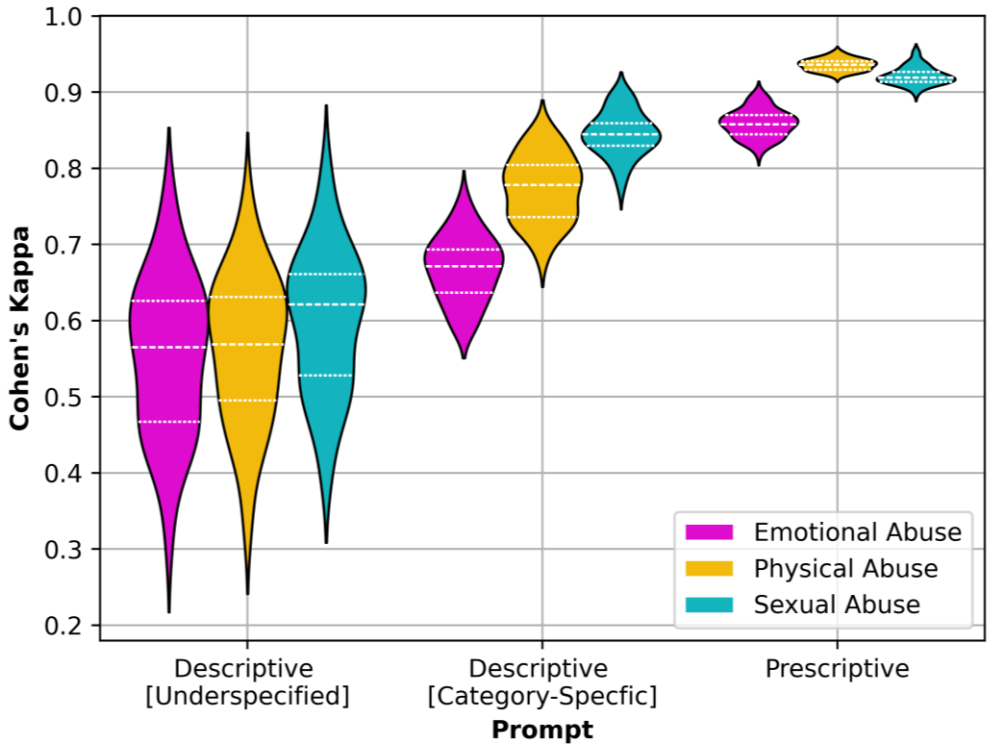
$$\kappa = P_o - \frac{P_e}{1 - P_e}, \kappa \in [-1, 1]$$

- P_o: P(Agreement)
- P_e: P(Random agreement)

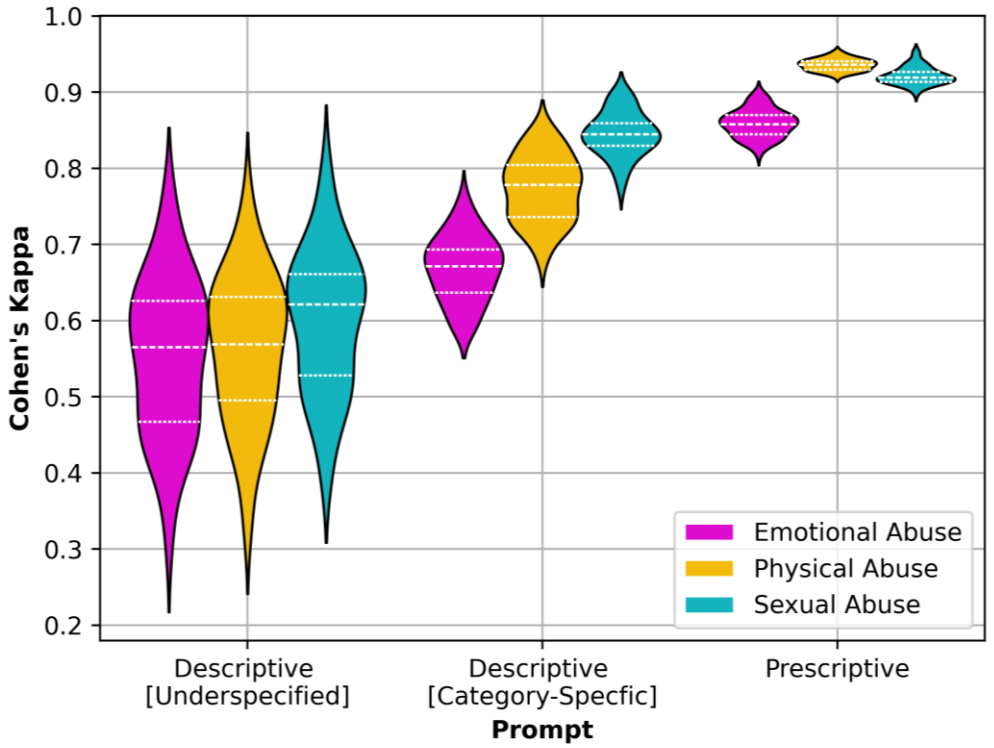
$$H_0: \mu_{\kappa_{Pres}} = \mu_{\kappa_{Desc}} \qquad H_A: \mu_{\kappa_{Pres}} > \mu_{\kappa_{Desc}}$$

