

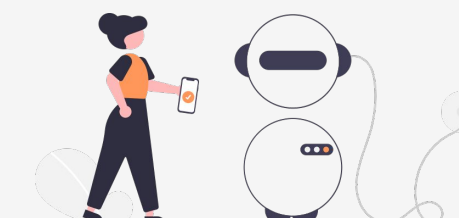
Assessing Large Language Model Effectiveness At Interpreting Misogyny, Harassment, and Cyberbullying on Online Extremist Platforms



James Stevenson [0009-0001-0224-0097] my19303@bristol.ac.uk

Matthew Edwards [0000-0001-8099-0646] matthew.john.edwards@bristol.ac.uk

Bristol Cyber Security Group, School of Computer Science, University of Bristol



Overview

Please note, due to the nature of this paper some quoted hate speech may be present throughout the slides.






Research Questions	01	Behaviour Correlations	05
Methodology	02	Conclusion	06
Data Labelling and Definitions	03		
Classifying Behaviours	04		

James

Stevenson



Background

-  9 years experience in security research and consultancy
-  Occasional conference speaker
-  Published and self-published author

PhD Focus

My PhD focuses on the intersection of computer science and social science, looking at ways machine learning and AI can be used to classify and predict violent extremism and related online harms.

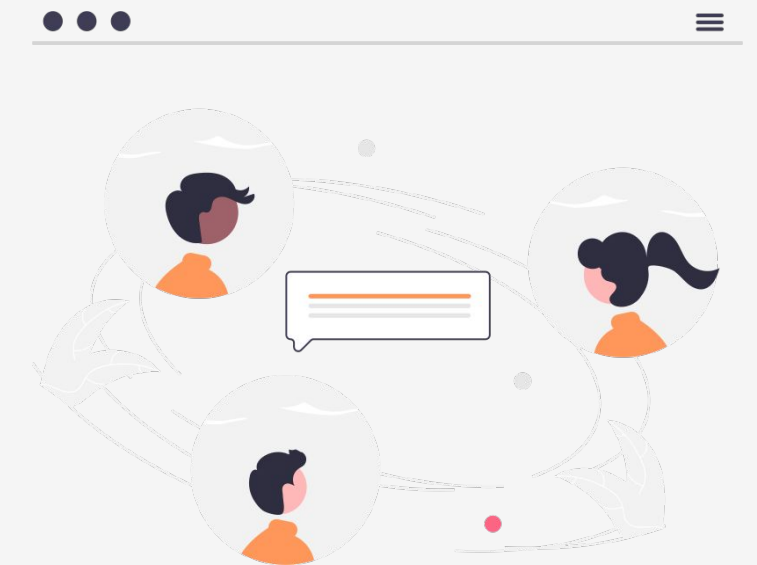
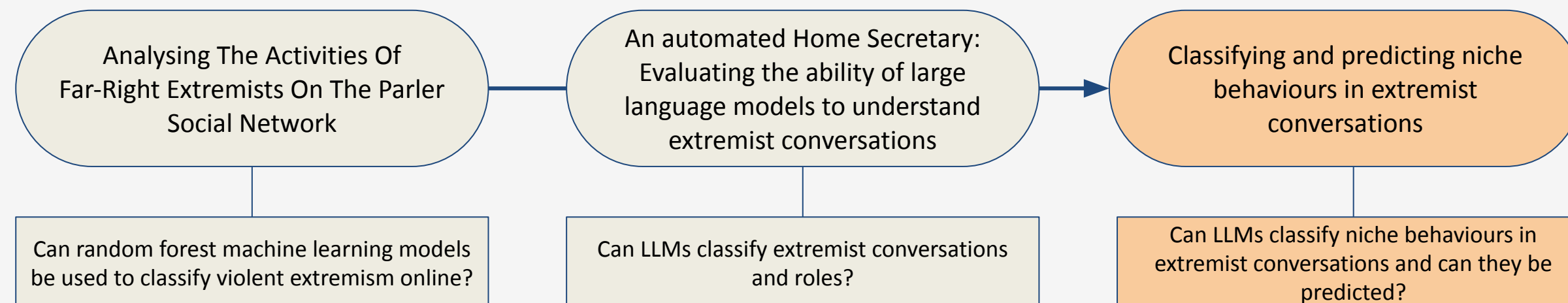


jamesstevenson@bsky.social

Previous Research



Our research fundamentally explores how we can use machine learning to classify, predict, and understand extremism online (with a specific focus on far-right extremism), this sits hand-in-hand with **protecting women and girls online**.

