CAN STATE-OF-THE-ART GENERATIVE LARGE LANGUAGE MODELS

IDENTIFY AND RATIONALISE EXTREME CONTENT AND ROLES IN CONVERSATIONS

James Stevenson and Luke Gassmann
my19303@bristol.ac.uk, luke.gassmann@bristol.ac.uk

OVER

80% Accuracy

Chat GPT 4.0 scored an 80.6% agreement score when compared with the ‘is extremism’ / ‘not extremism’ label provided by two human researchers.

www.linkedin.com/in/jamesstev

9 Labels

- **People Leader**: Directs, recruits and mobilises block members either virtually or in the real world.
- **Leader Influencer**: Directs the conversation as a knowledge source and/or gatekeeper, with other members reflecting their attitudes.
- **Engager Negator**: Negative or berating interactions in an attempt to reduce discussion or offer a counter argument against a fundamental principle groups agreement.
- **Engager Supporter**: Positively interacts with the topic, encouraging or promoting future further discussion and ideological success.
- **Engager Neutral**: Neutral topic interaction to learn or socially interact with members.
- **Bystander**: Does not engage with the main discussion but remains within the discussion block.
- **NATTC**: The user is at the beginning or end of a block (outside of the time boxed discussion) discussing a topic that the content is not engaged with.
- **Is Extremism** - The conversation meets the UK definition of extremism.
- **Non-Extremism** - The conversation does not meet the UK definition of extremism.

‘Conversations’ were broken into overlapping blocks of 10. Usernames were stripped and researchers/ LLMs were asked to label each user with the labels based on available posts in that block from the user.

Extremism Label

Model	Agreement Score
ChatGPT 4.0	0.806
mistralaimixtral-8x7b	0.757
Llama3 70b	0.75
ChatGPT 3.5	0.191

Role Label

Model	Agreement Score
Llama3 70b	0.376
mistralaimixtral-8x7b	0.329
ChatGPT 3.5	0.318
ChatGPT 4.0	0.192

In inital research, some models drastically underperfom due to hallucinating and not following prompt ‘rules’.

DATASET

Telegram

From 630 unique users and 127404 posts taken from a right-wing Telegram channel from ‘The Pushshift Telegram Dataset’ over 2018-03-24 to 2019-09-09. Of these 500 users were manually labelled by researchers for comparison by the LLMs.

LLM REASONING

LLMs were asked to explain their decision making, such as:

> User Bs content includes several instances of extremism. For example, statements like ‘Keep bothering with enlightenment ideas like voting or women in politics all the way to your extinction white man’ and ‘Get back in the kitchen’ indicate the promotion of intolerance and the negation of fundamental rights and freedoms of others. Promoting literature like ***** ****** supports the extremist label, as ***** is known for promoting violent and white supremacist ideologies.

NEW DEFINITION OF EXTREMISM (2024)

75% over 62%

The new definition of extremism (2024) outperformed the shorter older definition of extremism from the 2011 UK Prevent Strategy. As seen in the LLAMA70b model having an agreement score of 75% with the new definition and 62% with the older.

Future Research: Llama3 70b was reliably in the top performers, when compared to human research agreement, in all tasks. Some models suffered with hallucinating rules and adding invalid labels. Role labels should be reduced to key roles. Further research should build on developing custom LLM adapters trained off human labelled extremist content.

Full New definition of Extremism (2024)

Model	Agreement Score
Llama3 70b (long 2024 definition)	0.592

2011 UK Prevent Strategy Definition

Model	Agreement Score
Llama3 70b (2011 Definition)	0.623
mistralaimixtral-8x7b (2011 Definition)	0.463