Icons from Flaticon.com

Prescriptive prompting leads to significant increase in agreement



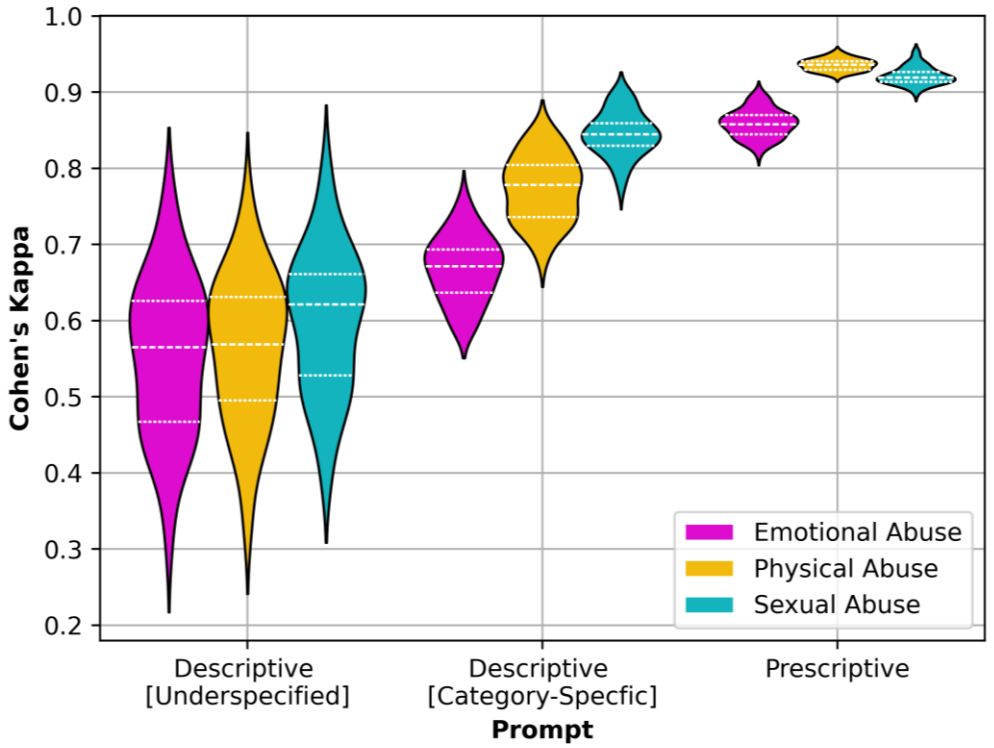
Prescriptive prompting leads to significant increase in agreement



		Emotional Abuse	Physical Abuse	Sexual abuse
$\mu_{\kappa_{Pres}} > \mu_{\kappa_{Desc[US]}}$	p-value	$5.27 * 10^{-34}$	$\mathbf{6.12 * 10^{-43}}$	$1.08 * 10^{-38}$
	Cohen's d	4.15	**5.46**	4.81
$\mu_{\kappa_{Pres}} > \mu_{\kappa_{Desc[CS]}}$	p-value	$\mathbf{6.16 * 10^{-45}}$	$1.11 * 10^{-41}$	$2.43 * 10^{-28}$
	Cohen's d	**5.79**	5.26	3.44
$\mu_{\kappa_{Desc[CS]}} > \mu_{\kappa_{Desc[US]}}$	p-value	$5.44 * 10^{-10}$	$6.63 * 10^{-23}$	$\mathbf{4.18 * 10^{-29}}$
	Cohen's d	1.47	2.82	**3.53**