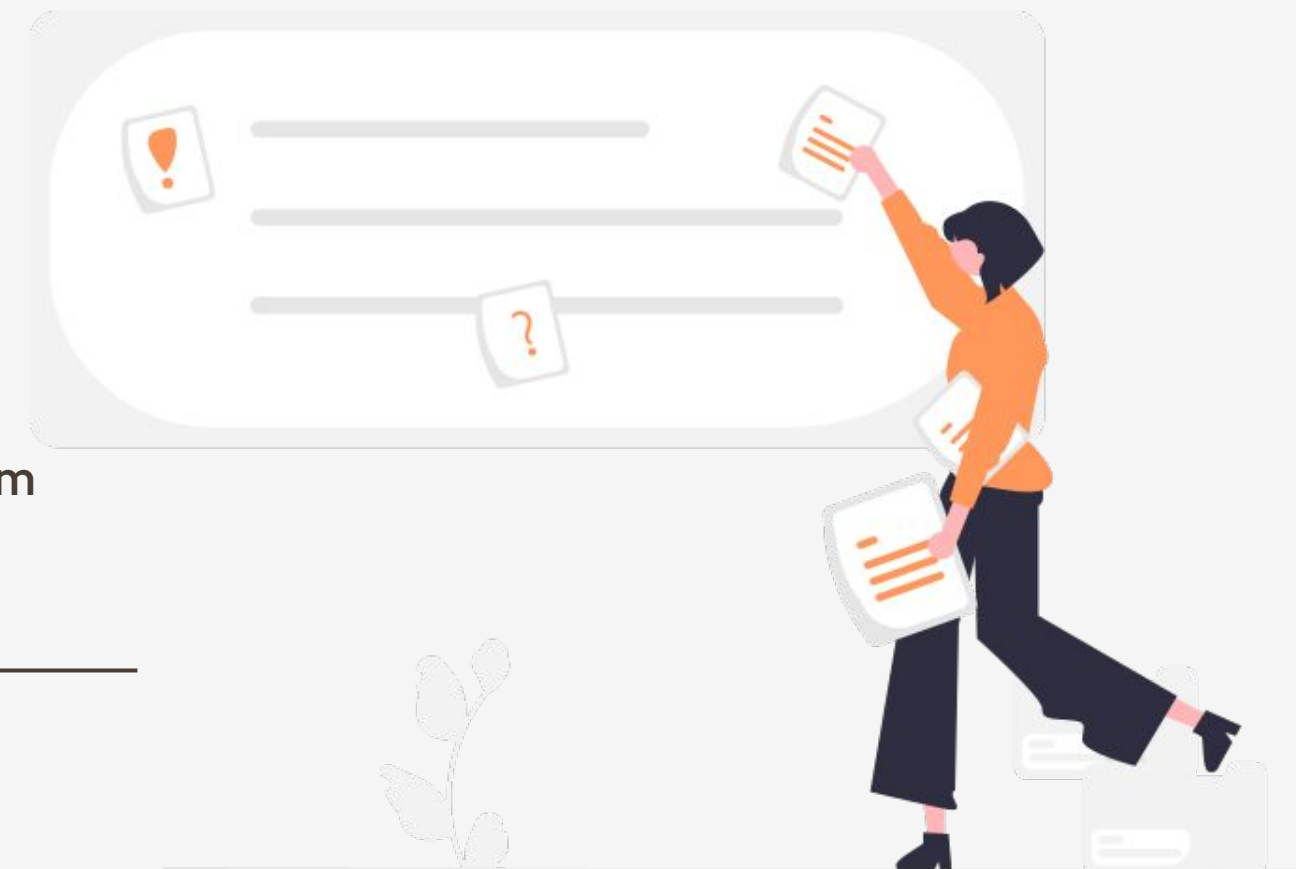


Research Questions

We test whether frontier LLMs can equal human accuracy when detecting extremism and extremism adjacent behaviours.

01 Can large language models **accurately detect an array of behaviours of interest in extremism conversations?**

02 Are there any **correlations** between behaviours seen across extremist conversations?



Methodology



1) Data Selection

Using the Pushshift Telegram dataset we take a candidate channel and split posts into blocks of ten consecutive posts - we later use these conversation blocks for analysis.



2) Data Labelling and Definitions

We define a set of behaviours such as: conspiracy theories dissemination, engaged in cyberbullying, engaged in harassment, etc - with a mix of legal and human definitions.



3) Classifying Behaviours

We take a subset of 110 days including 54 users and 21162 posts, where we use llama-4-maverick and gpt-oss-120b to label each chunk's behaviours.



4) User Behaviour Correlations

We review the correlations in behaviours observed in user activity across the 110 day window.

Data Selection

- Pushshift Telegram Dataset
- Radical Agenda channel - 127,404 posts from 630 unique users, spanning from 24 March 2018 to 9 September 2019.
 - **Initial subset: Timeframe: 2018-03-24 03:25:43 to 2018-07-12 23:29:25, 110 days, 21162 posts, and 54 users.**
- The channel was selected by extracting channels with a high occurrence of keywords (such as 'genocidal', 'civil war', and 'armed') in their 'about' sections. . The list of channels was then further filtered to channels flagged for violent content (i.e., channels violating Apple's App Store guidelines).
- The dataset was chunked into segments of 10 posts, with a sliding window of 3 posts to maintain context across overlapping conversations.



Data Chunking

...

1749277800	UserA	Morning coffee obtained, inbox battle commences. ☕	...
1749279600	UserA	@<USER>: Anyone know if the Northern line is still down?	...
1749279660	UserB	@<USER>: @<USER> It's moving again but expect delays.	...
1749279720	UserC	@<USER>: @<USER> Cheers, that saves me from freezing on the platform.	...
1749279780	UserD	@<USER>: @<USER> Bring a coffee, the queue is savage.	...
1749279840	UserD	@<USER>: @<USER> Already two flat whites in (send help).	...
1749279900	UserC	@<USER>: @<USER> @<USER> Might cycle instead, at least I control the chaos.	...
1749279960	UserD	@<USER>: @<USER> Bold move, traffic looks like a car park.	...
1749280020	UserC	@<USER>: @<USER> @<USER> Race you both. Loser buys lunch.	...
1749280080	UserD	@<USER>: Deal. Screenshot taken for evidence.	...
1749280140	...	@<USER>: Remember, contactless on the bus has a cap so lunch could cost more.	...

...

Conversation Block

Data Chunking

...

1749277800	...	Morning coffee obtained, inbox battle commences. ☕	...
1749279600	...	@<USER>: Anyone know if the Northern line is still down?	...
1749279660	...	@<USER>: @<USER> It's moving again but expect delays.	...
1749279720	UserA	@<USER>: @<USER> Cheers, that saves me from freezing on the platform.	...
1749279780	UserB	@<USER>: @<USER> Bring a coffee, the queue is savage.	...
1749279840	UserB	@<USER>: @<USER> Already two flat whites in (send help).	...
1749279900	UserA	@<USER>: @<USER> @<USER> Might cycle instead, at least I control the chaos.	...
1749279960	UserB	@<USER>: @<USER> Bold move, traffic looks like a car park.	...
1749280020	UserA	@<USER>: @<USER> @<USER> Race you both. Loser buys lunch.	...
1749280080	UserB	@<USER>: Deal. Screenshot taken for evidence.	...
1749280140	UserC	@<USER>: Remember, contactless on the bus has a cap so lunch could cost more.	...

...

Conversation Block

Data Chunking

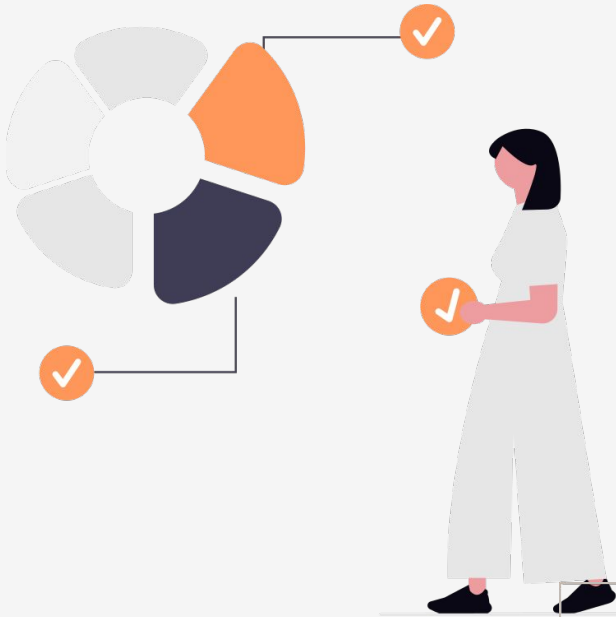
...