- Strongest effect per row marked in bold
- Strongest effect per category marked in respective color

Prescriptive prompting leads to significant increase in agreement



		Emotional Abuse	Physical Abuse	Sexual abuse
$\mu_{\kappa_{Pres}} > \mu_{\kappa_{Desc[US]}}$	p-value	$5.27 * 10^{-34}$	$\mathbf{6.12 * 10^{-43}}$	$1.08 * 10^{-38}$
	Cohen's d	4.15	**5.46**	4.81
$\mu_{\kappa_{Pres}} > \mu_{\kappa_{Desc[CS]}}$	p-value	$\mathbf{6.16 * 10^{-45}}$	$1.11 * 10^{-41}$	$2.43 * 10^{-28}$
	Cohen's d	**5.79**	5.26	3.44
$\mu_{\kappa_{Desc[CS]}} > \mu_{\kappa_{Desc[US]}}$	p-value	$5.44 * 10^{-10}$	$6.63 * 10^{-23}$	$\mathbf{4.18 * 10^{-29}}$
	Cohen's d	1.47	2.82	**3.53**

- Strongest effect per row marked in bold
- Strongest effect per category marked in respective color

Conclusions

- H_0 can be rejected for all categories and prompt pairs
- Physical abuse annotations benefit the most from a prescriptive prompt
- Agreements in sexual abuse annotations increases already noticeably with a category-specific prompt

Hypothesis 2:

Prescriptive annotation guidelines for trigger warnings lead to higher annotator agreements than descriptive guidelines

For all three categories, H_0 can be rejected

Conclusion


AVENGERS
EARTH'S MIGHTIEST HEROES?!
POWER MAN
SPIDER-MAN
WOLVERINE
DAZZLER
HAWKEYE
IRON MAN
INVISIBLE GIRL
DAREDEVIL
ANT-MAN
HULK
DOCTOR STRANGE
SHE-HULK
BLACK BOLT
SPIDER-WOMAN
SILVER SURFER
PICK TWO!
(ANSWER INSIDE!)
MARVEL
THOR VS. QUICK-SAND!
SUPER-SIZE

Findings and Outlook

Findings

- Warnings for abuse on A03 are used consistently with an expected vocabulary
  - Consistency is significant for the vocabulary as a whole; 38%-55% of individual terms are significant by themselves

Findings and Outlook

Findings

- Warnings for abuse on AO3 are used consistently with an expected vocabulary
  - Consistency is significant for the vocabulary as a whole; 38%-55% of individual terms are significant by themselves
- Prescriptive annotation prompts increase agreement in sociodemographic prompting

Findings and Outlook

Findings

- Warnings for abuse on AO3 are used consistently with an expected vocabulary
  - Consistency is significant for the vocabulary as a whole; 38%–55% of individual terms are significant by themselves
- Prescriptive annotation prompts increase agreement in sociodemographic prompting

Future Work

- Verify if effect of prescriptive annotation guidelines transfers to human annotators

Findings and Outlook

Findings

- Warnings for abuse on AO3 are used consistently with an expected vocabulary
  - Consistency is significant for the vocabulary as a whole; 38%-55% of individual terms are significant by themselves
- Prescriptive annotation prompts increase agreement in sociodemographic prompting

Future Work

- Verify if effect of prescriptive annotation guidelines transfers to human annotators
- Extend analysis to additional warnings beyond abuse

Findings and Outlook

Findings

- Warnings for abuse on AO3 are used consistently with an expected vocabulary
  - Consistency is significant for the vocabulary as a whole; 38%-55% of individual terms are significant by themselves
- Prescriptive annotation prompts increase agreement in sociodemographic prompting