1749277800	...	Morning coffee obtained, inbox battle commences. ☕	...
1749279600	...	@<USER>: Anyone know if the Northern line is still down?	...
1749279660	...	@<USER>: @<USER> It's moving again but expect delays.	...
1749279720	..	@<USER>: @<USER> Cheers, that saves me from freezing on the platform.	...
1749279780	...	@<USER>: @<USER> Bring a coffee, the queue is savage.	...
1749279840	...	@<USER>: @<USER> Already two flat whites in (send help).	...
1749279900	UserA	@<USER>: @<USER> @<USER> Might cycle instead, at least I control the chaos.	...
1749279960	UserB	@<USER>: @<USER> Bold move, traffic looks like a car park.	...
1749280020	UserA	@<USER>: @<USER> @<USER> Race you both. Loser buys lunch.	...
1749280080	UserB	@<USER>: Deal. Screenshot taken for evidence.	...
1749280140	UserC	@<USER>: Remember, contactless on the bus has a cap so lunch could cost more.	...

...

Conversation Block

Extremism "...promotion or advancement of an ideology based on violence..."	Mobilisation "...call to action... and operational rhetoric..."	Nostalgia "...based on indicators such as slogans invoking restoration, praise for historical regimes, idealization of past social hierarchies..."	Propaganda and Recruitment "...explicit invitations to join or support extremist groups, heroic portrayals of leaders, distribution of official extremist media, requests for money or logistical support..."
Threats of Violence "...Advocacy, glorification, incitement or detailed instruction for physical harm against persons or property..."	Perceived Grievances "...Statements asserting unfair treatment, injustice or victimisation of the speaker or their in-group..."	Engaged in Harassment "...Content that intentionally targets an individual or group with abusive, insulting or threatening language or behaviour aimed at causing distress..."	



Engaged in Stalking

“...persistently tracks, monitors or attempts to contact another user across posts, messages or platforms...”

Engaged in Cyberbullying

“...repeatedly mocks, shames or threatens a minor or vulnerable person online...”

Misogyny

“...expresses hatred, contempt or prejudice towards women or girls, including degrading stereotypes, calls for exclusion or violence...”

**Conspiracy theories
Dissemination**

“...promotes unfounded explanations attributing events to secret plots by powerful actors, urging acceptance or further spread...”

Terrorism Solicitation

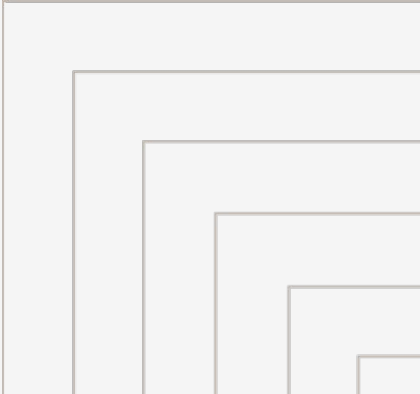
“solicits someone to commit or to contribute to terrorist offences...”

Terrorism Inciting

“incites or advocates terrorist offences...”

Terrorism Providing

“provides instruction on how to conduct terrorism attacks”

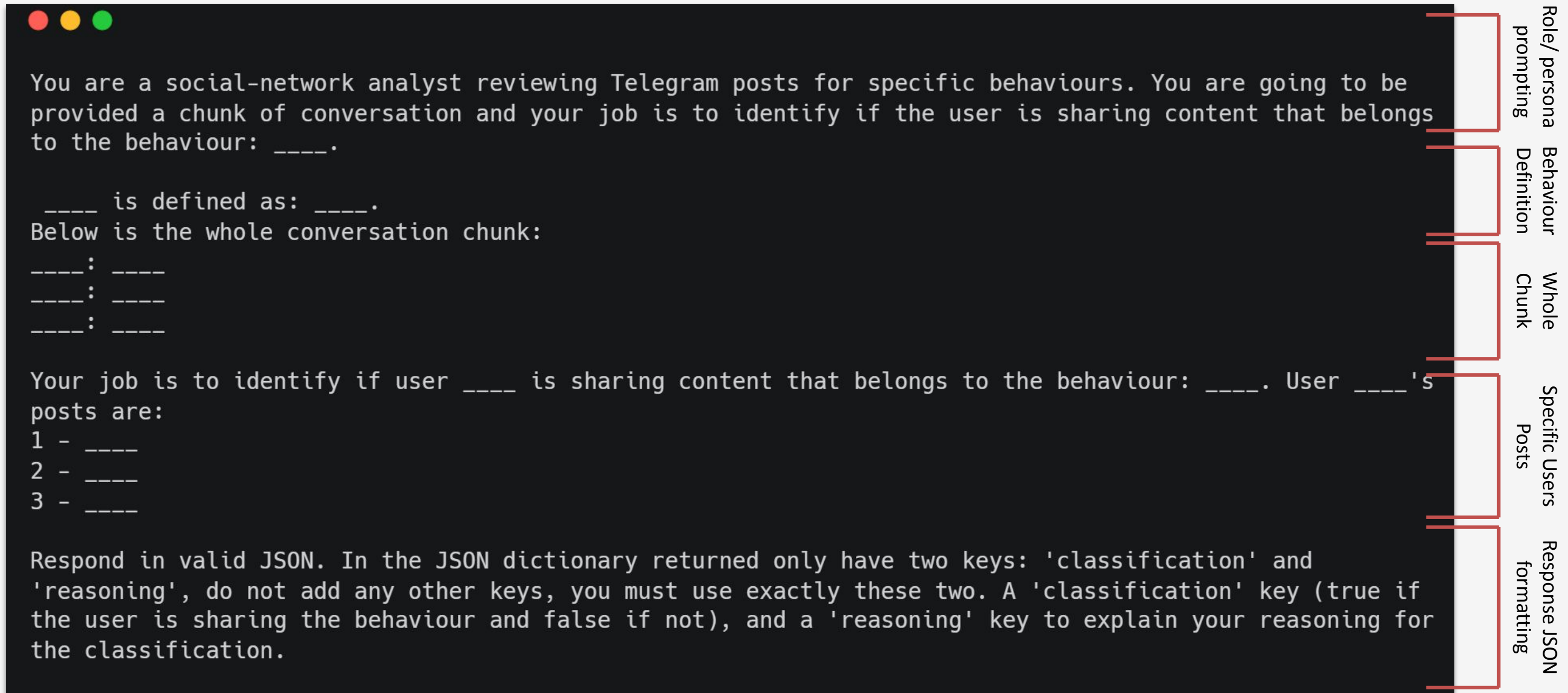


LLM Data Classification

Two models were used across the subset data for classification. Each model would be provided with a **definition of the behaviour** they were classifying, the full **conversation chunk**, and a second list of only the **posts from the user** being classified. The LLM was then asked to classify true or false on if the user's posts were portraying the behaviour.

- llama-4-maverick-17b-128e-instruct
- gpt-oss-120b

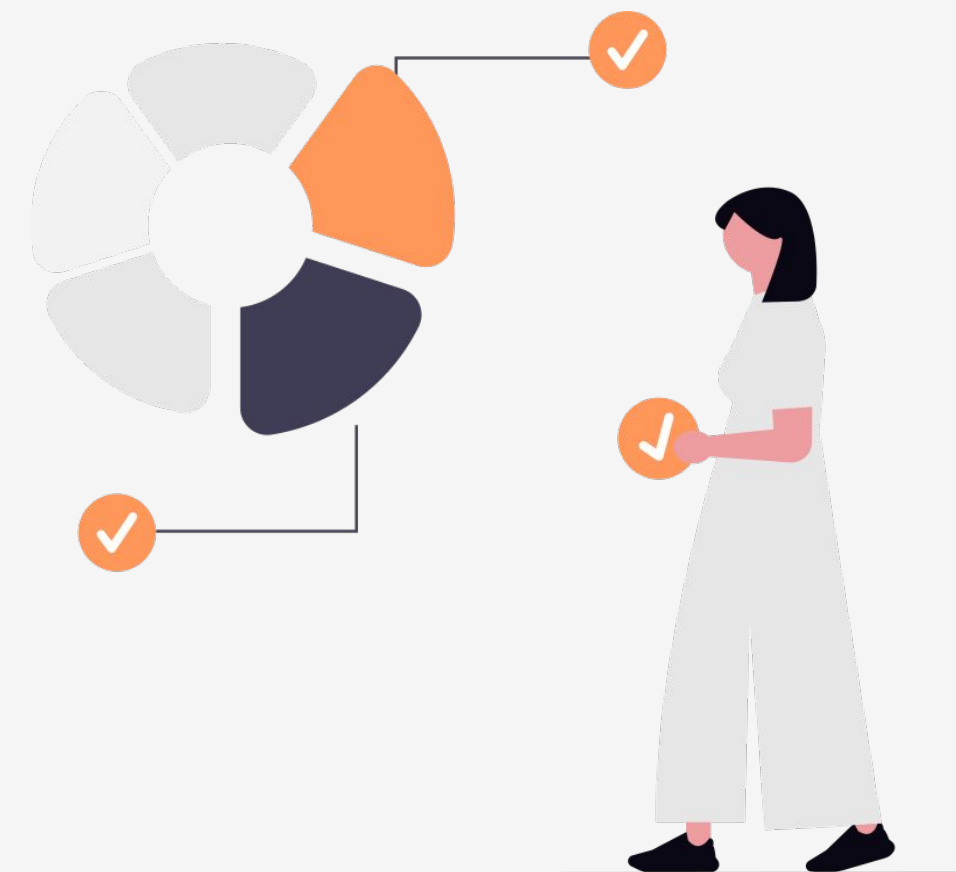


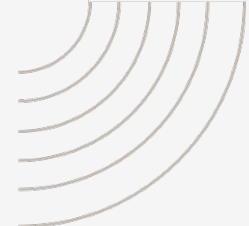


Human Labelling

For initial accuracy metrics a smaller subset of the data has been labelled by a human researcher who was provided the same definition and chunk information. Roughly 300 user conversations in chunks split roughly between:

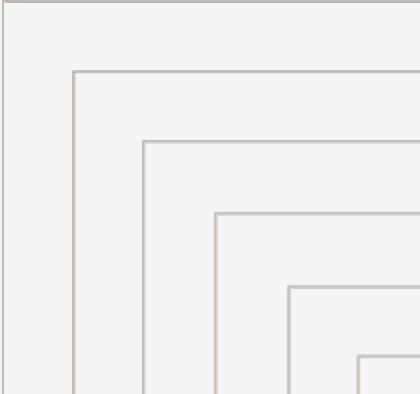
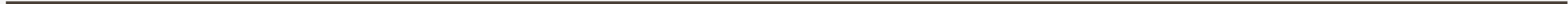
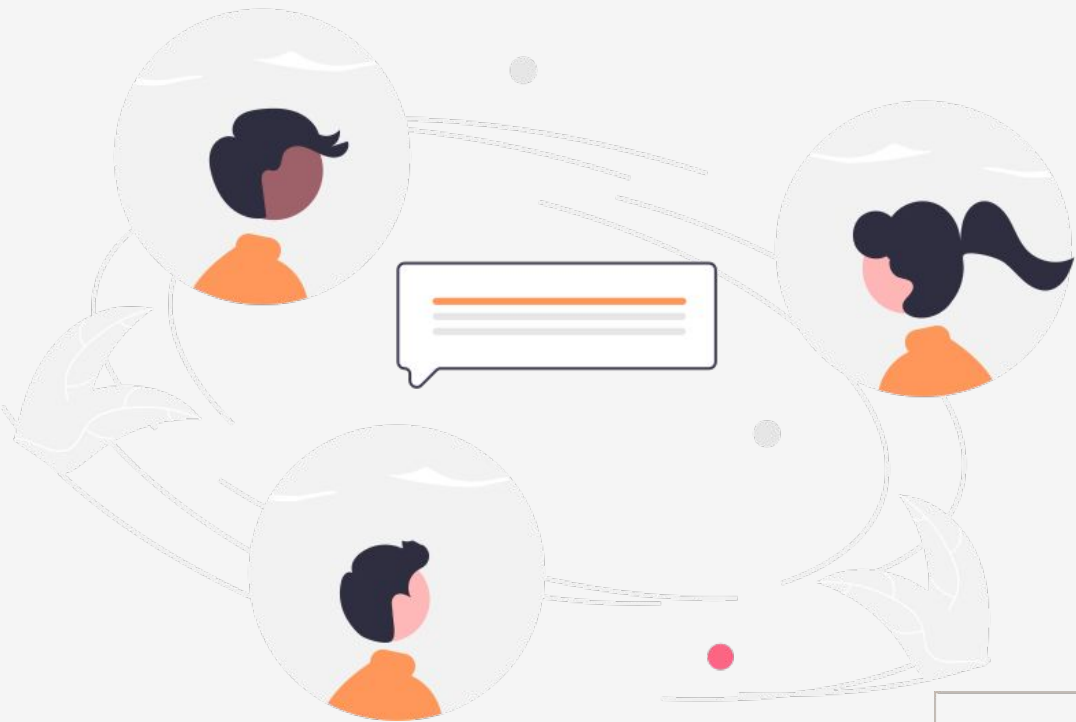
- engaged in cyberbullying
- engaged in harassment
- engaged in stalking
- extremism
- misogyny
- terrorism providing
- threats of violence