Future Work

- Verify if effect of prescriptive annotation guidelines transfers to human annotators
- Extend analysis to additional warnings beyond abuse
- Cluster tags in further granularity than warning categories by forming functional subcategories

Implied/Referenced Child Neglect Technically

Explicit Discussion of Sexual Abuse

Findings and Outlook

Findings

- Warnings for abuse on AO3 are used consistently with an expected vocabulary
  - Consistency is significant for the vocabulary as a whole; 38%-55% of individual terms are significant by themselves
- Prescriptive annotation prompts increase agreement in sociodemographic prompting

Future Work

- Verify if effect of prescriptive annotation guidelines transfers to human annotators
- Extend analysis to additional warnings beyond abuse
- Cluster tags in further granularity than warning categories by forming functional subcategories

Sources

- Beck et al. (2024)
  - Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting; https://aclanthology.org/2024.eacl-long.159
- Davani et al. (2021)
  - Hate Speech Classifiers Learn Human-Like Social Stereotypes; https://doi.org/10.48550/arXiv.2110.14839
- Hardie (2014)
  - Log Ratio – an informal introduction; https://cass.lancs.ac.uk/log-ratio-an-informal-introduction/
- Lijffijt et al. (2014)
  - Significance testing of word frequencies in corpora; https://doi.org/10.1093/llc/fqu064
- Röttger, P. et al.(2022)
  - Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks; https://doi.org/10.18653/v1/2022.naacl-main.13
- Wiegmann et al. (2023)
  - Trigger Warning Assignment as a Multi-Label Document Classification Problem; https://doi.org/10.18653/v1/2023.acl-long.676

Backup

Emotional Abuse: Highest z-scores and log ratios

Highest z-score (Vocabulary)

Term	POS-Tag	Z-Score	Log Ratio
hurt	VERB	43,55	0,32
gaslighting	NOUN	38,71	3,87
fear	NOUN	33,21	0,25
force	VERB	30,22	0,15
tear	NOUN	29,35	0,18
anxiety	NOUN	29,17	0,39
trust	VERB	28,57	0,20
panic	NOUN	27,15	0,26
gaslight	VERB	26,38	3,15
guilt	NOUN	25,64	0,26
manipulate	VERB	25,16	0,49
sob	VERB	24,29	0,30
anger	NOUN	24,24	0,17
dread	NOUN	23,68	0,47
scared	ADJ	23,22	0,19
punish	VERB	23,04	0,35
cry	VERB	22,86	0,13
angry	ADJ	21,69	0,15
punishment	NOUN	21,33	0,28
rage	NOUN	21,03	0,20

Highest log ratio (Vocabulary)

Term	POS-Tag	Z-Score	Log Ratio
gaslighting	NOUN	38,71	3,87
blaming	ADJ	2,83	3,33
ridiculing	ADJ	2,83	3,33
gaslight	VERB	26,38	3,15
manipulated	ADJ	5,63	3,02
blamed	ADJ	3,97	2,78
shaming	ADJ	2,02	2,48
prohibited	ADJ	3,42	2,43
sniveling	ADJ	3,42	2,43
gaslighting	ADJ	1,89	2,33
disregarding	ADJ	3,00	2,15
ridiculed	ADJ	1,98	1,75
berating	ADJ	2,93	1,63
hindering	ADJ	1,83	1,61
overbearance	NOUN	1,26	1,56
hurting	ADJ	2,00	1,43
invalidation	NOUN	2,24	1,35
ignored	ADJ	1,52	1,33
manipulating	ADJ	1,08	1,33
abandoned	ADJ	3,80	1,22

Highest z-score (Non-vocabulary)

Term	POS-Tag	Z-Score
sick	ADJ	34,68
flinch	VERB	34,14
lie	NOUN	33,97
try	VERB	31,63
wrong	ADJ	31,33
deserve	VERB	30,39
remember	VERB	30,38
fault	NOUN	30,10
matter	VERB	30,04
blink	VERB	29,86
suppose	VERB	29,58
swallow	VERB	29,51
memory	NOUN	29,28
awful	ADJ	29,12
sob	NOUN	29,08
lie	VERB	28,60
trust	VERB	28,57
comfort	NOUN	28,45
understand	VERB	28,29
safe	ADJ	27,88