Human Labelling

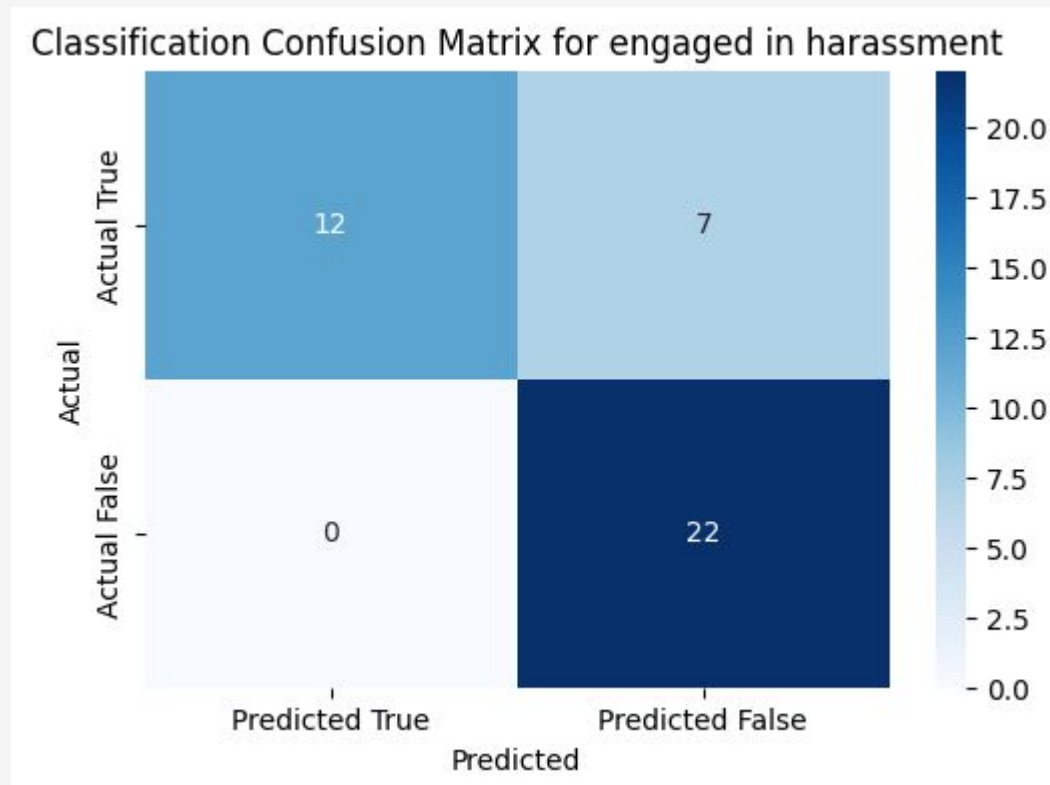
Behaviour	OSS Accuracy	Maverick Accuracy	Ensamble Consensus Accuracy
engaged in cyberbullying	84.09%	72.73%	84.09%
engaged in harassment	85.37%	82.93%	90.24%
engaged in stalking	95.35%	65.12%	95.35%
extremism	78.72%	63.83%	82.98%
misogyny	82.61%	73.91%	82.61%
terrorism providing	90.24%	73.17%	90.24%
threats of violence	93.33%	73.33%	93.33%



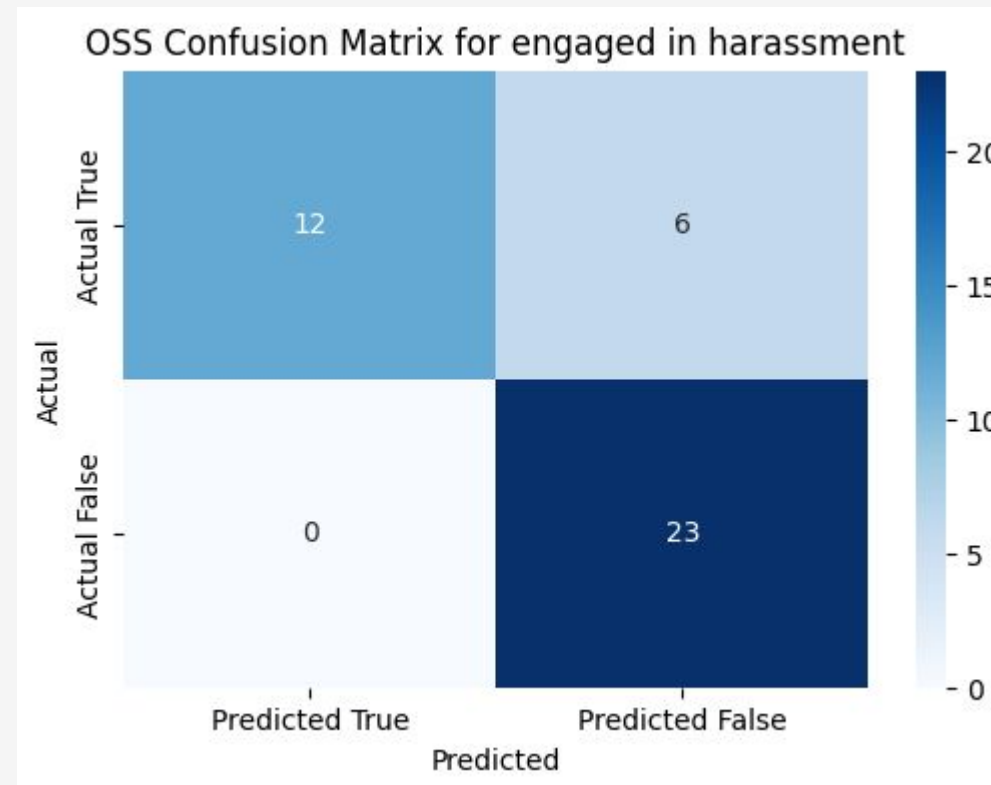
Reviewing LLM Accuracy



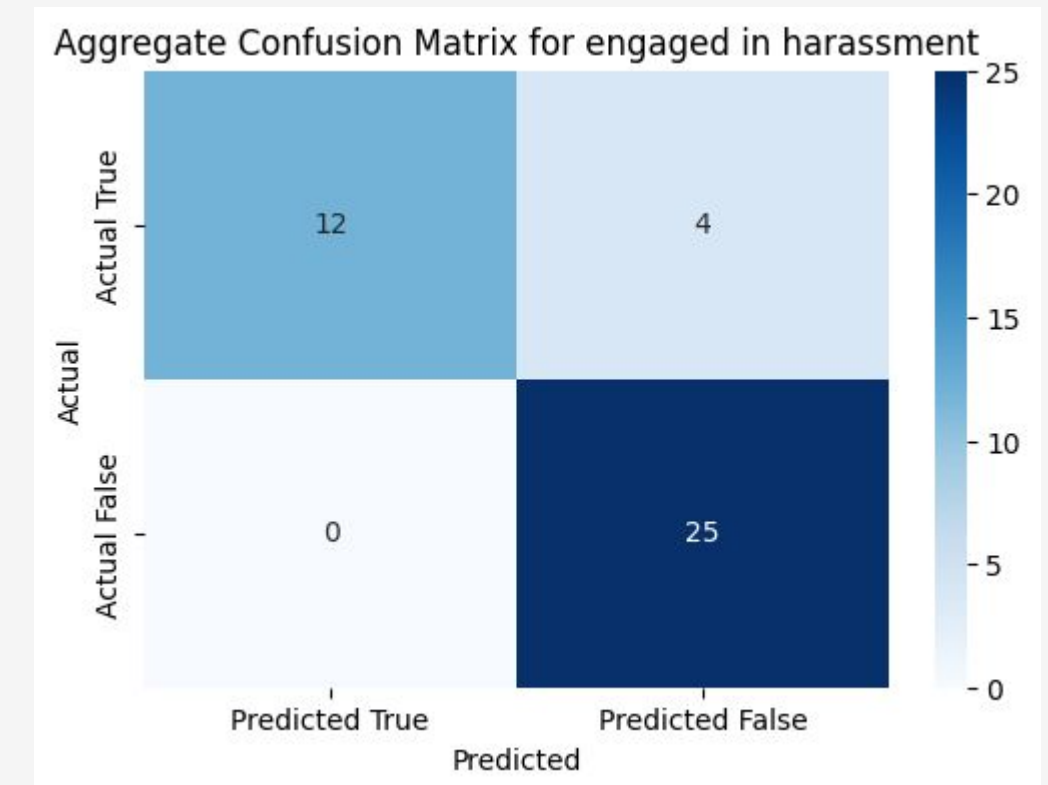
Confusion Matrix For Harassment Behaviour



Maverick Harassment Classification
Compared With Human



Gpt-oss Harassment Classification
Compared With Human



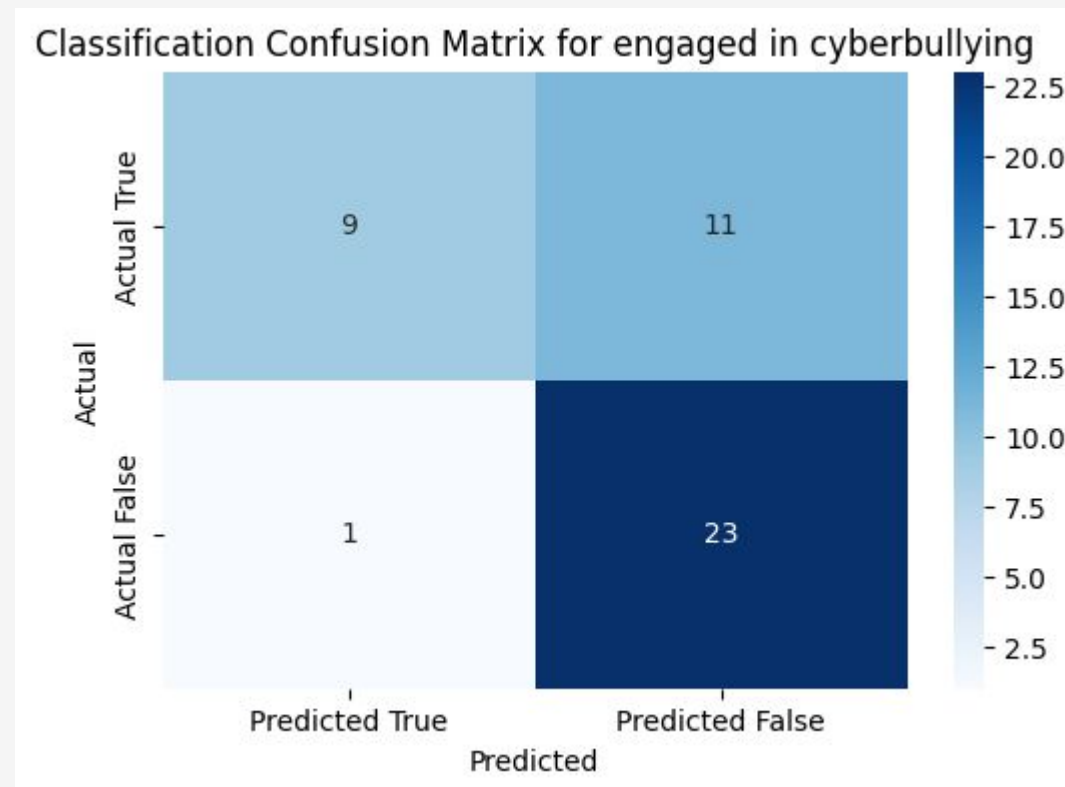
Ensamble Consensus
Harassment Classification Compared With
Human