Emotional Abuse: Lowest z-scores

Lowest z-score (Vocabulary)

Term	POS-Tag	Z-Score	Log Ratio
infringement	NOUN	-5,36	-1,78
curse	VERB	-3,80	-0,27
whimpering	ADJ	-1,09	-0,54
disoriented	ADJ	-1,03	-0,56
abase	VERB	-0,88	-0,54
harasser	NOUN	-0,66	-0,66
yelled	ADJ	-0,62	-0,24
fuming	ADJ	-0,60	-0,53
deceiving	ADJ	-0,60	-0,70
tyrannize	VERB	-0,56	-1,09
embarrassed	ADJ	-0,47	-0,21
swearing	NOUN	-0,46	-0,21
silenced	ADJ	-0,44	-0,84
bully	VERB	-0,43	-0,23
deserted	ADJ	-0,43	-0,24
coerced	ADJ	-0,41	-0,52
desert	VERB	-0,37	-0,21
mocked	ADJ	-0,32	-0,40
compelled	ADJ	-0,28	-0,44
abasement	NOUN	-0,15	-0,32

Physical Abuse: Highest z-scores and log ratios

Highest z-score (Vocabulary)

Term	POS-Tag	Z-Score	Log Ratio
bruise	NOUN	80,77	0,88
scared	ADJ	48,93	0,42
beat	VERB	47,30	0,35
beating	NOUN	46,53	0,93
flinch	VERB	43,61	0,36
bruise	VERB	36,46	0,47
punishment	NOUN	33,40	0,42
cut	NOUN	32,96	0,55
broken	ADJ	32,81	0,33
anxiety	NOUN	32,81	0,37
hit	VERB	32,28	0,16
punish	VERB	30,72	0,43
break	VERB	30,48	0,11
scar	NOUN	29,23	0,32
bruised	ADJ	28,85	0,75
injury	NOUN	25,66	0,29
bruising	NOUN	25,27	0,80
wound	NOUN	25,24	0,19
bleed	VERB	24,83	0,24
lock	VERB	24,60	0,15

Highest log ratio (Vocabulary)

Term	POS-Tag	Z-Score	Log Ratio
thump	VERB	2,05	4,07
overmedication	NOUN	3,17	3,10
scalding	ADJ	2,47	2,01
gnawed	ADJ	1,30	1,84
hitting	ADJ	2,82	1,70
overmedicate	VERB	3,26	1,65
slapped	ADJ	1,01	1,43
restraining	ADJ	1,26	1,26
gashed	ADJ	3,66	1,24
pushed	ADJ	2,19	1,17
kicked	ADJ	1,67	1,14
pummeled	ADJ	1,38	0,97
withheld	ADJ	1,58	0,94
beating	NOUN	46,53	0,93
bruise	NOUN	80,77	0,88
swelling	ADJ	0,98	0,84
wrest	NOUN	0,91	0,84
bruising	NOUN	25,27	0,80
caning	NOUN	3,12	0,79
contusion	NOUN	8,94	0,76

Highest z-score (Non-vocabulary)

Term	POS-Tag	Z-Score
tear	NOUN	82,37
abuse	NOUN	79,61
hurt	VERB	75,90
pain	NOUN	64,59
door	NOUN	58,44
cry	VERB	56,19
father	NOUN	53,29
sorry	ADJ	50,40
house	NOUN	50,39
okay	ADJ	49,00
sit	VERB	47,50
walk	VERB	47,10
fear	NOUN	47,05
shake	VERB	45,16
abuse	VERB	44,84
stay	VERB	44,77
car	NOUN	44,16
hospital	NOUN	44,15
sob	VERB	43,93
room	NOUN	43,68

Physical Abuse: Lowest z-scores

Lowest z-score (Vocabulary)