Low false positives but high false negatives

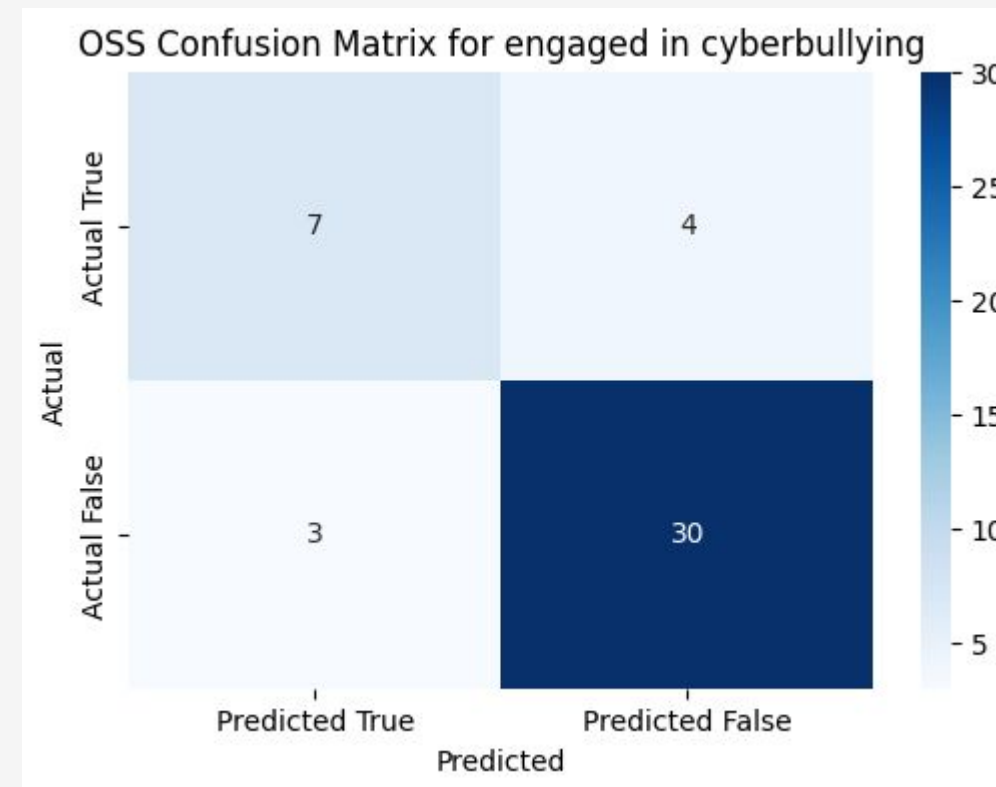
Reviewing LLM Accuracy



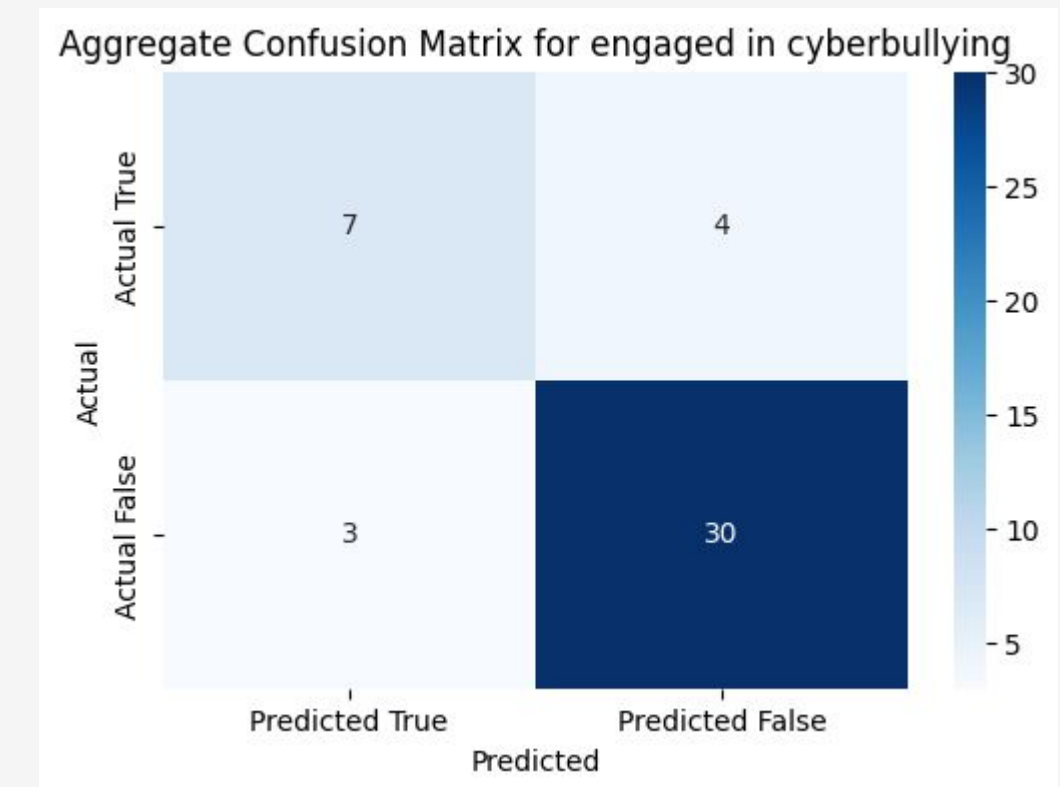
Confusion Matrix For Cyber Bullying



Maverick Cyber Bullying Classification
Compared With Human



Gpt-oss Cyber Bullying Classification
Compared With Human



Ensamble Consensus
Cyber Bullying Classification Compared
With Human

Low false positives but high false negatives

LLM Reasoning

LLMs were provided with a conversation chunk and asked to review the content of a user sharing several posts about death and violent threats.

Llama Maverick: (True) Threats-of-violence

"User B's posts contain phrases that can be classified as threats of violence or glorification of violence. Specifically, posts 3 and 4, 'David "Save a life, kill a nog" Hogg' and 'David "Smash the ZOG" Hogg', contain language that implies violence against specific groups ('nog' is a derogatory term often used against certain ethnic or religious groups, and 'ZOG' is a conspiracy theory term referring to a perceived oppressive force). The quotes around the phrases suggest that they are being used to promote or glorify violence, fitting the definition of threats of violence."

Gpt-oss-120b: (True) Threats-of-violence

"User B's posts include explicit violent language. Post 3 contains the phrase "kill a nog," which is a direct call to kill a person (or group), meeting the indicator of direct threats to kill. Post 4 says "Smash the ZOG," where "ZOG" is a known antisemitic slur and the phrase is a clear call to violent action against that out-group. Both posts constitute advocacy or incitement to physical harm, so the behavior is classified as threats of violence. The other posts are non-violent and do not affect the overall classification."

LLM Reasoning

LLMs were provided with a conversation chunk and asked to review the content of a user sharing several posts revolving around insulting women.

Llama Maverick: (True) Misogyny

"User A's statement 'Let people insult wives' is a clear expression of a misogynistic attitude, as it advocates for the degradation or disrespect towards women, specifically in the context of their marital status. This aligns with the definition of misogyny, which includes degrading stereotypes or contempt towards women."

Gpt-oss-120b: (True) Misogyny

"User A's statement 'Let people insult wives' encourages harassment and contempt toward women, which aligns with the definition of misogyny (expressing hatred, contempt, or prejudice toward women). This single instance is sufficient to classify the content as misogynistic, even though the other messages do not reference women."