Term	POS-Tag	Z-Score	Log Ratio
trap	NOUN	-11,27	-0,33
bind	VERB	-11,05	-0,25
push	NOUN	-9,71	-0,25
capture	VERB	-9,69	-0,19
spank	VERB	-9,45	-0,41
pull	NOUN	-7,84	-0,24
slapping	NOUN	-7,70	-0,47
denial	NOUN	-6,66	-0,26
spank	NOUN	-6,56	-0,76
attack	VERB	-6,20	-0,14
spanking	NOUN	-6,20	-0,27
swell	NOUN	-5,73	-0,25
rope	NOUN	-5,18	-0,18
bound	ADJ	-5,07	-0,55
manhandle	VERB	-4,93	-0,26
skeletal	ADJ	-4,79	-0,35
intoxicate	VERB	-4,70	-0,20
confine	NOUN	-3,96	-0,19
jab	NOUN	-3,89	-0,22
attacker	NOUN	-3,86	-0,26

Sexual Abuse: Highest z-scores and log ratios

Highest z-score (Vocabulary)

Term	POS-Tag	Z-Score	Log Ratio
rape	VERB	105,72	1,62
rape	NOUN	69,73	1,48
sexual	ADJ	52,38	0,71
scared	ADJ	51,38	0,45
touch	VERB	50,07	0,33
sex	NOUN	47,76	0,48
fear	NOUN	42,33	0,26
bruise	NOUN	40,85	0,45
molest	VERB	40,85	1,57
force	VERB	36,93	0,21
anxiety	NOUN	34,89	0,43
pedophile	NOUN	30,76	1,65
violate	VERB	29,35	0,69
bleed	VERB	25,53	0,25
touch	NOUN	23,66	0,15
vulnerable	ADJ	23,14	0,28
pregnant	ADJ	23,06	0,42
terrified	ADJ	22,71	0,39
suicide	NOUN	22,44	0,54
consensual	ADJ	22,28	0,98

Highest log ratio (Vocabulary)

Term	POS-Tag	Z-Score	Log Ratio
hiv	NOUN	3,51	4,65
violating	ADJ	3,51	4,65
depredate	VERB	1,70	2,23
molestation	NOUN	20,06	1,96
gonorrhea	NOUN	6,58	1,94
hypersexuality	NOUN	5,71	1,67
pedophile	NOUN	30,76	1,65
rape	VERB	105,72	1,62
syphilis	NOUN	5,53	1,62
molest	VERB	40,85	1,57
rape	NOUN	69,73	1,48
violated	ADJ	4,35	1,48
chlamydia	NOUN	6,18	1,33
rectal	NOUN	1,08	1,33
std	NOUN	15,32	1,18
victimize	VERB	10,36	0,99
consensual	ADJ	22,28	0,98
raped	ADJ	1,11	0,95
rectal	ADJ	6,82	0,93
promiscuous	ADJ	9,24	0,87

Highest z-score (Non-vocabulary)

Term	POS-Tag	Z-Score
abuse	NOUN	84,87
hurt	VERB	75,83
tear	NOUN	67,94
abuse	VERB	65,62
want	VERB	62,72
cry	VERB	60,95
okay	ADJ	55,29
sick	ADJ	54,19
trauma	NOUN	54,15
therapy	NOUN	51,59
feel	VERB	51,49
whore	NOUN	51,37
sob	VERB	49,97
pain	NOUN	49,23
bathroom	NOUN	49,02
safe	ADJ	48,36
bed	NOUN	47,75
nightmare	NOUN	43,88
baby	NOUN	43,56
shake	VERB	43,24

Sexual Abuse: Lowest z-scores

Lowest z-score (Vocabulary)

Term	POS-Tag	Z-Score	Log Ratio
stalk	VERB	-8,78	-0,26
thrust	VERB	-7,06	-0,20
concentration	NOUN	-6,60	-0,26
pussy	NOUN	-6,44	-0,53
obscene	ADJ	-5,62	-0,33
stalker	NOUN	-4,75	-0,34
flash	VERB	-4,22	-0,11
leak	VERB	-3,23	-0,19
exhibitionism	NOUN	-3,13	-0,74
creepy	ADJ	-3,11	-0,19
grope	NOUN	-1,91	-0,32
breast	NOUN	-1,77	-0,22
libido	NOUN	-1,75	-0,23
unwelcome	ADJ	-1,56	-0,16
indecency	NOUN	-1,54	-0,30
irritated	ADJ	-1,32	-0,14
disorganized	ADJ	-1,11	-0,21
subjugated	ADJ	-0,98	-1,50
ram	VERB	-0,90	-0,12
unease	NOUN	-0,84	-0,12