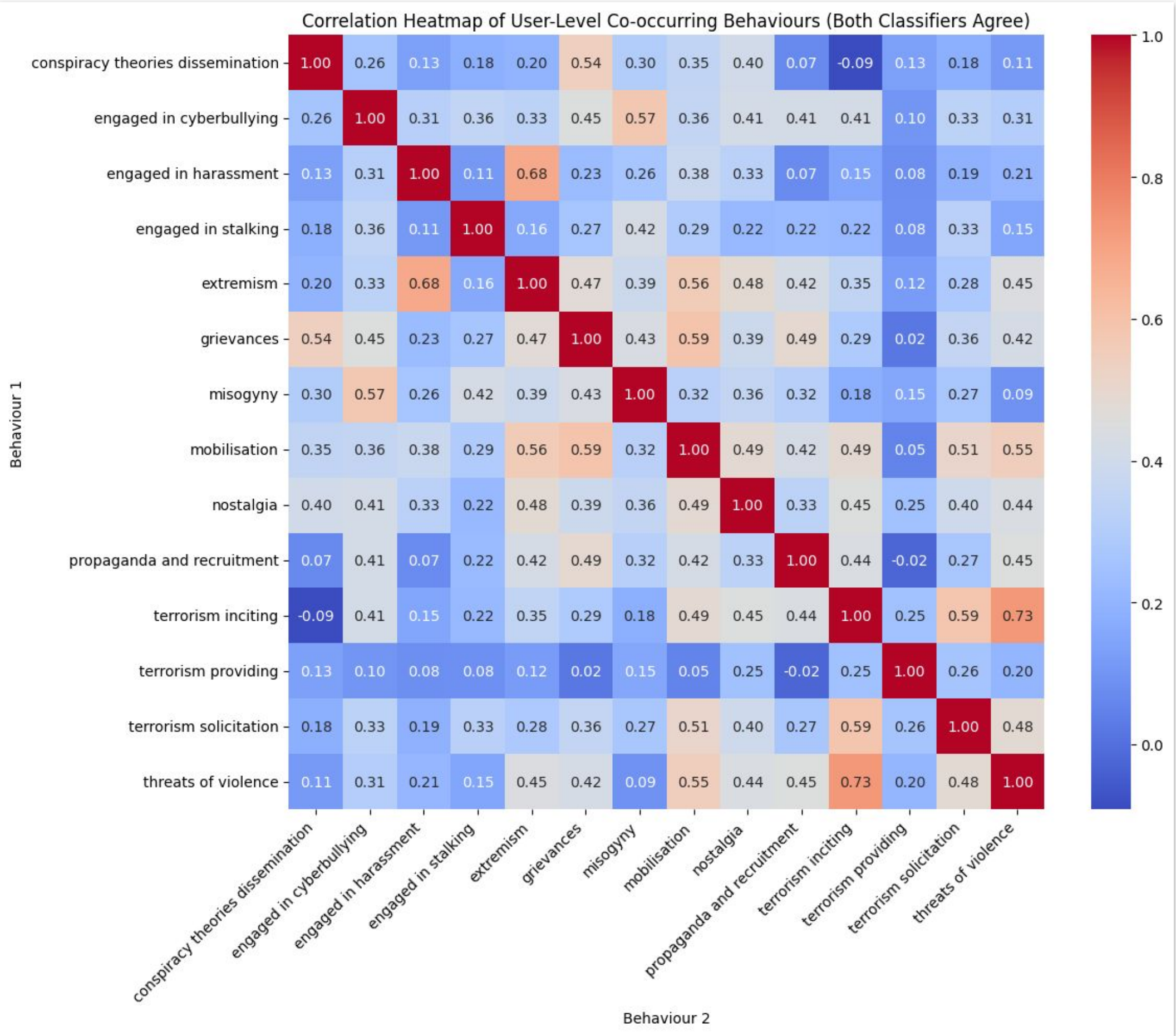
Behaviour Correlation

- As the Ensemble Consensus approach has the highest accuracy, we used this to define if a user is sharing a given behaviour.
- Across the sample dataset we aggregate all behaviours shared in conversations by each of the 54 users.
- Then we run a correlation coefficient across user behaviours to identify conversation behaviour patterns.



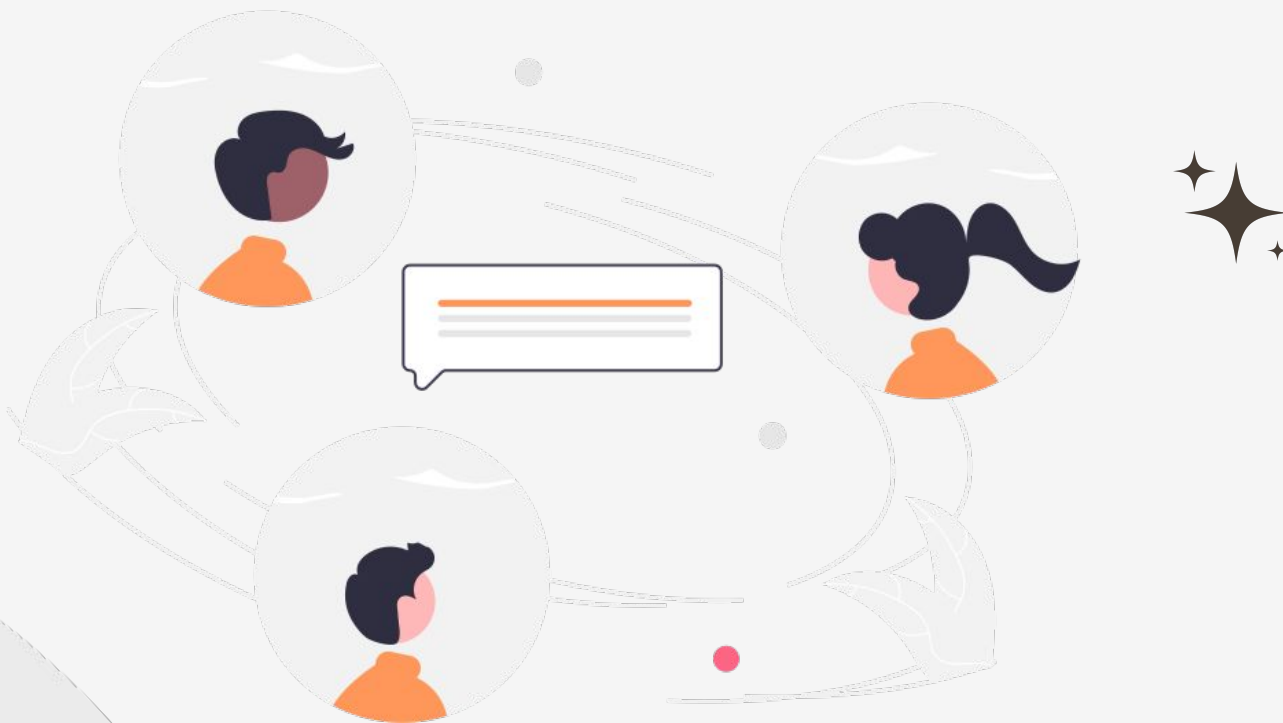
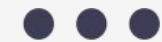
Correlation of user level co-occurring behaviours

- Extremism has a very strong correlation with harassment (0.68)
- terrorism inciting and threats of violence have a high correlation (0.73)
- Mobilisation co-occurs strongly with grievances (0.59) and threats of violence (0.55)
- Misogyny is moderately linked with cyberbullying (0.57) and stalking (0.42) and grievances (0.43) and extremism (0.39)
- Nostalgia shows moderate co-occurrence with mobilisation (0.49) and extremism (0.48)
- Conspiracy theories dissemination correlates with grievances (0.54)



Conclusion

LLMs show an initial ability to detect niche behaviours in online extremist conversations – with initial accuracies over 90%, however, with low false positives but high false negatives.



Behaviour Correlations

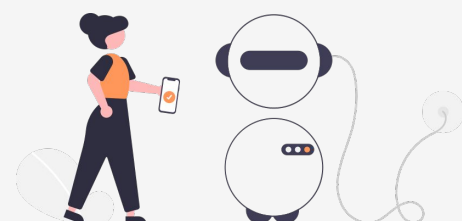
Extremism and related behaviours show strong links with harassment, threats, mobilisation, grievances, misogyny, nostalgia and conspiracy theories, highlighting significant overlaps between harmful attitudes and actions.



Next Steps

- Extending the sample dataset and increasing the user labelled sample.
- Fine tuning behaviour definitions/ prompts to improve accuracy.
- Utilising temporal graph networks or time series analysis to predict occurrence of such behaviours.

THANK YOU



James Stevenson [0009-0001-0224-0097] my19303@bristol.ac.uk

Matthew Edwards [0000-0001-8099-0646] matthew.john.edwards@bristol.ac.uk